UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

Deep Semantic Entity linking

"Documento Definitivo"

Doutoramento em Informática

Pedro Simões Ruas

Tese orientada por:

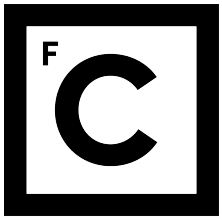
Professor Doutor Francisco José Moreira Couto

Documento especialmente elaborado para a obtenção do grau de doutor

2024

UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS



Ciências
ULisboa

Deep Semantic Entity linking

Doutoramento em Informática

Pedro Simões Ruas

Tese orientada por:

Professor Doutor Francisco José Moreira Couto

Júri:
Presidente:

- Doutor Manuel João Caneira Monteiro da Fonseca, Professor Associado com Agregação e Presidente do Departamento de Informática, da Faculdade de Ciências da Universidade de Lisboa.

Vogais:

- Doutora Carla Alexandra Teixeira Lopes, Professora Auxiliar da Faculdade de Engenharia da Universidade do Porto;
- Doutor Sérgio Guilherme Aleixo de Matos, Professor Associado com Agregação do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro;
- Doutor Francisco José Moreira Couto, Professor Associado com Agregação da Faculdade de Ciências da Universidade de Lisboa (orientador);
- Doutor André Nuno Carvalho Souto, Professor Auxiliar da Faculdade de Ciências da Universidade de Lisboa.

This work has been supported by FCT through Deep Semantic Tagger (DeST) Project under Grant PTDC/CCIBIO/28685/2017 (http://dest.rd.ciencias.ulisboa.pt/), in part by LASIGE Research Unit under Grants UIDB/00408/2020 (https://doi.org/10.54499/UIDB/00408/2020) and ref. UIDP/00408/2020 (https://doi.org/10.54499/UIDP/00408/2020), and in part by FCT through PhD Scholarship under Grant ref. 2020.05393.BD

Documento especialmente elaborado para a obtenção do grau de doutor

2024

Acknowledgements

My deepest gratitude to my supervisor, Francisco Couto, for his invaluable guidance and support throughout these years. His mentorship shaped not only my research but also my personal development. The freedom to explore that he gave me and our discussions were essential for pushing my work forward and preparing me for my next steps.

I would also like to thank my colleagues Sofia and Diana for their help, motivation, and the discussions that challenged me to improve my work. I am equally thankful to all the people with whom I've collaborated. A special thanks to Alexandra and Carla from LASIGE for their valuable assistance over the years.

I am grateful to Michael Schroeder for the opportunity to explore new directions and perspectives and to my colleagues in Dresden, whose company made my time there memorable.

I deeply appreciate the financial support provided by FCT and LASIGE and the resources provided by LASIGE, the Faculty of Sciences of the University of Lisbon and the University of Malaga, which were vital to completing this work.

I am grateful to my friends, who have been a constant source of encouragement and balance throughout this journey.

To my family—parents, grandmother, brother João, and sister Juliana—I am grateful for their encouragement and support. I'm thankful to Arménio, who was a pillar of invaluable wisdom.

My appreciation also goes to all the cats who inspired me during my work with their infinite sagacity, especially Cali.

Finally, I'm forever indebted to Filipa, the foundation of everything. Her endless patience, strength, and unwavering belief in me made this achievement possible. This accomplishment is as much hers as it is mine.

Abstract

Knowledge organization systems, such as ontologies and knowledge graphs, are essential for organizing biomedical and clinical information and data. However, the growing volume of available scientific literature raises challenges in their maintenance. Entity linking approaches assist humans in curation by mapping entities described in text to entries of the knowledge organization systems, but their lack of coverage originates unlinkable or NIL entities. Besides, the state-of-the-art depends on deep learning models trained on large amounts of human-annotated data, which is hard to acquire. The present research work focuses on tackling these limitations of human-annotated data in the biomedical entity linking task. First, it addresses the lack of coverage of biomedical knowledge organization systems by using relation extraction to find missing relations and focusing on the problem of the NIL entities. Relation extraction increases the semantic information available for graph-based entity linking approaches (REEL), and focusing on the partial mapping of NIL entities (i.e. NIL entity linking) also improves the performance of such approaches (NILINKER). Second, the research work proposes a new deep learning model trained on a large-scale training dataset generated through automatic methods. The model is part of the pipeline X-Linker integrating different entity linking models, providing more flexibility and performance. The pipeline achieved state-of-the-art performance in the biomedical entity linking task in several datasets (BC5CDR-Disease, BioRED-Chemical, NCBI Disease). The described approaches and several others focusing on related tasks, such as named entity recognition, text classification, and recommendation of biomedical entities, were applied to several case studies, including competitions, workshops and challenges.

Keywords: Biomedical Entity Linking, Text Mining, Natural Language Processing, Knowledge Organization Systems, Deep Learning

Resumo

Sistemas de organização do conhecimento, incluindo ontologias e grafos de conhecimento, são essenciais na organização de dados e informação biomédicos e clínicos. No entanto, a crescente quantidade de literatura científica disponível levanta desafios à manutenção destes sistemas. As abordagens automáticas de mapeamento de entidades ajudam os especialistas humanos no processo de curadoria através da associação de entidades descritas em texto com registos presentes em sistemas de organização de conhecimento, mas as limitações destes em relação à sua abrangência originam entidades não mapeáveis ou NIL e deficiências ao nível da informação contextual. Para além disso, o estado da arte depende de modelos de aprendizagem profunda treinados em grandes quantidades de dados anotados por humanos, que são difíceis de gerar. O trabalho de investigação aqui descrito foca-se em resolver estas limitações. Em primeiro lugar, aborda a abrangência limitada de sistemas biomédicos de organização do conhecimento através do recurso a abordagens automáticas para extração de relações para encontrar relações ausentes e através do foco no problema das entidades NIL. A extração de relações aumenta a informação semântica disponível em abordagens de mapeamento de entidades baseadas em grafos (REEL). O foco no mapeamento parcial de entidades NIL também aumenta o desempenho de abordagens de mapeamento de entidades (NILINKER). Em segundo lugar, o trabalho propõe uma nova abordagem para mapeamento de entidades baseada num modelo de aprendizagem profunda treinado num conjunto de dados de larga escala gerado automaticamente. O modelo é integrado no *pipeline* X-Linker que integra diferentes abordagens, o que leva a um aumento da flexibilidade e da performance. Este *pipeline* alcançou um desempenho estado da arte na tarefa de mapeamento de entidades biomédicas em vários conjuntos de dados (BC5CDR-Disease, BioRED-Chemical, NCBI Disease). As abordagens descritas e outras relacionadas, como reconhecimento de entidades, classificação de texto e recomendação de entidades biomédicas foram aplicadas em diferentes casos de estudo, incluindo competições e *workshops* e no desenvolvimento de uma ferramenta de anotação de texto biomédico com foco na usabilidade (BENT).

Palavras chave: Mapeamento de entidades biomédicas, Prospecção de texto, Processamento de linguagem natural, Sistemas de organização do conhecimento, Aprendizagem profunda

Resumo Alargado

Grande parte do conhecimento científico e tecnológico está hoje disponível sob o formato de texto. A literatura científica, sob a forma de artigos em particular, é essencial para o avanço científico, uma vez que é o principal meio de comunicação a que os cientistas recorrem para partilha de informação. Para além disso, as novas descobertas científicas são sempre interpretadas em relação à literatura já existente.

No entanto, o ritmo de publicação de texto científico nos vários formatos continua a aumentar consideravelmente, o que cria desafios no acesso a informação relevante.

Neste sentido, sistemas de organização do conhecimento, como ontologias ou grafos de conhecimento, desempenham um papel relevante na gestão eficiente e integração de grandes quantidades de dados, incluindo aqueles extraídos da literatura científica. Estes sistemas possibilitam a padronização e a partilha de conhecimento e, para além disso, permitem a aplicação de processos automáticos para descobrir novo conhecimento, uma vez que o conteúdo armazenado é simultaneamente acessível por utilizadores humanos e sistemas automáticos.

Estes sistemas de organização do conhecimento são normalmente curados por especialistas humanos, que convertem o conteúdo relevante presente em texto para o esquema lógico do sistema alvo, um processo que é lento, dispendioso e que requer conhecimentos especializados no domínio a ser modelado.

As abordagens de prospeção de texto facilitam processos de curadoria através da extração automática de padrões em grandes quantidades de texto. Dentro do *pipeline* de prospeção de texto, o mapeamento de entidades é um componente essencial, uma vez que é responsável pela ligação entre texto, normalmente expresso em linguagem natural por humanos, e sistemas de organização do conhecimento, que são repositórios estruturados com conteúdo acessível por programas automáticos. Várias aplicações dependem dos resultados que saem de abordagens de mapeamento de entidades, incluindo sistemas de recuperação de informação, motores de busca e sistemas de resposta automática a questões, outros componentes do *pipeline* de prospeção de texto, como abordagens automáticas para extração de relações e de eventos, sistemas para organização, gestão e prospeção de literatura científica, sistemas focados em aplicações clínicas, através da extração e mapeamento de entidades relevantes a partir de registos eletrónicos, no-

tas clínicas, prescrições e relatórios laboratoriais, que têm como objetivo melhorar a tomada de decisões clínicas, o atendimento ao paciente e a análise de dados na área da saúde.

A produção e publicação expressivas de novo conhecimento científico origina desafios ao desempenho de abordagens de mapeamento de entidades e à abrangência dos sistemas de organização do conhecimento, em particular na área biomédica. O processo de curadoria destes sistemas não acompanha o ritmo de publicação, o que origina entidades não mapeáveis ou NIL, isto é, entidades reconhecidas num determinado texto que não são mapeáveis para nenhum dos registos de um determinado sistema de organização de conhecimento. Esta abrangência limitada ou incompletude dos sistemas de organização do conhecimento faz com que informação semântica potencialmente relevante permaneça inacessível no texto. A existência de entidades NIL é um dos maiores desafios a abordagens de mapeamento de entidades, uma vez que impossibilita a concretização da sua tarefa última: associar entidades expressas num texto a um registo específico presente em sistemas de organização do conhecimento. As abordagens existentes que se concentram nas entidades NIL apenas tentam prever quais as entidades NIL de entre todas as entidades presentes num determinado documento ou no máximo, tentam agrupar as entidades NIL conhecidas.

Para além disso, a abrangência limitada ou incompletude dos sistemas de organização do conhecimento manifesta-se por vezes na ausência de relações semânticas entre registos de um determinado sistema de organização do conhecimento. A ausência destas relações reduz a informação que abordagens de mapeamento de entidades têm disponível, em particular, a informação contextual acerca de uma determinada entidade. Esta limitação traduz-se num desempenho reduzido deste tipo de abordagens.

Por outro lado, o facto de o processo de curadoria de dados ser moroso leva a que os conjuntos de dados de treino disponíveis sejam limitados em número e ao nível da representação de entidades. O desempenho de abordagens baseadas em modelos de aprendizagem profunda está intimamente dependente dos dados acessados durante a sua fase de treino, logo, se os conjuntos de dados de treino são limitados, o seu desempenho também o será. Deste modo, é necessário resolver esta limitação através de métodos automáticos de geração de dados ou métodos como supervisão distante ou *zero-shot*.

As limitações ao nível das abordagens de mapeamento de entidades traduzem-se numa pior gestão de sistemas de organização do conhecimento e, consequentemente, numa pior organização e partilha de dados e informação científicos. Para além disso, as aplicações que dependem dos resultados gerados por estas abordagens também são afetadas, como por exemplo, sistemas de recuperação de informação e prospeção de texto, aplicações para prospeção de literatura biomédica e sistemas relacionados com a área

clínica. As abordagens de mapeamento de entidades são essenciais para adicionar uma camada semântica ao texto, aumentando a capacidade de máquinas interagirem com a linguagem humana e de derivarem conhecimento a partir dela.

O presente trabalho de investigação foca-se na resolução dos dois desafios relationados com as limitações de dados anotados por humanos na tarefa de mapeamento de entidades: a insuficiente abrangência de sistemas de organização do conhecimento e a excessiva dependência de conjuntos de dados limitados para treino de abordagens baseadas em aprendizagem profunda.

Os objetivo último do trabalho é o desenvolvimento de abordagens de mapeamento de entidades biomédicas com um desempenho estado da arte. Os objetivos específicos incluem a resolução da insuficiente abrangência de sistemas de organização do conhecimento biomédico através do recurso a abordagens de extração de relações e a abordagens baseadas em aprendizagem profunda para mapear entidades NIL (objetivo 1) e a diminuição da dependência excessiva de conjuntos de dados anotados por humanos através do foco em abordagens baseadas em aprendizagem profunda treinadas em dados gerados automaticamente e na integração de vários tipos de abordagens para mapeamento (objetivo 2).

Para atingir o objetivo 1, o presente trabalho inclui duas estratégias.

A primeira consiste no uso de abordagens de extração de relações para suprir a carência de relações descritas em sistemas de organização do conhecimento biomédico. Abordagens de mapeamento de entidades baseadas em grafos tipicamente constroem grafos de desambiguação contendo todas as entidades ou candidatos num determinado documento: os nós representam candidatos e as conexões entre nós são baseadas na informação disponível no respetivo sistema de organização do conhecimento. No entanto, se a informação disponível no sistema de organização do conhecimento for insuficiente, as conexões do grafo de desambiguação também serão afectadas, levando a um pior desempenho da abordagem de mapeamento de entidades. O presente trabalho propõe uma abordagem designada por REEL, que é capaz de mapear entidades associadas a compostos químicos e doenças para vários sistemas de organização de conhecimento biomédico (ChEBI, MEDIC, CTD-Chemicals) usando grafos, o algoritmo Personalized PageRank e extração de relações para completar o grafo com relações extraídas a partir de artigos científicos. A inovação aqui consiste em usar extração de relações para melhorar o mapeamento de entidades e não o oposto, o que acontece tipicamente em *pipelines* de prospeção de texto.

A segunda estratégia consiste no desenvolvimento de uma abordagem para mapear parcialmente entidades NIL para sistemas de organização de conhecimento biomédico. O pressuposto é que será preferível efetuar um mapeamento parcial, ainda que imperfeito, em detrimento de ignorar por completo a entidade

NIL. O trabalho propõe uma abordagem baseada em aprendizagem profunda e no mecanismo de atenção para conseguir representar efetivamente as entidades NIL e gerar uma lista de candidatos possíveis a partir dos sistemas de organização de conhecimento biomédico alvo. A abordagem, designada por NILINKER, é inspirada no conceito de composicionalidade semântica para gerar representações para as entidades NIL. Para além disso, é demonstrado que a integração deste tipo de abordagens com a abordagem de mapeamento de entidades REEL descrita anteriormente leva a um desempenho acrescido desta última.

Para atingir o objetivo 2, a principal contribuição consiste numa abordagem de mapeamento de entidades baseada em aprendizagem profunda treinada num conjunto de dados de larga escala que é posteriormente integrada num *pipeline*. O modelo de aprendizagem profundo é adaptado a partir da tarefa de *extreme multilabel ranking*. Neste cenário, existem entidades a serem mapeadas e um grande número possível de candidatos, isto é, todos os registos disponíveis no sistema de organização de conhecimento, e objetivo é construir representações precisas das entidades e dos registos. O modelo é designado por PECOS-EL. O conjunto de dados de treino é gerado automaticamente a partir de anotações automáticas e informação armazenada em dois sistemas de organização de conhecimento biomédico, MEDIC e CTD-Chemical.

O *pipeline* proposto integra diferentes módulos para lidar com diferentes entidades: detetor de abreviaturas, um componente *string matcher* que gera candidatos com base na semelhança lexical entre entidades e candidatos, a abordagem PECOS-EL baseada em aprendizagem profunda e um modelo baseado em grafos e no algoritmo Personalized PageRank que consegue mapear entidades mais complexas que necessitem de informação contextual. O *pipeline* resultante é capaz de mapear de forma precisa entidades biomédicas associadas com compostos químicos e doenças aos sistemas de organização de conhecimento MEDIC e CTD-Chemical sem necessitar de dados anotados por humanos no processo de treino. O seu desempenho foi avaliado em diversos conjuntos de dados biomédicos, atingindo um desempenho estado da arte em três deles: BC5CDR-Disease, NCBI-Disease, BioRED-Chemical.

Para além das abordagens propostas, o presente trabalho também inclui uma revisão sistemática da tarefa de mapeamento de entidades e do panorama dos sistemas de organização do conhecimento nas áreas biomédica e clínica. A revisão permite uma melhor caracterização do estado da arte, da evolução da tarefa ao longo do tempo, bem como das limitações das abordagens atuais, contribuindo, deste modo para atingir os objetivos 1 e 2.

As outras contribuições descritas no presente trabalho são diversas e possibilitaram uma melhor compreensão das limitações, aplicações e estratégias de melhoramento do mapeamento de entidades. Estas

contribuem em menor grau para os objetivos 1 e 2. Incluem participações em competições e desafios para testar as abordagens desenvolvidas (NASA LitCoin NLP Competition, ChEMU, CANTEMIST, MESINESP2) e outras abordagens relacionadas, como reconhecimento de entidades, classificação de texto, recomendação de entidades biomédicas. Incluem também o desenvolvimento de uma ferramenta de anotação para texto biomédico designada por BENT. A ferramenta foi desenvolvida em Python e a sua prioridade é a usabilidade, de modo a que utilizadores não familiarizados com a área de prospeção de texto possam anotar os seus textos biomédicos.

Palavras Chave: Mapeamento de entidades biomédicas, Prospecção de texto, Processamento de linguagem natural, Sistemas de organização do conhecimento, Aprendizagem profunda

Contents

List of Figures

List of Tables

List of Abbreviations

ANN artificial neural network

BERT Bidirectional Encoders Representations from Transformers

ChEBI Chemical Entities of Biological Interest

CNN Convolutional neural network

CTD Comparative Toxicogenomics Database

DL Deep Learning

EL Entity Linking

GO Gene Ontology

GO-BP Gene Ontology - Biological process

GO-CC Gene Ontology - Cellular component

GO-MF Gene Ontology - Molecular function

GPUs Graphics processing units

HP Human Phenotype Ontology

IC Information Content

ICD International Classification of Diseases

KOS Knowledge Organization System

LLM Large Language Model

LSTM Long short-term memory neural network

MedDRA Medical Dictionary for Regulatory Activities

NCBI National Center for Biotechnology Information

NER Named Entity Recognition

NLP Natural Language Processing

OBO Open Biological and Biomedical Ontologies

OMIM Online Mendelian Inheritance in Man

OWL Web Ontology Language

PPR Personalized Page Rank

QA Question Answering

RDF Resource Description Framework

RE Relation Extraction

REEL Relation Extraction for Entity Linking

RNN Recurrent neural network

SNOMED-CT Systematized Nomenclature of Medicine - Clinical Terms

SOTA State-of-the-art

SSM Semantic Similarity Measure

TM Text Mining

UMLS Unified Medical Language System

W3C World Wide Web Consortium

XMR Extreme Multilabel Ranking

Chapter 1

Introduction

Scientific and technological knowledge forms the foundation of modern society, and a considerable portion is available in text format. Scientific literature, in particular, is essential for advancing scientific research. For example, when new researchers begin to work in a new field of study, they must understand past findings to develop new scientific hypotheses. Even for an experienced researcher, staying up to date with the latest discoveries published daily is essential. Additionally, scientific literature aids in interpreting research results by providing context and insights from similar studies.

However, the amount of knowledge in this format continues to proliferate, challenging its access. The number of scientific articles in repositories has been steadily increasing: only during 2023, 1,567,478 new citations were added to the *PubMed* collection[1], and since the beginning of 2024 until the end of August there were 158,079 new submissions to *arXiv*[2]. This increase is observed not only in the scientific literature but also in other text formats, such as patents, with 56,900 patents registered with the *World Intellectual Property Organization* during 2022, and clinical trials, where the number of records in the *International Clinical Trials Registry Platform* increased from 2,408 in 1999 to 744,100 in 2022[3]. Accordingly, Bornmann et al. [26] estimated that the annual overall growth rate of publication is 4.10%, leading to a doubling time of 17.3 years.

For instance, the surge in scientific publication during the COVID-19 pandemic underscores the im-

[1] `https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html`

[2] `https://arxiv.org/stats/monthly_submissions`

[3] `https://www.who.int/observatories/global-observatory-on-health-research-and-development/monitoring/number-of-trial-registrations-by-year-location-disease-and-phase-of-development/#data-sources`

portance of organizing scientific knowledge and making it readily accessible to researchers, clinicians, public health officials, and the general public.

A Knowledge Organization System (KOS), such as ontologies, databases, terminologies, knowledge bases and graphs, is essential for efficient data management since it can integrate and organize large amounts of data, information and knowledge, including those extracted from scientific literature [101]. One of the main goals of creating and maintaining KOSs is enhancing data standardization and sharing and reuse across a given community. Besides, KOSs facilitate automated processes since its content is readable by humans and automated systems, which fosters knowledge discovery through the identification of new patterns or insights. The MEDIC vocabulary [55] is an example of a biomedical KOS that focuses on disease, including a hierarchy of disease concepts and their relations. For example, its structure represents the entry "Pneumonia" (identifier D011014) as a child concept of the entry "Respiratory Tract Infections" (identifier D012141), which by its turn, is a child concept of "Infections" (identifier D007239)[4]. This information is structured in a format that can be injected into a Deep Learning (DL)-based approach for reasoning, for example.

KOSs typically include data curated by human experts, who manually analyze primary text sources and then translate relevant content into the logical schema of the target KOS. This curation process is slow and requires deep expertise in the specific domain. Text Mining (TM) approaches assist in the curation process by extracting relevant patterns and insights in extensive collections of text using automated methods [100, 261, 200, 92, 142, 10, 37].

Entity Linking (EL) is an essential component of the TM pipeline, as it bridges the gap between natural language, which is written and understood by humans, and KOSs, structured repositories whose content is accessible to computers: EL methods link identified entities in the text to corresponding entries in a target KOS [198]. Multiple downstream applications are depending on the output of EL approaches:

- information retrieval approaches, such as search engines [164, 21, 95] or Question Answering (QA) [224, 143] methods, where EL can link entities identified in user queries to provide more context which allows the return of more accurate results.

- downstream TM pipeline components, in which the output of EL approaches is fed into Relation

[4] As of September 13th, 2024.

Extraction (RE) [215, 5] or event extraction [144] approaches.

- applications for organizing, managing and mining scientific literature [201, 259, 260], which can ultimately lead to the discovery of new scientific knowledge and the generation of new hypotheses [280].

- clinical-focused applications, through the extraction and linking of relevant information, such as diseases, symptoms, treatments, and medication, among others, from electronic health records, clinical notes, prescriptions, and laboratory reports, to improve clinical decision-making, patient care and healthcare analytics [256, 257, 239, 227, 255, 159]. Also another application is the improvement of clinical trial matching systems, in which the goal is to identify eligible participants for a given clinical trial [120, 235].

- public health monitoring systems by using social media text as source [141, 167, 180], pharmacovigillance through the extraction of adverse drug reactions from clinical narratives [44] or from social media text [69, 56], and also drug repurposing (i.e. find new applications for existing drugs) through the building of knowledge graphs that organize information [109].

The rapid production of new scientific knowledge poses challenges to the performance of EL approaches and the coverage of available KOSs, especially in the biomedical field. The curation process relies heavily on human effort and needs to catch up with the fast-evolving literature. Consequently, this lack of coverage or incompleteness in KOSs leads to unlinkable entities, i.e., entities recognized in a text that cannot be linked to any entry in a given target KOS because no suitable representation exists. For instance, when the COVID-19 pandemic began and the term "COVID-19" was coined, the concept was initially absent from biomedical and clinical terminologies. However, preprints containing the term and describing the disease were already being published. In this case, it would be preferable to link the term to similar entries already represented in typical biomedical KOSs, such as "coronavirus" or "pneumonia", even if these terms only partially represented its meaning: if a term remains unlinked and locked in the respective source text it will not be accessible by automated approaches requiring more structured data as input. Semantic information will remain buried in the text. Dealing with NIL entities like this poses a significant challenge to the performance of EL approaches, as it undermines the core objective of

these approaches: adding a structured semantic layer to text written in natural language [65]. Existing approaches dealing with this problem focus on predicting whether an entity is NIL [32] or clustering the identified NIL entities in a document or a corpus [22], without attempting to connect them to a target KOS.

The lack of coverage in biomedical KOSs also manifests through the absence of relevant semantic relations between the existing entries. For example, the current version of the MEDIC vocabulary[5] represents the entry "COVID-19" (identifier D000086382) and its relation with the entry "Pneumonia, Viral" (identifier D011024). However, this KOS does not include the relations between this disease and the symptoms "loss of taste" and "loss of smell", although they appear in scientific literature [81]. If one only consulted the information stored in this KOS, one would not be aware of the connection between these conditions and "COVID-19", meaning that the KOS is outdated and lacks coverage. The MEDIC vocabulary also includes entries for "Ageusia"/loss of taste (identifier D000370) and for "Anosmia"/loss of smell (identifier D000086582), but they are not connected to "COVID-19". Even if certain relations are not explicitly described in a given KOS, the information may still be essential to provide context in a scenario involving linking an entity to a target KOS. EL approaches often require contextual information to link entities more accurately. However, if the target KOS is incomplete, the context will also be insufficient, leading to worse linking decisions and a consequent drop in performance.

On the other hand, the slow pace of the curation process also results in a persistent lack of human-annotated datasets for training and validating EL approaches, especially those based on DL. In particular, the datasets focusing on the biomedical domain have low annotation diversity and volume, as annotating large-scale datasets by human experts is costly and requires specific biomedical expertise [229, 217]. Besides, the evolution of the text sources originates new entities that are not represented in the existing curated datasets nor in the domain KOSs, as mentioned above. The performance of DL-based approaches is related to the information available during the training process, so if the datasets are limited, the performance will be limited as well. It is necessary to focus on approaches that are not impacted by the limitations of human-labelled data, such as distant supervision [186, 70, 133] or zero-shot methods [150, 277].

Without effective EL approaches, the curation process is slower and inefective, consequently, KOSs

[5] As of September 13th, 2024.

will not be maintained, hindering the organization and sharing of scientific data and information. Additionally, applications depending on the accuracy of the output provided by EL approaches, including the aformentioned information retrieval and TM methods, applications focusing on mining biomedical literature and systems related to clinical-decision and healthcare activities will be hindered.

By adding a structured semantic layer to text, EL approaches can ultimately enhance machines' ability to interact with human language and extract knowledge from it. Thus, to improve the performance of EL and related approaches, it is necessary to overcome the two main challenges arising from the limitations of human-annotated data: lack of coverage in KOSs and reliance on scarce, limited training datasets.

1.1 Objectives

The main challenges identified in the previous section hindering the development of the EL task are centered around the limitations of human-annotated data at both KOS and dataset levels. Thus, the central hypothesis of the present work is to overcome these limitations to improve the performance of EL approaches more concretely by addressing the problem of NIL entities and by reducing the need for human-annotated datasets through the exploration of automatic data annotation methods. The ultimate goal is to develop EL approaches with State-of-the-art (SOTA) performance in the biomedical domain. The research work included two specific objectives:

1. To tackle the lack of coverage in biomedical KOSs by:
    - using RE to bridge the information gaps stored in KOSs and developing a DL-based approach to handle NIL entities.
2. To reduce the reliance on human-annotated datasets by:
    - focusing on EL approaches that leverage DL architectures trained on automatically labelled data and on integrating various EL methods.

1.2 Methodology

The research work includes a diverse set of approaches to address each of the described objectives, including different methods for the EL task, such as graphs and DL models. Below is a detailed overview

of the methodology used to achieve the objectives.

1.2.1 Tackling the lack of coverage in biomedical KOSs

This work focused on graph-based and DL-based approaches to tackle the lack of coverage of biomedical KOSs.

1.2.1.1 Improving disambiguation graphs with RE

EL approaches based on collective inference usually resort to disambiguation graphs. These include the entity mentions to link in a given document and the respective candidates extracted from the target KOS as nodes. The edges between nodes are based on the information stored in the KOS, such as child-parent relations or more distant ones. A metric of coherence, such as the Personalized Page Rank (PPR) algorithm, is applied to the graph, ranking each node according to its relevance: the reasoning is that the nodes better connected to the rest of the graph have higher coherency compared to the nodes worse connected, so they are more relevant to the respective entity mentions [191, 131]. However, if the target KOS lacks coverage at the level of relations, the disambiguation graph will be incomplete, missing potential relevant edges and hindering node ranking. The central hypothesis here is that it is possible to overcome the lack of coverage of biomedical KOSs by directly extracting from scientific literature relevant relations that can complete the disambiguation graphs. Adding the extracted relations originates denser graphs containing more semantic information. Thus, the candidate ranking process is more accurate, and the performance of the EL approach increases. The main novelty here is using RE approaches to improve EL approaches when the opposite happens in the context of TM pipelines. Figure 1.1 showcases an example of this approach applied to candidates in the Chemical Entities of Biological Interest (ChEBI) ontology.

The entity mention “sodium deoxycholate” has only one homonymous ChEBI candidate. The other two entity mentions have five candidates each. Building a disambiguation graph with these candidates originates a graph with no edges between the nodes since there is no relation described in ChEBI between these entries. However, there is a relation described in the literature between the candidates “Cl-” (identifier 17996) and “trisodium vanadate” (identifier 35607), which can represent an edge between the respective nodes in the graph. When applied, the PPR algorithm will assign higher scores to these nodes. For each entity mention, the respective candidate node with the highest score is selected, and the entity

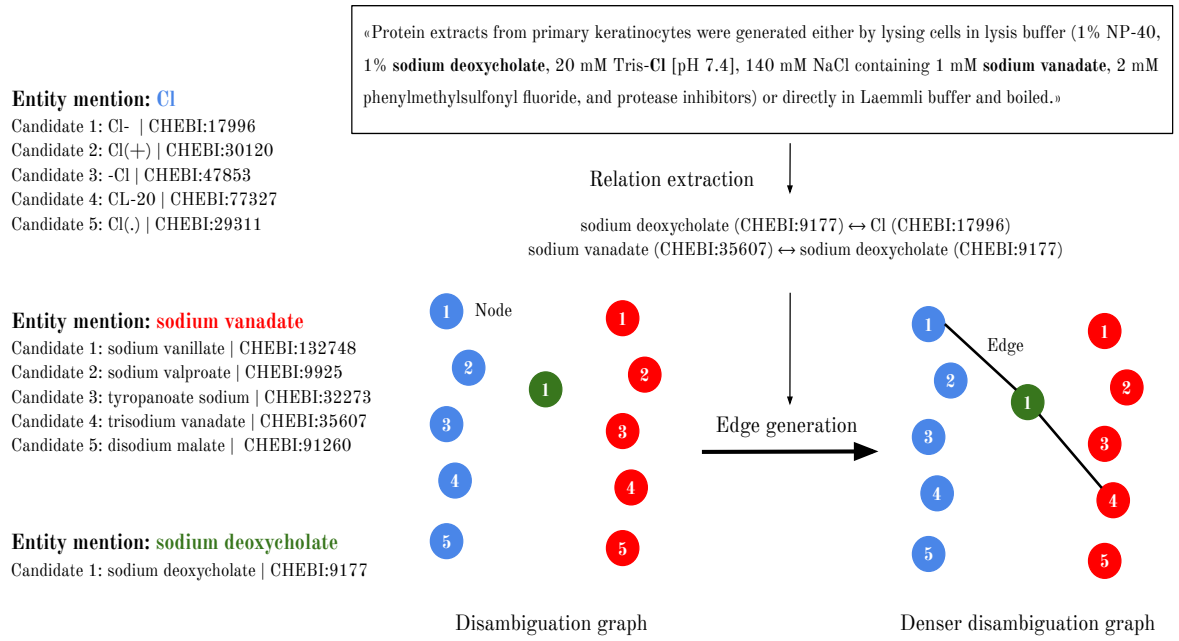


Figure 1.1: Disambiguation graph improved by extracted relations: Example showcasing the building of the disambiguation graph for three different entity mentions in a document of the CRAFT-ChEBI dataset and the further densification of the graph with extracted relations from the dataset. This figure is present in the Chapter 3 of the present document.

mentions “Cl” and ‘sodium vanadate” are respectively linked to “Cl-” (identifier 17996) and “trisodium vanadate” (identifier 35607). This way, RE approaches can overcome the lack of coverage in biomedical KOSs to improve EL performance.

1.2.1.2 Attention mechanism to deal with NIL entities

NIL entities arise from a lack of coverage in target KOS. Existing approaches focusing on these entities usually only predict their existence, cluster, or ignore them. The central hypothesis of this section is that it is possible to link NIL entities to target KOS partially, which brings improvements to EL approaches. This task is designated by NIL entity linking. The main novelty is the development of an approach focusing explicitly on linking NIL entities to target KOS and not just on predicting or clustering them.

To partially link the NIL entities to biomedical KOSs, the methodology involves developing a DL-

based approach inspired by the model proposed by Qi et al. [194] to model the semantic compositionality of multi-word expressions. In linguistics, the principle of semantic compositionality states that the meaning of a syntactically complex expression (e.g. an expression with two words) is determined by the meanings of its syntactic components and the way they are combined. Qi et al. [194] retrieved sememes, indivisible semantic units of meaning in human languages, for each word in a multi-word expression to generate embeddings that accurately capture the meaning of the entire multi-word expression. The resulting representations or embeddings were built according to the sememes of the words composing the expressions. Analogously, the meaning of a NIL entity can be captured by the KOS candidates associated with its constituent words.

The first step is to build a KOS-derived word-concept dictionary in which each word appearing in the name and synonym strings is associated with the respective entry identifiers. For example, the entry "Taste Disorders" in the MEDIC vocabulary with identifier D013651 is converted into two entries in the mentioned dictionary: {"taste": D013651, "disorders": D013651}. The first module of the proposed approach retrieves KOS candidates for the words of a given input NIL entity using the generated word-concept dictionary. BioWordvec [278] and node2vec [84] embeddings represent the retrieved words and the entire set of entries in the target KOS, respectively.

For example, consider the hypothetical NIL entity "losing taste" and assume that this entity is not present in the MEDIC vocabulary. First, the NIL entity undergoes steps of tokenization, lemmatization and normalization, resulting in the words "loss" and "taste": the first word, "loss", is associated with the candidates "Alveolar Bone Loss" (D016301), "Surgical Blood Loss" (D016063) and "Embryo Loss" (D020964), whereas the second word "taste" is associated with the candidate "Taste Disorders" (D013651).

The second module is a neural network that uses the attention mechanism to determine the relevance of each of the previously retrieved candidates to the entire input NIL entity. The problem of NIL entity linking is framed as a multi-classification problem. The attention model [76] includes keys or vectors corresponding to the KOS candidates' embeddings and the words that constitute the NIL entity, energy scores, the compatibility function, and the distribution function. The attention model assigns scores to each of the KOS candidates. The scores for the candidates for the second word are influenced by the embeddings of the first word and vice versa.

Considering the same example as before, this neural network would assign a higher attention score to the candidate "Taste Disorders" (D013651). Thus, the approach would partially link the NIL entity "Loss of taste" to "Taste disorders" (D013651).

As the existing biomedical datasets lacks annotations of NIL entities partially linked to target KOS, the methodology includes the generation of silver standard dataset using existing data in biomedical KOSs.

Graph-based approaches are prone to incomplete disambiguation graphs due to a lack of coverage in target KOS, as described in the previous subsection. The existence of NIL entities also creates incomplete disambiguation graphs. So the goal was to explore the combination of this approach by focusing on NIL entity linking with the previous EL approach based on the PPR algorithm and RE to generate candidates for the ignored NIL entities and, in turn, generate more edges between the other nodes in the graph to increase the overall performance in the EL task.

1.2.2 Reducing the reliance on human-annotated datasets for training EL approaches

The focus is on developing an EL approach that relies on DL architecture and automatic methods for training data generation to tackle this objective. Annotating training data for DL models is a slow and costly process, so the methodology involves automatically generating a large-scale training dataset using the data provided by the tool Pubtator3 [258] and the data already stored in biomedical KOSs. This research introduces the novelty of framing the EL problem as an Extreme Multilabel Ranking (XMR) task: the entities to annotate constitute the documents, and the list of entries in the target KOS is the list of pre-defined categories. This way, the PECOS framework [274], initially developed for the XMR problem, can be adapted for the EL task and trained in the large-scale generated dataset. The resulting PECOS model assigns relevant KOS candidates for each entity mention in a given document.

However, in some cases, integrating contextual information leads to better linking decisions. Graph-based EL approaches, where the EL problem is framed as a coherence maximization problem over a graph, bring advantages. Different entities require different EL approaches, so integrating different EL approaches improves performance.

The proposed EL pipeline includes different components: an abbreviation detector, a string matcher (which matches entity mentions to KOSs entries by calculating the respective lexical similarity between them), a DL-based model adapted from the XMR task (PECOS models) and a graph-based model ap-

plying the PPR algorithm that deals with the most complex cases requiring contextual information. The pipeline takes as input a set of entity mentions in a given document and applies an abbreviation detector. Then, it applies the string matcher to generate KOS candidates according to the lexical similarity and, simultaneously, the DL model trained on the large-scale training dataset. A score threshold of 0.1 filters the candidates to add to each candidate list: if the retrieved candidate by the PECOS model and the candidate retrieved by the string matcher have a score of 1.0, both join the candidate list for the respective entity mention; if both have a score lower than 1.0, the rule-based filter proceeds. If the candidate retrieved by the PECOS model has a score higher than the filtering threshold (0.1) it is added to the candidate list; if its score is lower than 0.1, the candidate retrieved by the PECOS model and the candidate retrieved by the string matcher candidate join the candidate list. In this case, the linking process requires contextual information, so the candidate list and the candidates for other entity mentions in the same document are used to build a disambiguation graph. The PPR algorithm is applied to the graph, assigning higher scores to the nodes that are more coherent with the graph. The pipeline then selects the highest-scoring candidate to link the input entity mention.

Consider an example with two disease mentions: "vasculitis" and "vasculitic". The pipeline first applies the abbreviation detector module to both mentions. The PECOS-based model and the string matcher retrieve the same candidate from the MEDIC vocabulary for "vasculitis": "Vasculities" (identifier D014657). The score of the PECOS model is 1.0, and the string matcher returns an exact match, so the entity mentioned is linked to this candidate. For the entity mentioned "vasculitic", the PECOS-based model returns the candidate "Congenital disorder" with a low score of 0.0964, and the string matcher returns the candidate "Vasculitis" with a score of 0.90. As the candidate retrieved by PECOS has a score lower than 0.10, both candidates join to the candidate list of "vasculitic". Then, the pipeline builds a disambiguation graph, including the candidates for both entity mentions in the document, and it applies the PPR algorithm to the graph. In this graph, the candidate "Vasculitis" (D014657) for the entity mention "vasculitic" is the most connected node, so the entity mention is linked to it.

The pipeline links biomedical entities (diseases and chemicals) to the target KOSs, MEDIC and Chemical vocabularies from the Comparative Toxicogenomics Database (CTD), without requiring new human-labelled data.

1.3 Contributions

The main contributions of the current work are the following:

- Systematic review of EL and the landscape of biomedical and clinical KOSs: providing an overview of the evolution of the EL task in the years 2013-2024, focusing on the types of approaches, datasets used, and KOSs explored. The work contributes to the central goal of developing an EL approach with SOTA performance and, specifically, to objectives 1 and 2.
- Relation Extraction for Entity Linking (REEL): EL approach based on graphs and the PPR algorithm, improved by RE. The work contributes to objective 1.
- NILINKER: DL-based approach for NIL entity linking that improves the performance of the REEL approach. The work contributes to objective 1.
- X-LINKER: pipeline combining different modules (abbreviation detector, string matcher, DL-based model and PPR-based module) to link disease and chemical entities to biomedical KOSs. The work contributes to Objective 2.

These contributions are detailed further in the following sections, categorized by type.

1.3.1 Book chapter and review

Chapter 2 integrates the content of a review article about EL approaches and biomedical and clinical KOSs and a book chapter addressing semantic similarity in the context of KOSs, which contribute to objectives 1 and 2:

- **Ruas, P.**[*], Conceição, S. I. R.[*], and Couto, F. M. (2024). **Systematic Review of Named Entity Linking and Knowledge Organization Systems in Biomedical and Clinical Domains**. Submitted to *ACM Computing Surveys* and currently under review. Available in Appendix A.
- Couto, F. M., Lamúrias, A., **Ruas, P.** (2024). **Semantic similarity definition**. In *Reference Module in Life Sciences*. DOI: `https://doi.org/10.1016/B978-0-323-95502-7.00085-3` [48].

[*]Authors contributed equally to this research

1.3.2 Approaches

Chapters 3, 4, and 5 correspond to research articles focusing on the development of approaches focused on the EL task, starting from the more straightforward graph-based approach REEL (objective 1), and going to the DL model based on the attention mechanism NILINKER to link NIL entities (objective 1) and to the X-Linker approach to link disease and chemical entities (objective 2). The combination of REEL with the NILINKER model improved the performance in the EL task, highlighting the relevance of dealing with NIL entities (objective 1). The evaluation of X-Linker in several commonly used datasets in the biomedical domain showed that it achieved higher performance in three of them: increases of 1.66, 0.38 and 2.47 p.p. in top-1 accuracy compared to a SOTA approach. X-Linker combines high performance with flexibility since it can be applied to new datasets or contexts without new annotations (objective 2).

- **Ruas, P.**, Lamurias, A., and Couto, F. M. (2020). **Linking chemical and disease entities to ontologies by integrating PageRank with extracted relations from literature**. In *Journal of Cheminformatics* (Q1 Scimago) , 12, 57. DOI: `10.1186/s13321-020-00461-4` [210]. Code repository publicly available.

- **Ruas, P.** and Couto, F. M. (2022). **NILINKER: Attention-based approach to NIL entity linking**. In *Journal of Biomedical Informatics* (Q1 Scimago), 132, 104137. DOI: `https://doi.org/10.1016/j.jbi.2022.104137` [207]. Code repository publicly available.

- **Ruas, P.**, Gallego, F., Veredas Navarro, F. J., & Couto, F. M. (2024). **X-Linker: Hybrid biomedical entity linking with XR-Transformer and automatically labelled data**. Submitted to *IEEE Transactions on Knowledge and Data Engineering* and currently under review [208]. Available as a preprint in `http://www.arxiv.org/abs/2407.06292`. Code repository publicly available.

1.3.3 Real-world Assessments

Chapter 7 describes participation in workshops, challenges and competitions and the development of a Python package (BENT) for entity annotation of biomedical text. These works provided valuable opportunities to assess the current state of the EL task through testing of the developed approaches in public benchmarks and through exploring adjacent tasks, such as Named Entity Recognition (NER), recommendation or text classification. Besides, these works allowed a better understanding of the limitations,

applications and strategies for improving of the EL approaches. Although to a smaller extent, they have contributed for objectives 1 and 2, and to the central goal of achieving SOTA performance in the biomedical EL task:

- **Ruas, P.** (2021). **Deep Semantic Entity Linking**. In *Advances in Information Retrieval. ECIR 2021. Lecture Notes in Computer Science* (CORE A), Vol. 12657, pages 682–687, Cham. DOI: `https://doi.org/10.1007/978-3-030-72240-1_81` [204]. Article available in Appendix B.

- **Ruas, P.** [*], Sousa, D. F.[*], Neves, A.[*], Cruz, C., & Couto, F. M. (2023). **LASIGE and UNICAGE solution to the NASA LitCoin NLP Competition**. Available as preprint in *arXiv*: `https://arxiv.org/abs/2308.05609` [212]. Article available in Appendix C.

- BENT: Python library for NER and EL in the biomedical domain. Package available in `https://pypi.org/project/bent/`

- **Ruas, P.**, Lamurias, A., & Couto, F. M. (2020). **LasigeBioTM team at CLEF2020 ChEMU evaluation lab: Named entity recognition and event extraction from chemical reactions described in patents using BioBERT NER and RE**. In *The workshop ChEMU: Named Entity Recognition and Event Extraction of Chemical Reactions from Patents (CLEF 2020 Working Notes)*. URL: `https://ceur-ws.org/Vol-2696/paper_175.pdf` [209]. Article available in Appendix D.

- **Ruas, P.**, Neves, A., Andrade, V. D. T., Couto, F. M., & Aragón, M. E. (2020). **LasigeBioTM at CANTEMIST: Named Entity Recognition and Normalization of Tumour Morphology Entities and Clinical Coding of Spanish Health-related Documents**. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with the 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020)*, pages 422–437. URL: `https://ceur-ws.org/Vol-2664/cantemist_paper11.pdf` [211]. Code repository publicly available: `https://github.com/lasigeBioTM/CANTEMIST-Participation`. Article available in Appendix E.

[*]Authors contributed equally to this research

- Barros, M.*, **Ruas, P.***, Sousa, D.*, Bangash, A. H., & Couto, F. M. (2021). **COVID-19 recommender system based on an annotated multilingual corpus**. In *Genomics & Informatics*, 19(3), e24. URL: `http://genominfo.org/journal/view.php?number=667`. DOI: `10.5808/gi.21008` [14]. Code repository publicly available. Article available in Appendix F.

- **Ruas, P.**, Andrade, V. D. T., & Couto, F. M. (2021). **LASIGE-BioTM at MESINESP2: Entity linking with semantic similarity and extreme multi-label classification on Spanish biomedical documents**. In *Proceedings of CLEF 2021*, pages 324–334. URL: `http://ceur-ws.org/Vol-2936/#paper-24` [205]. Code repository publicly available. Article available in Appendix G.

- **Ruas, P.**, Andrade, V., & Couto, F. (2021). **Lasige-BioTM at ProfNER: BiLSTM-CRF and contextual Spanish embeddings for Named Entity Recognition and Tweet Binary Classification**. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 108–111. DOI: `10.18653/v1/2021.smm4h-1.21` [206]. Code repository publicly available. Article available in Appendix H.

- Barros, M., Couto, F. M., Pato, M., & **Ruas, P.** (2021). **Creating recommender systems datasets in scientific fields**. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD '21)*, pages 4029–4030. DOI: `https://doi.org/10.1145/3447548.3470805` [13]. Code repository publicly available: Code repository publicly available. Article available in Appendix I.

1.4 Document Structure

In addition to the present introductory chapter, this document is structured into seven chapters as follows:

- **Chapter 2** (Biomedical Entity Linking) provides an overview of the key concepts to understand biomedical EL according to the main objectives established previously. It is associated with the contributions described in "1.3.1 Book chapter and review".

- **Chapter 3** (Linking chemical and disease entities to ontologies) presents the REEL system. It is associated with the contribution described in"1.3.2 Approaches".

*Authors contributed equally to this research

- **Chapter 4** (NILINKER: attention-based approach to NIL Entity Linking) showcases the NILINKER system. It is associated with the contribution "1.3.2 Approaches".
- **Chapter 5** (X-Linker: Hybrid Biomedical Entity Linking with XR-Transformer and automatically labelled data) addresses the X-Linker approach. It is associated with the contribution described in "1.3.2 Approaches".
- **Chapter 6** (Real-World Assessments) compiles all other research work conducted throughout this thesis by dividing each contribution into a section summarising its motivation and the work developed. It is associated with the contributions described in "1.3.3 Real-world assessments".
- **Chapter 7** (General Discussion and Conclusions) closes the thesis by presenting a general discussion, the main conclusions, and a discussion about future work.

Chapter 2

Biomedical entity linking

This chapter provides an overview of the key concepts to understand the work described in the present thesis. It is partially based on the following review article and, to a smaller extent, on the book chapter:

- **Ruas, P.**[*], Conceição, S. I. R.[*], and Couto, F. M. (2024). **Systematic Review of Named Entity Linking and Knowledge Organization Systems in Biomedical and Clinical Domains**. Submitted to *ACM Computing Surveys*. Available in Appendix A.
- Couto, F. M., Lamúrias, A., **Ruas, P.** (2024). **Semantic similarity definition**. In *Reference Module in Life Sciences*. DOI: `https://doi.org/10.1016/B978-0-323-95502-7.00085-3` [48].

2.1 Text mining and natural language processing

TM is the process of automatically extracting relevant knowledge, patterns or insights from free text [240]. The field of TM includes several tasks, which can be executed independently or sequentially, in a pipeline manner, according to the guiding objectives. Fleuren and Alkema [74] point out the following main applications of biomedical TM: information retrieval, named entity recognition and linking, relation extraction, and knowledge discovery. Far from an exhaustive list, the TM pipeline can include the following steps or tasks:

- **NER**: recognition of named entities in a text expressed in natural language, and their classification according with pre-defined categories.

[*]Authors contributed equally to this research

- **EL**: linking of a recognized entity in text to an entry of a given target KOS, e.g., ontology, terminology, vocabulary, knowledge base or graph, that accurately represents its semantics.
- **RE**: detection of semantic relations between two given entities in a text and classification of the relation according to pre-defined categories.
- **Text classification**: assignment of a given text to pre-defined categories according to its content or topics.
- **QA**: automatic answering to questions that are expressed in natural language.
- **Sentiment analysis**: identification of subjective information, for example, the opinion towards a product, in textual sources.
- **Document summarization**: generation of a shorter version of a given text, while maintaining the original content as possible.

Natural Language Processing (NLP) and TM are closely related fields. NLP arose from the intersection of artificial intelligence and linguistics, with a focus on understanding, interpreting and generating human language [177]. TM applications can resort to NLP techniques in order to achieve the goal of deriving knowledge from text expressed in natural language.

The present work focuses on the task of EL, an essential component of the TM pipeline.

2.2 Entity linking definition

The processing of a given text piece begins by performing NER, with the identification of each entity mention e within it. The collection of identified entities within the text or the corpus is denoted as E, and these identified entities are then categorised accordingly.

The goal of EL is to associate each $m \in M$ with the respective identifier of a target repository or KOS that accurately represents its meaning.

Given an input document I containing n entity mentions $M = \{m_1, m_2, ..., m_n\}$, and a target KOS K containing l entries $E = \{e_1, e_2, ..., e_l\}$, the goal is to assign each mention M_i to its corresponding entry e_j from the target K KOS:

$$EntityLinking(M, E) = \{(m_1, e_{j_1}), (m_2, e_{j_2}), ...(m_n, e_{j_n})\}$$

Where e_{j_i} represents the entry linked to mention m_i.

NER and EL are usually modelled as distinct problems, however, approaches that perform end-to-end entity extraction have also been proposed [49, 80, 163, 127, 29].

The typical EL approach includes two components: candidate generation and candidate disambiguation.

The candidate generation stage builds a list of candidate entities C_{m_i} from the target KOS K for each entity mention $m_i \in M$ (collection of recognized entities) through the function:

$$C_{m_i} = \text{GenerateCandidates}(m_i, E) = \{c_1, c_2, \ldots, c_n\}$$

This influences the subsequent stages of the EL approach.

The candidate disambiguation stage ranks and selects (or disambiguates) the highest-scoring candidate for each entity mention e through the function:

$$\text{Disambiguate}(m_i, C_{m_i}) = \{\text{argmax}_{c \in C}\, \text{score}(m_i, c_i)\}$$

2.3 Applications of entity linking

Applications of EL include the improvement of several tasks and systems, such as automated biocuration [196], automatic KOS population [65], question answering [224], recommender systems and digital assistants [236, 176, 105, 115], search engines and semantic search [164, 15, 96], automatic speech recognition [17], background linking of news [106], social media text [165, 117], web pages [268], clinical reports for enhancement of the reading experience [98], aid in the clinical decision process [246], identification of adverse drug-drug interactions [246], biosurveillance [39], among others.

2.4 Challenges in entity linking

The main challenges of the EL task are [198]:

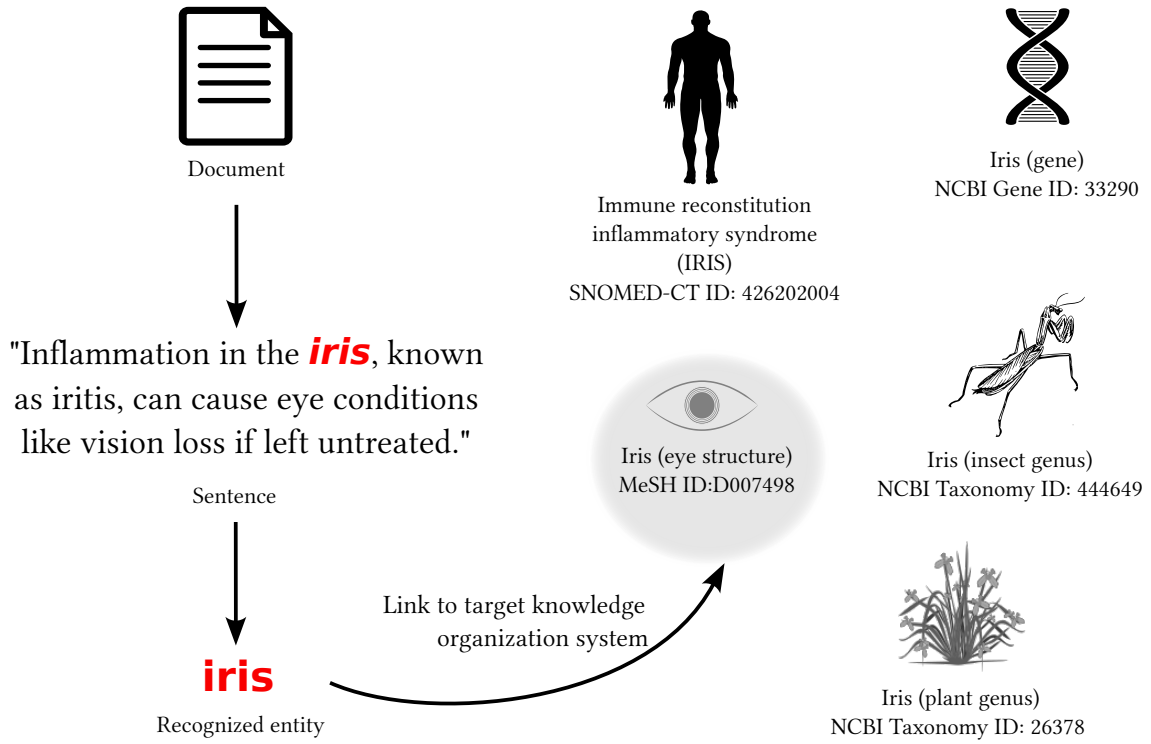


Figure 2.1: Example showcasing the linking of the entity mention "iris" to an entry in the target KOS Medical Subject Headings (MeSH).

- Entity name variations: the human gene "tumour necrosis factor" can be represented by different designations, such as "DIF", "TNFA", "TNFSF2", "TNLG1F", or 'TNF-alpha".
- Ambiguity: for example, the word "iris" can either refer to an anatomic part (structure in the eye) or a genus of plants (see Figure 2.1).
- Insufficient coverage of the target KOSs: it is necessary to update the existing KOSs with information that is newly published, however, manual curation is slow and costly.
- Scarcity of resources for non-English terminologies or highly specific domains.

Other challenges that are not exclusively related to the EL task but that also affect the performance in the task:

- Lack of training datasets, particularly of human-annotated biomedical datasets [282], which hinders the development of supervised approaches.
- Heavy computational cost associated with the training and the fine-tuning of language models with millions of parameters [218], which are currently the basis of SOTA approaches for several NLP tasks.

To overcome these challenges, a diverse set of approaches have been proposed over the years.

2.5 Categorization of entity linking approaches

To categorize the diverse types of EL approaches, the terminology proposed by Ferré and Langlais [73] was expanded. Note that a given approach can be categorized according to more than one category. This terminology is represented in Figure 2.2 and is described below:

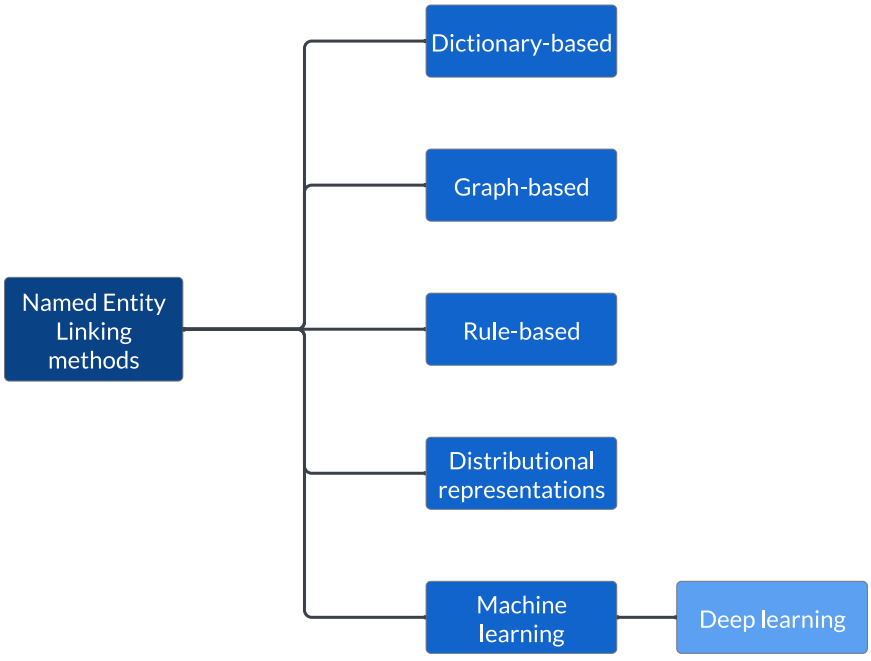


Figure 2.2: Entity Linking approaches.

- **Dictionary-based**: approaches based on matching the surface form of the mention with the surface forms of the candidates. The matching is based on lexical comparison.

- **Graph-based**: approaches involving the building of graphs, usually including every mention in a given document and the respective candidates. A measure of coherence or similarity is usually calculated between mentions/candidates to pick the best candidate.
- **Rule-based**: approaches based on symbolic rules crafted by human experts (e.g. "if-then" rules).
- **Distributional representations**: mentions and concepts are represented by embeddings instead of their surface forms. The similarity between mention and candidate embeddings can be used directly to pick the best candidate or the embeddings can be included in more complex approaches, such as a machine or DL models.
- **Machine learning**: approaches that utilize algorithms to learn patterns and relationships in data to accurately link named entities mentioned in text to their corresponding entries in a knowledge base, without requiring explicit programming instructions.
  - **DL**: approaches that leverage deep learning techniques, specifically artificial neural networks with multiple hidden layers, to model complex relationships and contextual information, enabling the identification and linking of named entities in text to their corresponding knowledge base entries with high precision.

Rule-based models and dictionary-based mainly use techniques like string or semantic matching algorithms based on similarity measure [220, 213, 6, 69, 138, 267, 44, 230], term frequency-inverse document frequency [163, 135], mapping systems [33, 60, 4, 202, 121, 116, 139, 125, 257, 255], distances such as Levenshtein [124, 154, 147] or probabilities [25, 234].

Although these approaches allowed more control due to their transparency they convey many limitations. The limitations that are most encountered in these systems include mismatches due to highly similar concepts, wrong string matching due to concept boundaries in the NER phase and the inability to handle synonyms, co-references and syntax-level processing.

For distributional representations, techniques such as vector space [242], k-nearest neighbours [286], semi-Markov models [136] and clustering [8] are employed. Limitations in this category include the confusion between entity types for static, non-contextualized embeddings, for example, confusing genes with chemicals or diseases with general biomedical vocabulary.

Graph-based approaches use techniques such as the PPR algorithm [131, 190] and node importance

and inter-node coherence [244]. Reported limitations consist of problems with a lack of edges between candidates and errors dealing with parent-child concepts.

Earlier ML approaches consisted in Naive Bayes [82], Recurrent neural network (RNN) [262, 91], Convolutional neural network (CNN) [91, 57, 259], Long short-term memory neural network (LSTM) [167, 158, 249] and n-gram models [189]. ML approaches also explored unsupervised techniques [282, 279, 122].

The rapid growth of ML became more apparent following the introduction of the Transformer architecture [252] and the development of models like Bidirectional Encoders Representations from Transformers (BERT) [59]. Research increasingly focused on using pre-trained models such as BioBERT [140], SciBERT [16], PubMedBERT [87], and ELMo [192]. These models are initially trained on large-scale text datasets, which can include web pages, Wikipedia pages, scientific articles, or other formats, in an unsupervised manner. This means that only unlabelled data is required. This pre-training step generates a language model that has learned the statistical relationships present in the training corpus. The pre-trained model can then be fine-tuned, i.e., further trained on a specific downstream task for which labelled data is available. Kalyan et al. [118] provides a comprehensive overview of biomedical transformer-based pre-trained language models.

The majority of these approaches are based on learning features to rank the candidates' concepts to improve entity disambiguation. Overall, these approaches adopt a more straightforward approach through which the data is fed into the model to automatically learn the natural features of a dataset, without any further input or manipulation [228, 283, 263, 237, 42, 251, 119, 19, 250, 113, 231, 245, 36, 180, 2, 277, 175, 61, 247, 232, 130, 153, 94, 52]. Other approaches attempt to learn representations by integrating certain KOS concepts or semantic types [266, 72, 222, 276], to generate deep contextualized embeddings [51, 62, 254, 50, 20, 34, 271, 149], to include graphs embeddings to better incorporate complex representations [253, 156] and, additionally, other approaches focus on low resource and zero-shot problems [181, 173]. A great deal of issues in the ML category are related to ambiguous annotations including entity overlap, hypernyms, hyponyms, failure to process abbreviations and difficulty to deal with complex phrases or expressions outside KOSs.

While currently pre-trained models are commonly used, there is still some research regarding more conventional ML techniques like LSTMs with Word2Vec and its similar such as BioWordVec, or CNNs

[185, 23, 270, 128, 223, 207, 129, 88]. A few models use a combination of approaches [40, 238, 235, 1].

Another common categorization is based on the features that inform the disambiguation decision. In this sense, there exist **local**, **global**, and **hybrid** approaches [199].

The first EL models explored the similarity between a given entity mention and its candidate concepts from the target KOS, in particular, at the lexical and morphological level. This means that each entity was linked (or disambiguated) independently of the remaining entities present in the same text or document (local models). Bunescu and Pasca [32] developed an end-to-end system including the detection of named entities (NER step) and their linking to the respective Wikipedia articles (EL step) that was based on local features, such as the lexical similarity between entities and Wikipedia titles, and on Wikipedia-specific features, such as disambiguation and redirect pages, categories, and hyperlinks. Mihalcea and Csomai [168] proposed the end-to-end system *Wikify!*, which detected important concepts in the input document and then linked the concepts to the corresponding Wikipedia article. The first step of the proposed pipeline included an unsupervised keyword extraction algorithm performing candidate extraction by comparing all possible n-grams from a parsed document with the n-grams present in the target KOS. The second step consisted of keyword ranking, where each keyword was classified according to its likelihood of being relevant, which is based on three different methods: TF-IDF (term frequency-inverse document frequency), χ^2 independence test, and *Keyphraseness*. The third step is the disambiguation, where each relevant keyword is associated with the respective Wikipedia article. The disambiguation algorithm is based on a local similarity metric between the keyword surrounding context and the Wikipedia definitions of the candidate articles, and on a Naive Bayes classifier that includes local and topic features.

At this point, local models leveraged mostly specific Wikipedia features, which were absent from other KOSs with different architectures, so these models were not transferable to different domains. Besides, the ambiguity associated with named entities means that two given named entities can present the same lexical forms but represent very different meanings according to the context where they appear. Consequently, it is necessary to consider the global context of the entities, i.e., the paragraph or document where the entity appears, the other entities present, etc. To overcome the limitations of local models, the trend changed towards the development of global models, in which the disambiguation of the entities is performed simultaneously across a given text. The goal of global models is thus to find a coherent set of disambiguations across a given document [199].

One common type of global model consist of graph-based approaches, where the KOS candidates for the entities present in a given document are the nodes of a disambiguation graph and the connections between the nodes are based on several criteria [3, 191, 287, 90]. Usually, the edges between nodes are based on the relations between KOS candidates in the respective KOS where they belong. For example, if concept A and concept B belong to the same KOS, and concept A is a child concept of B, in the disambiguation graph the node associated with concept A will be also linked to the node associated with the concept B. The key idea behind graph-based approaches is that the entities that appear in the same text or document share a degree of relatedness. After the disambiguation graph is built, a ranking approach determines the relevance of each node to the graph. For instance, Pershina et al. [191], Guo and Barbosa [90] apply ranking algorithms based on random walks. In general, nodes with higher degrees are more connected to the nodes of the graph, thus are assigned higher scores. The highest scored candidate for each entity disambiguates it.

More recently, ML-based models, in particular those based on DL, have also been explored in the EL task.

2.5.1 Artificial neural networks and DL

Some of the DL networks architectures explored in the EL task include CNNs [145], LSTMs [123], graph convolutional networks [35], but there are also other architectures [78, 195]. These approaches typically leverage global or local features or both. Similarly to the trend observed in other TM tasks, the current SOTA of the task also includes approaches based on pre-trained language models, in particular, BERT [273].

An artificial neural network (ANN) is used for ML in function approximation, classification, and data processing. The goal of an ANN is to learn relationships or patterns between input and output variables. Its structure is loosely inspired by the neurons of the human nervous system. It consists of layers of nodes, the computational equivalent of biological neurons, where each node receives inputs from external sources or other nodes and outputs a non-linear function of the sum of the inputs to other nodes. Each node computes a weighted sum of its inputs, adds a bias term b, and then applies an activation function to determine its output. The bias term helps shift the activation function, allowing the model to fit the data better. The activation function determines whether the node is activated based on its output value. There

are several types of layers in neural networks: **input layers** receive data from external sources, **hidden layers** receive and provide data within the neural network, and **output layers** send data out of the neural network. There are two main types of neural networks:

- **Feed-forward networks**: the data flow from input to output without feedback connections.
- **RNNs**: there are feedback connections since past information can influence the output of a given node. So, they are effective at handling sequential or time-series data.

There are two broad learning paradigms:

- **Supervised learning**: during the training process the ANN has access to input-target pairs, being the goal the discovery of a mapping function between input and output.
- **Unsupervised learning**: no input-output pairs are provided to the ANN during training, the network must find relevant features in the input data and develop its representation for input data.

Hornik et al. [104] argue that standard multi-layer feed-forward networks with only one hidden layer are universal approximators, i.e. can find a mapping function between any given pair of input and target outputs, and only fail if the learning process is not adequate, if there is an inadequate number of hidden units, or if it is stochastic instead of a deterministic relation between input and target.

Each node is a distinct computational unit that can be considered an independent linear regression model. It includes input, data, weights, bias, and output. The input data is a set of variables x_1, x_2, x_3, ..., x_m. A weight w measures the importance of a given input variable for the output of a given node. A connection between two given nodes j and k is defined by a weight w_{jk}, which represents the impact of the output or signal of the node j on unit k. A neuron receives the inputs from the nodes of the previous layer, sums them up along with the bias term, and then applies the activation function to the obtained sum, producing the output. There are several types of activation function, namely linear, binary step, and non-linear. Among the latter, commonly used in neural networks, there are the sigmoid, the hyperbolic tangent and the Rectified Linear Unit (ReLU). Thus, the output of a node can be obtained by an activation function f:

$$output = f\left(\sum_{i=1}^{m} w_i x_i + b\right)$$

where w_i are the weights, x_i are the input variables, and b is the bias term.

In supervised learning, the goal is that the network neurons find the combination of weights that minimise the difference between the predicted output(s) and the true output(s). Important features that positively contribute to error minimisation receive higher weights, important features with a negative impact receive large but negative weights.

The cost or error function C determines the error between the value $\hat{y}$ (or values) predicted by the neural network and the expected value y. It is generally obtained averaging the loss function or error C_x for each training instance:

$$C = \frac{1}{n} \sum_{x} C_x$$

There are different types of loss functions, such as Mean Squared Error, Mean Absolute Error, Binary Cross-Entropy, Hinge, Multi-Class Cross-Entropy. As the error function depends on the network weights, the goal is to find the combination of weights that translates into a global minimum value for the cost function. The weights of a network are usually initialised to random values. In each training iteration, input data flows through the network, and the **backpropagation** algorithm computes the gradient of the loss function associated with a given training instance concerning the weights of the network. **Gradient descent** attempts to find the local minimum of the cost function. The weights are updated, being the speed of the update determined by the **learning rate**. The general schema of a simple feed-forward neural network is represented in Figure 2.3.

The term **DL** is employed to designate neural networks that contain more than one hidden layer. DL methods learn representations for input data by including multiple processing layers that perform non-linear data transformations, which consecutively generates more abstract representations [137].

2.5.2 Attention mechanism

The attention mechanism is used in the context of neural networks to assign weights to different features or regions of the input, which reflects their importance for the task at hand. During the training process, the learning of the weights are used to enhance the ability of the neural network to capture complex patterns. This mechanism can improve the performance of the neural network by giving higher weights to relevant elements of the input, and, conversely, by giving lower weights for elements that decrease the

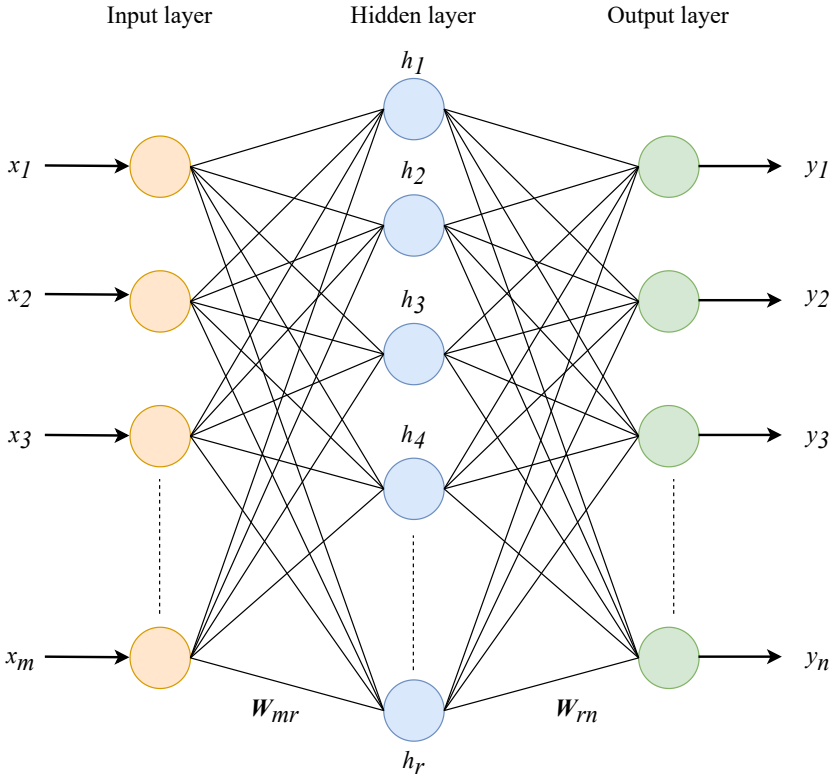


Figure 2.3: Feed-forward neural network. x_1, x_2, x_3, x_m are input variables, h_1, h_2, h_3, h_4 are hidden layers, y_1, y_2, y_3, y_n are output variables, W_{mr} is the weight matrix relative to the output of the input layer that are fed to the hidden layer, and W_{rn} is the weight matrix relative to the output of the hidden layer that are fed to the output layer.

performance. Besides, it can shed light on the behaviour of the neural network, as it helps in the interpretation of the output. Attention has been used in many TM tasks, such as text classification, language modelling, NER, EL, question answering, text summarization, among many others [76].

The core attention model, as defined by Galassi et al. [76], maps the input text sequence x, by representing as a matrix $\boldsymbol{K}$ of d_k vectors k_i, designated by keys, and a query $\boldsymbol{q}$ to a distribution $\boldsymbol{a}$ of d_k attention weights a_i, which highlights the relevance of each key. In the first step, the compatibility function f evaluates the relevance of $\boldsymbol{K}$ according to the query $\boldsymbol{q}$, which is a vector of dimensions n_q in which respect attention is calculated, and returns the vector $\mathbf{e}$ representing the energy scores:

$$\boldsymbol{e} = f(\mathbf{q}, \boldsymbol{K})$$

The distribution function g computes the vector of attention weights **a** based on the energy scores **e**:

$$\boldsymbol{a} = g(\mathbf{e})$$

The output of the core attention model is the vector $\boldsymbol{a}$ containing the attention weights, which are integrated into the remaining computations of the neural network. The schema for this model is represented in Figure 2.4.

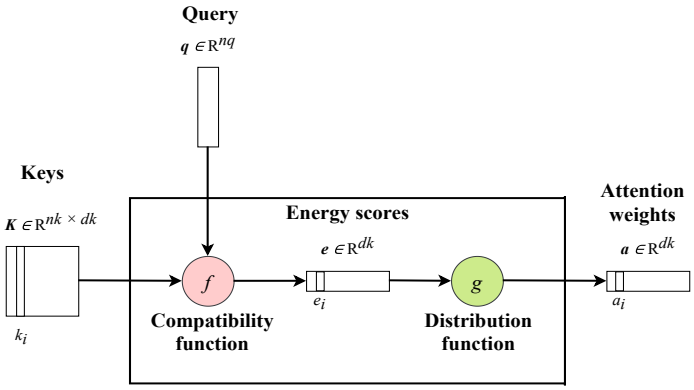


Figure 2.4: Core attention model.

The paper by Vaswani et al. [252] introduced the Transformer architecture, which uses the attention mechanism. The input of the target sequence flows through a stack of six encoders, each one includes a self-attention layer, an encoder-decoder attention layer, and a feed-forward neural network. The output of the encoder feeds the encoder-decoder attention layer. The main advance of this architecture is the use of the multi-head self-attention layer instead of the recurrent and convolutional layers used in RNNs and CNNs, respectively. The multi-head self-attention layer allows for the simultaneous determination of attention weights for each word in a sentence relative to all other words. This parallel computation of multiple attention heads enables efficient use of Graphics processing units (GPUs), significantly reducing computation time. The Transformer architecture is the basis of several pre-trained language models in the TM field that will be further described in the Subsection 2.5.3.

On a different line of research, Qi et al. [194] have explored the attention mechanism jointly with neural networks for modelling semantic compositionality. Semantic compositionality is a linguistic concept asserting that the meaning of a complex syntactic unit, for example, a multi-word expression with

two words, can be derived from the meaning of its constituent parts, in the previous example, the left and the right words. The authors represented each constituent word by the respective sememes. A sememe is the minimum semantic language unit of meaning, for example, the word "husband" is associated with the sememes "family", "human", "male". HowNet is a common-sense knowledge base that includes an extensive list of English and Chinese words associated with their respective sememes [63]. Qi et al. [194] proposed a model to determine the semantic compositionality of two-worded expressions that uses a mutual attention mechanism: the left and right words are associated with sememes, each sememe is assigned an attention weight based on the sememes for the opposite word. The output of the model consists of a vectorized representation (i.e. embeddings) of the expression that reflects its meaning. The obtained representations were then used in downstream tasks, such as multi-word similarity computation and multi-word sememe prediction.

2.5.3 Language models

Similarly to what happened across most TM tasks, a recent approach consists of performing **transfer learning** by leveraging pre-trained **language models**.

A language model corresponds to a probability distribution of words or other linguistic units determined over a given corpus in order to capture grammatical, syntactical and other features associated with natural language. A language model is able to learn the rules and structure of the language during training, so it can be used for next word prediction in a given text sequence. Currently, neural language models are generated by training a neural network through self-supervision over a large corpus of text: these models are trained to minimize the error when predicting the next word in a given piece of text present in the corpus. A Large Language Model (LLM) is generated by increasing the size of these neural networks. LLMs have enhanced abilities in natural language generation and understanding [281].

To build and apply effective supervised DL approaches requires extensive training data, which is not widely available for every task and domain, especially if it is the case of human-labelled data. However, it is possible to partially bypass this by performing transfer learning, i.e., by applying or fine-tuning a model previously trained for a specific task in a different target task [184]. Pre-trained language models are usually trained on large corpora and then fine-tuned to specific TM tasks, such as EL. The most common pre-trained language models have a Transformer-based architecture [252], like BERT [59] and GPT

[162]. BERT-based models are the current SOTA approaches in several TM tasks. As the name of the model BERT suggests, its architecture is a multi-layer bidirectional Transformer encoder. The framework behind BERT includes two steps: the pre-training, in which the model is trained on unlabeled data on two different tasks (masked language model and next sentence prediction), and then fine-tuning. In the first step, the pre-training data consists of BooksCorpus, with 800 million words, and English Wikipedia, with 2,500 million words. In the masked language model pre-training task, the authors masked a fixed percentage of the input tokens at random, and then make the model predict the masked tokens. In the next sentence prediction Task, the authors generated a large amount of sentences pairs, half of the pairs corresponding to true subsequent sentences, and the other half corresponding to sentences that are not subsequent, but random. The pre-training step allowed thus the generation of a large language model that captures the underlying semantics of the natural language, and that can be further fine-tuned to several downstream tasks, such as natural language understanding or question answering. Besides fine-tuning, BERT can also be used for generating contextualised word embeddings.

Since the original BERT model was trained on general corpora, its language understanding is lower in domain-specific text. For instance, the biomedical text presents specific challenges, such as the presence of rare or complex words that are not present in general domain text. This motivated the proposal of BERT-based models that are pre-trained on scientific corpora, such as BioBERT [140] (additionally pre-trained on PubMed articles), ClinicalBERT [7] (additionally pre-trained on clinical notes), or SciBERT [16] (additionally pre-trained on Semantic Scholar articles).

Recently, generative models and prompt-based learning gained popularity. Given its ability to efficiently incorporate context, prompt-based learning in the context of EL helps with the issues of ambiguity and name variations [275, 285, 284].

2.6 Knowledge organization systems

KOSs, originally proposed in the context of the Semantic Web, play an essential role in the EL task.

The idea behind the Semantic Web was to transform the World Wide Web into a format that is simultaneously readable by humans, but especially by machines, facilitating the understanding and interpretation of information online [18]. This involves enriching web content with metadata, annotations, and structured data, allowing machines to not only find and retrieve information but also grasp its context and

relationships. To achieve this, several knowledge representation languages were defined by the World Wide Web Consortium (W3C)[1], like the Resource Description Framework (RDF) and the Web Ontology Language (OWL). These are employed to formally represent metadata, including concepts, relationships, and entity categories, thereby enabling the embedding of semantics into data.

Simple Knowledge Organization System is a data model proposed by the W3C for representing semantic information in the Web, allowing the sharing and cross-linking of multiple KOSs. It states that *"concepts can be identified using URIs, labelled with lexical strings in one or more natural languages, assigned notations (lexical codes), documented with various types of note, linked to other concepts and organized into informal hierarchies and association networks, aggregated into concept schemes, grouped into labelled and/or ordered collections, and mapped to concepts in other schemes."* [2].

The essential function of a KOS is to organize knowledge and information, which includes their management and representation, by *"organiz[ing] documents, document representations, works and concepts"* [101] and retrieval, to serve *"as a bridge between the user's information need and the material in the collection"* [102].

There are plenty typologies of KOS, however, the exact distinction between them is not always clear. The categorization proposed by Hodge [102] describes three main types of KOSs:

- Term lists: focus on the terms, less structured hierarchy. Examples: authority files, glossaries, dictionaries, gazetteers.
- Classifications: include terms and a structured hierarchy, with a focus on subject sets. Examples: subject headings, taxonomies.
- Relationship lists: more complex and structured hierarchy, with a focus on the relations between terms. Examples: thesauri, semantic networks, ontologies.

One of the most extensive taxonomy of KOSs was proposed by Rocha Souza et al. [203]. The main division level is based on structure type:

- Unstructured texts: for example, abstracts.

[1] `https://www.w3.org/`
[2] `https://www.w3.org/TR/skos-reference/`

- Concepts, relationship and layout: examples include mind maps, data models, and entity-relationship models.
- Term and/or concept lists: are simple structures typically characterized by alphabetical displays, usually lacking hierarchical arrangements. Examples: dictionaries, gazetteers, glossaries.
- Concept and relationship structures: includes a variety of structures offering varying degrees of relationship expressiveness. Simpler structures consist of hierarchical arrangements with basic hyponym/hyperonym relationships, more complex systems like thesauri may incorporate meronomy (i.e. part-whole relationships) with ontologies providing the highest level of expressiveness. Examples: taxonomies, thesauri, information retrieval indexes, semantic networks, ontologies, and controlled vocabularies.

The secondary division level is related to the application domain and use cases, which encompass a heterogeneous and wide-ranging set of sixty divisions (some are mentioned in the examples above).

One of the main goals of creating and maintaining KOSs is enhancing data sharing and reuse across a given domain. The approach to knowledge representation is thus heavily shaped by the goals and language of the community that will use it. In the biomedical and clinical domains, many KOSs are categorized and described as **ontologies**. A widely cited definition of ontology in the context of computer and information sciences (i.e. ontology as a knowledge representation artefact and not a philosophy branch) is the one proposed by Gruber: *"[A]n ontology is an explicit specification of a conceptualization."* [86].

Guarino expands on this definition, defining that an ontology is an *engineering artefact* that includes a *"specific vocabulary used to describe a certain reality (...)"* and a *"(...) hierarchy of concepts related by subsumption relationships (...)"*, which can be complemented by *"axioms (...) to express other relationships between concepts and to constrain their intended interpretation"* [89]. A subsumption relationship, also referred to as a hyponym-hypernym, "is-a" or "SubClassOf" relationship, describes a connection between two classes: a specific class (hyponym) and a generic class (hypernym). In this relationship, instances of the specific class are encompassed within the broader generic class. This analogy draws a parallel to a child-parent relationship. Figure 2.5 shows an example of the concept "caffeine" represented in the hierarchy of the ChEBI ontology and the different types of relationships with other concepts.

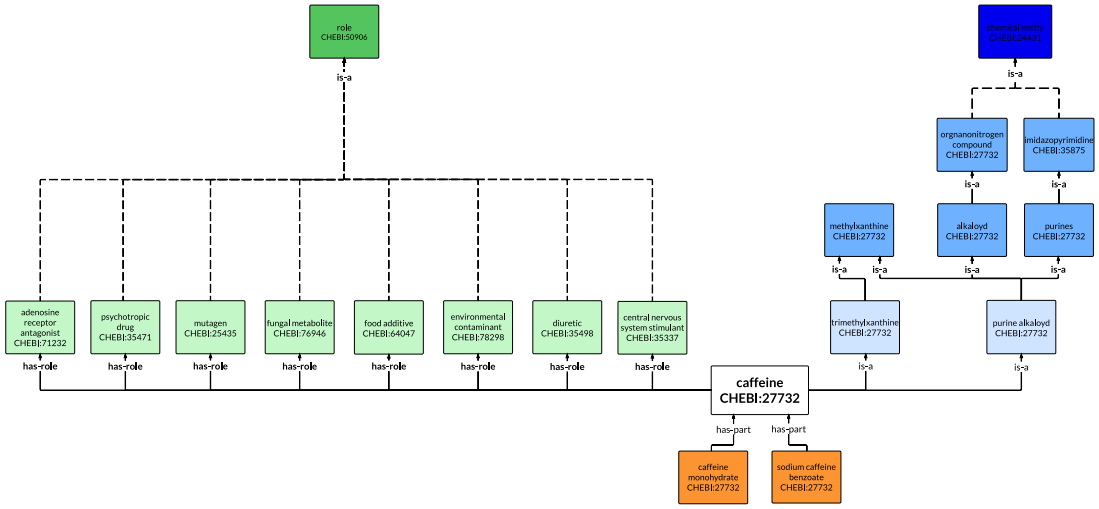


Figure 2.5: Context of the concept "caffeine" in the the ChEBI ontology (some relationships and concepts are not shown), Accessed in 24th June 2024. Each box includes the concept name and the respective ChEBI identifier. There are three types of relations involving "caffeine": "is-a" (blue boxes), "has-part" (orange boxes), "has-role" (green boxes). Solid lines correspond a to direct connection between two concepts (e.g. "caffeine is-a trimethylxanthine", whereas dashed lines correspond to indirect connections, meaning that one or more ontology levels are omitted (e.g. "organitrogen compound is-a chemical entity.", but "chemical entity is distant ancestor of 'organitrogen compound").

In graph theory, there are cyclic graphs, in which there is at least one closed loop or path that begins and ends in the same node, and acyclic graphs, where there are no closed loops. Usually, ontologies are directed acyclic graphs, since the relations or edges have an orientation (e.g. concept A is a child of concept B and not the other way around) and there are no cycles or closed loops, meaning that all the paths start in more specific terms and end in the root term of the ontology, the common ancestor to every term in the ontology.

During the last decade, the designation **knowledge graph** has been gaining traction in both academia and industry. Similarly to the confusion in defining different KOSs, the definition of what exactly consists of a knowledge graph remains contended. Hogan and coworkers define a knowledge graph *"as a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent potentially different relations between these entities*" [103], which, overpassing the terminology *nodes* and *edges*, is a definition indistinguishable from that of an ontology. By its turn, the definition by Noy et al. [182] assert that a knowledge graph is an attempt to model a certain knowledge domain by providing descriptions of domain entities, designated by *nodes* or *vertices*,

and pairwise relations between them, designated by *edges*. A graph can be directed, where an edge between two nodes has direction, or is undirected, in which case the edge is symmetrical.

A common model of knowledge graph consists in the **directed edge-labelled graph** or multi-relational graph: set of nodes representing entities, and the directed edges between the nodes represent the relations between the respective entities [103]. The RDF is a standard data model based on the concept of directed edge-labelled graphs. A directed edge-labelled graph is a tuple $G = (N, E, L)$ where N is a set of nodes, L is a set of edge labels (predicates, properties, or relation types), and $E \in V \times L \times V$ is a set of edges [117]. An edge $(s, p, o) \in E$ is a triplet with format: *subject (s), predicate (p), object (o)*.

Other types of knowledge graphs (non-exhaustive list) are **heterogeneous graphs**, where the graph nodes can have different types, **property graphs**, which have the flexibility to associate nodes and edges with property–value pairs and labels and, **graph datasets**, which include a set of named graphs and a default graph [103].

To amplify the confusion, in the scientific literature focused on the EL task the designation **knowledge base** is highly prevalent. However following the definition of Ehrlinger and Wöß [68], a knowledge base must include a semantic component and a reasoning component able to reason over the "facts" described in the semantic layer. So, one can argue that designating an ontology as a knowledge base is imprecise.

2.6.1 The biomedical and clinical landscape of knowledge representation

The current biomedical and clinical knowledge representation landscape is broadly characterized by two differing paradigms to ensure data integration and interoperability: the Unified Medical Language System (UMLS) [24] and the Open Biological and Biomedical Ontologies (OBO) Foundry [108].

Online resources to access biomedical and clinical KOSs include: OBO Foundry[3], BioPortal[4], OLS[5] and Ontobee[6].

The **UMLS** is a collection of resources, including a very large biomedical metathesaurus that provides a standard to connect a wide-ranging set of vocabularies in the biomedical and clinical domains. The metathesaurus includes 3,300,000 concepts, from 185 source vocabularies, spanning 28 languages[7].

[3] `https://obofoundry.org/`
[4] `https://bioportal.bioontology.org/`
[5] `https://www.ebi.ac.uk/ols/index`
[6] `https://ontobee.org/`
[7] 2023AB release

For some concepts, it provides synonyms and definitions, as well the relations between different concepts. Synonym terms are clustered together to create a concept, which then has relations with other concepts that are either inherited from the respective source vocabulary or are added by the metathesaurus editors. Along with the "Metathesaurus", the UMLS provides two other knowledge sources: "Semantic Network" and the "SPECIALIST Lexicon and Lexical Tools". The "Semantic Network" is a set of broad categories or semantic types and their interrelations. The semantic types categorize every concept present in the UMLS, working in practice as an upper ontology. The latest version (*2023AA*) includes 127 semantic types[8]. The SPECIALIST lexicon is a large syntactic English lexicon that includes both general and biomedical-specific terms[9]. To access the UMLS resources, users must sign a license agreement beforehand and comply with distribution rules. The UMLS includes several vocabularies that are typically used separately in EL. The International Classification of Diseases (ICD)-9, ICD-10, RxNorm, DrugBank, Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT), Medical Dictionary for Regulatory Activities (MedDRA), MeSH, Online Mendelian Inheritance in Man (OMIM), DrugBank, Gene Ontology (GO), Human Phenotype Ontology (HP), National Center for Biotechnology Information (NCBI) Taxonomy, NCI Thesaurus, Orphanet and Radlex are just a few examples.

The **OBO Foundry** includes 258 ontologies, 181 of them active[10], with a particular focus on basic research. The OBO foundry establishes a set of design patterns and common principles that every ontology must adhere to ensure interoperability. There are twenty principles[11], including:

> *P1) Open - The ontology MUST be openly available to be used by all without any constraint other than (a) its origin must be acknowledged and (b) it is not to be altered and subsequently redistributed in altered form under the original name or with the same identifiers.*
> *P2) Common Format - The ontology is made available in a common formal language in an accepted concrete syntax.*
> (...)
> *P10) Commitment To Collaboration - OBO Foundry ontology development, in common with many other standards-oriented scientific activities, should be carried out collaboratively.*

[8] `https://lhncbc.nlm.nih.gov/semanticnetwork/SemanticNetworkArchive.html`, accessed in 7 march 2024
[9] `https://lhncbc.nlm.nih.gov/LSG/Projects/lexicon/current/web/index.html`, accessed in 7 march 2024
[10] As of 5th March 2024
[11] `http://obofoundry.org/principles/fp-000-summary.html`

(...)

Thus, the requirements to access the resources of the OBO foundry are less restrictive compared to the UMLS. However, the ontologies must be available under the OBO format, which is a knowledge representation language extending the OWL to the biological sciences.

One key difference between the UMLS and the OBO Foundry is the focus of the latter on the development and maintenance of interoperable ontologies, whereas the former focuses on including a wide range of existing terminology and classification systems from various sources. The OBO foundry follows a bottom-up approach, i.e., it intends to guide the development of ontologies according to a set of common rules, whereas the UMLS follows a top-down approach, where the goal is to ensure the interoperability of already existing, heterogeneous KOSs.

One must note that the availability of a given KOS in either UMLS or OBO foundry is not mutually exclusive, i.e., several KOSs are accessible within both frameworks (e.g. GO, HPO).

Besides UMLS and OBO foundry, other models exist to ensure data standardization, particularly in the clinical domain. *The Observational Medical Outcomes Partnership* (OMOP) Common Data Model (CDM)[12] attempts to structure and standardize the storing of data from observational studies, and depends on *The Observational Health Data Sciences and Informatics* (OHDSI) vocabularies. It targets data coming from healthcare activities, which is relevant because there is a huge disparity in healthcare systems around the world. It is important to ensure the standardization of the data model and representation in clinical databases so it is possible to generate evidence from it. It provides guidelines for the construction of standardized vocabularies with targets in clinical entities, such as findings, drugs and health measurements. The Clinical Data Interchange Standards Consortium (CDISC)[13] focus on standardizing clinical data collection. It is guided by the CDISC Controlled Terminology (CT)[14] which comprises a set of code lists and respective values that CDISC-defined electronic datasets should follow.

There is no agreement on how to classify the existing semantic resources. For instance, Grabar et al. [83] point out that UMLS has been labelled as metathesaurus or domain-specific terminological system, as an ontology and as both, and the resource MeSH has been designated as a terminology, a thesaurus, an

[12] `https://ohdsi.org/data-standardization/`
[13] `https://www.cdisc.org/standards`
[14] `https://www.cdisc.org/standards/terminology/controlled-terminology`

ontology and a controlled vocabulary, besides the evident subject headings. In many cases, the adopted designation is heavily influenced by the use case and the application domain. For example, an ontology has a different definition whether you are a philosopher or a bioinformatician.

Despite this populated landscape, a major obstacle lies in the conversion of text into structured, useful, shareable data stored within a KOS. A resource is as good as the quality of the data being integrated. Continuously populating the structure with updated and relevant data from the ever-evolving biomedical and clinical fields is crucial to the relevance of KOS. Thus, information extraction pipelines, including EL approaches, assume a pivotal role in connecting text and KOSs [182].

2.6.2 Knowledge organization systems used in the entity linking task

Below is the description of some of the most commonly used KOSs in the EL task in the biomedical and clinical domains.

The MeSH[15] thesaurus is a controlled and hierarchically organised vocabulary that is used to index journals, cataloguing and search for biomedical articles and health-related information [155]. This thesaurus is actively maintained to modify and update the emerging new concepts while deprecating older ones. In general, these updates are divided into daily updates for supplemental records and annual updates for MeSH *Descriptors* and MeSH *Qualifiers*. The main unit of indexing and retrieval is the *Descriptors*, which are the main headings. *Qualifiers* are the subheadings and are combined with *descriptors* to index citations according to a specific aspect beyond the main headings. There are 16 *descriptor* categories such as *A* for anatomic terms, *B* for organisms, *C* for diseases, and *D* for drugs and chemicals, which in turn are subdivided into more categories. Each unique *descriptor* appears at least in one place in the trees and may appear in as many additional locations as necessary.

SNOMED-CT[16] focuses on healthcare-related concepts, to facilitate the organization and storage of clinical data extracted from electronic health records. It was developed by an extensive set of experts, including *"physicians, nurses, physician assistants, pharmacists, informaticians, medical technicians"*. It includes four components: Concept Codes, Descriptions, Relations, and Reference Sets (clustering of concepts into sets and cross-references to other KOSs). The latest version[17] includes 367,584 concepts

[15] `https://www.nlm.nih.gov/mesh/meshhome.html`, accessed in 7th March 2024
[16] `https://www.snomed.org/`, accessed in 7th March 2024
[17] Version: 2024-05-01, international edition

[226]. It is a proprietary resource, maintained and distributed by *SNOMED International*. and available in multiple languages.

MedDRA[18] is an international medical terminology developed by the *International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use* (ICH) [30]. It is focused on pharmacovigilance and clinical research, providing standardization for *"(...) adverse event information associated with the use of biopharmaceuticals and other medical products (e.g., medical devices and vaccines)"*[19].

The NCBI includes a comprehensive repertoire of relevant biomedical and clinical information, including several KOSs. For instance, the Gene database[20], contains information about known and predicted genes from all major taxonomic groups, ranging from viruses to eukaryotes [161, 31]. The gene records provide detailed information such as nomenclature, sequence, structure, function, variations and phenotypes. The Taxonomy database[21] [216], as its name suggests, provides a hierarchy of organism names arranged according to phylogenetic criteria. The highest level is the domain, followed by kingdom, phylum, class, order, family, genus, and species.

The OMIM database[22] contains information about human genes and genetic phenotypes, with an emphasis on the relationship between phenotype and genotype. This database is updated daily and contains information on all known Mendelian disorders.

The CTD[23] provides manually curated information about chemical-gene/protein interactions, chemical-disease and gene-disease relationships. This is a public database that focuses on the progress in the understanding of the influence of environmental exposures on human health [55]. It provides a generalised scope of the entities in diverse human health contexts due to its interconnection with other large vocabularies and KOSs. It provides a collection of data categories, being chemicals, diseases and genes the most commonly used for EL. CTD-Chemical combines a subset of the MeSH thesaurus along with information about their chemical structures, interaction with genes and proteins, disease relationships, enriched pathways and functional annotations. CTD-Disease, best known as MEDIC disease vocabulary, maps OMIM

[18] `https://www.meddra.org/`
[19] `https://www.meddra.org/faq`, accessed in 7th March 2024
[20] `https://www.ncbi.nlm.nih.gov/gene`
[21] `https://www.ncbi.nlm.nih.gov/taxonomy`
[22] `https://www.omim.org/`
[23] `https://ctdbase.org/`

diseases to the MeSH disease category. CTD-Genes, which includes symbols, names and synonyms, is a cross-species vocabulary that arises from the NCBI. It provides information about interactions with chemicals, disease relationships, associated pathways and functional annotations.

The advent of genome sequencing led to the generation of vast amounts of molecular data, unveiling the fact that the fundamental biological functions are shared across all eukaryotic organisms. This was the motivation to build the project GO[24] [12, 46]. Established as a collaborative effort among several model organism databases, including the FlyBase, the Mouse Genome Informatics and the Saccharomyces Genome Database (SGD), the GO Consortium aimed to develop a unified ontology to serve as a structured, universally applicable, and controlled vocabulary suited to describe the functions and roles of genes and gene products across diverse organisms. A gene product is either a ribonucleic acid (RNA) molecule or a protein originating in the expression of a gene. The process of tagging gene products with GO terms is designated by GO annotation.

The GO enables researchers to annotate genes and gene products with consistent descriptors, thereby promoting data integration and analysis. This standardized approach enhances cross-species comparisons and supports a deeper understanding of biological processes at the molecular level.

There are three categories for GO terms, which in practice work as three distinct ontologies:

- Gene Ontology - Biological process (GO-BP): terms that describe a biological process in which a gene or a gene product is involved, it can include one or more molecular functions.
- Gene Ontology - Molecular function (GO-MF): terms related to the biochemical activity of a gene or a gene product, such as “ligand” or “transporter”.
- Gene Ontology - Cellular component (GO-CC): terms associated with cell or extracellular sites (e.g. “nucleus”, “ribosome”, “membrane”).

The GO can be considered a knowledge graph as well: it includes an underlying ontology as the semantic backbone, whose concepts are used to annotate proteins across organisms and domains.

The release available at the time of writing (2024-9-18) of GO release includes 40,939 terms (26,552 biological process terms, 10,365 molecular function terms, 4,022 cellular component terms), 7,894,411

[24] `https://geneontology.org/`

annotations and 1,573,444 annotated gene products from 5,426 species[25]. This ontology is freely available as part of the OBO initiative.

The ChEBI ontology[26] [97] focuses on small chemical compounds, including metabolites, drugs, and other bioactive molecules that are involved in processes in living organisms. Three data sources were the basis for building ChEBI: the Integrated Relational Enzyme database of the *European Bioinformatics Institute* (IntEnz), KEGG COMPOUND and the Chemical Ontology.

As with similar ontologies, ChEBI is a graph-theoretic structure, with terms having the role of nodes, and the relations between terms having the role of edges. But, unlike other ontologies, it is a directed cyclic graph. ChEBI is a structured hierarchical, representation, including three high levels of categorization: *chemical entity*, *role* and *subatomic particle*. Two of these sub-ontologies classify compounds according to structural chemical features, namely *chemical entity* and *subatomic particle*, and the other one, *role*, categorizes compounds according to their biological and chemical role, as their application.

Besides the most common ones *is-a* (e.g. child-parent relation) and *is part of*, there are numerous, chemical types of relation: *is conjugate acid of*, *is conjugate base of*, *is tautomer of*, *is enantiomer of*, *has functional parent*, *has parent hydride*, *is substituent group from*.

Along with the classification of chemical compounds in the ontology, the ChEBI database stores other relevant information, such as different chemical names and synonyms, definitions, chemical information, (IUPAC names, SMILES representation), and also cross-references with other databases and KOSs, such as UniProt and PubChem. The data included in ChEBI is freely available as part of the OBO initiative.

The ICD[27] is a medical classification for healthcare, epidemiological and clinical purposes developed by the *World Health Organization*. The purpose of the ICD was to be the worldwide standard for tracking morbidity and mortality statistics, support coding tools in clinical settings and reimbursement systems, and facilitate automated decision-making in healthcare. The ICD structure includes clinical codes ranging from diseases and diagnostics to symptoms and findings. The latest version, ICD-11, has replaced ICD-10 after 2022, which was active since 1993[28]. ICD-11 includes about 80,000 entries complemented by 40,000 synonyms, each characterizing a disease, syndrome, or health-related phenomenon [93]. The

[25] `https://geneontology.org/stats`
[26] `https://www.ebi.ac.uk/chebi/`
[27] `https://www.who.int/standards/classifications/classification-of-diseases`
[28] `https://www.who.int/standards/classifications/classification-of-diseases`

ICD is provided as part of the UMLS.

In addition to the variety of KOSs employed in the EL task, it is crucial to consider the diverse approaches used for a comprehensive understanding of the task.

2.7 Evaluation in entity linking

The common approach for model evaluation consists of measuring its performance in public benchmarks or corpora. Several corpora that focus on biomedical-related text are suitable for training and evaluation of EL models. Most of the corpora are based on English text. Three of the most commonly used corpora are BC5CDR, NCBI Disease and MedMentions, which are described below.

The BC5CDR [146] was created in 2016 as part of the BioCreative V competition. It is an English gold standard created using titles and abstracts from PubMed having as tagged entities diseases and chemicals jointly with their relation annotations. It is composed of 1,500 articles with 4,409 annotated chemicals, 5,818 annotated diseases and 3,116 chemical-disease interactions linked to MeSH.

NCBI Disease [64] is an English gold standard created in 2014 using 793 titles and abstracts from PubMed, encompassing 6,891 disease mentions linked to MeSH and OMIM.

MedMentions [174] consists of 350,000 mentions of biomedical concepts linked to UMLS identifiers. It is a gold standard that was built in 2019 using 4,000 titles and abstracts from PubMed.

Assessing the performance of an approach in the EL task requires common standards to allow for comparison between approaches. The metrics more frequently used in the evaluation of EL models are *accuracy*, *precision*, *recall*, and *F1-score*. *True positives* (TP) is the number of entities correctly linked, *false positives* (FP) is the number of entities wrongly linked, and *false negatives* (FN) is the number of entities that the model is not able to link to any concept when there exists a concept to be linked. *Precision*, *recall*, *F1-score* are calculated by the following expressions:

$$precision = \frac{TP}{TP + FP} \qquad recall = \frac{TP}{TP + FN}$$

$$F1 - score = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Another metric commonly used is *accuracy@k*, which determines the proportion of cases where the

correct concept is among the top-k concepts predicted by a given approach.

There still exist several challenges that impact the performance of EL systems, therefore, it is necessary to tackle these limitations in order to improve the SOTA in the EL task.

Chapter 3

Linking chemical and disease entities to ontologies by integrating PageRank with extracted relations from literature

Pedro Ruas

This chapter tackles objective 1 by using RE to overcoming the lack of information stored in biomedical KOSs, corresponding to the following journal article with minor adaptations for consistency and clarity:

> **Ruas, P.**, Lamurias, A., and Couto, F. M. (2020). **Linking chemical and disease entities to ontologies by integrating PageRank with extracted relations from literature**. *Journal of Cheminformatics* (Q1 Scimago) , 12, 57. DOI: `10.1186/s13321-020-00461-4` [210]. Code repository publicly available: `https://github.com/lasigeBioTM/REEL`

Abstract. *Background*: EL systems are a powerful aid to the manual curation of digital libraries, which is getting increasingly costly and inefficient due to the information overload. Models based on the PPR algorithm are one of the state-of-the-art approaches, but these have low performance when the disambiguation graphs are sparse. *Findings*: This work proposes a EL framework designated by REEL that uses automatically extracted relations to overcome this limitation. Our method builds a disambiguation graph, where the nodes are the ontology candidates for the entities and the edges are added according to the relations established in the text, which the method extracts automatically. The PPR algorithm and the Information Content (IC) of each ontology are then applied to choose the candidate for each

entity that maximizes the coherence of the disambiguation graph. We evaluated the method on three gold standards: the subset of the CRAFT corpus with ChEBI annotations (CRAFT-ChEBI), the subset of the BC5CDR corpus with disease annotations from the MEDIC vocabulary (BC5CDR-Diseases), and the subset with chemical annotations from the CTD-Chemical vocabulary (BC5CDR-Chemicals). The F1-Score achieved by REEL was 85.8%, 80.9%, and 90.3% in these gold standards, respectively, outperforming baseline approaches. *Conclusions*: We demonstrated that RE tools can improve EL by capturing semantic information expressed in text missing in KOSs and using it to improve the disambiguation graph of EL models. REEL can be adapted to any text mining pipeline and potentially to any domain, as long as there is an ontology or other KOSs available.

3.1 Introduction

3.1.1 Background

There has been an intense growth in the amount of scientific literature available, mainly in the form of scientific articles, whose content is mostly expressed in natural language. For instance, there are more than 30 million articles in the PubMed repository[1], which is one of the most used libraries in the Life Sciences and the Biomedical domains. This information overload creates problems for researchers who want to retrieve information, because they need to spend more time and effort to find the relevant articles for their work. Simultaneously, the number of online resources of biological information has also been rising, as it is the case of the domain ontologies. Domain ontologies provide a coherent representation of the knowledge in a specific scientific field, allowing a standardised nomenclature to people from different backgrounds [11]. In order to keep these resources relevant, it is necessary to extract the information locked in scientific literature and transfer it to the ontologies, a highly complex task that is usually done by dedicated curators. With the increase of literature available, manual curation of these repositories gets more costly and inefficient. Text mining tools are thus essential to aid both researchers and curators in the extraction of relevant information from large amounts of text.

EL is a typical text mining task (also designated by named entity disambiguation or normalisation), and its goal is to link each named entity in a given text to an appropriate identifier in a KOS, i.e., to associate an entity mention with the KOS concept that best represents it. EL systems are a fundamental

[1] `https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html`, accessed 15 January 2020

component of most text mining pipelines, usually used after the NER systems and before the RE systems. Some obstacles associated with EL are the presence of entity name variations in the text, like abbreviations, acronyms, alternate spellings or synonyms, and the presence of entity ambiguity, since polysemous entity mentions can be linked to more than one KOS concept according to their context [198]. For example, the entity mention "toxicity" can refer to "toxicity test", a laboratory technique, or to "cardiac toxicity", an adverse reaction. Additionally, EL systems developed for scientific text have to deal with the high ambiguity arising from a lack of nomenclature standardisation and the high specificity of the language, which means that, in many cases, only an expert can understand the text content [282].

Many EL approaches rely on the semantic information provided by KOSs, but in many cases KOSs lack important information. EL is usually a preceding step of RE because it is useful to know the entities present in a given text before finding relations between them but the relations described in the text can also disclose semantic information that may not be expressed in the KOS. So, our hypothesis is that RE approaches can overcome the missing domain knowledge in KOSs and improve the performance of the EL models that are highly dependant of the information provided by KOSs.

We have previously developed the PPR-SSM model [131], a graph-based approach that applies the PPR algorithm and a given Semantic Similarity Measure (SSM) to perform the disambiguation of biomedical entity mentions to several ontologies. The model builds a disambiguation graph for each text, where the nodes are ontology candidates and the edges are based on the ontology structure. One of the main limitations we have detected is that, sometimes, the model creates incomplete or sparse disambiguation graphs, i.e., with too few edges between the nodes, which hampers the application of the PPR algorithm and impacts the overall precision of the model. As the edges in the disambiguation graph are added if the candidates are linked in the ontology, we can infer that the information provided by the ontology is not enough to build a dense graph.

The main contribution of the present work is a framework to improve the precision of graph-based EL models, which we designate by REEL. This framework leverages the output of RE systems to build dense disambiguation graphs and to perform the disambiguation of disease and chemical entities. REEL was evaluated in several gold standards: the subset of the CRAFT corpus with ChEBI annotations (CRAFT-ChEBI) and the BC5CDR corpus with disease (MEDIC vocabulary) and chemical (CTD-Chemical vocabulary) annotations. The F1-Score obtained for the disambiguation of ChEBI, disease and chemical

mentions was, respectively, 0.8577, 0.8086 and 0.9025. The comparison with two baseline approaches (a string matching technique and a modified version of PPR-SSM based solely on information provided by ontologies) shows that REEL can substantially improve the precision of graph-based EL models.

3.1.2 Related work

3.1.2.1 Local EL models

The first EL models relied on local approaches, i.e., assumed that each entity in a text should be disambiguated individually according to its lexical or semantic features. This approach is limited because many times the meaning of an entity varies according to the context where it appears. One example of this type of approach is Bunescu and Pasca [32], in which the authors explored the disambiguation of Wikipedia entities using Support Vector Machines (SVM).

3.1.2.2 Integrating global evidence in EL models

More recent models assume that the entities in the same document must be somehow related, which means that the disambiguation of an entity influences the disambiguation of the others entities. PageRank is a random walks algorithm that was initially developed to measure the relative importance of web pages [183]. PageRank acts as a centrality measure in graphs or networks [77] and has been successfully adapted to the EL task [3, 90, 191]. For example, Pershina et al. [191] proposed an approach based on the PPR algorithm that combines local and global features to assist in the disambiguation. For each document, this approach builds a disambiguation graph in which the nodes consist of Wikipedia candidates for the named entities and the edges are added according to the Wikipedia link structure. PPR is applied on the disambiguation graph and then ranks each node according to its contribution to the coherence of the graph. The model was evaluated on the dataset AIDA and achieved a disambiguation accuracy of 91.7%. Guo and Barbosa [90] described a EL method for Wikipedia entities that determines the local similarity between textual mentions and entities (using lexical and statistical features) and a disambiguation graph that maximises global coherence between the candidates for the entities in a document. The algorithm then performs random walks in the graph to derive the semantic similarity between every pair of entities. These two approaches share some similarities with our method, in the sense that both are graph-based models and both apply the random walks algorithm over a disambiguation graph to maximise the global

coherence between entities in a given document. Besides, the models also include features to determine local similarity between each textual mentions and entities. Nevertheless, our method has noticeable differences to those approaches, namely, the edge generation in the disambiguation graph and the definition of the scoring function for the candidates.

Other EL approaches consist in the application of different DL techniques, like Ganea and Hofmann [78], which proposed a DL model that integrates local and global evidence to disambiguate entities at document level. The model includes entity embeddings to capture semantic information, a neural attention mechanism that selects words around the entity to help the disambiguation and a collective disambiguation module that uses a conditional random field for global inference in the document. Pre-trained language models, like BERT [59], create contextualised representations for entity mentions and have been fine-tuned for the EL task [273, 269].

3.1.2.3 EL models for biomedical text

There are fewer EL models developed for biomedical text. Usually the community challenges are a good way to assess the state-of-the-art in the field. For instance, the BioCreativE (Critical Assessment of Information Extraction in Biology) challenge contains tasks related to biomedical digital curation, between them some related with EL of biomedical entities, such as genes, chemicals and diseases. The description of the participating models in the latest edition can be consulted in Arighi et al. [9]. There are models that perform both named entity recognition and named entity linking of disease and chemical entities, such as TaggerOne [136] and DNorm [134]. D'Souza and Ng [67] proposed an approach that performs the disambiguation of disease mentions and obtained an accuracy of 90.75% and 84.65% and in the ShARe/CLEF eHealth Challenge corpus and the NCBI Disease corpus. More recently, Ji et al. [112] fine-tuned the BERT model and two of its variants, ClinicalBert and BioBert, to the EL task, evaluated them in the ShARe/CLEF, NCBI Disease and TAC2017ADR (drug labels) datasets and obtained an accuracy of 91.10%, 89.06% and 93.22%, respectively.

J-REED [178] is a model able to perform both EL and RE. However, this approach performs both tasks sequentially and does not improve EL with RE, which is the main goal of the present work. To the best of our knowledge, our method is the first attempt to use the RE output to improve the performance of graph-based EL methods.

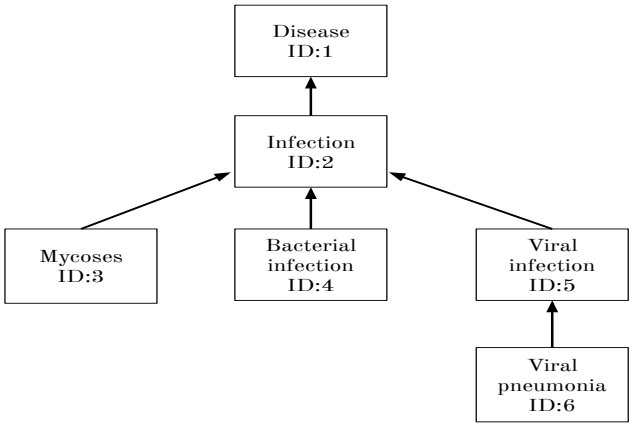


Figure 3.1: Subsumption relations: Example of a set of disease concepts and subsumption relations between them. The arrows denote the direction of the *is-a* relations.

3.2 Methods

3.2.1 Definition of the EL problem

The starting requirement for EL is a corpus containing documents with entity mentions already identified by a human annotator or by a NER system. The set of entity mentions in a given document is represented by E. The objective of EL is to link each entity mention e, with $e \in E$, to the concept in a KOS that best represents it. The output of a EL model consists of each entity mention associated with a KOS identifier. A KOS is a tuple $< C, R >$, where C is the set of concepts and R the set of relations between concepts. Each relation consists in a pair of concepts (c_1, c_2), with $c_1, c_2 \in C$. An ontology is a type of KOS that contains, among other types, subsumption or "is-a" relations between concepts (see Figure 3.1).

The EL task comprises two distinct steps:

- Candidate generation: Generation of the candidates list $CL(e) = \{c_e^1, ..., c_e^i | \forall c_e \in C\}$ for each entity mention e in set of document entity mentions E.

- Candidate ranking and disambiguation: Selection of the candidate c_e in the candidate list CL that best represents each entity mention e, i.e., the highest ranked candidate.

3.2.2 Candidate generation

3.2.2.1 Candidate list

The first step of EL is accomplished through a search in the KOS for each entity mentions (string matching technique). The candidates are ranked according to their lexical similarity which is determined by the edit distance, i.e, the minimum number of operations needed to convert one string into another. If there is an exact match between the entity mentions and any KOS concept, the mention is disambiguated with that concept and no candidates list is built. Otherwise, the first ten candidates are added to the candidates list. Additionally, if there are synonyms in the KOS for the candidates, they are also added to the list. The baseline model in this work selects for each entity mentions the candidate with more lexical similarity. This approach is very limited, as it ignores the document context where the mentions appear: the candidate with the most similar string is not always the correct disambiguation and, consequently, a global approach will be more accurate. For that it is necessary to build a disambiguation graph.

3.2.2.2 KOS-based disambiguation graph

The disambiguation graph G for a document is represented by $G = \{(e, c_e)|e \in E, c_e \in CL(e)\}$. Each node (e, c_e) in the graph is an entity mention/candidate pair and the edges between nodes are built according to the following link mode:

- `KOS-link`: Two nodes (e_1, c_{e_1}) and (e_2, c_{e_2}) are connected in the graph if the candidates c_{e_1} and c_{e_2} are directly connected by the KOS structure (the shortest path length between them is 1) and if $e_1 \neq e_2$. This latter constraint is to prevent the generation of noisy edges between nodes, as only one node/candidate per entity mention constitutes the correct disambiguation. For example, "Viral pneumonia" and "Viral infection" in the example of the Figure 3.1 are directly connected.

This method to build the disambiguation graph is the same used by our previous framework PPR-SSM [131] and other graph-based approaches [191, 90], in which the authors consider that an edge between two nodes or candidates occurs if the corresponding Wikipedia articles have at least one link between them.

The way the nodes are linked in the disambiguation graph directly affects the application of the PPR algorithm.

``After 48 h in fixative at 4 °C, brains were transferred to PBS, dehydrated in **alcohols**, treated with cedarwood oil and **methylsalicylate**, and embedded in paraffin for sectioning"

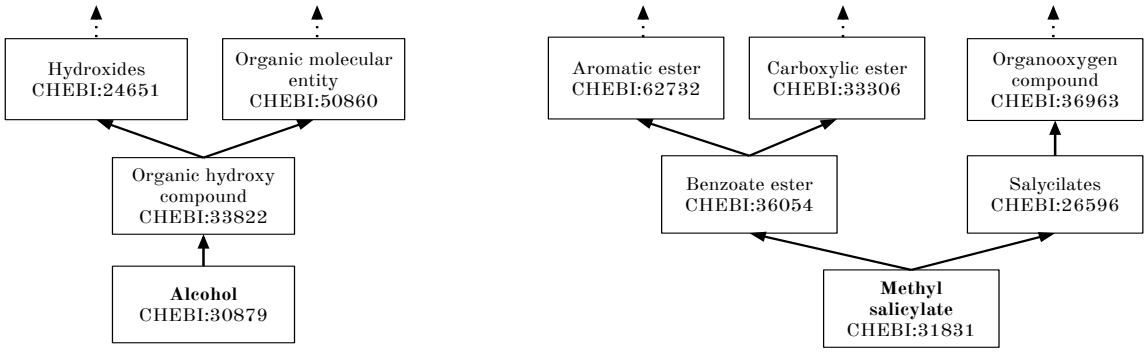


Figure 3.2: Relation between terms in the text: Example of a relation between the entity mentions "alcohols" and "methylsalicylate" that is described in an article, but it is not expressed in the ChEBI ontology structure. The closest ancestors of the respective ChEBI entities for these mentions, "Alcohol" and "Methyl salycilate", and the respective ChEBI identifiers are also shown.

3.2.2.3 Improvement of the disambiguation graph with extracted relations

The lack of domain knowledge in the building of the disambiguation graphs is responsible for some limitations, such as the scarcity of edges between nodes. The PPR algorithm is a measure of centrality in the graph, so the calculation of the score for each node is directly related with the number of edges that traverse the nodes. Consequently, if the disambiguation graph has few edges between its nodes, the PPR algorithm will assign scores according to other criteria than the number of edges, like node degree in the KOS, which does not properly assess the node contribution for the disambiguation graph coherence. In many cases, the simple inclusion of KOS relations between concepts is not enough to generate an adequate number of edges between nodes in the disambiguation graph. To overcome this problem, we propose to include information about entity relations described in a given corpus to generate the edges in the disambiguation graph. For that, two additional link modes between nodes in the disambiguation graph are defined:

- `Corpus-link`: Two nodes (e_1, c_{e_1}) and (e_2, c_{e_2}) are connected if the candidates c_{e_1} and c_{e_2} appear in a relation described in the text of the documents in the corpus and if $e_1 \neq e_2$. In Figure 3.2, the entity mentions "alcohols" and "methylsalicylate" are not linked in the structure of the ChEBI ontology, but there is an corpus document describing a relation between these two entities.

- `KOS-Corpus-link`: Two nodes (e_1, c_{e_1}) and (e_2, c_{e_2}) are connected if either appear in a relation described in text or if they are connected in the KOS.

Besides the relations explicitly described in text that can be extracted by an RE tool, we also include in our approach human annotations of chemical-disease interactions whenever these are available in the corpus. In this case, we assume that there is a relation between any two given disease entities if the same chemical entity plays a role in both. Conversely, two chemical entities are related if they are involved with the same disease entity.

With the disambiguation graph already built, it is thus necessary to compute the weights for each node/candidate according to their relevance to the entities.

3.2.3 Candidate ranking and disambiguation

The second step of EL is the disambiguation of each entity mention e to a candidate c_e, which is determined by the function:

$$disambiguate(e) = arg_{c_e} max\{score(e, c_e)\} \tag{3.1}$$

The $score$ function above determines the likelihood of the candidate c_e being the correct disambiguation for entity mention e. The PageRank algorithm performs random walks in a graph and returns a probability distribution of reaching each node after a given number of iterations. In each iteration, there is a teleport probability ϵ of jumping to a random node in the graph, and a probability of 1 - ϵ (also called damping factor) of following an outgoing edge of the current node. In this way, the algorithm ranks each node, which can be considered a measure of centrality in the graph. When the teleports are not random but adjusted to the same source node, the PageRank algorithm is designated by PPR. In the context of the EL task, considering a given source node s and a given node n in the graph G, the PageRank score of the relation $PPR(s \rightarrow n)$ measures the relevance of node n for node s. The contribution of node n

to the global coherence of the disambiguation graph G due to the presence of node s is expressed by the following equation:

$$coherence_s(n) = PPR(s \rightarrow n) \tag{3.2}$$

Thus, to determine the overall contribution of the node n to the global coherence it is necessary to sum all the contributions of the node related with the presence of the other source nodes (except the nodes representing candidates competing for the same entity mention as n):

$$coherence(n) = \sum_{s \in G} coherence_s(n) \tag{3.3}$$

The $coherence$ expression in equation 3.3 constitutes the $score$ function in the Equation 3.1. Intuitively, for each entity mention, the node/candidate that contributes the most for the global coherence of the disambiguation graph will be chosen to disambiguate the entity.

In order to add a layer of differentiation for the nodes in the disambiguation graph, the PageRank of each node n in relation to a source node s is multiplied by the Information Content (IC) of the node n:

$$coherence_s(n) = PPR(s \rightarrow n) \cdot IC(n) \tag{3.4}$$

The IC of a concept is a measure of its "rareness": rare concepts will have higher information content. In the present work, we use the extrinsic IC definition, as described by Couto et al. [48], in which the IC of a concept is associated with the frequency of its instances in an external dataset (for example a corpus). Pershina et al. [191] and Guo and Barbosa [90] do not include the IC in their approach. Instead, Pershina et al. [191] use the Freebase popularity, which is a score based on the in-edges and the out-edges of pages in Wikipedia and Freebase.

3.2.4 Models

To determine the impact of the relation extraction in the EL performance, we evaluated the following models described below in several datasets:

- "String matching": first baseline approach. For each entity to disambiguate, this model selects the candidate with highest lexical similarity (lowest edit distance) through string matching.

- "PPR-IC": second baseline, corresponds roughly to PPR-SSM [131] without using semantic similarity measures. Consists in the application of the PPR algorithm with the inclusion of the IC for each node. The edges in the disambiguation graph are exclusively based on KOS relations between concepts (link mode `KOS-link`). The scoring function is equation 3.4.

- "REEL(Corpus)": the application of the PPR-IC model, but using only relations described by the text to build the edge structure of the disambiguation graph (link mode `Corpus-link`). The scoring function is Equation 3.4.

- "REEL(KOS+Corpus)": the application of the PPR-IC using both relations described by the text (link mode `Corpus-link`) or KOS relations (link mode `KOS-link`) to build the edge structure of the disambiguation graph. The scoring function is Equation 3.4.

3.2.5 Data Description

The data used in this work consist of datasets/corpora and ontology files. The datasets contain the surface form of disease and chemical entities, and the respective ontology identifiers. The ontology files include information about the ontology concepts, as well the semantic relations between them.

3.2.5.1 Datasets

The CRAFT ("Colorado Richly Annotated Full-Text") corpus is a set of 67 full-text biomedical articles from PubMed Central Open Access subset [43]. This gold standard contains, among others, 4,548 manual annotations of ChEBI entities. The set of the corpus with ChEBI annotations will be further designed as "CRAFT-ChEBI". In this work, we used the version 3.0 of this corpus [43].

To demonstrate that REEL can easily be adapted to include relations extracted by different tools, we evaluated the performance of the model on the BC5CDR corpus openly available [146]. This gold standard was developed for the disease named entity recognition (DNER) task and the chemical-induced disease (CID) RE task in the for BioCreative V. This corpus consists of 1,500 PubMed abstracts annotated with 4,409 chemicals, 5,818 diseases and 3,116 chemical-disease interactions. The chemical annotations contain the respective MeSH unique ID from the "Chemicals and Drugs" category in the MeSH vocabulary, whereas the disease annotations contain the MeSH unique ID from "Diseases" category. The

set of the corpus with disease annotations is further designated by "BC5CDR-Diseases" and the set of the corpus with chemical annotations by "BC5CDR-Chemicals". We evaluated the models in the train, development and test sets and in a set containing all the corpus documents which we designate by "All".

3.2.5.2 Ontologies

The first ontology we used was the ChEBI ontology, which represents low-molecular weight chemical entities with biological relevance for living organisms [97]. As of 1 September, 2019 this repository contained 56,090 annotated entries[2]. In the experiments described in this work we used the data from the release 179[3].

The second ontology was the MEDIC Disease vocabulary from the CTD (Comparative Toxicogenomics Database), which is an hierarchical vocabulary that represents descriptors from the "Diseases" category of MeSH controlled vocabulary and genetic disorders from OMIM (Online Mendelian Inheritance in Man) repository [53]. As of May, 2020, this vocabulary contained 7,246 entries representing distinct diseases entities[4] and in the experiments of this work we used the data from the referred month release[5].

The third ontology was the Chemicals vocabulary also from CTD, an hierarchical vocabulary representing descriptors from the "Chemicals and drugs" category of MESH. As of May, 2020, this vocabulary contained 16,313 entries representing distinct chemical entities and in the experiments of this work we used the data from referred month release[6].

3.2.6 Evaluation Metrics

In each document of the corpus, repeated instances of an entity mention with the same surface form count as a unique entity. True positives (tp) refer to the number of entities correctly disambiguated, false positives (fp) to the number of entities wrongly disambiguated and false negatives (fn) to the number of entities that the model does not disambiguate. The performance of each model was evaluated in each dataset through the determination of the precision, recall and micro-averaged F1-score:

[2] `https://www.ebi.ac.uk/chebi/statisticsForward.do`, accessed 1 October 2019

[3] `ftp://ftp.ebi.ac.uk/pub/databases/chebi/archive/rel179/ontology/`, ChEBI ontology files, release 179, accessed 1 October 2019

[4] `http://ctdbase.org/about/dataStatus.go`, accessed 7 May 2020

[5] `ctdbase.org/reports/CTD_diseases.obo.gz`, accessed 2 May 2020

[6] `ctdbase.org/reports/CTD_chemicals.tsv.gz`, accessed 2 May 2020

$$precision = \frac{tp}{tp + fp} \cdot 100 \tag{3.5}$$

$$recall = \frac{tp}{tp + fn} \cdot 100 \tag{3.6}$$

$$F1 - score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3.7}$$

3.2.7 Implementation

3.2.7.1 Pre-processing

The module dedicated to the pre-processing of corpus documents was implemented in Python 3.6.8. This module generates the candidates lists through the FuzzyWuzzy Python library, i.e., for each entity mention the module obtains the first ten ontology concepts with more lexical similarity with the mention (and the synonyms for the candidates) and discards the candidates without a valid ontology identifier. The module also converts the ontologies into graph objects with the Networkx Python Library, which further allows the determination of the ontology relations between concepts. Alternatively, this module can also include information about relations between entities described in text, either from corpus annotations or from the output of any RE tool. In the end, the module creates a candidates file for each original document in the corpus, files that contain all the information necessary to build the disambiguation graph (nodes and edges).

3.2.7.2 BO-LSTM

To investigate the hypothesis that relations described in text can improve the edge generation in the disambiguation graph, we integrated the information obtained by BO-LSTM [132] in our EL method. This RE tool applies a model based on a recurrent neural network with LSTMs to detect and classify relations between entities in text. In BO-LSTM multi-channel architecture the information used in the detection and classification of relations differs with the specific "channel" considered: shortest dependency path (SDP), WordNet classes or ChEBI ancestors. In the ChEBI ancestors channel, this model first links each entity mention to a ChEBI identifier through string matching and then builds a vector with the respective

ChEBI ancestors. BO-LSTM was trained on the "SemEval 2013: Task 9 DDI extraction corpus" [99], that contains annotations of pharmacological substances and drug-drug interactions at the sentence level. It was later applied to the documents of CRAFT corpus using all the described channels in order to detect relations between every pair of ChEBI entities in a sentence. The output was a file with classification of each entity pair: "effect", if there is an interaction or relation between the entities or "noeffect", otherwise. For the link modes `Corpus-link` and `KOS-Corpus-link` this information is in the candidates files and, consequently, in the edge structure of the respective disambiguation graph. For a more detailed description of BO-LSTM implementation, please refer to the original publication [132].

3.2.7.3 PPR

The input to this part of the model is the candidates files from pre-processing stage. The model uses the PPR implementation proposed by Pershina et al. [191]. The PPR algorithm was computed according to the Monte Carlo algorithm proposed by Fogaras and Rácz [75]. We decided to maintain the same values for the PPR parameters described by Pershina et al. [191]: initialisation with 2,000 random walks for each source node, 5 steps of PPR and probability of jump to the source node (or teleport probability) of 0.2.

3.3 Results and discussion

The evaluation results for the models in the different datasets are available in the Table 3.1.

In the CRAFT-ChEBI dataset, the two REEL models achieved the same performance, a F1-Score of 85.8%, which is an improvement of 2.5 p.p. and 7.9 p.p. comparing with the "PPR-IC" and "String matching" baseline approaches. The precision achieved was 91.3%, an increase of 4.2 p.p and 13.5 p.p. comparing with the "PPR-IC" and "String matching" models.

In the BC5CDR-Diseases dataset, the model "REEL(Corpus)" achieved the highest F1-Score, 80.9%, which represents an improvement of 2.0 p.p. and 2.4 p.p. comparing with the "PPR-IC" and "String matching" models. The precision achieved was 86.9%, an increase of 3.5 p.p. and 4.2 p.p. comparing with the "PPR-IC" and "String matching" models. In the BC5CDR-Chemicals dataset, the models "REEL(Corpus)" and "REEL (KOS+Corpus)" obtained the highest F1-Score, 90.3%, but the increase from the baseline approaches was lower: 0.2 p.p. and 1.1 p.p. comparing with the "PPR-IC" and "String

Table 3.1: Evaluation results in the CRAFT-ChEBI dataset (top), BC5CDR-Diseases (middle) and BC5CDR-Chemicals (bottom). For the datasets BC5CDR-Diseases and BC5CDR-Chemicals the results for each subset are shown. "All" refer to the entire corpus and "Train", "Dev" and "Test" refer to the train, development and test sets, respectively.

	CRAFT-ChEBI		
Model	P	R	F1
String matching	77.8	78.0	77.9
PPR-IC	87.1	79.9	83.3
REEL(Corpus)	**91.3**	**80.9**	**85.8**
REEL(KOS+Corpus)	**91.3**	**80.9**	**85.8**

	BC5CDR-Diseases											
Model	All			Train			Dev			Test		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
String matching	82.7	74.7	78.5	81.7	74.1	77.7	82.7	72.3	77.2	83.6	77.5	80.4
PPR-IC	83.4	74.9	78.9	84.1	74.6	79.1	86.2	73.0	79.1	87.2	78.2	82.5
REEL(Corpus)	**86.9**	**75.6**	**80.9**	87.8	75.4	81.1	87.9	**73.5**	**80.1**	**89.0**	**78.6**	**83.5**
REEL(KOS+Corpus)	86.6	75.5	80.7	**88.2**	**75.5**	**81.4**	**88.0**	**73.5**	**80.1**	88.8	78.5	83.3

	BC5CDR-Chemicals											
Model	All			Train			Dev			Test		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
String matching	94.9	84.1	89.2	93.7	84.4	88.8	95.3	85.2	90.0	95.3	82.6	88.5
PPR-IC	96.7	84.4	90.1	97.6	84.9	90.8	97.6	85.5	91.2	98.0	83.0	89.9
REEL(Corpus)	**97.0**	**84.4**	**90.3**	**98.0**	**85.0**	**91.0**	**98.3**	**85.6**	**91.5**	**98.4**	**83.0**	**90.0**
REEL(KOS+Corpus)	**97.0**	**84.4**	**90.3**	**98.0**	**85.0**	**91.0**	98.2	**85.6**	**91.5**	**98.4**	**83.0**	**90.0**

matching" models. The precision increased by 0.3 p.p. and 2.1 p.p. from the precision of the "PPR-IC" and "String matching" models.

The two REEL models ("REEL(Corpus)" and "REEL(KOS+Corpus)") consistently achieved the best F1-Score in all datasets and respective sets. The higher F1-Score comparing with the baseline approaches is directly related with increases in the precision, as the recall did not substantially differ across models. These results suggest that the initial hypothesis of improving the precision of graph-based EL models through RE is true.

The use of a RE tool (BO-LSTM) and the inclusion of chemical-disease interactions of the BC5CDR corpus overcame the lack of domain knowledge in the KOS and originated denser disambiguation graphs, which by its turn, improved the performance of the PPR algorithm. The results obtained by REEL are explained by the fact that there is semantic information encoded in text and not expressed in the ontologies structure. For example, in the following sentence of document 14737183 of the CRAFT-ChEBI dataset

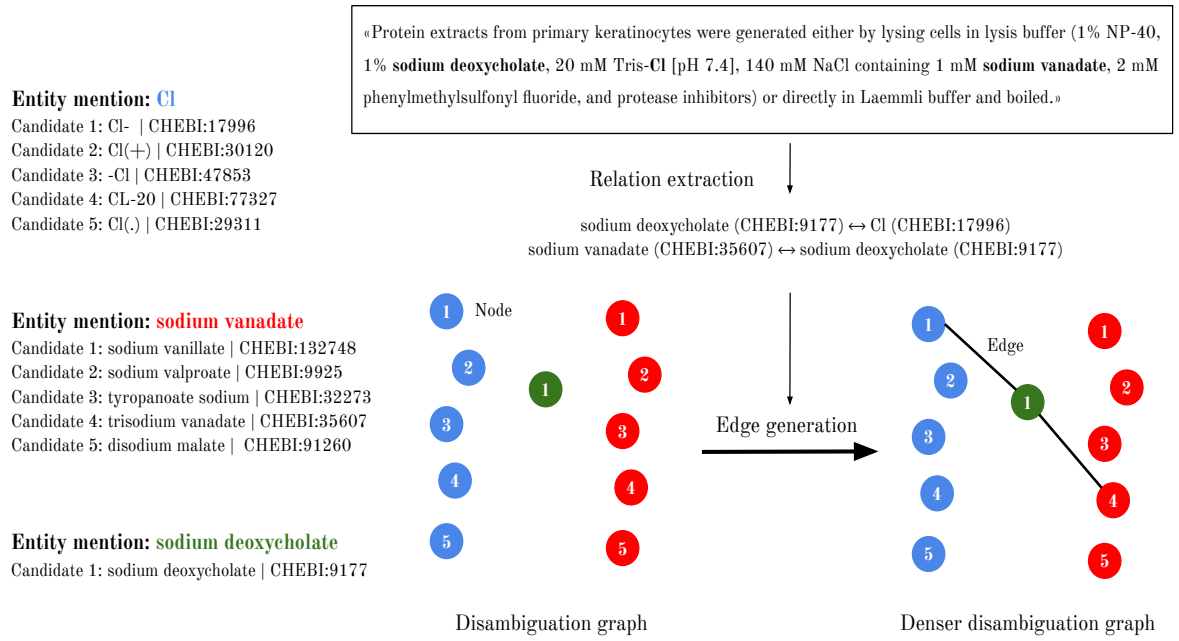


Figure 3.3: Disambiguation graph improved by extracted relations: Example showcasing the building of the disambiguation graph for three different entity mentions in a document of the CRAFT-ChEBI dataset and the further densification of the graph with extracted relations from the dataset.

"(...) skin from at/at mice reveal an abrupt dorsoventral transition of DOPA staining, which probably reflects the additive effects of reduced melanin content" there are two entities mentions: "DOPA", annotated with the identifier "CHEBI:49168", and "melanin", annotated with the identifier "CHEBI:25179". In the ChEBI structure these entities are not linked, but BO-LSTM can infer that the content of "melanin" affects the "DOPA staining" through contextual features.

Another example, more complete, is shown in Figure 3.3. In the document 15630473 of the CRAFT-ChEBI dataset there were, among others, the entity mentions "Cl", "sodium vanate" and "sodium deoxycholate". Only "sodium deoxycholate" had an exact match in ChEBI ontology, the homonymous concept with the identifier "CHEBI:9177", the other two entity mentions had each one a candidate list with several ChEBI candidates (in the figure these candidate lists are abbreviated). Without the RE output and relying only in KOS links, the disambiguation graph formed with the candidates for these entity mentions had no edges between the nodes. But BO-LSTM was able to extract relations between some of the can-

didates expressed elsewhere in the corpus and the addition of these relations to the disambiguation graph originated two new edges. The PPR algorithm assigns more weight to the nodes with higher degree (i.e. more interconnected) and in this case the nodes with higher degree (CHEBI:17996 and CHEBI:35607) corresponded to the correct disambiguation for the entity mentions.

The EL performance of REEL is indirectly related with the performance of the RE tool that is used. The RE performance of BO-LSTM is lower than the inclusion of the gold standard annotations present in the BC5CDR corpus: BO-LSTM is not able to extract all the relations between entities present in the CRAFT-ChEBI dataset, contrarily to what happens in the BC5CDR corpus, where all the relations are known. The use of a RE tool is a more realistic scenario than the inclusion of gold standard annotations, because not always these are available, so we measured the EL performance by REEL in these two different scenarios. According to Table 3.1, we can conclude that the EL performance increases comparing with baseline approaches in these two different scenarios, so we can conclude that this increase is independent of the source of the extracted relations.

In the ChEBI ancestors channel of BO-LSTM, the model links each entity mention with a ChEBI identifier through string matching, which limits the performance of the tool when extracting the relations in text. So, the performance of REEL benefits with the BO-LSTM output, but could by its turn be used to improve BO-LSTM performance, more concretely by replacing the string matching method and improving its disambiguation component.

The REEL framework can be potentially adapted to any domain, as long as there is an ontology or other structured KOSs available. This feature is specially relevant for the biomedical and life sciences domains, where there is a lack of text mining tools, but there are many digital libraries with scientific information available. REEL could be used in any gold standard dataset as long it contains annotations of biomedical ontology concepts and in any biomedical text where the entities are already recognized by a NER tool. The relations extracted in the CRAFT-ChEBI corpus (or in the BC5CDR corpus) could be included in REEL to improve its performance in that different gold standard/text. The framework only needs labelled data if it is necessary to train the RE tool, but if the tool is already trained or the relations are available, that need disappears.

3.3.1 Error analysis

Despite the positive results achieved by REEL, there were some errors that prevented an even higher performance.

One type of error is associated with the presence of composite mentions in the BC5CDR-Diseases dataset. For example, the disease mention "detrimental effect on memory and cognition" is annotated with two different gold labels: "detrimental effect on memory" (D008569) and "detrimental effect on cognition" (D003072). REEL is not adapted to recognise and deal with this type of entity mentions because only one candidate is selected per entity.

The second type of error is related with the candidate generation step, as many entity mentions did not have the correct disambiguation in the respective candidates list, which impacted mainly the recall of the model. The string matching technique to generate candidates is useful to restrict the field of possible ontology candidates but sometimes leaves out correct candidates with little lexical overlap with the entity mention.

Another type of error is due to the presence of few entity mentions in a document. The PPR algorithm has higher performance in bigger and denser disambiguation graphs, but in certain cases, there is not enough entity mentions in a document to build a disambiguation graph with these characteristics.

3.4 Conclusion

We developed REEL, which leverages extracted relations described in the text to build dense disambiguation graphs and then applies the PPR algorithm for candidate ranking and disambiguation. The framework was evaluated on three different gold standards, the CRAFT-ChEBI, the BC5CDR-Diseases and the BC5CDR-Chemicals datasets and achieved a F1-Score of 85.8%, 80.9% and 90.3%, respectively, which represents an improvement comparing with two baseline approaches. This improvement is due to increases in the precision.

The results show that REEL can be used to mitigate the problems associated with the application of PPR for EL using sparse disambiguation graphs. Our framework improved the performance of EL when the output of a deep-learning RE tool (BO-LSTM) is included and also when relations annotated in a gold standard (BC5CDR corpus) are included, which demonstrates that the framework has the flexibility to

easily integrate relations provided by any source.

For future work, we intend to explore pre-trained language models, like BioBERT [140], to further improve the determination of local similarity between mentions and ontology concepts. Besides, it would be interesting to adapt this method to other types of KOS other than the ontologies, like Wikipedia, which are less formally defined.

Availability of data and materials

- Project name: REEL
- Project home page: `https://github.com/lasigeBioTM/REEL`
- Operating system: Ubuntu 18.04 LTS
- Programming language: Python $\geq$ 3.6, Bash
- Other requirements: JDK $\geq$ 11.0.6, BO-LSTM
- License: Apache License 2.0

‘

The data supporting the results and the code necessary to reproduce the results are openly available on the referred GitHub repository, as well a Dockerfile to properly setup the environment to run the code.

The data used in this work is available in the following links:

- ChEBI ontology file
- CTD MEDIC Disease vocabulary and CTD Chemical vocabulary files are available under conditions on `ctdbase.org/reports/CTD_diseases.obo.gz` and on `ctdbase.org/reports/CTD_chemicals.tsv.gz`, respectively
- CRAFT Corpus
- BC5CDR corpus

Alternatively, the script “get_data.sh” in the GitHub repository automatically downloads all the necessary data.

Chapter 4

NILINKER: attention-based approach to NIL Entity Linking

Pedro Ruas

This chapter tackles objective 1 by overcoming the lack of coverage in target biomedical KOSs through the development of a DL approach to handle NIL entities, corresponding to the following journal article with minor adaptations for consistency and clarity:

> **Ruas, P.** and Couto, F. M. (2022). **NILINKER: Attention-based approach to NIL entity linking**. *Journal of Biomedical Informatics* (Q1 Scimago), 132, 104137. DOI: `https://doi.org/10.1016/j.jbi.2022.104137` [207]. Code repository publicly available: `https://github.com/lasigeBioTM/NILINKER`

Abstract. The existence of unlinkable (NIL) entities is a major hurdle affecting the performance of EL approaches, and, consequently, the performance of downstream models that depend on them. Existing approaches to deal with NIL entities focus mainly on clustering and prediction and are focused on general entities. However, other domains, such as the biomedical sciences, are also prone to the existence of NIL entities, given the growing nature of scientific literature. We propose NILINKER, a model that includes a candidate retrieval module for biomedical NIL entities and a neural network that leverages the attention mechanism to find the top-k relevant concepts from target KOSs (MEDIC, CTD-Chemical, ChEBI, HPO, CTD-Anatomy and GO-BP) that may partially represent a given NIL entity. We also make available a new evaluation dataset designated by EvaNIL, suitable for training and evaluating models focusing on

the NIL entity linking task. This dataset contains 846,165 documents (abstracts and full-text biomedical articles), including 1,071,776 annotations, distributed by six different partitions: EvaNIL-MEDIC, EvaNIL-CTD-Chemical, EvaNIL-ChEBI, EvaNIL-HPO, EvaNIL-CTD-Anatomy and EvaNIL-GO-BP. NILINKER was integrated into a graph-based EL model (REEL) and the results of the experiments show that this approach is able to increase the performance of the EL model.

4.1 Introduction

EL, also known as disambiguation or normalisation, is the task of mapping entities present in free text to entries in a target KOS that describe their meaning [198]. EL tools bridge the gap between natural language and computer-friendly structures, like a KOS, which provide a semantic representation simultaneously readable by humans and machines. These tools play a crucial role in the automatic population and curation of KOSs [110, 65] and, with the growing amount of text present in the Web, a large part of it expressed in natural language, the importance of these applications is only increasing. At the same time, more data means that keeping KOSs updated becomes harder and their coverage decreases, so the probability that a given entity has a correspondent KOS concept will also decrease. Besides, these tools also directly influence the performance of many other systems and tasks, such as search engines [164], question answering systems [224], electronic health records-based phenotyping [237], pharmacovigillance [44], among others.

In many cases, it is possible to link a recognized entity to a KOS concept that is able to express its meaning in the context of the text. This type of entity is designated by in-KOS. However, for other entities, there are simply no KOS concepts that accurately convey their original meaning, which are then designated by NIL, out-of-KOS, unlinkable or CUI-less entities [219]. For example, if we want to link the entity mention "bioinformatician" to the DBpedia ontology[1], at the moment of writing there is no concept that truly captures its meaning. The closest concept would be "Scientist", which is not specific enough but represents a substantial portion of the semantics of the entity, and its child concepts, "Biologist", "Entomologist", "Medician" and "Professor", related concepts that do not entirely represent the meaning of the entity. Dredze et al. [65] consider that the absence of KOS entries able to represent entities in the text is one of the main challenges that EL models face. First, it limits their recall, because

[1] `https://wiki.dbpedia.org/services-resources/ontology`

it defeats the main purpose of the EL task, which is precisely to link all entity mentions in the text to KOS entries. Since this type of system plays an important role in text mining pipelines, the performance of downstream tools that depend upon correct linked entities will also be hindered. Additionally, SOTA tools are injecting knowledge in DL models, but if KOSs are incomplete that process gets hindered.

Existing approaches to deal with NIL entities focus mainly on predicting them among the entire set of entities to link [32] or in their clustering [22]. For instance, a common approach considers a given entity as NIL if the score of the top candidate in the respective candidates' list is below a predefined threshold value. For instance, Chen et al. [41] proposed a model that learns a threshold value from the training set or instead sets it to zero if there are no NIL entities in the dataset. These approaches limit the semantic information that is possible to obtain from the text, since they just mark those entities as NIL which does not provide semantic value for other text mining tasks. Another approach detects emergent entities, which, according with the definition by Färber et al. [71], correspond to entities that are *trending* and *notable for the first time* in the context of tasks like semantic media monitoring or automatic speech recognition. Emerging entities are usually associated with real-time events, in which their number of mentions suddenly increase in news articles. Emerging entities correspond to NIL entities. The authors proposed a machine learning-based model to identify emerging entities in news articles among a set including entities that are linkable to Wikipedia and also unlinkable entities. The detected emergent entities with higher confidence score, consisting mainly of living and dead people and politicians, are then suggested for inclusion in Wikipedia. Other approaches [148] use linkable Wikipedia entities to help in the semantic typing of NIL entities. However, approaches focusing on biomedical NIL entities that attempt to leverage NIL entity linking to improve EL are still scarce. One example is DILBERT [166], which is a neural-based model for medical entity linking that includes a module for out-of-KOS prediction.

Similarly to what happens with news articles, scientific articles about diverse topics are constantly being published on large volumes due to the ever changing nature of the scientific landscape. Bornmann and Mutz [27] estimated an increase of 8%-9% in the number of new scientific articles from the end of the Second World War up until 2010. For instance, PubMed added new 952,919 citations only in 2020[2]. There are events that trigger a sudden increase in the number of publications, such as the COVID-19

[2] `https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html`

pandemic. This creates a challenge since many of the entities mentioned in the newly published scientific literature are not yet part of biomedical KOSs due to their novelty and due to the effort of manually curating a large volume of the text, so the performance of EL models decrease. Besides, biomedical KOSs miss Wikipedia-specific features. For example, Färber et al. [71] leverage Wikipedia page views to detect emergent entities, but this is not possible to replicate in the biomedical domain. At last, there is a lack of evaluation datasets which limits the training and the benchmarking of approaches developed for NIL entity linking.

In the present work, we address the following research questions:

1. How to develop an approach to link NIL entities to biomedical KOSs concepts?
2. How to improve the performance of biomedical EL approaches by linking NIL entities to KOSs concepts?

To tackle the aforementioned challenges, this work proposes an approach to link NIL entities to target biomedical KOSs designated by NILINKER, more concretely, by generating KOS candidates for the input NIL entity through a word-concept dictionary and then by applying a neural network using the attention model to determine the relevance of each KOS candidate. In the last step, the entity is associated with the highest-scoring candidates. For example, this approach would theoretically link the entity mention "bioinformatician" above to the KOS concept "Scientist", which would allow the expansion of the target KOS through the creation of a new subconcept of "Scientist". Besides, as the existence of NIL entities hinders the performance in the EL task, NILINKER was integrated jointly with a EL model and the impact of this integration was assessed on several evaluation datasets. The main contributions of this work are thus:

- The novel approach called **NILINKER** that, instead of detecting and clustering NIL entities, attempts to link them to available KOS concepts. The model associates the NIL entity with the top-k relevant KOS concepts.
- The dataset **EvaNIL** to train and benchmark NIL entity linking models.
- The model **REEL-NILINKER**, that performs biomedical EL leveraging NILINKER to deal with NIL entities.

The main novelties of this work are the explicit focus on developing a model to link the NIL entities and not just predict them and the proposed approach to train such model: taking entity mentions in EL datasets and converting them to "artificial NIL entities", forcing the model to link them to the direct ancestor of the original target KOS concept. This way, every entity mention appearing in a given document is leveraged and the available semantic information increases.

The code associated with the models and the performed experiments is fully available[3]. Relevant prior work are described in the next section.

4.2 Related work

4.2.1 Biomedical Named entity Linking

Besides NIL entities, there are other challenges in biomedical EL, such as entity name variations (e.g. the symbols "DIF", "TNFA", "TNFSF2", "TNLG1F", or "TNF-alpha" refer to the same gene, "tumour necrosis factor") and ambiguity (e.g. "iris" can be both an anatomical structure of the eye or a genus of plants).

Proposed approaches for the EL task can be broadly divided into local or global. Local models attempt to link each entity mention to the most apppropriate concept in the target KOS using local evidences (e.g. lexical similarity between mention and candidates), independently of of the global context (i.e. the other entity mentions appearing in the same document) [134, 67]. On the other hand, global models attempt to maximise the coherence of several linking decisions, i.e., the linking decision of a given entity mention is influenced by the linking of the other entities present in the same document and vice-versa [210, 41].

Besides this division, it is possible to further classify the models into rule-based, graph-based or machine learning/DL-based. One example of rule-based model was proposed by D'Souza and Ng [67], which includes a set of ten consecutive rules to link a given entity mention (exact match with a concept in the target KOS, abbreviation expansion, replacement with synonyms, among others). Graph-based models are global models that build a graph usually including the entity mentions and respective KOS candidate concepts and then score each candidate according to their role in the graph (for example, candidates associated with nodes with higher degree receive higher score) [210]. In terms of machine learning, there is a focus on models based on neural networks and, more recently, on large pre-trained language

[3] `https://github.com/lasigeBioTM/NILINKER`

models based on the Transformers architecture. Models like BioBERT [140], PubMedBERT [87], SciBERT [16] are pre-trained on large biomedical corpora through unsupervised tasks and then fine-tuned for specific tasks, including EL. Examples of these DL approaches include [112], BioSyn [237], BERN2 [238] (partially based on BioSyn), or DILBERT [166]. These models achieved SOTA performance in several EL benchmarks. These categories are not strict, since there are approaches that combine aspects of different categories. For instance, BioSyn [237] applies rules for abbreviation expansion, composite mention resolution, lowercasing and punctuation removal before generating the representation for the entity mentions with a DL model.

4.2.2 Predicting, clustering and typing NIL entities

Some recent studies proposed EL models that include modules for the prediction of NIL entities among the entire set of entity mentions (also called out-of-KOS prediction). Dredze et al. [65] proposed a model to link entities of several types (organizations, geopolitical and person entities) to entries of a KOS derived from Wikipedia, which also included a module to predict NIL entities. The model applied the Ranking SVM algorithm to rank the candidates for each mention according to a set of features (based on lexical, Wikipedia, popularity and document properties) and then adapted the SVM ranker to predict which entities were NIL according to predefined features. The ranker considered an entity as NIL if, for example, there was low string similarity between the entity and the respective KOS candidates and if the KOS candidates were not associated with any highly-ranked Wikitology pages (an ontology derived from the Wikipedia category system). DILBERT [166] integrates a module to predict out-of-KOS entity mentions in clinical trials by applying three strategies based on the distance of false-positive instances that differ on the threshold value: threshold value equivalent to the minimum distance of false positives, threshold value equivalent to the maximum distance and threshold value equivalent to the average of the two previous values.

The Knowledge Base Population track at the "Text Analysis Conference" (TAC) 2011 [110] introduced the requirement to cluster NIL entities (or "NIL queries") after the EL step. Wu et al. [264] divided the approaches to NIL entity clustering into three categories: string matching, hierarchical agglomerative clustering and graph-based. String matching based methods calculate a similarity measure between the surface forms of the predicted NIL entities and then cluster the entities sharing high similarity. The hier-

archical agglomerative clustering algorithm assigns the NIL entities to random clusters and iterates until the calculated distances between clusters are lower than a predefined threshold. Graph-based methods attempt to factor in the semantic relations between NIL entities. These methods usually create a semantic graph of NIL entities and then apply similarity measures or the hierarchical agglomerative clustering algorithm to create the clusters. More recently, Blissett and Ji [22] proposed a RNN architecture to perform cross-lingual NIL entity clustering.

Another approach consists of classifying unlinkable entities according to pre-defined types, which is designated by typing. Lin et al. [148] focus on the unlinkable noun phrase problem: if a noun phrase cannot be linked to any Wikipedia article or entity, the goal is first to determine if it is an entity and, if so, then classify it according to 1,339 Freebase semantic types. The authors used noun phrases extracted from the ClueWeb09Web corpus that are involved in relations described in the text, linked the noun phrases to Wikipedia, obtained a dataset with linked and unlinkable entities, and developed a classifier to distinguish between the two types (entities, non-entities) based on the usage patterns across time. Then they propagated the semantic types of the linked entities to the unlinkable entities by developing an algorithm that leverages, among other features, the relations described in text: the main idea is that two entities, one linkable and the other unlinkable, that appear in the same type of relations must also share the respective semantic types.

The task of emerging entity recognition or detection is a closely related one. The third edition of the *Workshop on Noisy User-generated Text* (WNUT2017) focused precisely on this task [58]. The evaluation dataset consisted of documents including mainly emergent or rare entities belonging to several types, such as person, location, corporation, product, creative-work and group.

There are differences between our work and the models previously described. They focus exclusively on NIL entities present in general text, whereas our focus was to develop an approach to deal with NIL entities in the biomedical domain. Biomedical KOSs have fewer available features compared to Wikipedia, thus specific approaches are required. For instance, the approach proposed by Lin et al. [148] explores Wikipedia-specific features, such as Freebase semantic types associated with an article, and leverages entities that are linked to Wikipedia to propagate their semantic types to unlinkable entities through relations described in text, whilst our approach does the opposite, i.e, uses NIL entity linking to improve EL. Färber et al. [71] also focus on Wikipedia entities, but only on emergent entities, which

are mostly of the type person. None directly determines the impact that the task of linking NIL entities can have on the broader EL task. So, our modelling of the problem is different, as well as training and evaluation methodologies, as it will be further described.

4.2.3 Attention models

The review by Galassi et al. [76] provides a comprehensive overview of attention mechanisms used in neural networks in the context of natural language processing and define a "core attention model", which contains elements that, according to the authors, are present in most attention architectures. In essence, an attention function maps an input sequence of keys (such as word or character embeddings) to a distribution of attention weights. A compatibility function returns the energy scores calculated based on the relevance of the keys with respect to an input query, and then the distribution function computes the attention weights for the keys according to the energy scores.

The seminal paper by Vaswani et al. [252] introduced the Transformer architecture, which represented an improvement over SOTA approaches in language modelling, until then based on RNNs. Transformer-based models are able to compute representations for input and output data using a self-attention mechanism. A recent trend in natural language processing consists of pre-training language representation models on large corpora and then fine-tuning them to specific downstream tasks. It is the case of BERT [59], which, as the name suggests, leverages a multi-layer bidirectional Transformer encoder and then it performs two unsupervised tasks (masked language model and next sentence prediction) on large corpora as training objectives. Other BERT-based models have also been trained on domain-specific corpora [140, 7, 16]. The self-attention mechanism in BERT allows the fine-tuning of the pre-trained model to several downstream tasks, including EL [112, 144, 273]. Besides BERT-related models, other attention-based models for EL have been proposed [179].

For the present work, the "Semantic compositionality with Mutual Sememe Attention" (SCMSA) model proposed by Qi et al. [194] for modelling semantic compositionality of multi-word expressions is of great relevance. The principle of semantic compositionality asserts that the meaning of a syntactically complex expression can be expressed as a function of the meaning of its syntactic parts [188]. Qi et al. [194] retrieved sememes, indivisible semantic units of meaning in human languages, for each word from an external KOS, HowNet [63]. This is a common-sense KOS including Chinese and English words

and a limited list of sememes, which can be combined in different ways to express the meaning of all words belonging to the KOS vocabulary. The main goal of the authors in the referred article was to obtain meaningful vector representations for multi-word expressions, more concretely, expressions that contain two different words. To accomplish that, they proposed an attention-based that assigns different weights for the sememes associated with the constituent words of a multi-word expression in order to build an embedding representation that reflects the relative importance of each sememe. This reasoning was the motivation to build NILINKER, as will be further described.

4.3 Methodology

4.3.1 Problem definition

The problem of linking a NIL entity to a KOS can be viewed as the selection of the most relevant KOS concept from the respective candidates' list, assuming that a perfect concept to describe the entity does not exist in the KOS. NIL entity linking is thus formulated in the following way: for each entity defined as NIL $NIL_e \in E$, being E the set of named entities mentioned in a given document, there is a candidates' list $CL(NIL_e) = \{c_1, ..., c_i\}$ retrieved from the KOS, and the goal is to find the candidate $c_i \in CL$ that partially disambiguates NIL_e. In the present work we will model the problem of NIL entity linking as a multi-class classification, since each instance will be assigned the most relevant concept or class from a list including multiple classes, i.e., all the concepts belonging to the considered KOS. However, in some cases, it may be helpful to obtain a set of several probable candidates instead of a single one, so the model needs to have the flexibility to return a variable number of KOS candidates. One of the criteria to define a given entity as NIL is the fact that common methods for candidate retrieval from a KOS return an empty list or below an acceptable confidence threshold, so it is necessary to find an alternative approach to build the candidates' list for NIL entities.

4.3.2 Candidate retrieval for NIL entities

In the context of the present work, NIL entities are the equivalent to the multi-word expressions considered in Qi et al. [194]. The authors resorted to HowNet to build the candidate sememes list for the words in the expression, however, this is not possible to replicate for NIL entities using HowNet for two reasons. First, the goal is to find the relative importance of candidate KOS concepts for NIL entities, and

not of sememes for multi-word expressions, so HowNet is not useful. Second, HowNet only includes common English words, missing domain-specific words that appear in biomedical text, the focus of this work.

To overcome this, it is proposed the construction of a KOS-derived word-concept dictionary: each word appearing in concept names and synonyms fields is associated with the respective concept ids. The approach includes the following steps:

1. Tokenization of the text present in *Name* and *Synonyms* fields for all concepts in the KOS, exclusion of stop words which are not relevant (e.g. "and", "or, "with"), and generation of a list with the remaining words.

2. Lemmatization and normalization of words present in the list. For example, the adjective "Parkinsonian" would be converted into the normalized lemma "parkinson".

3. Addition of the resulting words (keys) and the respective concept ids (values) where they appear to the KOS word-concept dictionary.

For the case of a NIL entity constituted by two words, w_1 and w_2, the respective candidates' list is expressed by:

$$CL(NIL_e) = CL(w_1) + CL(w_2) \tag{4.1}$$

Where $CL(w_1) = \{c^1_{w_1}, ..., c^i_{w_1}\}$ and $CL(w_2) = \{c^1_{w_2}, ..., c^i_{w_2}\}$.

An input NIL entity is tokenized and, for each word of its words, it is retrieved the respective candidates' list from the word-concept dictionary. Then, it is necessary to generate vectorized representations for both words and candidates to allow their input to the attention-based disambiguation model.

4.3.3 Representation of words and KOS concepts with embeddings

4.3.3.1 Word embeddings

Word embeddings are dense, low-dimensional, distributed representations of words in a vector space, in which similar words are grouped. The Word2vec [169] method is based on Skip-gram, which is an unsupervised technique to learn embeddings from large amounts of unstructured text. Word2vec learns

one vector per word, without considering the internal structure of words, which may decrease the quality of the representations in domains where it is common the existence of out-of-vocabulary and rare words, such as the biomedical one. BioWordVec [278] attempts to overcome these limitations, by using a subword embedding model to learn word representations from MeSH[4] terms and biomedical literature text. In the present work, we used pre-trained Word2vec and BioWord2Vec embeddings through Gensim[5] Python library. Considering an example of a NIL entity with two different words, the embeddings relative to word 1 and to word 2 are denoted as $\mathbf{w1}, \mathbf{w2} \in \mathbb{R}^d$, being d the dimension of embeddings. In cases where the NIL entity has only one constituent word, it is assumed that the NIL entity has two repeated words, i.e., the constituent word is simultaneously word 1 and word 2, and the procedure is the same. If the entity has more than 2 words, the models truncates it and only considers the first two words.

4.3.3.2 Concept embeddings

node2vec [84] is an algorithm to build feature vector representations for the nodes and edges of a network. A KOS is a network, where its concepts are the nodes and the relations between nodes are the edges. We used the code provided by the authors[6] to build embeddings for the concepts belonging to several KOS. Relations between concepts were in the input file with the format $(child\_concept) \rightarrow (parent\_concept)$, one relation per line. The output file contains a 200-dimensioned vector for each KOS concept. The parameters values are: "dimensions"=200 (128), "directed" ("undirected"), default values for the rest of the parameters. Considering the same example of a NIL entity with two different words, the set of embeddings relative to the KOS candidates associated with word 1 are defined as $w'_1 = \{\mathbf{c^1_{w_1}}, ..., \mathbf{c^i_{w_1}}\}$, with $c^i_{w_1} \in CL(w_1)$ and $\mathbf{c^i_{w_1}} \in R^d$. Conversely, the set of embeddings relative to the KOS candidates associated with word 2 are defined as $w'_2 = \{\mathbf{c^1_{w_2}}, ..., \mathbf{c^i_{w_2}}\}$, with $c^i_{w_2} \in CL(w_2)$ and $\mathbf{c^i_{w_2}} \in R^d$.

4.3.4 NILINKER: disambiguation of NIL entities

Qi et al. [194] proposed an approach to model semantic compositionality using sememes and have applied their model in a downstream sememe prediction task, which, as its name suggests, consists in the discovery of the most relevant sememes to represent the meaning of a multi-word expression. Similarly, the

[4] https://meshb.nlm.nih.gov/search
[5] https://radimrehurek.com/gensim/
[6] https://github.com/aditya-grover/node2vec

idea is that NIL entities can be viewed as multi-word expressions, and their meaning can be expressed through the most relevant KOS concepts. In the context of this work, the KOS-derived word-concept dictionary is the equivalent of HowNet.

Following the definition by Galassi et al. [76], the attention mechanism learns the relevance of the different parts of the input, so the development of an attention-based model contributes to the aforementioned goal. The core attention model defined by Galassi et al. [76] includes the following elements: *K*, keys or vectors which are the target of the attention weights, *e*, the energy scores, *f*, the compatibility function, and *g*, the distribution function, being the output a set of attention weights.

Considering the case of a NIL entity including word 1 and word 2, the attention weight for a KOS candidate associated with word 1 is obtained through the following distribution function:

$$a_{2,i} = \frac{exp(\mathbf{c^i_{w_1}} \cdot \mathbf{e_1})}{\sum_{\mathbf{c_j} \in w'_2} exp(\mathbf{c^j_{w_2}} \cdot \mathbf{e_1})} \tag{4.2}$$

Where $\mathbf{c^i_{w_1}}$ is the vector of candidate i for word 1, $\mathbf{c^j_{w_2}}$ is the vector of candidate j for word 2, and $\mathbf{e_1}$ is the vector of energy scores for word 1. This is obtained through the compatibility function:

$$e_1 = tanh(\mathbf{W_a w_1} + \mathbf{b_a}) \tag{4.3}$$

Where $\mathbf{W_a} \in R^{d \times d}$ is the weight matrix, $\mathbf{w_1} \in R^d$ is the vector of word 1, and $\mathbf{b_a} \in R^d$ is the bias vector. Conversely, it is possible to calculate $a_{1,j}$ for each $j \in CL(w_2)$, as well e_2.

The aggregated embeddings for word 1 candidates are denoted by w'_1 and the aggregated embeddings for word 2 candidates are denoted by w'_2. A single-layer perceptron classifies the KOS candidates:

$$\hat{y}_p = \sigma(\mathbf{W}_{KB} \cdot \mathbf{p}) \tag{4.4}$$

In which σ is the softmax function, $\mathbf{W}_{KB} = w'_1 + w'_2$, $\mathbf{p} = w_1 + w_2$, and $\hat{y}_p$ is the probability distribution for the KOS concepts. The higher the score, the higher the probability that the respective candidate is the correct disambiguation, consequently, the highest scoring candidate disambiguates the NIL entity. The full overview of the proposed model is available in Figure 4.1. The parameter `top_k` determines the number of of top-k candidates according to their probabilities to be returned for a given input entity.

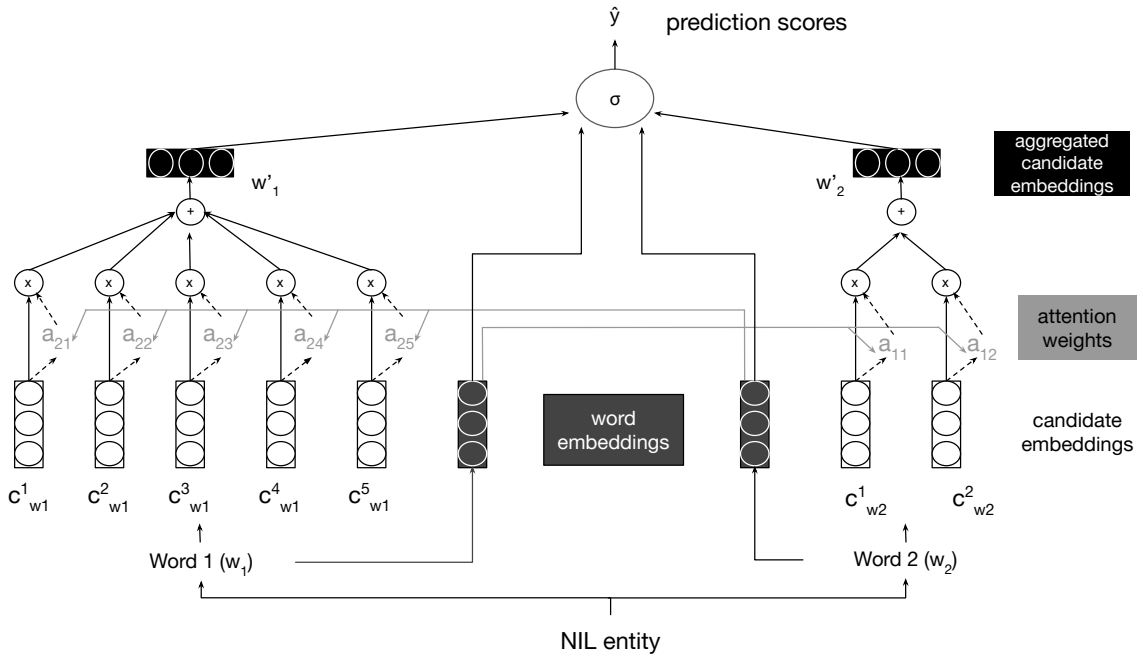


Figure 4.1: Architecture of the proposed NILINKER model to find the most relevant KOS concept for a NIL entity.

To calculate the training loss we use Softmax Cross Entropy Loss H:

$$H(y, \hat{y}_p) = -\sum_{x} y(x) log \hat{y}_p(x) \tag{4.5}$$

Where y is the true distribution of labels for given instance x (the true label corresponds to "1", the remaining to "0"), $\hat{y}_p$ is the estimated probability distribution outputted by the model. Therefore, the equation returns the distance between expected and model prediction.

Additionally, we modified the loss function by adding class weighting to further reduce the impact of class imbalance using the scikit-learn implementation `compute_class_weight` [187]. The vector including the weight for each label is obtained by:

$$ClassWeights = \frac{n_{samples}}{n_{classes} * B} \tag{4.6}$$

Where $n_{samples}$ corresponds to the total number of values present in y, $n_{classes}$ corresponds to the total number of classes/labels present in y, B corresponds to the set including the frequency values associated

with each label present in y (e.g. the index 0 corresponds to the label 0 and the value associated with the index 0 corresponds to the number of times that the label 0 is present in y). This means that frequent classes/labels in the dataset are assigned smaller weights, consequently have less impact on the loss value, whereas rare classes are assigned higher weights and have more impact on the loss.

One of the challenges hindering the development and implementation of the proposed model was the lack of training datasets, which was solved through the generation of a new dataset, as it is further described.

4.3.5 Data

Data used in the context of the present work consisted of files associated with commonly used biomedical KOSs and of annotated datasets belonging to the same domains.

4.3.5.1 KOS files

The files related with the following KOS were used: MEDIC vocabulary (version: 2022-04-04) [54], CTD-Anatomy (CTD-Anat) (version: 2022-04-04) [54], CTD-Chemical vocabulary (version: 2022-04-04) [54], ChEBI ontology (version: 13-Oct-2021 14:23) [97], GO (2021-09-01 release) [45], and HPO (2021-10-10 release) [126].

4.3.5.2 Datasets

Dataset	Set	Docs	Annots	1 word	2 words	3 words	4 words	5 words
BC5CDR-Disease [146]	Test	500	4,287	2,226 (51.9244 %)	1,264 (29.4845 %)	619 (14.4390 %)	107 (2.4959 %)	71 (1.6562 %)
	Test-Refined	322	692	171 (24.7110 %)	288 (41.6185%)	134 (19.3642 %)	56 (8.0925 %)	43 (6.2139 %)
BC5CDR-Chemical [146]	Test	500	5,015	4,144 (82.6321 %)	668 (13.3200 %)	93 (1.8544 %)	75 (1.4955 %)	35 (0.6979 %)
	Test-Refined	230	447	277 (61.9687 %)	111 (24.8322 %)	26 (5.8166 %)	15 (3.3557 %)	18 (4.0268 %)
NCBI Disease [64]	Test	100	960	239 (24.8958 %)	378 (39.3750 %)	198 (20.6250 %)	88 (9.1667 %)	57 (5.9375 %)
	Test-Refined	71	208	25 (12.0192 %)	76 (36.5385 %)	45 (21.6346 %)	32 (15.3846 %)	30 (14.4231 %)
GSC+ [85, 151]	All	228	2,773	1,079 (38.9109 %)	1,020 (36.7832 %)	403 (14.5330 %)	131 (4.7241 %)	140 (5.0487 %)
CHR [214]	Test	3,611	30,556	27,235 (89.1314 %)	3,236 (10.5904 %)	50 (0.1636 %)	31 (0.1015 %)	4 (0.0131 %)
	Test-Refined	233	526	444 (84.4106 %)	73 (13.8783 %)	7 (1.3308 %)	2 (0.3802 %)	0 (0 %)
PHAEDRA [243]	Test	115	1,633	1,527 (93.5089 %)	85 (5.2051 %)	18 (1.1023 %)	2 (0.1225 %)	1 (0.0612%)
	Test-Refined	37	177	150 (84.7458 %)	16 (9.0395 %)	9 (5.0847 %)	1 (0.5650 %)	1 (0.5650 %)

Table 4.1: Statistics for the evaluation EL datasets. Docs: Documents, Annots: Annotations, 1 word: Entities with 1 word, 2 words: Entities with 2 words, 3 words: Entities with 3 words, 4 words: Entities with 4 words, 5 words: Entities with 5 words

We have used the gold standard EL datasets that are described in Table 4.1. Based on the work by Tutubalina et al. [248], for evaluation we used the original test set and, to remove the bias, a refined test set that excludes entity mentions that are present in the training and development sets. The source of the datasets BC5CDR-Disease, BC5CDR-Chemical and NCBI Disease was the repository `https://github.com/insilicomedicine/Fair-Evaluation-BERT`. For the remaining datasets we have generated the refined test set excepting GSC+: we used the whole dataset since it does not have separated sets.

In order to train the NILINKER model, it is necessary a specific dataset including NIL entities associated with KOS concepts. The datasets used to build the EvaNIL dataset (see Subsection 4.3.5.3) were: PubMedDS [251] (silver standard, 13,197,430 PubMed abstracts with 57,943,354 biomedical annotations, MeSH, including MEDIC, CTD-Chemical and CTD-Anatomy concepts), CRAFT corpus [28] (gold standard, 97 PubMed articles with chemical and biological process entities, including ChEBI and GO concepts), and MedMentions [174] (gold standard, 4,392 PubMed abstracts with biomedical annotations, UMLS concept with cross-links to HPO concepts).

4.3.5.3 EvaNIL: large silver standard for NIL entity linking evaluation

As there are no datasets for evaluation of the NIL entity linking task and the development of gold standard corpora from scratch is both time and effort-intensive, we generated a large dataset from existing annotated datasets for EL evaluation. These datasets contain annotations with entities and the respective KOS concepts that disambiguate them, with format (entity, concept ID), however, they usually do not include NIL annotations, which is not representative of a real application scenario where there are always unlinkable entities.

To leverage the "linkable" entities already present in the annotations of those corpora, we assumed that the associated concept is absent from the respective KOS and, consequently, that the entity then should be linked to the direct ancestor of the annotated concept. This way, linkable entities were transformed into NIL entities. Applying this strategy, annotations present in several corpora (CRAFT corpus, MedMentions, and PubMedDS) were converted into the format (entity, direct ancestor concept ID). UMLS ids present in MedMentions corpus were mapped to HPO ids since this is a freely available ontology.

The final dataset includes six different sets, which aggregate annotations containing concept ids from

Partition	Subset	Docs	Annots	1 word	2 words	3 words	4 words	5 words
MEDIC	Train	294,926	379,041	242,978 (64.10%)	115,961 (30.59 %)	18,176 (4.80 %)	1,663 (0.44 %)	263 (0.07 %)
	Dev	62,000	80,143	1,347 (64.07 %)	24,292 (30.31 %)	4,122 (5.14 %)	330 (0.41 %)	52 (0.06 %)
	Test	62,000	80,050	51,734 (64.63 %)	24,175 (30.20 %)	3,788 (4.73 %)	311 (0.39 %)	42 (0.05 %)
	Test-Refined	147	151	54 (35.76 %)	77 (50.99 %)	14 (9.27 %)	4 (2.65 %)	2 (1.32 %)
CTD-Chemical	Train	10,749	15,755	12,701 (80.62 %)	2,560 (16.29 %)	420 (2.67 %)	66 (0.42 %)	8 (0.05 %)
	Dev	4,000	5,655	4,613 (81.57 %)	834 (14.75 %)	165 (2.92 %)	40 (0.71 %)	3 (0.05 %)
	Test	4,000	5,922	4,800 (81.05 %)	1,002 (16.92 %)	100 (1.69 %)	16 (0.27 %)	4 (0.07 %)
	Test-Refined	329	356	233 (65.45 %)	107 (30.06 %)	12 (3.37 %)	3 (0.84 %)	1 (0.28 %)
CTD-Anat	Train	282,153	343,262	223,452 (65.10 %)	109,917 (32.02 %)	9,346 2.72 %)	113 (0.03 %)	434 (0.13 %)
	Dev	62,000	75,578	49,437 (65.41 %)	23,906 (31.63 %)	2,130 (2.82 %)	17 (0.02 %)	88 (0.11 %)
	Test	62,000	75,925	49,871 (65.68 %)	24,002 (31.61 %)	1,937 (2.55 %)	29 (0.04 %)	86 (0.11 %)
	Test-Refined	78	79	19 (24.05 %)	55 (69.62 %)	4 (5.06 %)	1 (1.27 %)	0 (0 %)
ChEBI	Train	69	1,080	904 (83.70 %)	151 (13.98 %)	21 (1.94 %)	2 (0.19 %)	2 (0.19 %)
	Dev	14	196	157 (80.10 %)	24 (12.24 %)	8 (4.08 %)	7 (3.57 %)	0 (0 %)
	Test	14	204	162 (79.41 %)	37 (18.14 %)	5 (2.45 %)	0 (0 %)	0 (0 %)
	Test-Refined	13	61	38 (62.30 %)	19 (31.15 %)	4 (6.56 %)	0 (0 %)	0 (0 %)
HPO	Train	1,501	4,052	2,395 (59.11 %)	1,144 (28.23 %)	387 (9.55 %)	101 (2.49 %)	25 (0.62 %)
	Dev	321	850	483 (56.82 %)	248 (29.18 %)	104 (12.24 %)	10 (1.18 %)	5 (0.59 %)
	Test	321	847	492 (58.09 %)	241 (28.45 %)	93 (10.98 %)	18 (2.13 %)	3 (0.35 %)
	Test-Refined	160	296	101 (34.12 %)	133 (44.93 %)	50 (16.89 %)	10 (3.38 %)	2 (0.68 %)
GO-BP	Train	69	2,413	1,400 (58.02 %)	525 (21.76 %)	206 (8.54 %)	176 (7.29 %)	106 (4.39 %)
	Dev	14	377	206 (54.64 %)	87 (23.08 %)	44 (11.67 %)	29 (7.69 %)	11 (2.92 %)
	Test	14	426	262 (61.50 %)	100 (23.47 %)	29 (6.81 %)	19 (4.46 %)	16 (3.76 %)
	Test-Refined	14	139	38 (27.34 %)	48 (34.53 %)	25 (17.99 %)	13 (9.35 %)	15 (10.79 %)

Table 4.2: Statistics for the EvaNIL dataset. Docs: Documents, Annots: Annotations, 1 word: Entities with 1 word, 2 words: Entities with 2 words, 3 words: Entities with 3 words, 4 words: Entities with 4 words, 5 words: Entities with 5 words

the following KOSs: GO, ChEBI, MEDIC, CTD-Chemical, HPO, CTD-Anat. To reduce the bias, we also generated a test-refined set for each partition that excludes entity mentions appearing in the respective training and development sets. To reduce the class imbalance in the final dataset, repeated annotations in the same document were eliminated. The final dataset is available for download[7] and its statistics are shown in Table 4.2.

4.3.6 Experimental setup

The experimental setup was divided in two phases: evaluation on the EvaNIL dataset and evaluation of the impact in the EL task. An overview of the experimental setup is shown in Fig 4.2.

[7] `https://zenodo.org/record/5849231`

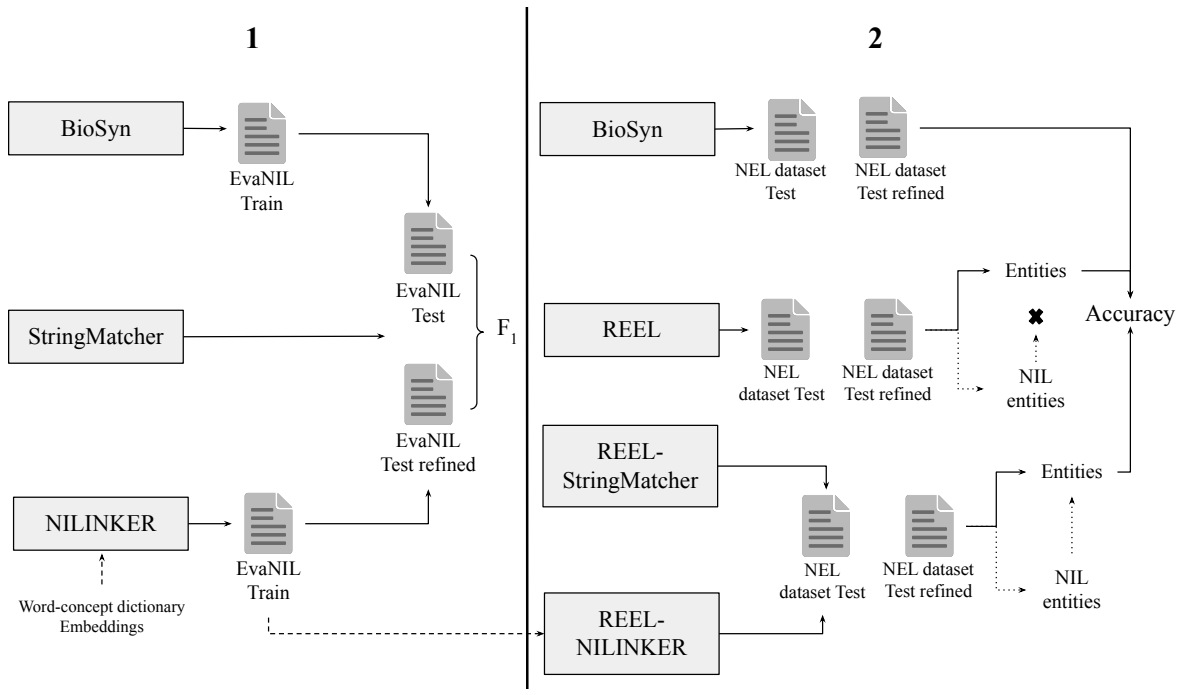


Figure 4.2: Experimental setup: 1) evaluation on the EvaNIL dataset, 2) determination of the impact in the EL task of integrating the trained NILINKER model with a EL model.

4.3.6.1 Evaluation on the EvaNIL dataset

The following models were evaluated on each partition of the EvaNIL dataset, more concretely on the test and on the test-refined sets:

- StringMatcher: a simple baseline that associates each NIL entity with the most similar concept in the respective target KOS according to lexical distance.
- BioSyn: model with SOTA performance in the EL task [237] it learns representations for biomedical entity mentions using synonym marginalization and iterative candidate retrieval. The original BioBERT-based implementation was trained using the same parameters used by the authors.
- NILINKER

The models BioSyn and NILINKER were trained and validated on the training and development sets and evaluated on the test and test-refined sets of the respective EvaNIL partition. For the model BioSyn,

we used the default values described by its authors [237]: number of top candidates k of 20, mini-batch size of 16, learning rate of $1e^{-5}$, dense ratio of 0.5, training epochs equal to 10.

The model StringMatcher is only evaluated on the test and test-refined sets. In terms of evaluation metrics, True positives (tp) refer to the number of entities correctly linked, false positives (fp) to the number of entities wrongly linked and false negatives (fn) to the number of entities that the model does not link. As the EvaNIL dataset is class imbalanced, the performance of each model was evaluated through $precision$, $recall$ and the micro-averaged F_1 $score$.

We tested for statistical significance between the baseline model (StringMatcher) and the alternative models (either NILINKER or BioSyn) using the paired Wilcoxon Signed-Rank test, following the recommendation of Dror et al. [66], and, in case of detected statistical significance effect size, we also determined its magnitude through the Wilcoxon effect size (r) [233]. Concerning the paired Wilcoxon Signed-Rank test, for each pair of models being tested, we considered the following null (H_0) and alternative (H_1) statistical hypotheses:

- H_0: There is no difference between the performance of the baseline model and the performance of the alternative model.
- H_1: The performance of the baseline approach is different than the performance of the alternative model.

A difference is statistically significant when the p-value p is smaller than the significance level α ($p < \alpha$), being $\alpha = 0.05$ in the context of this work. We developed a R script to perform the statistical analysis based on the functions `wilcox_test` and `wilcox_effsize` of the `rstatix` package[8].

4.3.6.2 Impact in the EL task

The goal of this phase was to evaluate the trained NILINKER models on downstream tasks. Given the impact that NIL entities have on the performance of EL models, we studied the effect of integrating the NILINKER model with a EL model. We used the model REEL previously developed by our group (Chapter 3), which links biomedical entities (chemicals and diseases), to several biomedical KOSs, more concretely, ChEBI, MEDIC, and CTD-Chemical.

[8] `https://cran.r-project.org/web/packages/rstatix/index.html`

REEL includes two components: candidate generation (retrieval from the KOS of the most relevant candidates for a given entity) and candidate disambiguation (selection of the most relevant candidate). The candidates for the entities present in a given document form a disambiguation graph: each candidate is a node, the edges between the candidates are added according to existing relations in the structure of the KOS and, if available, to relations extracted directly from the respective text. The candidate disambiguation component is based on the PPR algorithm and on IC. An entity was considered to be NIL a) if the respective candidates' list was empty, b) if the correct candidate for the entity is not present in the retrieved candidate list or c) if there is no KOS id associated with it in the respective corpus. As REEL is a graph-based approach that is based on the maximisation of the coherence between the nodes/candidates we consider that it is useful to study the impact of adding nodes by the NIL entity linking models. This experiment evaluated the following models:

- BioSyn: model with SOTA performance in the EL task [237]. We used the SapBERT-based models already trained on the training and development sets of the BC5CDR-Disease[9], the BC5CDR-Chemical[10] and the NCBI Disease[11] datasets.
- REEL: baseline approach, ignores NIL entities.
- REEL-StringMatcher: StringMatcher associates each NIL entity with the most similar concept in the respective target KOS according to lexical distance.
- REEL-NILINKER: for each NIL entity, NILINKER returns the `top-k` most probable KOS concepts. The parameter `top-k` was optimized directly on each evaluation dataset, contrary to standard practice which represents a risk of overfitting.

We evaluated the models in gold standard datasets focusing on disease entities, such as BC5CDR-Disease, GSC+ and NCBI Disease and on chemical entities, such as BC5CDR-Chemical, CHR, PHAEDRA. The model BioSyn was evaluated on the datasets BC5CDR-Disease, BC5CDR-Chemical and NCBI Disease since these are popular datasets, which allows the comparison with other SOTA approaches, and because they include two of the most important types of biomedical entities, chemicals

[9] `https://huggingface.co/dmis-lab/biosyn-sapbert-bc5cdr-disease`
[10] `https://huggingface.co/dmis-lab/biosyn-sapbert-bc5cdr-chemical`
[11] `https://huggingface.co/dmis-lab/biosyn-sapbert-ncbi-disease`

and disease. The performance of each model is represented by the returned accuracy at 1 ($acc@1$) in each dataset, a common metric used in the EL tasks that allows the comparison with SOTA models. This metric calculates the proportion of cases where the correct KOS concept is among the top-1 concepts predicted.

We performed a statistical analysis of the performance of the three models in the evaluation datasets. We applied the same testing for statistical significance, the Wilcoxon Signed-Rank test. The baseline approach was the REEL model and the alternative approach was either the REEL-StringMatcher, REEL-NILINKER or the BioSyn models.

4.3.7 Implementation

The NILINKER model was implemented using Tensorflow 2.5.1 and Python 3.8.6. The two phases of the experimental setup were executed on a Tesla M10 GPU. For each partition of the EvaNIL dataset, we performed a manual hyperparameter optimization before training, focusing on the parameters `learning_rate`, `optimizer`, `train_batch_size`, and `test_batch_size`. The training and evaluation times for the final versions of the NILINKER model were approximately 0.2, 1.4, 5.2, 10, 1.2 and 1.7 hours for the partitions HPO, ChEBI, MEDIC, CTD-CHEM, GO-BP and CTD-ANAT, respectively.

4.4 Results and discussion

4.4.1 Results

The results relative to the evaluation on the EvaNIL dataset are shown in Table 4.3. On the original test sets, the best result was achieved by NILINKER in the MEDIC partition (0.9401 $F_1\ score$) and the worst result by the BioSyn model in the ChEBI partition (0.0194 $F_1\ score$) . On the test-refined sets, the best result was achieved by NILINKER in the MEDIC partition (0.7149 $F_1\ score$) and the worst result by the BioSyn model in the GO-BP partition (0.0284 $F_1\ score$).

The results relative to the second evaluation, the integration of NILINKER with the REEL model, are shown in Table 4.4. The best model in all datasets was BioSyn. As expected, the results for the test-refined sets were worse than the results for the original Test sets. In three of the datasets (BC5CDR-Disease, NCBI-Disease and PHAEDTA) the performance of the REEL-NILINKER model was higher compared with the baseline model REEL, although the difference in the performance was only statistically

Partition	Model	Test			Test-refined		
		P	R	F1	P	R	F1
HPO	StringMatcher	0.0295	**1.0000**	0.0573	0.0439	**1.0000**	0.0841
	BioSyn	0.0473	0.9524	0.0902*** (0.1330)	0.0578	0.8947	0.1086^{ns} (0.1160)
	NILINKER	**0.4274**	**1.0000**	**0.5988**** (0.6310)	**0.1351**	**1.0000**	**0.2381**** (0.3020)
ChEBI	StringMatcher	0.0343	**1.0000**	0.0664	0.0492	**1.0000**	0.0938
	BioSyn	0.0098	**1.0000**	0.0194* (0.1570)	0.0164	**1.0000**	0.0323^{ns} (0.1810)
	NILINKER	**0.3971**	**1.0000**	**0.5684**** (0.6020)	**0.0820**	**1.0000**	**0.1515**^{ns} (0.1810)
MEDIC	StringMatcher	0.0450	**1.0000**	0.0861	0.0530	**1.0000**	0.1006
	BioSyn	0.0533	0.8364	0.1002**** (0.0879)	0.0530	**1.0000**	0.1006^{ns} (0)
	NILINKER	**0.8869**	**1.0000**	**0.9401**** (0.9180)	**0.5563**	**1.0000**	**0.7149**** (0.7090)
CTD-Chemical	StringMatcher	0.0336	**1.0000**	0.0650	0.0674	**1.0000**	0.1263
	BioSyn	0.0111	**1.0000**	0.0220^{ns} (0.0225)	0.0253	**1.0000**	0.0493*** (0.2050)
	NILINKER	**0.3242**	**1.0000**	**0.4897**** (0.5390)	**0.0955**	**1.0000**	**0.1744**** (0.1680)
GO-BP	StringMatcher	0.0117	**1.0000**	0.0232	0.0360	**1.0000**	0.0694
	BioSyn	0.0047	**1.0000**	0.00934^{ns} (0.0839)	0.0140	**1.0000**	0.0284^{ns} (0.1470)
	NILINKER	**0.4742**	**1.0000**	**0.6433**** (0.6800)	**0.1079**	**1.00000**	**0.1948**** (0.2680)
CTD-ANAT	StringMatcher	0.0323	**1.0000**	0.0626	0.0759	**1.0000**	0.1412
	BioSyn	0.0311	0.9692	0.0603**** (0.0348)	0.1039	0.8000	0.1839^{ns} (0.1590)
	NILINKER	**0.9156**	**1.0000**	**0.9560**** (0.9400)	**0.4937**	**1.0000**	**0.6610**** (0.6460)

Table 4.3: Evaluation results on each partition of the EvaNIL dataset for the NIL entity linking task, along with the significance levels obtained from the paired Wilcoxon Signed-Rank test (comparison with the baseline performance) and the Wilcoxon effect size (between parenthesis) in the F_1 columns. In the NIL entity linking task, we assume that the gold label does not exist in the target KOS so the goal is to associate the NIL entity with the direct ancestor of the original gold label. Highest value by metric for each subset (test and test-refined) is highlighted. P: Precision; R: Recall; F1: F1-score; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.0001$; ****: $p \approx 0$.

significant in two of those datasets (BC5CDR-Disease and NCBI-Disease). The optimal values for the parameter `top_k` in the datasets BC5CDR-Disease, GSC+, NCBI Disease, BC5CDR-Chemical, CHR and PHAEDRA are 2, 6, 2, 5, 20 and 5, respectively.

4.4.2 Discussion

The results of the evaluation on the EvaNIL dataset show that the NILINKER model still have room for improvement, in particular in the linking of ChEBI, CTD-Chemical and HPO entities. The number of documents used for training the models NILINKER-CTD-Chem (10,749), NILINKER-HPO model (1,501), NILINKER-ChEBI (69) and NILINKER-GO-BP (69) is substantially lower than the documents used for training the NILINKER-MEDIC model (294,926) and the NILINKER-CTD-ANAT model (282,153).

Dataset	Model	Test	Test-refined
BC5CDR-Disease	REEL	0.7462	0.5665
	REEL-StringMatcher	0.7464^{ns} (0.0153)	0.5751* (0.0931)
	REEL-NILINKER	0.7667**** (0.1430)	0.5882*** (0.1470)
	BioSyn	**0.9384**** (0.4380)	**0.7861**** (0.4690)
BC5CDR-Chemical	REEL	0.9378	0.8456
	REEL-StringMatcher	0.9386^{ns} (0.0282)	0.8479^{ns} (0.0473)
	REEL-NILINKER	0.9376^{ns} (0.0141)	0.8434^{ns} (0.0473)
	BioSyn	**0.9665*** (0.1250)	**0.8810*** (0.1250)
NCBI Disease	REEL	0.7250	0.4904
	REEL-StringMatcher	0.7396*** (0.1210)	0.4760^{ns} (0.1200)
	REEL-NILINKER	0.7333** (0.0913)	0.5096^{ns} (0.1390)
	BioSyn	**0.9219**** (0.4440)	**0.7692**** (0.5280)
GSC+	REEL	**0.6012**	-
	REEL-StringMatcher	0.5943**** (0.0828)	-
	REEL-NILINKER	0.6004^{ns} (0.0269)	-
CHR	REEL	0.6647	0.4106
	REEL-StringMatcher	**0.7280**** (0.2520)	0.4125^{ns} (0.0436)
	REEL-NILINKER	0.6647^{ns} (0.0057)	0.4106^{ns} (0)
PHAEDRA	REEL	0.7740	**0.1864**
	REEL-StringMatcher	0.7722^{ns} (0.0429)	0.1695^{ns} (0.1300)
	REEL-NILINKER	**0.7746**^{ns} (0.0247)	**0.1864**^{ns} (0)

Table 4.4: Impact in the EL task of applying a NIL entity linking model: $acc@1$ for the models BioSyn, REEL, REEL-StringMatcher, REEL-NILINKER in six different evaluation datasets (Test and Test-refined sets, along with the significance levels obtained from the paired Wilcoxon Signed-Rank test (comparison with the baseline performance) and the Wilcoxon effect size (between parenthesis). *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.0001$, ****: $p \approx 0$

So, we believe the relative lower performance of the former models is due to the smaller size of the training datasets. Since CTD-Chemical and ChEBI are large KOSs, we had to reduce the size of the training datasets due to memory constraints. So, expanding the training datasets while solving the memory issues may improve the performance of these models, and, consequently, increase their impact on downstream tasks.

BioSyn obtained a low performance in the NIL entity linking evaluation executed on the EvaNIL

dataset which was expected since BioSyn was originally developed for the EL task, whereas the NIL entity linking task has a distinct goal: associate a given NIL entity not with the most appropriate KOS concept but with its direct ancestor.

Counterintuitively, the performance of the BioSyn and of the StringMatcher models was lower in the test-refined sets of the EvaNIL dataset than the original test sets. This is partially related with the imbalance of the dataset. In the test set of the EvaNIL-MEDIC partition, the 10 most common annotations and respective frequencies are: “hypertension”, (2,576), “stroke” (1,821), “inflammation” (1,820), “schizophrenia” (1,613), “pain” (1,376), “heart failure” (1,258), “body weight” (1,089), “recurrence” (1,037), “insulin resistance” (1,018), “multiple sclerosis” (920). By its hand, in the test-refined set of the same partition, the 10 most common annotations and respective frequencies are: “LYMPHEDEMA” (4) “Munchausen Syndrome” (3) “CONSTIPATION” (2), “anticholinergic syndrome” (2), “SARCOPENIA” (2), “Granulomatous Mastitis” (2), “HYPEREOSINOPHILIC SYNDROME” (2), “calcium metabolism disorders” (2), “Neglected Diseases” (2), “No-Reflow Phenomenon” (2). So, the class imbalance in the test-refined set is substantially lower than the original test set. The performance of StringMatcher and BioSyn in the original test set is not as high as expected, but in the test-refined sets is comparatively higher due to the lower repetition of annotations. For example, BioSyn wrongly predicts the gold label of the annotation “hypertension”, the most common annotation in the test set: the top prediction is the label “MESH:D000075222”, but the gold label in the dataset is “MESH:D014652”. This corresponds to 2,576 wrong predictions, whereas in the test-refined set this annotation is not even present. This means that, in the test-refined set, the weight of the wrong predictions is lower than the weight of the wrong predictions in test set.

Although the difference in the performance of the NILINKER model compared with the baseline REEL is small (effect sizes of 0.143 in the BC5CDR-Disease dataset and 0.0913 in the NCBI Disease dataset), the difference is consistent since it was observed in evaluation datasets with different characteristics. We did our best to present an extensive evaluation as possible, so the results are probably generalizable to other entities in the biomedical domain.

We performed an extensive analysis to find the factors behind the impact of NILINKER in the EL task. The analysis focused on the following questions:

1. Is the impact of NILINKER in the EL task associated with the percentage of NIL entities in the

evaluation dataset?

2. Is the impact of NILINKER in the EL task associated with its performance determined on the respective partition of the EvaNIL dataset (determined in the test set)?

3. Does the presence of more or less general entities in the evaluation dataset influence NILINKER's impact in the EL task? NILINKER was trained to link a given entity to more general entities in the target KOS, but if general entities were already present in a dataset, NILINKER would just link the entities to the root concept of a KOS. Some entities were considered as NIL by the REEL model, but they are still associated with a KOS identifier in the respective dataset, which generally does not correspond to the root concept identifier.

4. How exactly NILINKER impacts the EL task? In each evaluation dataset, there are entities that REEL consider as NIL, even if they are associated with a given KOS id. NILINKER retrieves the top-k candidates for those entities. Does the correct KOS ID for the NIL entities is present in the retrieved candidates' list?

Table 4.5 includes the result of the analysis done to answer the questions above. The impact of the NILINKER model is not associated with the number of NIL entities in the evaluation dataset. However, the higher the $F_1 score$ obtained on the EvaNIL evaluation, the higher the increase of performance in the EL task. The analysis also suggest that the impact of the NILINKER model is independent of the "specificity" (average number of ancestors) of the entities in the evaluation datasets.

With respect to the question 4, the results suggest that applying NILINKER increases the number of cases where the correct candidates are present in the candidates' lists of NIL entities. The datasets where the number of correct candidates increased more (BC5CDR-Disease and NCBI Disease) were also the datasets where the model REEL-NILINKER improved more the performance compared with the baseline REEL. One example showing the impact of the integrating NILINKER with the REEL model is shown in Figure 4.3.

The impact of NILINKER in the EL task is also indirect, i.e., the retrieved candidates are added to the disambiguation graph, this will include more candidates/nodes, but, more crucially, the graph will now include more edges between all the nodes/candidates. The number of edges in the graph is essential for

the application of the PPR algorithm. The score of each node/candidate is mostly based on the number of edges it is involved in the graph. The retrieved candidates, even if none of them corresponds to the correct candidate for a NIL entity, will increase the semantic information stored in the graph, which by its turn helps in the disambiguation of the other linkable entities.

In the EL task, the inclusion of a large number of candidates outputted by NILINKER for the NIL entities can increase the noise in the disambiguation graph, which leads to performance loss. The number of candidates proposed by NILINKER is specified by the parameter `top-k`, so it is crucial to find its optimal value for each dataset that NILINKER is being applied: a candidate list with too many candidates can increase noise, but a list with too few can also miss relevant semantic information.

Dataset	Target KOS	Entities	NILs	% NILs	Avg ancestors	F_1	Δ	NIL model	Solution in list %
BC5CDR-Disease	MEDIC	4,287	1,194	27.86 %	3.5	0.9401	$+0.0205$	None	72.13
								NILINKER	73.97
GSC+	HPO	2,773	1,005	36.24 %	5.0	0.5988	-0.0008	None	63.76
								NILINKER	64.05
NCBI Disease	MEDIC	960	314	32.71 %	3.7	0.9401	$+0.0083$	None	67.30
								NILINKER	67.92
BC5CDR-Chemical	CTD-CHEM	5,381	1,547	28.75 %	4.7	0.4897	-0.0002	None	71.25
								NILINKER	71.27
CHR	CHEBI	30,556	12,532	41.01 %	8.0	0.5684	$+0.0000$	None	58.99
								NILINKER	58.99
PHAEDRA	CTD-CHEM	1,633	434	26.58 %	5.0	0.4897	$+0.0006$	None	73.42
								NILINKER	73.42

Table 4.5: Analysis of the results. Total entities and NILs count in each evaluation dataset (identified by the REEL model), as well the relative percentage of NILs to the number of unique entities present. F_1 refers to the $F_1 score$ value obtained by the NILINKER model on the respective EvaNIL partition. Δ refers to the difference in performance of the REEL-NILINKER model compared with the baseline model REEL in given dataset in terms of $acc@1$. Solution in list indicates the percentage of cases where REEL retrieved a non-empty candidates' list in given dataset

If an entity has more than two words, NILINKER truncates the entity and only considers the first two words, which leads to a loss of information of unknown impact. One way of improving the performance could pass by adapting the architecture of the model to accept any input entity size without resorting to truncation, thus preserving the input data. However, this adaptation would not be trivial, which puts it outside the scope of the present work, so we leave that task to future work.

For simplicity, we only included annotations with exact one direct ancestor in the EvaNIL dataset, which leaves out of the dataset a large part of the concepts of each KOS. This relative low annotation diversity hinders the training of the several NILINKER models, and consequently, the impact of these in

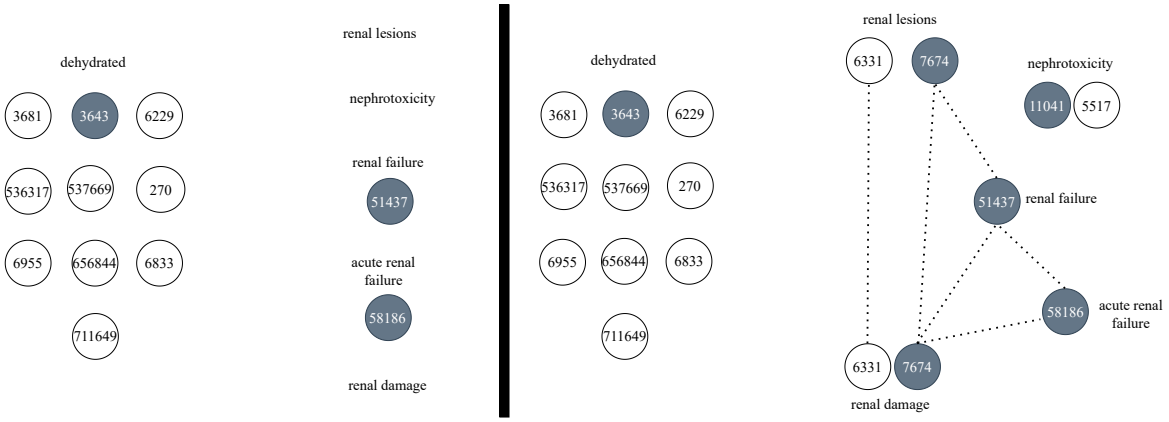


Figure 4.3: Example of the impact of NILINKER-MEDIC in the disambiguation graph of REEL (document "931801" of the dataset BC5CDR-Disease). In the left, it is present the disambiguation graph produced by the REEL model including the MEDIC candidates (circles) for three linkable entity ("dehydrated", "renal failure" and "acute renal failure") and also several NIL entities ("renal lesions", "nephrotoxicity", "renal damage") that have an empty candidates' list. Gray circles correspond to the highest scored candidates for the respective entity. In the right, it is present the disambiguation graph after applying the model REEL-NILINKER: the previous NIL entities now have associated candidates, which by its turn will be involved in more edges (dashed lines), increasing the semantic information of the graph. In this case the entities "renal lesions" and "renal damage" are correctly disambiguated to the candidate 7674 (MESH:D007674).

downstream tasks.

We used a simple rule to tokenize the entities in the generation of the word-concept dictionary and in the preprocessing of input entities that was based on white spaces and also on hyphens for chemical entities. This is not optimal for biomedical entities since these are frequently composed by complex rare words.

The EL datasets used in the evaluation only included in-KOS or linkable entities, but for future work it would be interesting to assess the performance of the REEL-NILINKER model in other datasets including out-of-KOS entities, such as the one proposed by Miftahutdinov et al. [166].

Another limitation of NILINKER is that it uses non-contextual embeddings to represent the words that are part of the input entities. The model encodes two entities sharing their string with the same vector, but their context, thus their meaning, may be different. Both limitations impact the first step of the proposed approach, candidate retrieval, and consequently, the following steps. Candidate retrieval is essential because the attention weights are assigned according to the candidates associated with the

entity, and the attention weights will influence the prediction scores.

4.5 Conclusion

In this work, we propose NILINKER, a model which can be used to partially link biomedical NIL entities to concepts belonging to several KOSs (MEDIC, HPO, CTD-Chemical, ChEBI, GO-BP and CTD-ANAT) and to improve the performance of EL models. NILINKER includes several components, namely a candidate retrieval from a word-concept dictionary, and a neural network leveraging the attention mechanism to determine the relevance of each KOS concept to the input NIL entity. We also propose a new evaluation dataset specifically developed for the NIL entity linking task (EvaNIL), and we provide the results of the evaluation of several models. The best results were obtained for the partition EvaNIL-MEDIC ($F_1 score$ of 0.9401). On the other hand, we integrated NILINKER with REEL, a graph-based EL model, and determined if the integration translated into an increase in performance in the EL task. The results of the evaluation done in several EL datasets show that integrating NILINKER with a EL model leads to an increase of performance. Thus, we can conclude that the answers to the initial research questions are positive.

We focused on answering the initial research questions, but there is still room to improve NILINKER. So, future work should focus on solving the limitations that NILINKER still presents. Possible directions include the use of contextual embeddings to improve the representation layer, the adaptation of the model to accept entities with more than two words, the use of tokenizers trained on biomedical text and, crucially, the improvement of the training dataset EvaNIL, by increasing the diversity of annotations that are present (for example, each entity could be associated with more distant ancestors in the respective KOS hierarchy). It would be also interesting to develop a pipeline for end-to-end Entity Extraction integrating NER, EL, and NIL entity linking.

Chapter 5

X-Linker: Hybrid Biomedical Entity Linking with XR-Transformer and automatically labelled data

Pedro Ruas

This chapter tackles objective 2 by describing a DL-based EL approach that reduces the need for human-annotated data in training. It corresponds to the following article submitted to a journal and currently under review with minor adaptations for consistency and clarity:

> **Ruas, P.**, Gallego, F., Veredas Navarro, F. J., & Couto, F. M. (2024). **X-Linker: Hybrid biomedical entity linking with XR-Transformer and automatically labelled data**. Submitted to *IEEE Transactions on Knowledge and Data Engineering* and currently under review [208]. Available as a preprint in `http://www.arxiv.org/abs/2407.06292`. Code repository publicly available: `https://github.com/lasigeBioTM/X-Linker`

Abstract. SOTA DL EL approaches require large amounts of costly human-labeled data. Existing datasets are limited in scale, resulting in low coverage of biomedical concepts and reduced performance of supervised EL models. Applying these models to different data leads to data drift, causing decreased linking accuracy. In this work we resort to automatic data to generate large-scale training datasets, which allows the exploration of approaches originally developed for the task of XMR in the biomedical EL task. We propose the hybrid X-Linker pipeline, that includes different modules to link disease and chemical entity mentions to concepts in the MEDIC and in the CTD-Chemical vocabularies, respectively. X-Linker was evaluated in several biomedical datasets, more concretely, BC5CDR-Disease, BioRED-

Disease, NCBI-Disease, BC5CDR-Chemical, BioRED-Chemical and NLM-Chem, reaching a top-1 accuracy of 0.8307, 0.7969, 0.8271, 0.9511, 0.9248, 0.7895, respectively. The source code of X-Linker and its associated data are publicly available for performing biomedical EL without the need for labelled entity mentions with identifiers from target KOSs.

5.1 Introduction

EL is the task of linking an entity mention in a given piece of text to an entry in a target KOS, such as an ontology, a knowledge base or graph, a terminology, etc. The entry must accurately represent the meaning of the linked entity. EL plays an essential role in text mining and natural language processing pipelines, since it connects text expressed in natural language to semantic, computer-friendly representations. For example, in the biomedical field, a large amount of information is stored in the form of clinical notes, expressed in natural language, which contain entities that must be standardized using ontologies such as SNOMED-CT or UMLS.

Challenges in the biomedical EL task include name variations (synonyms, acronyms), ambiguity (where the same name can denote different entities) [219], and the highly specialised language, which obstructs the utilization of complex resources typically available for general EL approaches, such as Wikipedia. The challenge of ambiguity is illustrated by the entity mention "iris", which can have several possible meanings for it: an eye-related anatomical structure, an insect or a plant taxonomic genus, a disease's acronym (immune reconstitution inflammatory syndrome) or a gene. Searching for "iris" in NCBI-Gene returns multiple homonymous results: "[Drosophila melanogaster (fruit fly)]" - `Gene ID: 33290`), "Iris iris [Tribolium castaneum (red flour beetle)]" - `Gene ID: 103314968`, and "Iris iris [Dalotia coriaria]" - `Gene ID: 135789998`[1]. Besides, the insufficient coverage of the target KOS results in outdated information and unlinkable entity mentions [207].

Another challenge lays down on the fact that current SOTA approaches resort to supervised, DL-based approaches that require abundant quality annotated data [219]. Large human-labelled datasets are expensive and hard to build, since its creation require biomedical expertise [229, 217]. The performance of the DL models is bounded by the information accessed during their training. If most of the datasets are small-scale, the applicability will be similarly restricted. To unlock the vast amount of biomedical text

[1] Search performed in June 5th, 2024

available and improve the performance of the task, it is necessary to go beyond supervised approaches trained on limited human-annotated data.

To evaluate the true generalization capability of EL approaches, Tutubalina et al. [248] underscored the significance of assessing these methods, in particular supervised ones, on refined test sets. These test sets exclude annotations that are concurrently present in both the training and development sets. The performance of a supervised approach is heavily reliant on the dataset, which does not align with a real setting where these approaches are employed for inference without artificial partitions of the available data into training and test sets.

To mitigate the requirement for costly human-labelled training data, emphasis should be placed on domain-independent approaches trained on domains with large amounts of labelled data and then applied in domains with limited labelled data available [217]. To achieve this objective, distant supervision [186, 70, 133] and zero-shot methods [150, 277] have been investigated in the context of the EL task.

Distant supervision consists of the generation of training data using only a limited amount of human-labelled data, for example, information such as concept names, relations present in a curated KOS [170, 133]. The entity names and synonyms described in the target KOS can be matched with unlabelled text to label annotations instances [133]. Zero-shot methods focus on generalising an approach to new domains and entities, which were not accessed during the training stage. The key idea is to develop an approach able to link the entities by having only access to the descriptions of the entities belonging to the target domain [217].

Biomedical KOSs typically represent a large number of concepts, however not every concept is represented in the datasets used to evaluate the EL task. The BC5CDR dataset [146] includes 4,424 disease annotations, associated with 674 MEDIC vocabulary concepts, and 5,385 chemical annotations associated with 676 CTD-Chemical vocabulary concepts (check Table 5.2) [55], which corresponds to a KOS concept coverage of 5.1 % and 0.38% in the dataset, respectively[2]. To develop EL approaches capable of handling a large number of concepts in the target KOS, relying solely on evaluation datasets is insufficient.

In our work, we expand on the idea of distant supervision and zero-shot in order to build a large-scale training dataset and an EL approach that is able to link disease and chemical entities without the need of

[2]MEDIC and CTD-Chemical vocabularies version:Feb 28 2024 10:59 EST

retraining with human-labelled data. To effectively train a DL-based model in the generated large-scale dataset, we frame the EL task an extreme multi-label XMR problem [38], where there are a large amount of source texts to label as well a large set of target labels, and adapt them to the EL task. We explore the hypothesis of applying XMR approaches to the biomedical domain jointly with several types of EL approaches that have proven effective in the task. The contributions of this work include:

- EL Pipeline X-Linker including different modules that can be applied to biomedical KOSs, more concretely, to link disease and chemical entities.
- Large-scale training datasets with automatic entity annotations.
- Source code publicly available to allow experiment reproducibility and further improvements: `https://github.com/lasigeBioTM/X-Linker`

5.2 Related Work

In the past decade, varied types of approaches have been proposed to address the problem of EL in the biomedical domain, ranging from heuristics [67] to the most recent DL-based architectures [150, 275, 277].

Rule-based approaches offer the advantage of bypassing the requirement for a large volume of labelled data, albeit at the cost of performance. For example, D'Souza and Ng [67] proposed a multi-pass sieve approach for EL in clinical records and scientific articles. This work shares some similarities to our work in the sense that it outlines a rule-based pipeline that includes multiple entity processing steps, including string matching the input mentions to the target KOS, abbreviation expansion, identification of composite mentions and several syntactic transformations, such as stemming, hyphenation or dehyphenation, suffixation and the replacement of numbers by their extended form.

Machine learning-based approaches improve the performance in the task, but require human-labelled data. For instance, TaggerOne [136] is a machine learning-based approach that models jointly NER and EL. The EL component is a supervised semantic indexer that generates vectorized representations for both input token and candidate KOS entities and then it assesses the correlation between tokens and target KOS entities using a semi-Markov model.

Supervised approaches work better for domains with plenty of labelled data available, such as the general domain that has Wikipedia. Therefore, more recent approaches to the biomedical EL task focus on DL architectures that require less annotated data.

For instance, BioSyn [237] focuses on learning sparse and dense representations for entities using the synonym marginalization technique. The approach applies an iteractive candidate retrieval with the goal of maximising the marginal likelihood of the synonyms being present in the top candidates.

The recently proposed BELHD [79] expands on BioSyn by focusing on homonym entities. The approach replaces homonym entities by a disambiguated version included in the target KOS and then introduces candidate sharing and a new objective function to train the BioSyn model.

BERN2 [238] shares similarities with our work as it adopts a hybrid approach: initially employing a rule-based module to link entities, then applying the DL BioSYN model for more challenging cases. However, BERN2 is a supervised approach that uses annotations from the training sets of evaluation datasets to fine-tune BioSYN.

Several zero-shot approaches have been proposed to tackle the EL task but these mostly focus on the general domain [152, 272, 265, 241]. In the biomedical domain, two zero-shot approaches have achieved SOTA performance: SapBERT [150] and KRISSBERT [277]. SapBERT [150] represents an unsupervised approach with a focus on learning representations for entities within the target KOS. The method involves pretraining a Transformer-based model on UMLS data using a self-alignment objective. Initially, it clusters synonyms of UMLS entries, after which a BERT-based model learns a mapping function between names and their corresponding Concept Unique Identifiers (CUIs). KRISSBERT [277] introduces a self-supervised method for EL aimed at mitigating the scarcity of annotated data for model training. The approach generates entity annotations by matching UMLS entity names with unlabeled PubMed documents. Subsequently, it employs contrastive learning to train a contextual encoder. This involves creating positive pairs, where two entity mentions are associated with the same UMLS CUI, and negative pairs, where two entity mentions are linked to different CUIs. The encoder is trained to map mentions of the same entity closer together and mentions of different entities further apart, thereby generating distinct representations for UMLS entities.

Recently, Jiang et al. [114] introduced an XMR-based approach for general EL, demonstrating the adaptability of techniques initially developed for XMR to the EL task. There are several key differences

between their approach and ours: their entity retriever utilizes beam search, while ours employs a BERT-based matcher (XR-Transformer); their method is applied to datasets annotated with Wikipedia entities, whereas we focus on the biomedical domain. Additionally, we implement a hybrid pipeline that incorporates modules for various types of EL, whereas they solely apply the XMR-based model. We further provide an extensive description of our proposed approach to the EL task.

5.3 Methods

5.3.1 EL definition

Let T be a text document containing a set of entity mentions $M = \{m_1, m_2, \ldots, m_n\}$, and KOS a knowledge organization system including a set of entities or concepts $E = \{e_1, e_2, \ldots, e_k\}$. The goal of EL is to map each mention $m_i \in M$ recognized in a given document to its corresponding entity $e_j \in E$. Ideally, input entity mentions are linked to entities that accurately represent their semantic meaning.

There are two essential phases in the EL task:

- **Candidate Generation**: the goal is, for each mention m_i, to generate a set of candidate entities $C_i \subseteq E$.
- **Candidate ranking and disambiguation**: the goal is to rank the candidate entities C_i based on their similarity scores $\text{sim}(m_i, e_j)$. Depending on the approach, the similarity can be calculated based on the features of the individual mentions (**local approach**), the features of other mentions within the same document (**global approach**), or a combination of both. The entity e_i^* with the highest similarity score in the candidates set is selected:

$$e_i^* = \arg\max_{e_j \in C_i} \text{sim}(m_i, e_j)$$

The final mapping $\mathcal{M}$ from mentions M to entities E is given by:

$$\mathcal{M}(m_i) = e_i^* \quad \forall m_i \in M$$

There are various methods for candidate generation and ranking in the EL task. We demonstrate in this work that no single approach is universally optimal for all entities. Instead, a combination of different

approaches yields better performance.

5.3.2 Entity Linking as a string similarity problem

The candidate generation is achieved using a string similarity function. One commonly employed string similarity function is the edit distance also designated by Levenshtein distance. The Levenshtein distance between two given strings represents the minimum number of single-character edits (insertions, deletions, or substitutions) necessary to convert one string into the other. Defining the distance as d, the distance between a given mention $m_i \in M$ and an entity $e_i \in E$, the goal is to compute the distance between a mention and every entity and then choose the entity the smallest distance to link the mention to. The limitation is that it relies solely on individual features of the input entity mention, and these features are strictly string-based, lacking consideration of contextual information.

5.3.3 Entity Linking as an eXtreme Multilabel Ranking problem: PECOS-EL

In this work we investigate framing of the biomedical EL task as an XMR problem, which to the best of our knowledge, has been only recently attempted in the general domain by Jiang et al. [114]. Given an input entity mention, the goal is to return the most relevant labels or identifiers from a large set of labels included in target KOS. We used PECOS [274], a framework originally designed for Information Retrieval approaches. In the EL task, the input mention serves as the text and the set of entities E in the KOS represents the target labels. The PECOS framework encompases three stages:

1. **Semantic Label Indexing**: the set of KOS entities E is partitioned into K clusters.
2. **Matching**: an entity mention is mapped into relevant clusters through a learned scoring function.
3. **Ranking**: a ranker assigns scores to the candidate entities present in the matched clusters.

In semantic label indexing, labels/entities from a target KOS are grouped into clusters reducing the search space. Representations for each entity $z_e : e \in E$ are obtained by aggregating the feature vectors of the training instances associated with the entity. The clustering algorithm maps assigns each entity to a cluster: $c_e \in Cl^E$, where c_e denotes the index of the cluster containing the entity e. The clustering is represented by the clustering matrix $Cl^E \in \{0,1\}^{E \times K}$ with E representing the entities in the target KOS and K representing the number of entity clusters.

During the matching stage, a general matcher function $g(x, k)$ determines the relevance between an instance x and the k-th entity cluster. The top-b clusters in Cl are identified through $g_b(x)$:

$$g_b(x) = \arg \max_{S \subset Cl: |S|=b} \sum_{k \in S} g(x, k)$$

The functions $g_b(x)$ attempts to find the subset S of size b included in Cl that maximises the function g evaluated at each S for a given x. The deep text vectorizer is a pre-trained Transformer model, specifically BioBERT. Although we briefly explored other BERT-based models (BERT, SciBERT, BioBERT, PubMedBERT), but we found the differences to be minimal.

After the matching stage, the ranker $h(x, e)$ models the relevance between x and each candidate entity belonging to the clusters previously identified by the matcher function $g_b(x)$.

We trained two PECOS models for two entity types, "Disease" and "Chemical". We further describe the generated training data.

5.3.4 Generation of training data with automatic labelling

DL-based approaches that focus on specific tasks usually require a vast amount of human-labelled data, which is scarce in the biomedical domain. The annotation process is a bottleneck in the development of such approaches, since it is slow, costly and it requires biomedical expertise. To overcome this obstacle, we generated training datasets for the model PECOS-EL that include automatic annotations obtained from Pubtator3 [258] and the target KOSs.

5.3.4.1 Pubtator3 data

Pubtator3 is a resource for exploring the biomedical literature present in PubMed[3], providing information retrieval and extraction utilities. Pubtator3 also includes DL-based tools for entity recognition and linking focusing on six common biomedical entity types: *Gene*, *Chemical*, *Disease*, *CellLine*, *Species*, *Variant*. These entity types are linked, respectively, to the resources: NCBI Gene database, MeSH, MeSH, Cellosaurus, NCBI Taxonomy, and the NCBI Single Nucleotide Polymorphism (dbSNP) database. The PubTator3 FTP site[4] provides bulk downloads of all PubMed abstracts and the full-texts associated with arti-

[3] `https://pubmed.ncbi.nlm.nih.gov/`
[4] `https://ftp.ncbi.nlm.nih.gov/pub/lu/PubTator3/`

cles from the PMC Open Access Subset (PMC-OA), as well the respective entity and relation annotations. Pubtator3 applies TaggerOne [136] and NLM-Chem [107] to link disease and chemical entities, respectively. The micro-averaged F1-score of Pubtator3 determined in the BioRED dataset is 0.7917 and 0.8192 for disease and chemical entities (data extracted from "Supplementary Table 3" in Wei et al. [258]). The latest version of the Pubtator3 includes 135,861,884 chemical annotations (file "chemical2pubtator3.gz") and 154,124,935 disease annotations (file "disease2pubtator3.gz"). We applied a pre-processing pipeline to convert each Pubtator file into useful training data:

1. Document removal: deletion of annotations associated with documents present in the evaluation datasets (see Table 5.2).
2. Lowercasing the text of each annotation.
3. Deduplication of annotations.
4. Removal of obsolete target KOS identifiers and conversion of the identifier into numerical indexes.
5. Sorting the annotations associated with each KOS identifier according to their frequency in the Pubtator3.0 set of annotations.

5.3.4.2 KOSs data

The target KOSs consisted of the following curated data retrieved from the CTD [55] (MDI Biological Laboratory, Salisbury Cove, Maine, and NC State University, Raleigh, North Carolina, URL: `http://ctdbase.org/`): *Disease Vocabulary* (also called *MEDIC*)[5], *Chemical Vocabulary* [6]. Both *MEDIC* and *CTD-Chemical* include entities associated with the respective MeSH identifiers, which allows us to integrate both KOSs and Pubtator3 information in a common data space for training.

5.3.4.3 Training datasets

We generated files with training data for each entity type: "Chemical" and "Disease". Each entity type has several dataset versions. Table 5.1 summarizes the information for all generated files. As a baseline, the training files "Disease-KOS" and "Chemical-KOS" only include names and synonyms extracted from

[5]Version:Feb 28 2024 10:59 EST
[6]Version:Feb 28 2024 10:59 EST

the target KOSs, more concretely, the MEDIC and the CTD-Chemical vocabularies. For the *Chemical* training file "Chemical-All", all the annotations present in the set provided by Pubtator3 of the type "Chemical" that had a valid MeSH identifier (i.e. an identifier present in *CTD-Chemical* version used in this work) were included. For the "Disease" training files, given the higher number of annotations that were available, we generated several versions of the dataset, setting as threshold the number of maximum instances allowed per KOS entity: "Disease-100", "Disease-200", "Disease-300", "Disease-400". We also generated the file "Disease-All" including all the Pubtator3 annotations and KOS names and syonyms. The structure of the training file includes two columns: a numerical index associated with the respective KOS identifier and the string associated with the annotation or the KOS canonical name or synonym.

Table 5.1: Versions of the generated training files per entity type and respective number of instances. For the "Disease" training data, the training data included at most 100, 200, 300 and 400 per KOS entity, as well one version with all instances

Type	**Name**	**Description**	**Instances**
Chemical	Chemical-KOS	KOS labels	452,318
	Chemical-All	KOS labels + Pubtator (all)	1,123,842
Disease	Disease-KOS	KOS labels	89,465
	Disease-100	KOS labels + Pubtator (100)	828,163
	Disease-200	KOS labels + Pubtator (200)	1,402,332
	Disease-300	KOS labels + Pubtator (300)	1,873,901
	Disease-400	KOS labels + Pubtator (400)	2,275,258
	Disease-All	KOS labels + Pubtator (All)	9,497,985

The training data incorporates alternate names for each entity. Nevertheless, in certain instances, integrating information about the context of the entity into the linking decision can enhance performance.

5.3.5 Entity linking as collective coherence maximization problem: Personalized PageRank

One of the main obstacles in the EL task is the presence of homonym entities, i.e., entities sharing the same string but with highly different meanings [79]. One way to diminish the impact of such cases is by applying a global approach, which takes into account the document context to perform the linking process. In this type of approach, a given entity mention is linked according to how the other entity

mentions present in the same document are linked. We previously demonstrated how the PPR algorithm can be integrated into such global approach [131, 210].

In a given document T, for each entity mention $m_i \in M$, the approach generates a set of candidate entities $C_i \subseteq E$. Using these candidates entities, the approach builds a graph disambiguation G, represented as $G(N, V)$, with N as the set of nodes in the graph and V as the set of vertices or edges connecting the nodes. Each node $n \in N$ corresponds to a pair consisting of an entity mention and its respective KOS candidate. The graph can be described as $G = \{(m, c) \mid m \in M, c \in C\}$. The edges between candidate nodes are based on the direct edges defined in the target KOS, for instance, on *is-a* relationships. The original PageRank algorithm [183] simulates random walks on a graph, and in each walk there is a teleport probability e on going to a random node and a $1 - e$ probability of going to a node connected with the current one. In the PPR [191] variation, the teleports are always performed to some predefined source node .The stationary distribution resulting from these walks assign scores or weights to each node in the graph. The PPR algorithm calculates the coherence of each node in the graph G, i.e., how well the node fits into the set of all nodes. The algorithm starts by measuring the pairwise coherence of a source node s and a target node t:

$$coherence_s(t) = PPR(s \rightarrow t)$$

Following the previous approach developed by our group [131], we enhance the coherence by multiplying it by the IC of the node t. This adjustment encourages the algorithm to select more specific entries within the KOS at the expense of more general ones:

$$\text{coherence}_s(t) = \text{PPR}(s \rightarrow t) \cdot \text{IC}(t)$$

We opted for the intrinsic definition for IC, in which the IC of a KOS entity e is given by its frequency in the respective KOS [48]:

$$\text{IC}(e) = -\log(p(e))$$

where p is the probability of the entity e and is represented as

$$p(e) = \frac{\text{Desc}(e) + 1}{|E|}$$

Where $Desc$ correspond to the number of child entities or direct descendants of the entity e in the structure of the target KOS, and $|E|$ is the set of every entity represented in the target KOS.

After calculating all the pairwise coherences for node s, the global coherence of t is given by the sum of its coherence with each source node s:

$$coherence(t) = \sum_{s \in G} coherence_s(t)$$

One drawback of this approach is its vulnerability to noise propagation. In certain scenarios, there might be multiple entity mentions with "imperfect" candidate lists, meaning the list either lacks the correct candidate or contains candidates that are highly unrelated to the initial mention. However, if these unrelated candidates integrate well into the graph, the PPR algorithm may assign them a high score, even though they are not the correct linking decision. The impact of this error type amplifies with the number of entity mentions featuring "imperfect" candidate lists.

Different entities require different linking approaches, hence it is essential to combine different approaches to minimize the drawbacks of each one.

5.3.6 X-Linker: pipeline for EL

To deal with different entities, we explore the combination of the previous approaches into a single pipeline, designated by *X-Linker*. *X-Linker* is a heuristic approach that employs abbreviation detection, string matching, to the PECOS-EL model and to the PPR-based model according to the entity being linked. The overview of the X-Linker pipeline is shown in Figure 5.1 and the pseudo-code is shown in Algorithm 1.

The algorithm starts by taking a set of entity mentions M and initializing a score threshold for filtering matches outputted by the respective PECOS-EL model. For each mention m, it applies an abbreviation detector to convert m to its long form $long\_m$. Then, it retrieves candidate matches using a string matcher and the PECOS-EL model, storing results in $string\_matches$ and $pecos\_matches$, respectively. If the top candidate from $string\_matches$ or $pecos\_matches$ has a perfect score (1.0), it is added to the

Algorithm 1 X-Linker pipeline

Input: M
Initialize: $threshold$
for each $m \in M$ **do**
  $long\_m \leftarrow$ apply_abbreviation_detector(m)
  $C \leftarrow []$
  $string\_matches \leftarrow$ apply_string_matcher($long\_m$)
  $pecos\_matches \leftarrow$ apply_pecos_el($long\_m$)
  if $string\_matches$.top_candidate['score'] $== 1.0$ **then**
    C.append($string\_matches$.top_candidate)
  end if
  if $pecos\_matches$.top_candidate['score'] $== 1.0$ **then**
    C.append($pecos\_matches$.top_candidate)
  else
    if $pecos\_matches$.top_candidate['score'] $>= threshold$ **then**
      C.append($pecos\_matches$.top_candidate)
    else
      C.append($pecos\_matches$.top_candidate)
      C.append($string\_matches$.top_candidate)
    end if
  end if
end for
Initialize: G
$G \leftarrow$ build_disambiguation_graph(C)
$PPR\_scores \leftarrow$ apply_ppr_model(G)
for each entity mention $m \in M$ **do**
  pick_highest_scoring_candidate($PPR\_scores(m)$)
end for

candidate list C. If the top candidate from $pecos\_matches$ has a score above the threshold, it is also added to C. If the score is below the threshold, both the top candidate from $pecos\_matches$ and the top candidate from $string\_matches$ are added to C. Once candidate lists for all mentions are completed, a disambiguation graph G is built based on these candidates. The PPR algorithm is then applied to G to compute scores for each candidate. Finally, for each mention m, the highest-scoring candidate from the PPR results is selected to disambiguate the mention.

A description of the implementation of the X-Linker pipeline is available in the Appendix "Implementation" .

5.4 Experiments

Table 5.2: Description of the evaluation datasets.

KOS	Entity type	Dataset	Total	NIL		Used
				No ID	Obsolete	
MEDIC	Disease	BC5CDR [146]	4,424	11	61	4,352
		BioRED [157]	917	12	0	905
		NCBI-Disease [64]	960	83	0	877
CTD-Chemical	Chemical	BC5CDR [146]	5,385	280	30	5,075
		BioRED [157]	754	27	0	727
		NLM-Chem [107]	11,772	893	0	10,879

The datasets used for evaluation of the several approaches are described in Table 5.2. We selected datasets including annotations of the type "Disease" or "Chemical" that are commonly used[7]: BC5CDR (819 citations), BioRED (74 citations), NCBI-Disease (843 citations), NLM-Chem (52 citations). From each dataset, we removed the NIL annotations, including annotations associated with no KOS identifiers (whose identifier is "-1" or "-"), but also obsolete annotations, i.e., annotations with KOS identifiers that are not present in the KOS version used in our experiments. The performance of an approach in the target evaluation dataset is assessed through the calculation of the top-k accuracy, which is defined as:

[7]The source for the number of citations is Google Scholar (`https://scholar.google.com/`) and the search was performed on June 18th, 2024

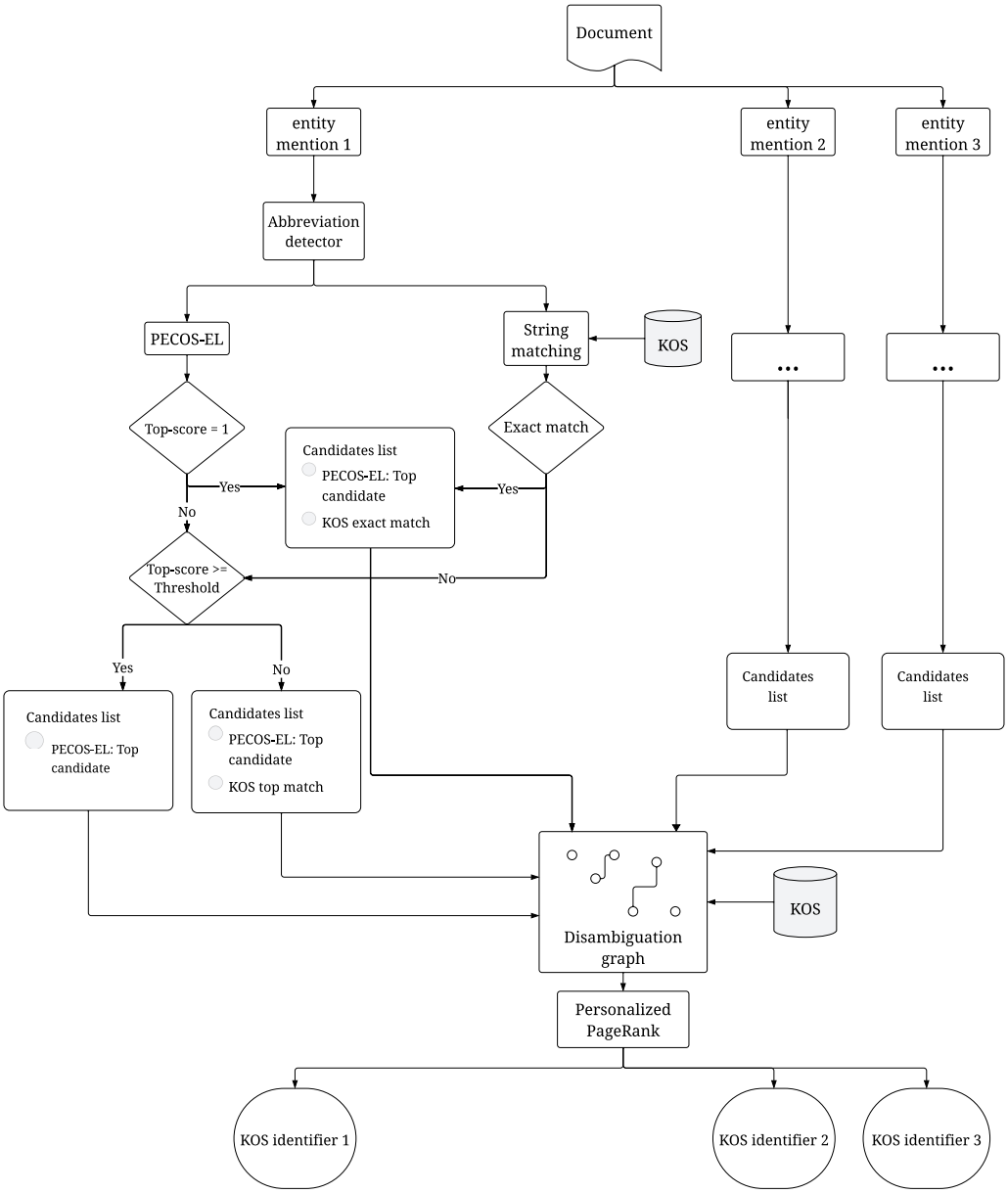


Figure 5.1: X-Linker pipeline to link biomedical entities to target KOS.

$$\text{Top-k Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{y_i \in \{\hat{y}_{i,1}, \hat{y}_{i,2}, \ldots, \hat{y}_{i,k}\}\}$$

where:

- N is the total number of evaluation instances.

- y_i is the true KOS identifier for the i-th instance.
- $\hat{y}_{i,1}, \hat{y}_{i,2}, \ldots, \hat{y}_{i,k}$ are the top k predicted identifiers (ranked by confidence) for the i-th instance.

Besides the baselines defined in our work, we used the SOTA approach SapBERT [150] for relative comparison of the performance.

5.5 Results and discussion

5.5.1 Impact of training data in the PECOS-EL model

Table 5.3: Top-1 and top-5 accuracy of the PECOS-EL Disease model trained on different training datasets and applied to the evaluation datasets BC5CDR-Disease, BioRED-Disease and NCBI-Disease.

Dataset	BC5CDR-Disease		BioRED-Disease		NCBI-Disease	
-	top-1	top-5	top-1	top-5	top-1	top-5
Disease_KOS	0.6473	0.7238	0.6114	0.6681	0.6519	0.7281
Disease_100	0.7682	0.8495	0.6681	0.8286	0.5961	0.8464
Disease_200	0.7803	0.8787	0.6790	0.8515	0.6394	0.8294
Disease_300	**0.7870**	0.8817	0.6987	0.8515	0.6837	0.8476
Disease_400	0.7746	**0.8847**	**0.7380**	**0.8854**	**0.7292**	**0.8737**

To assess the impact of the size of the training data and of the addition of Pubtator3 annotations, we evaluated the performance of the model PECOS-EL-Disease trained on different versions of the training data as described in Table 5.1. Since the training dataset "Disease-All" is very large-sized (9,497,985), we were not able to train the PECOS-EL-Disease model in this dataset due to the out-of-memory error. The performance of the Disease PECOS-EL model when applied to the different dataset versions is shown in Table 5.3. For the PECOS-EL-Disease model, incorporating Pubtator annotations into the training data enhances performance in the EL task. Moreover, as the number of Pubtator annotations increases, performance improves accordingly. However, there are some caveats. In the NCBI-Disease dataset, the addition of Pubtator annotations to the "Disease-100" training dataset decreases the top-1-accuracy to 0.6519 from 0.5961, which was obtained when training PECOS-EL-Disease in the "Disease-KOS" dataset. Training the model in the dataset "Disease-200" increases the top-1-accuracy to 0.6394, still below the performance of the model trained in the dataset in the "Disease-KOS" dataset. It's only when PECOS-EL-Disease is trained in the dataset "Disease-300" that the top-1-accuracy surpasses the baseline

(0.6837). The highest top-1-accuracy is obtained when the model is trained in the dataset "Disease-400": 0.7292. In the BioRED evaluation dataset, the performance of the PECOS-EL-Disease increases with the number of Pubtator annotations in the training data, reaching a maximum of 0.7380. In the BC5CDR-Disease dataset, the performance of the model PECOS-EL-Disease also increases with the number of Pubtator annotations in the training data, peaking when the model is trained in the dataset "Disease-300" with a top-1-accuracy of 0.7870 and decreasing with the model training in "Disease-400". This contradictory result may be explained by the nature of the Pubtator annotations present in the training data, more concretely, it can be attributable to the fact that there are Pubtator annotations sharing the same string, but associated with different KOS identifiers. For an explanation of this, check the next Subsection 5.5.2. Observing the top-5 accuracy, the performance increases with the higher number of instances in the dataset. Also the annotation performance of Pubtator3 is not 100%, so we can safely assume that there will be errors present in the training data which further decrease the downstream performance in the evaluation of the EL task in the selected datasets.

5.5.2 Is PECOS-EL a zero-shot entity linker?

Table 5.4: Overview of overlapping strings with the evaluation datasets and the training data

Type	Train instances	KOS concepts	Dataset	Used	Overlap	
					Train file	KOS
Disease	1,402,332 (Disease-200) 2,275,258 (Disease-400)	13,292	BC5CDR	4,352	4,116 (94.58 %)	2,815 (64.68 %)
			BioRED	905	868 (95.91 %)	531 (58.67 %)
			NCBI-Disease	877	823 (93.84 %)	523 (59.64 %)
Chemical	1,123,842	176,444	BC5CDR	5,075	4,922 (96.99 %)	4,067 (80.14 %)
			BioRED	727	675 (92.85 %)	497 (68.36 %)
			NLM-Chem	10,879	9,472 (87.07 %)	5,171 (47.53 %)

We analysed for each evaluation dataset the percentage of annotations with string that are also present in the data used to train the PECOS-EL model, as seen in Table 5.4. Following the strict definition for zero-shot evaluation, i.e., an EL approach must be able to link entities that were not seen during training using only the entity descriptions, the PECOS-EL models in fact are not zero-shot entity linkers [217]. We followed the refined evaluation method recommended by the authors in Tutubalina et al. [248]. Specifically, we removed any annotations present in the test sets of the evaluation datasets from

the training datasets. Besides, the training data was gathered from the natural distribution of entities in biomedical literature. Therefore, we assume that the performance of X-Linker is robust since it is not dependent on a specific evaluation dataset. The only drawback is that the training data is biased towards the past, in the sense that is based on text already existing. There is no assurance that the same entities will continue to appear in biomedical text in the future. However, the X-Linker approach can be updated with new training data and, in the event of new entities emerging, there is the potential to employ an approach that specifically handles NIL or unlinkable entities. This type approach helps prevent the loss of semantic information and mitigates decreases in performance by EL approaches [207, 197].

5.5.3 Impact of abbreviation detection

Table 5.5: Impact of adding different modules to the X-Linker pipeline

Module	Disease			Chemical		
	BC5CDR	BioRED	NCBI-Disease	BC5CDR	BioRED	NLM-Chem
PECOS	0.7803	0.7380	0.7292	0.8051	0.7729	0.6592
+abbrv_detect	0.8079	0.7664	0.7952	0.8564	0.8345	0.7164
+abbrv_detect+SM	0.8228	0.7937	**0.8271**	0.9492	**0.9248**	0.7850
+abbrv_detect+SM+PPR	**0.8307**	**0.7969**	**0.8271**	**0.9511**	**0.9248**	**0.7895**

As shown in Table 5.5, adding an abbreviation detection module greatly improves the performance of PECOS-EL. PECOS-EL relies solely on variations in the text of an entity for training and does not consider its context, thus its performance is highly dependent on the input mention text. Jiang et al. [114] showed that considering the mention's context makes the approach more robust to text variations, but the resources required to train such a model leave that work for future exploration.

5.5.4 Impact of string matching and of the rule-based filter

Table 5.5 shows the impact of adding the string matcher module to the X-Linker pipeline, which showed advantageous in all evaluation datasets. Analysing the data shown in 5.6, the overlap of string between training and evaluation data ranges from 96.99% in the BC5CDR-Chemical dataset to 87.07% in the NLM-Chem dataset. However, some of the strings in the training data are associated with more than one KOS identifier. Moreover, in some cases the identifier for a given string in the training dataset is not the

Table 5.6: Overview of overlapping strings in the training and evaluation datasets and respective correctness of the associated KOS identifiers according to the evaluation datasets annotation.

Type	Dataset	Train set overlap		Correct KOS ID in the list		
		Total	Correct KOS ID in the list	Incorrect	Exact match	Ambiguous
Disease	BC5CDR	4,116 (94.58 %)	3,779 (91.81%)	337 (8.92%)	2,743 (72.59%)	1,036 (27.41%)
	BioRED	868 (95.91 %)	775 (89.29%)	93 (12.0%)	381 (49.16%)	394 (50.84%)
	NCBI-Disease	823 (93.84 %)	751 (91.25%)	72 (9.59%)	362 (48.2%)	389 (51.8%)
Chemical	BC5CDR	4,922 (96.99 %)	4,811 (97.74%)	111 (2.31%)	3,862 (80.27%)	949 (19.73%)
	BioRED	675 (92.85 %)	665 (98.52%)	10 (1.5%)	485 (72.93%)	180 (27.07%)
	NLM-Chem	9,472 (87.07 %)	8,789 (92.79%)	683 (7.77%)	5,094 (57.96%)	3,695 (42.04%)

same identifier associated with the same string in the evaluation dataset (check column "Incorrect" in Table 5.6). For example, 12.0% of the strings present in the Disease training dataset that are also present in the BioRED-Disease dataset are associated with different identifiers. Even if the KOS identifier that appears associated with a given string in the evaluation dataset is also associated with the string in the training dataset, there is a relevant part of ambiguity (check column "Ambiguous" in Table 5.6), i.e., there are more than one identifier for the string. For example, in the BioRED-Disease dataset, 89.29% of the strings in the training dataset are associated with the correct identifier as defined in the evaluation dataset, but only 49.16% of those strings have only one identifier. The remaining 50.84% strings have more than one associated identifier.

As shown in Table 5.6, there are annotations in the evaluation datasets associated with entity names or strings that have an exact match in the respective target KOS. However, not always the identifier associated with the exact matching is the same as the identifier chosen to annotate the entities in the evaluation datasets. This highlights the inherent ambiguity of the annotation process, but also that the task EL has not an universal definition. Quite the contrary, the annotation criteria is strictly associated with the scope of the motivation. For example, in the context of an annotation project centred on rare diseases, the annotation guidelines will instruct annotators to prioritise selecting more specific diseases. However, if the project encompasses various entity types simultaneously, such as chemicals, anatomical parts, cell types, etc., the annotation guidelines may not necessitate the same level of specificity as in the case of rare diseases. In such instances, a broader categorization may be sufficient to fulfil the project's objectives. Evaluation datasets are useful to straightforwardly assess the performance of EL approaches, which can be then complemented by more extensive and realistic evaluations, as for example user testing.

A rule-based pipeline such as X-Linker is essential to diminish the impact of these disparities.

5.5.5 Document context improves the performance

Table 5.5 shows the impact of adding the PPR algorithm-based module to the X-Linker pipeline. With the previously mentioned modules, the PECOS-EL module jointly with the abbreviation detector and the string matcher is able to deal with a large part of the entities present in the evaluation datasets. However, context in the EL task is very relevant, since the same entity string can have multiple meanings according with the surrounding entities. For that, establishing a measure of coherence between a given entity and the other entities present in the same document can help to disambiguate decisions, as shown in the literature [191, 193].

The Figure 5.2 shows an example of how the X-Linker pipeline links two entity mentions present in the document with PubMed ID 19263707 from the BC5CDR-Disease dataset to entries in the MEDIC vocabulary: "vasculitic" and "vasculitis". As a first step, X-Linker applies abbreviation detection to each mention. Then, the model PECOS-EL-Disease predicts the candidates "Congenital Disorder" with identifier `D009358` and "vasculitis" with identifier `D014657` for the mentions "vasculitic" and "vasculitis," respectively. Concurrently, the string matcher retrieves the candidate "vasculitis" (`D014657`) from MEDIC for both mentions. In the disambiguation process, for "vasculitic," PECOS-EL-Disease scores low (0.0964), and the string matcher finds a close candidate (score 0.9). Both are added to the candidate list due to the low PECOS-EL-Disease score. For "vasculitis," PECOS-EL-Disease scores 1.0, and the string matcher confirms an exact match (MEDIC). Only "vasculitis" (`D014657`) is listed. Both lists feature "vasculitis," linked in the disambiguation graph by MEDIC relations. The PPR "vasculitis" (`D014657`) as the top candidate, resolving ambiguity.

5.5.6 Comparison with SapBERT

The SOTA EL approach SapBERT exhibits a high top-1 accuracy across all evaluated datasets, particularly for chemical entities. Like the PECOS-EL model, SapBERT relies on the mention text. Therefore, we also present its results after applying a pre-processing step of abbreviation detection for a fairer comparison. X-Linker achieves higher performance in three of the evaluation datasets: BC5CDR-Disease, NCBI-Disease and BioRED-Chemical. SapBERT's performance is higher in the remaining three evalua-

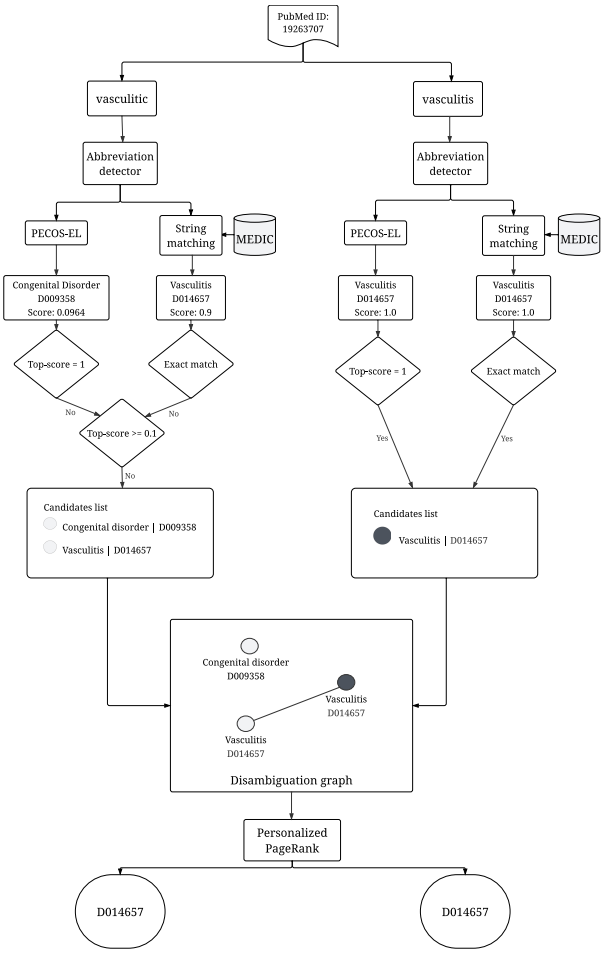


Figure 5.2: Example of the application of the X-Linker pipeline to the BC5CRD dataset involving the entity mentions “vasculitis” and “vasculitic”.

tion datasets: BioRED-Disease, BC5CDR-Chemical and NLM-Chem. For diseases, SapBERT’s performance is higher in a smaller dataset (BioRED-Disease), whereas for chemicals SapBERT’s performance is higher in the larger datasets (BC5CDR-Chemical and NLM-Chem). X-Linker’s performance is higher than the performance of PECOS-EL, which highlights the importance of combining different types of EL approaches.

5.5.7 Error analysis

One type of error is related with the specificity of the annotations. For instance, the entity “liver neoplasms” (document 26033014 in the BC5CDR dataset) is annotated with the MEDIC concept “Liver

Table 5.7: Top-1 Accuracy of the X-Linker approach compared to PECOS-EL and the baseline SOTA SapBERT.

Model	Disease			Chemical		
	BC5CDR	BioRED	NCBI-Disease	BC5CDR	BioRED	NLM-Chem
SapBERT	0.7824	0.7434	0.7845	0.8664	0.7661	0.6678
+abbrv. detection	0.8141	**0.8177**	0.8233	**0.9559**	0.9001	**0.7950**
PECOS-EL	0.7803	0.7380	0.7292	0.8051	0.7729	0.6592
+abbrv. detection	0.8079	0.7664	0.7952	0.8564	0.8345	0.7164
X-Linker (best)	**0.8307**	0.7969	**0.8271**	0.9511	**0.9248**	0.7895

neoplasms" (identifier `D008113`) and X-Linker correctly links the entity mention to the referred concept. However, in the same document, the entity mention "liver cancer" has the candidates "Liver neoplasms" (`D008113`) and "Carcinoma, hepatocellular" (`D006528`) and X-Linker links the entity mention to the child concept (`D006528`) instead of the correct one, the parent concept `D008113`. The same happens with the entity mention "cognitive impairment" (document 24802403 in BC5CDR dataset) which X-Linker links the entity mention to the parent concept "Cognitive dysfunction" (D060825) instead of the correct parent concept "Cognition disorders" (D003072). This relates to the implementation of the PPR algorithm, which considers the IC of each concept to score the candidates. As a result, more specific terms are preferred over more general ones. Neverthless, the opposite also happens: the entity mention "Deterioration of vision" is linked to the parent concept "vision disorders" (`D014786`) instead of the correct child concept "Vision, Low" (`D015354`).

In other cases, the X-Linker approach is unable to produce a candidates list with the correct candidate. The entity mention "AL" (document 24040781 of the BC5CDR dataset) is an abbreviation of "Amyloidosis", so it should be linked to the concept "Amyloidosis" (`D000686`). However the generated candidates are "Mousa Al din Al Nassar syndrome" (`C536989`), "Pallor" (`D010167`) and Abetalipoproteinemia (`D000012`). The abbreviation detector fails to identify the abbreviation, and X-Linker generates wrong candidates. In other case, the entity mention "mania" (document 19447152 of the BC5CDR dataset) is linked to the concept "Mania" (`D000087122`) instead of the concept "Bipolar Disorder" (`D001714`). In the Disease dataset used to train the PECOS-EL-Disease model the string "mania" is annotated with the identifier (`D000087122`) so the model outputted this identifier.

Other type of error is related with composite mentions, since X-Linker fails to deal with these mentions. The entity mention "hemorrhagic strokes" (document 19293073 of the BC5CDR dataset) is annotated with the identifiers `D020300` ("Intracranial Hemorrhages") and `D020521` ("Stroke"), but X-Linker links the mention to the concept "Hemorrhagic Stroke" (`D000083302`).

5.5.8 Limitations

There are several limitations associated with X-Linker. First, the performance of the PECOS-EL models is influenced by the accuracy of the automatic annotations provided by PubTator3, which were used for training. Any discrepancies arising from the automatic annotation will lead to downstream lower performance in the evaluation process using datasets. The matcher component of the PECOS-EL model uses BioBERT as the encoder model, meaning any biases associated with BioBERT may affect the results. Due to memory constraints, we did not train the PECOS-EL-Disease model on the entire training dataset and both PECOS-EL models were trained solely on the entity text, without incorporating the respective context. Training the PECOS-EL models requires significant GPU resources, so we did not perform extensive hyperparameter optimization, which may have resulted in suboptimal performance compared to fully optimized models.

5.6 Conclusion

We generated large-scale training datasets including automatic annotations to train a DL-based XMR approach adapted to the biomedical EL designated by PECOS-EL. This module was integrated in the hybrid pipeline X-Linker, an EL approach including different modules to link disease and chemical entities to the MEDIC and CTD-Chemical vocabularies without the need of human-labelled data. We carried out an extensive evaluation of the X-Linker approach, resulting in top-1 accuracy values of 0.8307, 0.7969, 0.8271, 0.9511, 0.9248, 0.7895 in the datasets BC5CDR-Disease, BioRED-Disease, NCBI-Disease, BC5CDR-Chemical, BioRED-Chemical and NLM-Chem, respectively.

In future work, we plan to enhance entity linking using X-Linker to connect mentions to the UMLS. While our current study focused on smaller KOSs due to computational limits, future directions include adapting PECOS-EL to utilize the UMLS with lightweight BERT-based matchers. Additionally, we'll explore integrating NCBI Gene and Taxonomy data from Pubtator3 for generating training datasets. Cur-

rently, PECOS-EL employs a modified K-means algorithm based on string representations of KOS entities, so we aim to boost model performance by exploring different clustering approaches that incorporate KOS information and metadata.

5.7 Appendix

5.7.1 Implementation

Our approach includes as a first step the rule-based abbreviation detector Ab3P created by Sohn et al. [221]. To implement X-Linker we used the PECOS framework [274], with the code available at `https://github.com/amzn/pecos`. Model training was done in two setups: (1) a server including 2 Intel(R) Xeon(R) Silver 4114 CPU @ 2.20GHz and 8 Tesla M10 GPUs; Total CPU memory: $\approx$ 64 GB. Total GPU memory: $\approx$ 64 GB; (2) an HPC cluster with 8 nodes with each 2 x AMD EPYC 7742 processors/node x 64 cores 128 cores, 1024 GB RAM, 40 GB VRAM each GPU, 4 GPU NVIDIA A100 (only 80GB RAM were used for training). Training time varied according to the entity type and the number of instances: Disease-400 (the large file with Disease entities) $\approx$ 8 hours in the HPC cluster; Chemical $\approx$ 16 hours. In the X-Linker pipeline, the threshold for candidate filtering is set to 0.1 as the default.

Chapter 6

Real-Word Assessments

This chapter compiles all other research work conducted throughout this thesis by dividing each contribution into a section summarising its motivation and the work developed. This work includes solo and group participation in workshops, challenges, a doctoral consortium, a Python Package, a human-annotated dataset, and other adjacent journal contributions corresponding to real-world assessments of the approaches and topics explored in this thesis.

6.1 Deep Semantic Entity Linking

The 2021 edition of the *European Conference on Information Retrieval* (ECIR) 2021 organized a Doctoral Consortium. At this venue, PhD students can publish their PhD proposals and receive mentoring from experienced and specialized researchers. The proposal associated with the current research work was publicly presented and thoroughly discussed by the participants, which led to improvements in the planning of the research work. The published paper is available in Appendix C:

> **Ruas, P.** (2021). **Deep Semantic Entity Linking**. In *Advances in Information Retrieval. ECIR 2021. Lecture Notes in Computer Science* (CORE A), Vol. 12657, pages 682–687, Cham. DOI: `https://doi.org/10.1007/978-3-030-72240-1_81` [204]. Article available in Appendix B.

6.2 LASIGE and UNICAGE solution to the NASA LITCOIN NLP competition

The 2022 edition of the *LitCoin NLP Challenge*[1] was part of the *NASA Tournament Lab*, hosted by the *National Center for Advancing Translational Sciences* (NCATS) and the *National Library of Medicine* (NLM). In the mentioned edition, our team "LasigeUnicage" was awarded the 7th Prize out of 232 participating teams, reflecting a successful collaboration between academia (LASIGE) and industry (Unicage).

The motivation for this challenge was the large amount of biomedical information still absent from open data repositories, making it hard to find the connections between available scientific discoveries described in the literature. TM approaches can expand the coverage of the open repositories, including entity extraction models. They can discover gaps in clinical research by identifying entities in free text and the respective relations between them and then translate this information into data-driven knowledge graphs integrating heterogeneous medical scientific data.

The competition aimed at creating approaches to leverage the vast volume of published biomedical articles to discover new knowledge and research hypotheses.

The specific goal of the challenge was to extract scientific entities from scientific articles (Part 1), connect them by generating knowledge assertions, and then classify them as novel findings or trivial information (Part 2). The organization provided a dataset including scientific resaerch articles and knowledge assertions between entities present in the article.

Our team has developed an approach integrating data engineering approaches for efficient data processing with state-of-the-art TM solutions from an academic research environment. For the first part of the challenge, the approach consisted of an ensemble of six NER models obtained by fine-tuning the base model PubMedBERT [87] in a custom training dataset specially developed for the competition. These models were able to recognize in text entities of six different types: *DiseaseOrPhenotypicFeature*, *ChemicalEntity*, *OrganismTaxon*, *GeneOrGeneProduct*, *SequenceVariant*, *CellLine*. For the second part, the approach relied on the BiOnt system [225] to extract relations between the recognized entities and on a post-processing module to identify wether the relations corresponded to novel knowledge.

The preprint describing the approach is available in Appendix C:

[1] `https://ncats.nih.gov/funding/challenges/winners/litcoin-nlp/details`

Ruas, P. [*], Sousa, D. F.[*], Neves, A.[*], Cruz, C., & Couto, F. M. (2023). **LASIGE and UNICAGE solution to the NASA LitCoin NLP Competition**. Available as preprint at *arXiv*, `https://arxiv.org/abs/2308.05609` [212]. Article available in Appendix C.

6.3 BENT Python Package

BENT is a Python Library for NER and EL in the biomedical domain: `https://pypi.org/project/bent/`. It is an annotation tool for biomedical text with a focus on usability for users without specific TM or NLP expertise.

The NER module was built by fine-tuning the base model PubMedBERT [87] in several NER datasets (complete list in the below links). BENT includes the following NER models to recognize several types: chemicals, diseases, genes and proteins, organisms, anatomical entities, cell types, cell components, cell lines, biological processes (as defined by GO-BP sub-ontology), variant entities (a variant entity is a DNA-level or protein-level mutation as defined by the *Human Genome Variation Society* nomenclature).

The EL module is a graph-based approach based on the REEL approach (see Chapter 3) and the NILINKER approach (see Chapter 4). The following KOSs are included in the current version[2] (the respective entity types being linked in parentheses): MEDIC and Disease ontology (*disease*), ChEBI and CTD-Chemical (*chemical*), NCBI Gene and CTD-Gene (*gene*), NCBI Taxonomy (*organism*), GO-BP (*biological process*), CTD-Anatomy, UBERON and Foundation model of Anatomy (*anatomical*), GO-CC (*cell component*), Cell Ontology (*cell type*), Cellosaurus (*cell line*). A custom KOS can be integrated into BENT if the use requires it; it just requires two files specifying the list of names for the entries and the respective direct relation between them.

6.4 BiOrange: augmenting BioRED dataset by annotating NIL entities and n-ary relations

After the mentioned approaches, the research also focused on improving evaluation in the EL task. There is an ongoing project aiming at expanding the BioRED dataset [157] for NIL entity linking and n-ary relation extraction (i.e. extraction of relations involving more than two entities). The project's first

[*] Authors contributed equally to this research
[2] Version 1.0

stage involved the development of annotating guidelines focusing on NIL entities and n-ary relations. The second stage involved three rounds of manual annotation based on the developed guidelines. The resulting dataset (not publicly available yet) is designated as BiORANGE and includes four additional entity types (*CellTypeOrAnatomicConcept*, *NILGene*, *NILDis*, *NILChem*), totalling 346 new annotations across 31 documents (PubMed titles and abstracts).

6.5 LasigeBioTM team at CLEF2020 ChEMU evaluation lab: Named Entity Recognition and Event extraction from chemical reactions described in patents using BioBERT NER and RE

The *Cheminformatics Elsevier Melbourne University* (ChEMU) evaluation lab 2020, part of the *Conference and Labs of the Evaluation Forum* 2020 (CLEF2020), focused on developing TM approaches for chemical patents.

Chemical patents represent a vast wealth of information about new chemical compounds, which can play an essential role in drug discovery. However, the specific linguistic properties of patents and the large amount of available documents require developing specific approaches.

The ChEMU 2020 lab included two information extraction tasks starting from a dataset composed of chemical reaction processes described in chemical patents: recognition of entities and respective roles in the context of the reactions (corresponding to the NER and semantic role labelling tasks) and also the extraction of events between the recognized entities. The target entities included the reaction products, catalysts, reagents, solvents, and conditions related to the reaction, such as temperature, time, and yield-related entities.

The approach developed by our team for the first task was based on fine-tuning BioBERT [140] in the provided training dataset, which achieved the third position out of 11 teams. The article describing the approach is available in Appendix D:

> **Ruas, P.**, Lamurias, A., & Couto, F. M. (2020). **LasigeBioTM team at CLEF2020 ChEMU evaluation lab: Named entity recognition and event extraction from chemical reactions described in patents using BioBERT NER and RE**. In *The workshop ChEMU: Named Entity Recognition and Event Extraction of Chemical Reactions from Patents (CLEF 2020 Working Notes)*. URL:

https://ceur-ws.org/Vol-2696/paper_175.pdf [209]. Article available in Appendix D.

6.6 LasigeBioTM at CANTEMIST: Named Entity Recognition and Normalization of Tumour Morphology Entities and Clinical Coding of Spanish Health-related Documents

The goal of the *CANcer TExt Mining Shared Task – Tumor Named Entity Recognition* (CANTEMIST) competition [Miranda-Escalada et al.] was to develop TM and NLP techniques for extracting information from Spanish clinical records related to the histological types of neoplasms. The ultimate objective was to discover new knowledge about cancer treatments to improve patient healthcare.

CANTEMIST included three subtasks focused on extracting tumor morphology entities from Spanish health-related documents: CANTEMIST-NER, CANTEMIST-NORM, and CANTEMIST-CODING.

The motivation behind this task was to develop TM approaches capable of extracting valuable information from electronic health records to enhance doctor-patient interactions and biomedical research. Since healthcare professionals often communicate in their native languages, a significant amount of data and information is available in non-English languages. Therefore, developing comprehensive TM approaches that can handle multiple languages is crucial.

For CANTEMIST-NER, we generated Spanish biomedical Flair embeddings on PubMed abstracts and then trained a BiLSTM+CRF NER tagger[3] on the CANTEMIST corpus using the trained embeddings.

For the CANTEMIST-NORM, our approach adapted the graph-based approach REEL (see Chapter 3) to link the recognized entities to concepts in the target terminology, the Spanish version of the terminology developed by the *World Health Organization International Classification of Diseases for Oncology* (ICD-O).

For CANTEMIST-CODING, we framed the task as an XMR problem and adapted the X-Transformer model (see Chapter 5) to assign relevant codes from the target terminology to each of the clinical cases.

The article describing the approach is available in Appendix E:

Ruas, P., Neves, A., Andrade, V. D. T., Couto, F. M., & Aragón, M. E. (2020). **LasigeBioTM at CANTEMIST: Named Entity Recognition and Normalization of Tumour Morphology Enti-**

[3] akbik2019flair

ties and Clinical Coding of Spanish Health-related Documents. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with the 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020)*, pages 422–437. URL: `https://ceur-ws.org/Vol-2664/cantemist_paper11.pdf` [211]. Code repository publicly available. Article available in Appendix E.

6.7 COVID-19 recommender system based on an annotated multilingual corpus

The *Biomedical Linked Annotation Hackathon* (BLAH)[4] is a series of annual hackathon events focused on open collaboration to develop resources for biomedical literature annotation and mining. The 2021 edition focused on COVID-19. The goal was to work on projects attempting to tackle the fast pace of publication through TM approaches. Our group identified a fundamental limitation on existing resources: the lack of annotated multilingual datasets focusing on COVID-19-related entities and relations. Our group's proposed pipeline included a module to retrieve COVID-19-related articles. A set of parallel documents (titles and abstracts) were obtained for each language in the study: English, Portuguese and Spanish. Then, a module to perform NER and EL based on the rule-based MER tool [47] was applied, followed by a module to perform RE. Finally, a recommendation algorithm leveraged the entities and relations previously identified to recommend relevant entities for COVID-19, such as chemical compounds. 20 English and 20 Portuguese abstracts were sampled for manual validation of the annotations.

The project's contributions include a TM pipeline for document retrieval, entity and relation extraction, and recommendation, and a set of multilingual parallel datasets (English/Portuguese/Spanish) related with COVID-19 that allow the development and evaluation of similar pipelines. The article describing the project is available in Appendix F:

Barros, M.[*], **Ruas, P.**[*], Sousa, D.[*], Bangash, A. H., & Couto, F. M. (2021). **COVID-19 recommender system based on an annotated multilingual corpus**. *Genomics & Informatics*, 19(3), e24. URL: `http://genominfo.org/journal/view.php?number=667`. DOI: `10.5808/gi.21008` [14]. Code repository publicly available. Article available in Appendix F.

[4] `https://blah7.linkedannotation.org/`

[*] Authors contributed equally to this research

6.8 LASIGE-BioTM at MESINESP2: entity linking with semantic similarity and extreme multi-label classification on Spanish biomedical documents

The *Medical Semantic Indexing in Spanish 2* (MESINESP2) task was organized in the context of the BioASQ/ CLEF 2021 Challenge, and its motivation was the development of TM and NLP approaches to improve access to health and biomedical-related documents to healthcare professionals, researchers and decision-makers. The goal was to semantic index Spanish health-related documents, such as articles, clinical trials and summaries of healthcare projects, to the MeSH-derived vocabulary *Descriptores en Ciencias de la Salud* (DeCS). The approach developed by our group was a pipeline with two modules. The first one was an EL approach based on REEL (see Chapter 3) that linked the recognized entities in the texts to entries in the DeCS vocabulary and then applied a relevance filter based on semantic similarity. The second module was based on the X-Transformer model (see Chapter 5), which classified each document using the most relevant terms from the DeCS vocabulary. The participation was described in an article presented in the BioASQ workshop. The article is available in Appendix G:

> **Ruas, P.**, Andrade, V. D. T., & Couto, F. M. (2021). **LASIGE-BioTM at MESINESP2: Entity linking with semantic similarity and extreme multi-label classification on Spanish biomedical documents**. In *Proceedings of CLEF 2021*, pages 324–334. URL: `http://ceur-ws.org/Vol-2936/#paper-24` [205]. Code repository publicly available. Article available in Appendix G.

6.9 Lasige-BioTM at ProfNER: BiLSTM-CRF and contextual Spanish embeddings for Named Entity Recognition and Tweet Binary Classification

The track *ProfNER-ST: Identification of professions and occupations in Health-related Social Media* [172] was part of the *Social Media Mining for Health Applications* (#SMM4H) Shared Task 2021 [160] and involved the extraction of occupation-related information from social media. This information can potentially be used to improve data-driven decisions, sociodemographic analysis and real-time monitoring of risk groups. The specific goals of the track were text classification of Spanish tweets and recognition of professions and occupations. The approach developed by our group first performed tweet bi-

nary classification to determine if a given Twitter contained a relevant entity. If so, the relevant entities would be recognized by a module consisting of a pre-processing step of data augmentation followed by a model based on the BiLSTM-CRF architecture. The approach was presented in the #SMM4H workshop co-located with the NAACL conference (CORE A). The article describing the approach is available in Appendix H:

> **Ruas, P.**, Andrade, V., & Couto, F. (2021). **Lasige-BioTM at ProfNER: BiLSTM-CRF and contextual Spanish embeddings for Named Entity Recognition and Tweet Binary Classification**. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 108–111. DOI: `10.18653/v1/2021.smm4h-1.21` [206]. Code repository publicly available. Article available in Appendix H.

6.10 Creating Recommender Systems Datasets in Scientific Fields

Our research group developed a tutorial on using recommendation systems in scientific fields presented in the 7th *ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (CORE A*). The goal was to demonstrate the process of generating a COVID-related recommendation dataset, including modules such as NER and EL, which play an essential role in annotating entities. The links between the recognized entities in text and the target KOS influence the recommendation process down the pipeline. The participation in this conference originated an article available in Appendix I

> Barros, M., Couto, F. M., Pato, M., & **Ruas, P.** (2021). **Creating recommender systems datasets in scientific fields**. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD '21)*, pages 4029–4030. DOI: `https://doi.org/10.1145/3447548.3470805` [13]. Code repository publicly available. Article available in Appendix I.

Chapter 7

General Discussion and Conclusions

The amount of scientific literature and other types of scientific text is expected to keep increasing. Thus, EL approaches and other components of the TM pipeline will still be necessary to organize and structure the available text and translate relevant insights into KOSs.

The initial hypothesis of this thesis was to increase the performance in the EL task, focusing on overcoming the limitations of human-annotated data by addressing the problem of NIL entities and reducing the need for human-annotated datasets. The described research shows that focusing on these two limitations improves the performance of the EL approaches. This thesis tackled the limitations of human-annotated data on two fronts. On the one hand, it attempted to overcome the lack of coverage in biomedical KOSs by integrating information extracted by RE into a graph-based EL approach (REEL) and by developing a DL model to link NIL entities to target KOSs partially (NILINKER). On the other hand, it attempted to decrease the reliance on human-annotated datasets for training DL-based models with the proposal of the X-Linker approach.

The present work opens two main research lines.

First, the work lays the foundation for a new task related to the classical EL task: NIL entity linking. This task aims to partially link biomedical NIL entities to target KOSs, even if a perfect match does not exist. The contributions at this level span both data creation and model development stages: the new dataset EvaNIL allows the training of approaches focusing on NIL entities, and the model NILINKER constitutes a solid baseline for anyone looking to build models to solve the task. The improvements in NIL entity linking will ultimately lead to EL approaches with increased performance and better curation

tools for KOSs.

Second, the approach X-Linker shows that it is possible to obtain SOTA performance without using supervised methods and human-annotated data and that an ensemble of different types of approaches provides more flexibility and performance. The combination of approaches leverages the strengths of each type of approach, more concretely, string-based, graph-based and DL-based models while minimizing the drawbacks associated with each type. Human-curated datasets can still be helpful for quick performance validation, but their usual scale is not enough for large-scale training of DL-based approaches. So, the work showcases the utility of automatic dataset generation methods with the assistance of data stored in KOS.

Approaches that do not require human-annotated data do not eliminate the need for effective curation pipelines with human intervention since it is still essential to ensure that KOSs store quality data and information. The field of NLP is receiving more attention from the general public with the advent of LLMs and respective applications with increased usability. LLMs can decrease the barrier entry for using NLP tools, but LLMs alone are still insufficient to deal with biomedical data's complexities. KOSs are still essential to ground the output of such LLM-based applications to handle the impact of the hallucination problem [111]. KOSs represent a reliable source of quality data that can be used to improve automatic methods. Maintaining the existing KOSs ecosystem is essential, and TM pipelines and EL approaches can be valuable assistance in such a process.

7.1 Summary of Contributions

The following subsections present a final perspective on the contributions in the context of the objectives defined in the introduction: to tackle the lack of coverage in biomedical KOSs (objective 1) and to reduce the reliance on human-annotated datasets for training EL approaches (objective 2).

The research work first involved overcoming the limitations associated with the lack of coverage in biomedical KOSs (objective 1) with the proposal of the approaches designated by REEL (Chapter 3) and NILINKER (Chapter 4).

The REEL approach is an EL model based on graphs, the PPR algorithm and RE. Graph-based EL approaches perform global or collective disambiguation of entities in a document by building a disambiguation graph that includes semantic information about the target KOS candidates for the entities.

Edges between nodes or candidates are based on the relations represented in the target KOS, but in some cases, relations are missing from the KOS and are only explicitly defined in scientific articles. This leads to sparse disambiguation graphs with incomplete semantic information, which decreases the candidate ranking process and, consequently, the performance of EL approaches. In the REEL approach, a DL RE model was used to extract relations between entities described in scientific articles and to add this semantic information to the disambiguation graph, thus overcoming the limitations of the target KOS. The resulting approach achieved F1-scores of 85.8%, 80.9%, and 90.3% in the gold standards CRAFT-ChEBI, BC5CDR-Disease, BC5CDR-Chemical, respectively, outperforming baseline appproaches. Usually, EL applications improve the performance of downstream tasks, such as RE, but the REEL approach showcased that the opposite is also beneficial to the EL task: to improve EL with the help of a RE.

After REEL, the research focused on overcoming biomedical KOSs limitations by handling unlinkable or NIL entities. The fast pace of publication in the biomedical domain leads to the surge of unlinkable or NIL entities, which biomedical KOSs do not represent. These entities decrease the performance of EL approaches. The approach proposed to deal with NIL entities, NILINKER, is a DL-based model based on the attention mechanism able to partially link biomedical entities to target KOSs, even when a perfect entry does not exist. This task was designated by NIL entity linking. NILINKER includes a candidate retrieval module for biomedical NIL entities and a neural network relying on the attention mechanism to find the most relevant KOS entries for the input NIL entities. Besides the model, this work proposed a new silver standard designated by EvaNIL, suitable for training and evaluating models focusing on the NIL entity linking problem. The NILINKER model was integrated into the REEL approach, and the experiments demonstrated that dealing with NIL entities improves the performance in the classical EL task.

In the second part, the work focused on achieving objective 2 of reducing the reliance on human-annotated data. SOTA DL-based EL approaches rely on extensive amounts of expensive human-labeled data, yet the available datasets are limited. Chapter 5 describes a possible solution to this problem, which is based on the automatic generation of a large-scale training dataset from existing datasets and KOSs and on a hybrid EL pipeline designated by X-Linker, which integrates different types of EL models to achieve SOTA performance in the biomedical domain. The generated dataset was used to train a DL-based XMR approach adapted for the first time to the biomedical EL task designated by PECOS-EL.

This model was integrated into the pipeline X-Linker, which includes several modules for dealing with different types of entities: an abbreviation detector, a string matcher for lexical comparison of entities and KOS candidates, a module based on the REEL approach relying on the PPR algorithm. This hybrid approach offers greater flexibility in handling biomedical entities with diverse features and leverages each method's strengths while minimizing their drawbacks. X-Linker achieved the highest performance in three evaluation datasets when compared with the SOTA approach SapBERT: BC5CDR-Disease, NCBI Disease and BioRED-Chemical

Other explorations and real-world assessments contributed to both objectives and are described in Chapter 6. These include the adaptation and evaluation of the described approaches in several competitions, challenges, and workshops, as well as the development of approaches for EL-related tasks, such as NER and recommendation. This chapter also describes the development of a new Python Package for text annotation designated by BENT. The future dataset BiORANGE will include NIL entity annotations, providing a human-curated resource to validate and compare the performance of approaches developed explicitly for NIL entity linking. The current version of this new dataset (not yet publicly available) includes four additional entity types (*CellTypeOrAnatomicConcept*, `NILGene`, `NILDis`, `NILChem`) totalling 346 new annotations across 31 documents, more concretely, PubMed titles and abstracts. The real-world assessments allowed the exploration of EL approaches using scientific articles as source text and varied sources, including chemical patents, social media text and clinical records.

7.2 Future Work

One of the future research directions will be exploring approaches focused on NIL entities to curate KOSs and to track the evolution of entities, i.e., entities may change over time, and it is important that KOSs can accurately represent the changes.

The most significant trend will be towards EL approaches relying on ever less human-annotated data. Current approaches still lack generalization ability; thus, the focus on zero-shot approaches has the potential to achieve this.

Another direction will be the integration of LLMs into mixed pipelines. LLMs are helpful for fast prototyping and scenarios where the available annotated data is scarce.

While LLMs are powerful in generating and interpreting natural language, supervised DL models

still provide more efficiency, accuracy, reliability and privacy in discriminative tasks, particularly in specialized fields as the biomedical one. Comprehensive TM and NLP applications will likely combine both types of approaches, LLMs and EL, to provide more flexibility, better performance, and also more explainability.

Another critical challenge is better dealing with text in languages besides English, which is particularly important in the clinical domain, where most available text is non-English.

While the large-scale deployment and widespread availability of approaches based on LLMs offer potential, they also risk increasing the volume of low-quality automatically generated text and, at the same time, worsen the issue of lack of explainability. This raises challenges for existing EL approaches but also opens opportunities since EL will potentially play a role in grounding the output of LLM models to reliable data stored in KOSs.

SOTA DL EL approaches, including X-Linker, are resource-intensive regarding computational power and energy consumption. Developing and applying DL-based approaches require specialized hardware, such as GPUs or Tensor Processing Units (TPUs). The large-scale adoption of LLMs will also worsen the problem. One promising direction is the development of more efficient approaches while maintaining comparable performance.

It would be interesting to move towards more realistic evaluation scenarios. Instead of relying on static datasets, the approaches should be tested through human-based validation involving real end users, which would be particularly relevant in applications focused on the clinical domain.

The issue of reproducibility in the field should also be the focus. There should be an emphasis on publicly sharing datasets and models, which is the more effective way to advance the EL task and the fields of TM and NLP.

References

[1] Abdulnazar, A., Kreuzthaler, M., Roller, R., and Schulz, S. (2023). SapBERT-based medical concept normalization using SNOMED CT. In *Caring is Sharing–Exploiting the Value in Data for Health and Innovation*, pages 825–826. IOS Press. 24

[2] Agarwal, D., Angell, R., Monath, N., and McCallum, A. (2022). Entity Linking via Explicit Mention-Mention Coreference Modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4644–4658, Seattle, United States. Association for Computational Linguistics. 23

[3] Alhelbawy, A. and Gaizauskas, R. (2014). Graph ranking for collective Named Entity Disambiguation. In *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, volume 2, pages 75–80, Baltimore, Maryland, USA, June 23-25 2014. 2014 Association for Computational Linguistics. 25, 48

[4] Allones, J., Martinez, D., and Taboada, M. (2014). Automated mapping of clinical terms into SNOMED-CT. An application to codify procedures in pathology. *Journal of medical systems*, 38:1–14. 22

[5] Almeida, T., Jonker, R. A. A., Antunes, R., Almeida, J. R., and Matos, S. (2024). Towards discovery: an end-to-end system for uncovering novel biomedical relations. *Database*, 2024:baae057. 3

[6] Alnazzawi, N., Thompson, P., and Ananiadou, S. (2016). Mapping phenotypic information in heterogeneous textual sources to a domain-specific terminological resource. *PLoS One*, 11(9):e0162287. 22

[7] Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., and McDermott, M. (2019). Publicly Available Clinical BERT Embeddings. In *Proceedings ofthe 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota. Association for Computational Linguistics. 31, 72

[8] Angell, R., Monath, N., Mohan, S., Yadav, N., and McCallum, A. (2021). Clustering-based Inference for Biomedical Entity Linking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2598–2608, Online. Association for Computational Linguistics. 22

[9] Arighi, C., Hirschman, L., Lemberger, T., Bayer, S., Liechti, R., Comeau, D., and Wu, C. (2017). Bio - ID Track Overview. In *Proceedings of the BioCreative VI Challenge Evaluation Workshop*, pages 14–19. 49

[10] Arnaboldi, V., Raciti, D., Van Auken, K., Chan, J. N., Müller, H.-M., and Sternberg, P. W. (2020). Text mining meets community curation: a newly designed curation platform to improve author experience and participation at WormBase. *Database*, 2020:baaa006. 2

[11] Arp, R., Smith, B., and Spear, A. D. (2015). *Building ontologies with basic formal ontology*. 46

[12] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29. 40

[13] Barros, M., Couto, F. M., Pato, M., and Ruas, P. (2021a). Creating Recommender Systems Datasets in Scientific Fields. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '21, page 4029–4030, New York, NY, USA. Association for Computing Machinery. 14, 124

[14] Barros, M., Ruas, P., Sousa, D., Bangash, A. H., and Couto, F. M. (2021b). COVID-19 recommender system based on an annotated multilingual corpus. *Genomics Inform*, 19(3):e24–. 14, 122

[15] Bast, H., Buchhold, B., and Haussmann, E. (2016). Semantic search on text and knowledge bases. *Foundations and Trends in Information Retrieval*, 10(2-3):119–271. 19

[16] Beltagy, I., Lo, K., and Cohan, A. (2020). SCIBERT: A pretrained language model for scientific text. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics (ACL). 23, 31, 70, 72

[17] Benton, A. and Dredze, M. (2015). Entity linking for spoken language. In *NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 225–230. Association for Computational Linguistics. 19

[18] Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web: A New Form of Web Content That is Meaningful to Computers Will Unleash a Revolution of New Possibilities. *Scientific American*. 31

[19] Bhowmik, R., Stratos, K., and de Melo, G. (2021). Fast and Effective Biomedical Entity Linking Using a Dual Encoder. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 28–37, online. Association for Computational Linguistics. 23

[20] Bitton, Y., Cohen, R., Schifter, T., Bachmat, E., Elhadad, M., and Elhadad, N. (2020). Cross-lingual Unified Medical Language System entity linking in online health communities. *Journal of the American Medical Informatics Association*, 27(10):1585–1592. 23

[21] Blanco, R., Ottaviano, G., and Meij, E. (2015). Fast and Space-Efficient Entity Linking for Queries. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 179–188, New York, NY, USA. Association for Computing Machinery. 2

[22] Blissett, K. and Ji, H. (2019). Cross-lingual NIL Entity Clustering for Low-resource Languages. In *Proceedings of the 2nd Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2019)*, pages 20–25, Minneapolis, USA, June 7, 2019. 2019 Association for Computational Linguistics. 4, 67, 71

[23] Bo, M. and Zhang, M. (2021). Learning Dynamic Coherence with Graph Attention Network for Biomedical Entity Linking. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. 24

[24] Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Reserach*, 32(Database issue):D267–70. 35

[25] Boguslav, M., Cohen, K. B., Jr., W. A. B., and Hunter, L. E. (2018). *Improving precision in concept normalization*. 22

[26] Bornmann, L., Haunschild, R., and Mutz, R. (2021). Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1):224. 1

[27] Bornmann, L. and Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222. 67

[28] Bretonnel Cohen, K., Verspoor, K., Fort, K., Funk, C., Bada, M., Palmer, M., and Hunter, L. (2016). The Colorado Richly Annotated Full Text (CRAFT) Corpus: Multi-Model Annotation In The Biomedical Domain. In *Handbook of Linguistic Annotation*. 79

[29] Broscheit, S. (2019). Investigating Entity Knowledge in BERT with Simple Neural End-To-End Entity Linking. In *Proceedings ofthe 23rd Conference on Computational Natural Language Learnin*, pages 677–685, Hong Kong, China. Association for Computational Linguistics. 19

[30] Brown, E. G., Wood, L., and Wood, S. (1999). The medical dictionary for regulatory activities (MedDRA). *Drug Saf.*, 20(2):109–117. 39

[31] Brown, G. R., Hem, V., Katz, K. S., Ovetsky, M., Wallin, C., Ermolaeva, O., Tolstoy, I., Tatusova, T., Pruitt, K. D., Maglott, D. R., et al. (2015). Gene: a gene-centered information resource at NCBI. *Nucleic Acids Research*, 43(D1):D36–D42. 39

[32] Bunescu, R. and Pasca, M. (2006). Using Encyclopedic Knowledge for Named Entity Disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for the*

Association for Computational Linguistics (EACL-06), number April, pages 9–16, Trento, Italy. 4, 24, 48, 67

[33] Campos, D., Matos, S., and Oliveira, J. L. (2013). A modular framework for biomedical concept recognition. *BMC bioinformatics*, 14:1–21. 22

[34] Cao, Y., Fang, L., and Zheng, Z. (2022). Enriching Pre-Trained Language Model with Multi-Task Learning and Context for Medical Concept Normalization. In *Proceedings of the 2022 International Conference on Intelligent Medicine and Health*, pages 79–83. 23

[35] Cao, Y., Hou, L., Li, J., and Liu, Z. (2018). Neural Collective Entity Linking. 25

[36] Cenikj, G., Petelin, G., Koroušić Seljak, B., and Eftimov, T. (2022). SciFoodNER: Food Named Entity Recognition for Scientific Text. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 4065–4073. 23

[37] Chambers, B. A., Basili, D., Word, L., Baker, N., Middleton, A., Judson, R. S., and Shah, I. (2024). Searching for LINCS to Stress: Using Text Mining to Automate Reference Chemical Curation. *Chemical Research in Toxicology*, 37(6):878–893. 2

[38] Chang, W.-C., Yu, H.-F., Zhong, K., Yang, Y., and Dhillon, I. S. (2020). Taming pretrained transformers for eXtreme multi-label text classification. In *KDD 2020*. 96

[39] Chapman, W. W., Fiszman, M., Dowling, J. N., Chapman, B. E., and Rindflesch, T. C. (2004). Identifying Respiratory Findings in Emergency Department Reports for Biosurveillance using MetaMap. In *Studies in Health Technology and Informatics*, chapter Volume 107, pages 487–491. IOS Press. 19

[40] Chen, L., Fu, W., Gu, Y., Sun, Z., Li, H., Li, E., Jiang, L., Gao, Y., and Huang, Y. (2020). Clinical concept normalization with a hybrid natural language processing system combining multilevel matching and machine learning ranking. *Journal of the American Medical Informatics Association*, 27(10):1576–1584. 24

[41] Chen, L., Varoquaux, G., and Suchanek, F. M. (2021). A Lightweight Neural Model for Biomedical Entity Linking. In *AAAI*. 67, 69

[42] Cho, H., Choi, D., and Lee, H. (2021). Re-Ranking System with BERT for Biomedical Concept Normalization. *IEEE Access*, 9:121253–121262. 23

[43] Cohen, K. B., Verspoor, K., Funk, C., Bada, M., Palmer, M., and Hunter, L. E. (2017). The Colorado Richly Annotated Full Text (CRAFT) Corpus : Multi-Model Annotation in the Biomedical Domain The Colorado Richly Annotated Full Text (CRAFT) Corpus : Multi-Model Annotation In The Biomedical Domain. In *The Handbook of Linguistic Annotation*, volume June. 55

[44] Combi, C., Zorzi, M., Pozzani, G., Arzenton, E., and Moretti, U. (2019). Normalizing Spontaneous Reports Into MedDRA: Some Experiments With MagiCoder. *IEEE Journal of Biomedical and Health Informatics*, 23(1):95–102. 3, 22, 66

[45] Consortium, T. G. O. (2018). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1):D330–D338. 78

[46] Consortium, T. G. O., Aleksander, S. A., Balhoff, J., Carbon, S., Cherry, J. M., Drabkin, H. J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N. L., Hill, D. P., Lee, R., Mi, H., Moxon, S., Mungall, C. J., Muruganugan, A., Mushayahama, T., Sternberg, P. W., Thomas, P. D., Van Auken, K., Ramsey, J., Siegele, D. A., Chisholm, R. L., Fey, P., Aspromonte, M. C., Nugnes, M. V., Quaglia, F., Tosatto, S., Giglio, M., Nadendla, S., Antonazzo, G., Attrill, H., dos Santos, G., Marygold, S., Strelets, V., Tabone, C. J., Thurmond, J., Zhou, P., Ahmed, S. H., Asanitthong, P., Luna Buitrago, D., Erdol, M. N., Gage, M. C., Ali Kadhum, M., Li, K. Y. C., Long, M., Michalak, A., Pesala, A., Pritazahra, A., Saverimuttu, S. C. C., Su, R., Thurlow, K. E., Lovering, R. C., Logie, C., Oliferenko, S., Blake, J., Christie, K., Corbani, L., Dolan, M. E., Drabkin, H. J., Hill, D. P., Ni, L., Sitnikov, D., Smith, C., Cuzick, A., Seager, J., Cooper, L., Elser, J., Jaiswal, P., Gupta, P., Jaiswal, P., Naithani, S., Lera-Ramirez, M., Rutherford, K., Wood, V., De Pons, J. L., Dwinell, M. R., Hayman, G. T., Kaldunski, M. L., Kwitek, A. E., Laulederkind, S. J. F., Tutaj, M. A., Vedi, M., Wang, S.-J., D'Eustachio, P., Aimo, L., Axelsen, K., Bridge, A., Hyka-Nouspikel, N., Morgat, A., Aleksander, S. A., Cherry, J. M., Engel, S. R., Karra, K., Miyasato, S. R., Nash, R. S., Skrzypek, M. S., Weng, S., Wong, E. D., Bakker, E., Berardini, T. Z., Reiser, L., Auchincloss, A., Axelsen, K., Argoud-Puy, G., Blatter, M.-C., Boutet, E., Breuza, L., Bridge, A., Casals-Casas, C., Coudert, E., Estreicher, A., Livia Famiglietti, M., Feuermann, M., Gos, A., Gruaz-Gumowski, N., Hulo, C., Hyka-Nouspikel, N., Jungo, F., Le Mercier, P., Lieberherr,

D., Masson, P., Morgat, A., Pedruzzi, I., Pourcel, L., Poux, S., Rivoire, C., Sundaram, S., Bateman, A., Bowler-Barnett, E., Bye-A-Jee, H., Denny, P., Ignatchenko, A., Ishtiaq, R., Lock, A., Lussi, Y., Magrane, M., Martin, M. J., Orchard, S., Raposo, P., Speretta, E., Tyagi, N., Warner, K., Zaru, R., Diehl, A. D., Lee, R., Chan, J., Diamantakis, S., Raciti, D., Zarowiecki, M., Fisher, M., James-Zorn, C., Ponferrada, V., Zorn, A., Ramachandran, S., Ruzicka, L., and Westerfield, M. (2023). The Gene Ontology knowledge base in 2023. *Genetics*, 224(1):iyad031. 40

[47] Couto, F. M. and Lamurias, A. (2018). MER: a shell script and annotation server for minimal named entity recognition and linking. *Journal of Cheminformatics*, 10(1):58. 122

[48] Couto, F. M., Lamurias, A. L., and Ruas, P. (2024). Semantic Similarity Definition. In *Reference Module in Life Sciences*. Elsevier. 11, 17, 54, 103

[49] Cucerzan, S. (2007). Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In *Proceedings ofthe 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, number June, pages 708–716. 19

[50] Cuffy, C., French, E., Fehrmann, S., and McInnes, B. T. (2022). Exploring Representations for Singular and Multi-Concept Relations for Biomedical Named Entity Normalization. In *Companion Proceedings of the Web Conference 2022*, pages 823–832. 23

[51] Dai, J., Zhang, M., Chen, G., Fan, J., Ngiam, K. Y., and Ooi, B. C. (2018). Fine-grained concept linking using neural networks in healthcare. In *Proceedings of the 2018 International Conference on Management of Data*, pages 51–66. 23

[52] Dai, R., Zhang, X., Li, F., and Li, C. (2024). Research on Normalization of Chinese Clinical Terms Based on Keyword Extraction and Data Augmentation Technology. In *Proceedings of the 2023 4th International Symposium on Artificial Intelligence for Medicine Science*, ISAIMS '23, page 1291–1298, New York, NY, USA. Association for Computing Machinery. 23

[53] Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., McMorran, R., Wiegers, J., Wiegers, T. C., and Mattingly, C. J. (2019). The Comparative Toxicogenomics Database: Update 2019. *Nucleic Acids Research*, 47(D1):D948–D954. 56

[54] Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., Wiegers, J., Wiegers, T. C., and Mattingly, C. J. (2020). Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Research*. gkaa891. 78

[55] Davis, A. P., Wiegers, T. C., Johnson, R. J., Sciaky, D., Wiegers, J., and Mattingly, C. (2022). Comparative Toxicogenomics Database (CTD): update 2023. *Nucleic Acids Research*, 51(D1):D1257–D1262. 2, 39, 95, 101

[56] De Rosa, M., Fenza, G., Gallo, A., Gallo, M., and Loia, V. (2021). Pharmacovigilance in the era of social media: Discovering adverse drug events cross-relating Twitter and PubMed. *Future Generation Computer Systems*, 114:394–402. 3

[57] Deng, P., Chen, H., Huang, M., Ruan, X., and Xu, L. (2019). An ensemble CNN method for biomedical entity normalization. In *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, pages 143–149, Hong Kong, China. Association for Computational Linguistics. 23

[58] Derczynski, L., Nichols, E., van Erp, M., and Limsopatham, N. (2018). Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition. In *Proceedings ofthe 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark, September 7, 2017. Association for Computational Linguistics. 71

[59] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. 23, 30, 49, 72

[60] Divita, G., Zeng, Q. T., Gundlapalli, A. V., Duvall, S., Nebeker, J., and Samore, M. H. (2014). Sophia: a expedient UMLS concept extraction annotator. In *AMIA Annual Symposium Proceedings*, volume 2014, page 467. American Medical Informatics Association. 22

[61] Dong, H., Chen, J., He, Y., Liu, Y., and Horrocks, I. (2023). Reveal the Unknown: Out-of-Knowledge-Base Mention Discovery with Entity Linking. In *Proceedings of the 32nd ACM Inter-*

national Conference on Information and Knowledge Management, CIKM '23, page 452–462, New York, NY, USA. Association for Computing Machinery. 23

[62] Dong, H., Suárez-Paniagua, V., Zhang, H., Wang, M., Whitfield, E., and Wu, H. (2021). Rare disease identification from clinical notes with ontologies and weak supervision. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2294–2298. IEEE. 23

[63] Dong, Z. and Dong, Q. (2003). HowNet - A hybrid language and knowledge resource. In *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE) 2003*, pages 820–824. 30, 72

[64] Doğan, R. I., Leaman, R., and Lu, Z. (2014). NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10. 42, 78, 106

[65] Dredze, M., Mcnamee, P., Rao, D., Gerber, A., and Finin, T. (2010). Entity Disambiguation for Knowledge Base Population. In *Proceedings of the 23rd International Conference on Computational Linguistics*, number August, pages 277–285, Beijing, August 2010. 4, 19, 66, 70

[66] Dror, R., Baumer, G., Shlomov, S., and Reichart, R. (2018). The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. In *Proceedings ofthe 56th Annual Meeting ofthe Association for Computational Linguistics (Long Papers)*, pages 1383–1392, Melbourne, Australia, July 15 - 20, 2018. Association for Computational Linguistics. 82

[67] D'Souza, J. and Ng, V. (2015). Sieve-Based Entity Linking for the Biomedical Domain. In *Proceedings ofthe 53rd Annual Meeting ofthe Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, pages 297–302. 49, 69, 96

[68] Ehrlinger, L. and Wöß, W. (2016). Towards a Definition of Knowledge Graphs. In *International Conference on Semantic Systems*. 35

[69] Emadzadeh, E., Sarker, A., Nikfarjam, A., and Gonzalez, G. (2018). Hybrid Semantic Analysis for Mapping Adverse Drug Reaction Mentions in Tweets to Medical Terminology. In *AMIA Annual Sym-*

posium Proceedings, pages 679–688. American Medical Informatics Association. PMID: 29854133; PMCID: PMC5977584. 3, 22

[70] Fan, M., Zhou, Q., and Zheng, T. F. (2015). Distant Supervision for Entity Linking. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 79–86, Shanghai, China. 4, 95

[71] Färber, M., Rettinger, A., and El Asmar, B. (2016). On Emerging Entity Detection. In *Lecture Notes in Computer Science, vol 10024*, chapter Knowledge, pages 223–238. Springer, Cham. 67, 68, 71

[72] Ferré, A., Bossy, R., Ba, M., Deléger, L., Lavergne, T., Zweigenbaum, P., and Nédellec, C. (2020). Handling Entity Normalization with no Annotated Corpus: Weakly Supervised Methods Based on Distributional Representation and Ontological Information. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1959–1966, Marseille, France. European Language Resources Association. 23

[73] Ferré, A. and Langlais, P. (2023). An analysis of entity normalization evaluation biases in specialized domains. *BMC Bioinformatics*, 24. 21

[74] Fleuren, W. W. and Alkema, W. (2015). Application of text mining in the biomedical domain. *Methods*, 74:97–106. 17

[75] Fogaras, D. and Rácz, B. (2004). Towards Scaling Fully Personalized PageRank. In *Algorithms and Models for the Web-Graph*, volume 3243. 58

[76] Galassi, A., Lippi, M., and Torroni, P. (2020). Attention in Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–18. 8, 28, 72, 76

[77] Ganapathiraju, M. K. and Orii, N. (2013). Research prioritization through prediction of future impact on biomedical science: a position paper on inference-analytics. *GigaScience*, 2(1). 2047-217X-2-11. 48

[78] Ganea, O.-E. and Hofmann, T. (2017). Deep Joint Entity Disambiguation with Local Neural Attention. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*,

pages 2619–2629, Copenhagen, Denmark, September 7–11, 2017. Association for Computational Linguistics. 25, 49

[79] Garda, S. and Leser, U. (2024). BELHD: Improving Biomedical Entity Linking with Homonoym Disambiguation. 97, 102

[80] Gattani, A., Lamba, D. S., Garera, N., Tiwari, M., Chai, X., Das, S., Subramaniam, S., Rajaraman, A., Harinarayan, V., and Doan, A. (2013). Entity Extraction , Linking , Classification , and Tagging for Social Media : a Wikipedia-Based Approach. In *Proceedings of the VLDB Endowment*, number 11, pages 1126–1137. 19

[81] Gentilotti, E., Gorska, A., Cecchini, M. P., Mirandola, M., Meroi, M., De Nardo, P., Sartori, A., De Toffoli, C. K., Kumar-Singh, S., Zanusso, G., Monaco, S., Tacconelli, E., Guedes, M. N. P., Maccarrone, G., Canziani, L. M., Davies, R. J., Vitali, S., Tomassini, G., Barana, B., Pezzani, M. D., Sibani, M., Mazzaferri, F., Savoldi, A., Righi, E., Franchina, G., Mongardi, M., Sorbello, S., Emiliani, M., Cordioli, R., Esposito, A., Sciammarella, C., Rosini, G., Perlini, C., Puviani, F. C., Fasan, D., Visentin, A., Dall'O', S. H., Zanchi, C., Armellini, M., Gibbin, E., Rovigo, L., Tavernaro, L., Rocchi, M., Scardellato, R., Luca, F., Castelli, A., Lattanzi, F., Cutone, C., Salvadori, A. G., Bonato, L., Del Piccolo, L., Marcanti, M., Zonta, M. P., Calì, D., Mason, A., Perlini, C., Konnova, A., Gupta, A., Smet, M., Hotterbeekx, A., Malhotra-Kumar, S., Scipione, G., Rossi, E., Cataudella, S., Casa, C. D., Chandramouli, B., Gioiosa, S., Naranjo, J. M., Ortali, M., Cecchetto, R., Gibellini, D., and the ORCHESTRA-UNIVR Study Group (2024). Chemosensory assessment and impact on quality of life in neurosensorial cluster of the post COVID 19 syndrome. *Scientific Reports*, 14(1):20951. 4

[82] Gobbel, G. T., Reeves, R., Jayaramaraja, S., Giuse, D., Speroff, T., Brown, S. H., Elkin, P. L., and Matheny, M. E. (2014). Development and evaluation of RapTAT: A machine learning system for concept mapping of phrases from medical narratives. *Journal of Biomedical Informatics*, 48:54–65. 23

[83] Grabar, N., Hamon, T., and Bodenreider, O. (2012). Ontologies and terminologies: Continuum or dichotomy? *Applied Ontology*, 7:375–386. 37

[84] Grover, A. and Leskovec, J. (2016). Node2Vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864, San Francisco, USA, August 13 - 17, 2016,. ACM. 8, 75

[85] Groza, T., Köhler, S., Doelken, S., Collier, N., Oellrich, A., Smedley, D., Couto, F. M., Baynam, G., Zankl, A., and Robinson, P. N. (2015). Automatic concept recognition using the Human Phenotype Ontology reference and test suite corpora. *Database*, 2015:1–13. 78

[86] Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2):199–220. 33

[87] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2021). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare*, 3(1). 23, 70, 118, 119

[88] Guan, F. and Tezuka, T. (2022). A medical Q&A system with entity linking and intent recognition. In *2022 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 820–829. IEEE. 24

[89] Guarino, N. (1998). Formal Ontologies and Information Systems. 33

[90] Guo, Z. and Barbosa, D. (2018). Robust named entity disambiguation with random walks. *Semantic Web*, 9(4):459–479. 25, 48, 51, 54

[91] Gustafson, E., Pacheco, J., Wehbe, F., Silverberg, J., and Thompson, W. (2017). A Machine Learning Algorithm for Identifying Atopic Dermatitis in Adults from Electronic Health Records. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 83–90. 23

[92] Gutiérrez-Sacristán, A., Bravo, □., Portero-Tresserra, M., Valverde, O., Armario, A., Blanco-Gandía, M., Farré, A., Fernández-Ibarrondo, L., Fonseca, F., Giraldo, J., Leis, A., Mané, A., Mayer, M., Montagud-Romero, S., Nadal, R., Ortiz, J., Pavon, F. J., Perez, E. J., Rodríguez-Arias, M., Serrano, A., Torrens, M., Warnault, V., Sanz, F., and Furlong, L. I. (2017). Text mining and expert curation to develop a database on psychiatric diseases and their genes. *Database*, 2017:bax043. 2

[93] Harrison, J. E., Weber, S., Jakob, R., and Chute, C. G. (2021). ICD-11: an international classification of diseases for the twenty-first century. *BMC Medical Informatics and Decision Making*, 21(6):206. 41

[94] Hartendorp, F., Seinen, T., van Mulligen, E. M., and Verberne, S. (2024). Biomedical Entity Linking for Dutch: Fine-tuning a Self-alignment BERT Model on an Automatically Generated Wikipedia Corpus. 23

[95] Hasibi, F., Balog, K., and Bratsberg, S. E. (2016). Exploiting Entity Linking in Queries for Entity Retrieval. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, page 209–218, New York, NY, USA. Association for Computing Machinery. 2

[96] Hasibi, F., Balog, K., and Bratsberg, S. E. (2017). Entity Linking in Queries: Efficiency vs. Effectiveness. In *Advances in Information Retrieval: 39th European Conference on IR Research, ECIR 2017*, chapter Entity Lin, pages 40–53. Springer International Publishing. 19

[97] Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., and Steinbeck, C. (2016). ChEBI in 2016 : Improved services and an expanding collection of metabolites. *Nucleic Acids Research*, 44(October 2015):1214–1219. 41, 56, 78

[98] He, J., de Rijke, M., Sevenster, M., van Ommering, R., and Qian, Y. (2011). Generating Links to Background Knowledge: A Case Study Using Narrative Radiology Reports. In *CIKM'11*, page 1867, October 24–28, 2011, Glasgow, Scotland, UK. ACM. 19

[99] Herrero-Zazo, M., Segura-Bedmar, I., Martínez, P., and Declerck, T. (2013). The DDI corpus: An annotated corpus with pharmacological substances and drug-drug interactions. *Journal of Biomedical Informatics*, 46(5):914–920. 58

[100] Hirschman, L., Burns, G. A. P. C., Krallinger, M., Arighi, C., Cohen, K. B., Valencia, A., Wu, C. H., Chatr-Aryamontri, A., Dowell, K. G., Huala, E., Lourenço, A., Nash, R., Veuthey, A.-L., Wiegers, T., and Winter, A. G. (2012). Text mining for the biocuration workflow. *Database*, 2012:bas020. 2

[101] Hjørland, B. (2008). What is Knowledge Organization (KO)? *Knowledge Organization*, 35. 2, 32

[102] Hodge, G. (2000). Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. Technical report, The Digital Library Federation, Council on Library and Information Resources, 1755 Massachusetts Avenue, NW, Suite 500, Washington, DC 20036. 32

[103] Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., and Zimmermann, A. (2021). Knowledge Graphs. *ACM Comput. Surv.*, 54(4). 34, 35

[104] Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366. 26

[105] Iovine, A., Narducci, F., and Semeraro, G. (2020). Conversational Recommender Systems and natural language: A study through the ConveRSE framework. *Decision Support Systems*, 131(June 2019). 19

[106] Irrera, O. and Silvello, G. (2021). Background linking: Joining entity linking with learning to rank models. *CEUR Workshop Proceedings*, 2816:64–77. 19

[107] Islamaj, R., Leaman, R., Kim, S., Kwon, D., Wei, C.-H., Comeau, D. C., Peng, Y., Cissel, D., Coss, C., Fisher, C., Guzman, R., Kochar, P. G., Koppel, S., Trinh, D., Sekiya, K., Ward, J., Whitman, D., Schmidt, S., and Lu, Z. (2021). NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature. *Scientific Data*, 8(1):91. 101, 106

[108] Jackson, R., Matentzoglu, N., Overton, J. A., Vita, R., Balhoff, J. P., Buttigieg, P. L., Carbon, S., Courtot, M., Diehl, A. D., Dooley, D. M., Duncan, W. D., Harris, N. L., Haendel, M. A., Lewis, S. E., Natale, D. A., Osumi-Sutherland, D., Ruttenberg, A., Schriml, L. M., Smith, B., Stoeckert Jr., C. J., Vasilevsky, N. A., Walls, R. L., Zheng, J., Mungall, C. J., and Peters, B. (2021). OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. *Database*, 2021:baab069. 35

[109] Jeynes, J. C. G., James, T., and Corney, M. (2024). *Natural Language Processing for Drug Discovery Knowledge Graphs: Promises and Pitfalls*. Springer US, New York, NY. 3

[110] Ji, H., Grishman, R., Dang, H. T., Griffitt, K., and Ellis, J. (2011). Overview of the TAC 2011 Knowledge Base Population Track. In *TAC 2011 Proceedings Papers*, pages 1–25. 66, 70

[111] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 55(12). 126

[112] Ji, Z., Wei, Q., and Xu, H. (2019). BERT-based Ranking for Biomedical Entity Normalization. 49, 70, 72

[113] Ji, Z., Xia, T., Han, M., and Xiao, J. (2021). A Neural Transition-based Joint Model for Disease Named Entity Recognition and Normalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing Volume 1: Long Papers*, pages 2819–2827, Online. Association for Computational Linguistics. 23

[114] Jiang, J.-Y., Chang, W.-C., Zhang, J., Hsieh, C.-J., and Yu, H.-F. (2024). Entity Disambiguation with Extreme Multi-label Ranking. In *Proceedings of the ACM on Web Conference 2024*, WWW '24, page 4172–4180, New York, NY, USA. Association for Computing Machinery. 97, 99, 110

[115] Joko, H., Hasibi, F., Balog, K., and de Vries, A. P. (2021). *Conversational Entity Linking: Problem Definition and Datasets*, volume 1. Association for Computing Machinery. 19

[116] Jonnagaddala, J., Jue, T. R., Chang, N.-W., and Dai, H.-J. (2016). Improving the dictionary lookup approach for disease normalization using enhanced dictionary and query expansion. *Database*, 2016:baw112. 22

[117] Kalloubi, F., Nfaoui, E. H., and El Beqqali, O. (2014). Graph based tweet entity linking using DBpedia. *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA*, 2014:501–506. 19

[118] Kalyan, K. S., Rajasekharan, A., and Sangeetha, S. (2022). AMMU: A survey of transformer-based biomedical pretrained language models. *Journal of Biomedical Informatics*, 126. 23

[119] Kalyan, K. S. and Sangeetha, S. (2021). Bertmcn: Mapping colloquial phrases to standard medical concepts using bert and highway network. *Artificial Intelligence in Medicine*, 112:102008. 23

[120] Kang, T., Zhang, S., Tang, Y., Hruby, G. W., Rusanov, A., Elhadad, N., and Weng, C. (2017). EliIE: An open-source information extraction system for clinical trial eligibility criteria. *Journal of the American Medical Informatics Association*, 24(6):1062–1071. 3

[121] Karadeniz, İ. and Özgür, A. (2015). Detection and categorization of bacteria habitats using shallow linguistic analysis. *BMC bioinformatics*, 16:1–14. 22

[122] Karadeniz, I. and Özgür, A. (2019). Linking entities through an ontology using word embeddings and syntactic re-ranking. *BMC bioinformatics*, 20:1–12. 23

[123] Kartsaklis, D., Pilehvar, M. T., and Collier, N. (2018). Mapping Text to Knowledge Graph Entities using Multi-Sense LSTMs. 25

[124] Kate, R. J. (2016). Normalizing clinical terms using learned edit distance patterns. *Journal of the American Medical Informatics Association*, 23(2):380–386. 22

[125] Kim, D., Lee, J., So, C. H., Jeon, H., Jeong, M., Choi, Y., Yoon, W., Sung, M., and Kang, J. (2019). A Neural Named Entity Recognition and Multi-Type Normalization Tool for Biomedical Text Mining. *IEEE Access*, 7:73729–73740. 22

[126] Köhler, S., Gargano, M., Matentzoglu, N., Carmody, L. C., Lewis-Smith, D., Vasilevsky, N. A., Danis, D., Balagura, G., Baynam, G., Brower, A. M., Callahan, T. J., Chute, C. G., Est, J. L., Galer, P. D., Ganesan, S., Griese, M., Haimel, M., Pazmandi, J., Hanauer, M., Harris, N. L., Hartnett, M. J., Hastreiter, M., Hauck, F., He, Y., Jeske, T., Kearney, H., Kindle, G., Klein, C., Knoflach, K., Krause, R., Lagorce, D., McMurry, J. A., Miller, J. A., Munoz-Torres, M. C., Peters, R. L., Rapp, C. K., Rath, A. M., Rind, S. A., Rosenberg, A. Z., Segal, M. M., Seidel, M. G., Smedley, D., Talmy, T., Thomas, Y., Wiafe, S. A., Xian, J., Yüksel, Z., Helbig, I., Mungall, C. J., Haendel, M. A., and Robinson, P. N. (2021). The human phenotype ontology in 2021. *Nucleic Acids Research*, 49(D1):D1207–D1217. 78

[127] Kolitsas, N., Ganea, O.-E., and Hofmann, T. (2018). End-to-End Neural Entity Linking. 19

[128] Kraljevic, Z., Searle, T., Shek, A., Roguski, L., Noor, K., Bean, D., Mascio, A., Zhu, L., Folarin, A. A., Roberts, A., Bendayan, R., Richardson, M. P., Stewart, R., Shah, A. D., Wong, W. K., Ibrahim, Z., Teo, J. T., and Dobson, R. J. (2021). Multi-domain clinical natural language processing with

MedCAT: The Medical Concept Annotation Toolkit. *Artificial Intelligence in Medicine*, 117:102083. 24

[129] Lai, T., Ji, H., and Zhai, C. (2021). BERT might be Overkill: A Tiny but Effective Biomedical Entity Linker based on Residual Convolutional Neural Networks. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1631–1639, Punta Cana, Dominican Republic. Association for Computational Linguistics. 24

[130] Lai, T. M., Zhai, C., and Ji, H. (2023). KEBLM: Knowledge-Enhanced Biomedical Language Models. *J. of Biomedical Informatics*, 143(C). 23

[131] Lamurias, A., Ruas, P., and Couto, F. M. (2019a). PPR-SSM: personalized PageRank and semantic similarity measures for entity linking. *BMC Bioinformatics*, 20(1):534. 6, 22, 47, 51, 55, 103

[132] Lamurias, A., Sousa, D., Clarke, L. A., and Couto, F. M. (2019b). BO-LSTM: Classifying relations via long short-term memory networks along biomedical ontologies. *BMC Bioinformatics*, 20(10). 57, 58

[133] Le, P. and Titov, I. (2019). Distant Learning for Entity Linking with Automatic Noise Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4081–4090, Florence, Italy. Association for Computational Linguistics. 4, 95

[134] Leaman, R., Dogan, R. I., and Lu, Z. (2013). DNorm: Disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917. 49, 69

[135] Leaman, R., Khare, R., and Lu, Z. (2015). Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics*, 57:28–37. 22

[136] Leaman, R. and Lu, Z. (2016). TaggerOne: joint named entity recognition and normalization with semi-Markov Models. *Bioinformatics*, 32(18):2839–2846. 22, 49, 96, 101

[137] Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444. 27

[138] Lee, H.-C. and Kao, H.-Y. (2017). CDRnN: A high performance chemical-disease recognizer in biomedical literature. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 374–379. IEEE. 22

[139] Lee, J., Song, H.-J., Yoon, E., Park, S.-B., Park, S.-H., Seo, J.-W., Park, P., and Choi, J. (2018). Automated extraction of Biomarker information from pathology reports. *BMC medical informatics and decision making*, 18:1–11. 22

[140] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240. 23, 31, 63, 70, 72, 120

[141] Lee, K., Hasan, S. A., Farri, O., Choudhary, A., and Agrawal, A. (2017). Medical Concept Normalization for Online User-Generated Texts. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 462–469. 3

[142] Lever, J., Jones, M. R., Danos, A. M., Krysiak, K., Bonakdar, M., Grewal, J. K., Culibrk, L., Griffith, O. L., Griffith, M., and Jones, S. J. M. (2019). Text-mining clinically relevant cancer biomarkers for curation into the CIViC database. *Genome Medicine*, 11(1):78. 2

[143] Li, B. Z., Min, S., Iyer, S., Mehdad, Y., and tau Yih, W. (2020). Efficient One-Pass End-to-End Entity Linking for Questions. 2

[144] Li, F., Jin, Y., Liu, W., Rawat, B. P. S., Cai, P., and Yu, H. (2019). Fine-tuning bidirectional encoder representations from transformers (BERT)–based models on large-scale electronic health record notes: An empirical study. *Journal of Medical Internet Research*, 21(9). 3, 72

[145] Li, H., Chen, Q., Tang, B., Wang, X., Xu, H., Wang, B., and Huang, D. (2017). CNN-based ranking for biomedical entity normalization. *BMC Bioinformatics*, 18(Suppl 11). 25

[146] Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C. H., Leaman, R., Davis, A. P., Mattingly, C. J., Wiegers, T. C., and Lu, Z. (2016). BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016:1–10. 42, 55, 78, 95, 106

[147] Lin, S.-J., Yeh, W.-C., Chiu, Y.-W., Chang, Y.-C., Hsu, M.-H., Chen, Y.-S., and Hsu, W.-L. (2022). A BERT-based ensemble learning approach for the BioCreative VII challenges: full-text chemical identification and multi-label classification in PubMed articles. *Database*, 2022:baac056. 22

[148] Lin, T., Mausam, and Etzioni, O. (2012). No noun phrase left behind: Detecting and typing unlinkable entities. In *EMNLP-CoNLL 2012 - 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Proceedings of the Conference*, number July, pages 893–903, Jeju Island, Korea, 12–14 July 2012. Association for Computational Linguistics. 67, 71

[149] Lin, T.-M., Hung, M.-C., and Lee, L.-H. (2024). Leveraging Dual Gloss Encoders in Chinese Biomedical Entity Linking. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 23(2). 23

[150] Liu, F., Shareghi, E., Meng, Z., Basaldella, M., and Collier, N. (2021). Self-Alignment Pretraining for Biomedical Entity Representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics. 4, 95, 96, 97, 108

[151] Lobo, M., Lamurias, A., and Couto, F. M. (2017). Identifying human phenotype terms by combining machine learning and validation rules. *BioMed Research International*, 2017. 78

[152] Logeswaran, L., Chang, M.-W., Lee, K., Toutanova, K., Devlin, J., and Lee, H. (2019). Zero-Shot Entity Linking by Reading Entity Descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460, Florence, Italy. Association for Computational Linguistics. 97

[153] López-García, G., Jerez, J. M., Ribelles, N., Alba, E., and Veredas, F. J. (2023). Explainable clinical coding with in-domain adapted transformers. *J. of Biomedical Informatics*, 139(C). 23

[154] Lou, Y., Zhang, Y., Qian, T., Li, F., Xiong, S., and Ji, D. (2017). A transition-based joint model for disease named entity recognition and normalization. *Bioinformatics*, 33(15):2363–2371. 22

[155] Lowe, H. J. and Barnett, G. O. (1987). MicroMeSH: a microcomputer system for searching and exploring the National Library of Medicine's Medical Subject Headings (MeSH) Vocabulary. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 717. American Medical Informatics Association. 38

[156] Lu, W., Zhang, G., Peng, X., Guan, H., and Wang, S. (2024). Medical Entity Disambiguation with Medical Mention Relation and Fine-grained Entity Knowledge. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11148–11158, Torino, Italia. ELRA and ICCL. 23

[157] Luo, L., Lai, P.-T., Wei, C.-H., Arighi, C. N., and Lu, Z. (2022). BioRED: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282. 106, 119

[158] Luo, Y.-F., Sun, W., and Rumshisky, A. (2018). A Hybrid Method for Normalization of Medical Concepts in Clinical Narrative. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 392–393. 23

[159] Madan, S., Julius Zimmer, F., Balabin, H., Schaaf, S., Fröhlich, H., Fluck, J., Neuner, I., Mathiak, K., Hofmann-Apitius, M., and Sarkheil, P. (2022). Deep Learning-based detection of psychiatric attributes from German mental health records. *International Journal of Medical Informatics*, 161:104724. 3

[160] Magge, A., Klein, A., Miranda-Escalada, A., Ali Al-Garadi, M., Alimova, I., Miftahutdinov, Z., Farre, E., Lima López, S., Flores, I., O'Connor, K., Weissenbacher, D., Tutubalina, E., Sarker, A., Banda, J., Krallinger, M., and Gonzalez-Hernandez, G. (2021). Overview of the Sixth Social Media Mining for Health Applications (#SMM4H) Shared Tasks at NAACL 2021. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 21–32, Mexico City, Mexico. Association for Computational Linguistics. 123

[161] Maglott, D., Barrett, T., Murphy, T., Feolo, M., Wagner, L., and Agarwala, R. (2013). Genes and Gene Expression. In *The NCBI Handbook [Internet]. 2nd edition*. National Center for Biotechnology Information (US). 39

[162] Matthew E. Peters, Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of NAACL-HLT 2018*, pages 2227–2237, New Orleans, Louisiana, June 1 - 6, 2018. 31

[163] Mehryary, F., Hakala, K., Kaewphan, S., Björne, J., Salakoski, T., and Ginter, F. (2017). End-

to-End System for Bacteria Habitat Extraction. In *Proceedings ofthe BioNLP 2017 workshop*, pages 80–90. Association for Computational Linguistics. 19, 22

[164] Meij, E., Balog, K., and Odijk, D. (2014). Entity linking and retrieval for semantic search. In *WSDM 2014 - Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, number February 2014, page 683, New York, New York, USA. 2, 19, 66

[165] Meij, E., Weerkamp, W., and de Rijke, M. (2012). Adding semantics to microblog posts. *Proceedings of the fifth ACM international conference on Web search and data mining - WSDM '12*, page 563. 19

[166] Miftahutdinov, Z., Kadurin, A., Kudrin, R., and Tutubalina, E. (2021). Medical Concept Normalization in Clinical Trials with Drug and Disease Representation Learning. *Bioinformatics*, (07):1–9. 67, 70, 90

[167] Miftahutdinov, Z. and Tutubalina, E. (2017). End-to-end deep framework for disease named entity recognition using social media data. In *2017 IEEE 30th Neumann Colloquium (NC)*, pages 000047–000052. 3, 23

[168] Mihalcea, R. and Csomai, A. (2007). Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242, Lisbon, Portugal. ACM New York, NY, USA ©2007. 24

[169] Mikolov, T., Sutskever, I., Chen, K. C., Corrado, G. S., and Dean, J. J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, volume 26, pages Pages 3111–3119. Curran Associates Inc.57 Morehouse LaneRed HookNYUnited States. 74

[170] Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics. 95

[Miranda-Escalada et al.] Miranda-Escalada, A., Farré, E., and Krallinger, M. Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. 121

[172] Miranda-Escalada, A., Farré-Maduell, E., Lima-López, S., Gascó, L., Briva-Iglesias, V., Agüero-Torales, M., and Krallinger, M. (2021). The ProfNER shared task on automatic recognition of occupation mentions in social media: systems, evaluation, guidelines, embeddings and corpora. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 13–20, Mexico City, Mexico. Association for Computational Linguistics. 123

[173] Mohan, S., Angell, R., Monath, N., and McCallum, A. (2021). Low resource recognition and linking of biomedical concepts from a large ontology. In *Proceedings of the 12th ACM conference on bioinformatics, computational biology, and health informatics*, pages 1–10. 23

[174] Mohan, S. and Li, D. (2019). MedMentions: A Large Biomedical Corpus Annotated with {UMLS} Concepts. In *Automated Knowledge Base Construction (AKBC)*. 42, 79

[175] Mrini, K., Nie, S., Gu, J., Wang, S., Sanjabi, M., and Firooz, H. (2022). Detection, Disambiguation, Re-ranking: Autoregressive Entity Linking as a Multi-Task Problem. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1972–1983, Dublin, Ireland. Association for Computational Linguistics. 23

[176] Musto, C., Semeraro, G., Lops, P., and de Gemmis, M. (2014). Combining distributional semantics and entity linking for context-aware content-based recommendation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8538:381–392. 19

[177] Nadkarni, P. M., Ohno-Machado, L., and Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551. 18

[178] Nguyen, D. B., Theobald, M., and Weikum, G. (2017). J-REED: Joint Relation Extraction and Entity Disambiguation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17*, pages 2227–2230. 49

[179] Nie, F., Cao, Y., Wang, J., Lin, C. Y., and Pan, R. (2018). Mention and entity description co-attention for entity disambiguation. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 5908–5915. 72

[180] Ningtyas, A. M., Hanbury, A., Piroi, F., and Andersson, L. (2022). Data Augmentation for Layperson's Medical Entity Linking Task. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '21, page 99–106, New York, NY, USA. Association for Computing Machinery. 3, 23

[181] Noh, J. and Kavuluru, R. (2021). Joint learning for biomedical NER and entity normalization: encoding schemes, counterfactual examples, and zero-shot evaluation. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–10. 23

[182] Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A., and Taylor, J. (2019). Industry-scale knowledge graphs: lessons and challenges. *Commun. ACM*, 62(8):36–43. 34, 38

[183] Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank Citation Ranking : Bringing Order to the Web. In *The Web Conference*. 48, 103

[184] Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359. 30

[185] Pape-Haugaard, L. et al. (2020). Clinical concept normalization on medical records using word embeddings and heuristics. *Digital Personalized Health and Medicine: Proceedings of MIE*, 2020(270):93. 24

[186] Pattisapu, N., Anand, V., Patil, S., Palshikar, G., and Varma, V. (2020). Distant supervision for medical concept normalization. *Journal of Biomedical Informatics*, 109:103522. 4, 95

[187] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. 77

[188] Pelletier, F. J. (1994). The Principle of Semantic Compositionality. *Topoi*, 13(1):11–24. 72

[189] Pérez, A., Atutxa, A., Casillas, A., Gojenola, K., and Sellart, Á. (2018). Inferred joint multigram models for medical term normalization according to ICD. *International journal of medical informatics*, 110:111–117. 23

[190] Perez, N., Accuosto, P., Bravo, À., Cuadros, M., Martínez-Garcia, E., Saggion, H., and Rigau, G. (2020). Cross-lingual semantic annotation of biomedical literature: experiments in Spanish and English. *Bioinformatics*, 36(6):1872–1880. 22

[191] Pershina, M., He, Y., and Grishman, R. (2015). Personalized Page Rank for Named Entity Disambiguation. In *Human Language Technologies: The 2015 Annual Conference ofthe North American Chapter ofthe ACL*, number Section 4, pages 238–243, Denver, Colorado, May 31 – June 5, 2015. Association for Computational Linguistics. 6, 25, 48, 51, 54, 58, 103, 112

[192] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics. 23

[193] Phan, M. C., Sun, A., Tay, Y., Han, J., and Li, C. (2019). Pair-Linking for Collective Entity Disambiguation: Two Could Be Better Than All. *IEEE Transactions on Knowledge and Data Engineering*, 31(7):1383–1396. 112

[194] Qi, F., Huang, J., Yang, C., Liu, Z., Chen, X., Liu, Q., and Sun, M. (2019). Modeling Semantic Compositionality with Sememe Knowledge. In *Proceedings ofthe 57th Annual Meeting ofthe Association for Computational Linguistics*, pages 5706–5715, Florence, Italy, July 28 - August 2, 2019. Association for Computational Linguistics. 8, 29, 30, 72, 73, 75

[195] Raiman, J. and Raiman, O. (2018). DeepType: Multilingual entity linking by neural type system evolution. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 5406–5413. 25

[196] Rak, R., Batista-navarro, R. T., Rowley, A., Carter, J., and Ananiadou, S. (2014). Text-mining-assisted biocuration workflows in Argo. *Database*, pages 1–14. 19

[197] Ran, C., Shen, W., Gao, J., Li, Y., Wang, J., and Jia, Y. (2023). Learning Entity Linking Features for Emerging Entities. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7088–7102. 110

[198] Rao, D., McNamee, P., and Dredze, M. (2013). Entity Linking: Finding Extracted Entities in a Knowledge Base. In *Multi-source, Multilingual Information Extraction and Summarization. Theory and Applications of Natural Language Processing.*, pages 93–115. Springer, Berlin, Heidelberg. 2, 19, 47, 66

[199] Ratinov, L., Roth, D., Downey, D., and Anderson, M. (2011). Local and Global Algorithms for Disambiguation to Wikipedia. In *HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1375–1384, Portland, Oregon — June 19 - 24, 2011. Association for Computational Linguistics Stroudsburg, PA, USA ©2011. 24

[200] Ravikumar, K. E., Wagholikar, K. B., and Liu, H. (2014). Challenges in adapting text mining for full text articles to assist pathway curation. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '14, page 551–558, New York, NY, USA. Association for Computing Machinery. 2

[201] Ren, J., Li, G., Ross, K., Arighi, C., McGarvey, P., Rao, S., Cowart, J., Madhavan, S., Vijay-Shanker, K., and Wu, C. H. (2018). iTextMine: integrated text-mining system for large-scale knowledge extraction from the literature. *Database*, 2018:bay128. 3

[202] Ren, K., Lai, A. M., Mukhopadhyay, A., Machiraju, R., Huang, K., and Xiang, Y. (2014). Effectively processing medical term queries on the UMLS Metathesaurus by layered dynamic programming. *BMC Medical Genomics*, 7(1):S11. 22

[203] Rocha Souza, R., Tudhope, D., and Almeida, M. (2012). Towards a Taxonomy of KOS: Dimensions for Classifying Knowledge Organization Systems. *Knowledge Organization*, 39:179–192. 32

[204] Ruas, P. (2021). Deep Semantic Entity Linking. In *Advances in Information Retrieval. ECIR 2021. Lecture Notes in Computer Science*, volume 12657, pages 682–687. Springer, Cham. 13, 117

[205] Ruas, P., Andrade, V., and Couto, F. M. (2021a). LASIGE-BioTM at MESINESP2: Entity Linking with Semantic Similarity and Extreme Multi-Label Classification on Spanish Biomedical Documents. In *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum*, pages 324–334, Bucharest, Romania. BioASQ: Large-scale biomedical semantic indexing and question answering. 14, 123

[206] Ruas, P., Andrade, V., and Couto, F. M. (2021b). Lasige-BioTM at ProfNER: BiLSTM-CRF and contextual Spanish embeddings for Named Entity Recognition and Tweet Binary Classification. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 108–111, Mexico City, Mexico. Association for Computational Linguistics. 14, 124

[207] Ruas, P. and Couto, F. M. (2022). NILINKER: Attention-based approach to NIL Entity Linking. *Journal of Biomedical Informatics*, 132:104137. 12, 24, 65, 94, 110

[208] Ruas, P., Gallego, F., Veredas, F. J., and Couto, F. M. (2024). Hybrid X-Linker: Automated Data Generation and Extreme Multi-label Ranking for Biomedical Entity Linking. 12, 93

[209] Ruas, P., Lamurias, A., and Couto, F. M. (2020a). LasigeBioTM team at CLEF2020 ChEMU evaluation lab: Named entity recognition and event extraction from chemical reactions described in patents using BioBERT NER and RE. In *The workshop ChEMU: Named Entity Recognition and Event Extraction of Chemical Reactions from Patents (CLEF 2020 Working Notes)*. 13, 121

[210] Ruas, P., Lamurias, A., and Couto, F. M. (2020b). Linking chemical and disease entities to ontologies by integrating PageRank with extracted relations from literature. *Journal of Cheminformatics*, 12(1):1–11. 12, 45, 69, 103

[211] Ruas, P., Neves, A., Andrade, V. D., Couto, F. M., and Aragón, M. E. (2020c). LasigeBioTM at CANTEMIST: Named Entity Recognition and Normalization of Tumour Morphology Entities and Clinical Coding of Spanish Health-related Documents. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020)*, IberLEF, pages 422–437. 13, 122

[212] Ruas, P., Sousa, D. F., Neves, A., Cruz, C., and Couto, F. M. (2023). LASIGE and UNICAGE solution to the NASA LitCoin NLP Competition. 13, 119

[213] Rudniy, A., Song, M., and Geller, J. (2014). Mapping biological entities using the longest approximately common prefix method. *BMC Bioinformatics*, 15(1):187. 22

[214] Sahu, S. K., Christopoulou, F., Miwa, M., and Ananiadou, S. (2019). Inter-sentence relation extraction with document-level graph convolutional neural network. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 4309–4316, Florence, Italy, July 28 - August 2, 2019. Association for Computational Linguistics (ACL). 78

[215] Sarol, M. J., Hong, G., Guerra, E., and Kilicoglu, H. (2024). Integrating deep learning architectures for enhanced biomedical relation extraction: a pipeline approach. *Database*, 2024:baae079. 3

[216] Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., and Karsch-Mizrachi, I. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database*, 2020. 39

[217] Sevgili, Ö., Shelmanov, A., Arkhipov, M., Panchenko, A., and Biemann, C. (2022). Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13:527–570. 3. 4, 94, 95, 109

[218] Sharir, O., Peleg, B., and Shoham, Y. (2020). The Cost of Training NLP Models: A Concise Overview. 21

[219] Shen, W., Wang, J., and Han, J. (2015). Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460. 66, 94

[220] Sohn, S., Clark, C., Halgrim, S. R., Murphy, S. P., Chute, C. G., and Liu, H. (2014). MedXN: an open source medication extraction and normalization tool for clinical text. *Journal of the American Medical Informatics Association*, 21(5):858–865. 22

[221] Sohn, S., Comeau, D. C., Kim, W., and Wilbur, W. J. (2008). Abbreviation definition identification based on automatic precision estimates. *BMC Bioinformatics*, 9(1):402. 116

[222] Sohrab, M. G., Duong, K., Miwa, M., Topić, G., Masami, I., and Hiroya, T. (2020). BENNERD: A Neural Named Entity Linking System for COVID-19. In *Proceedings of the 2020 Conference on*

Empirical Methods in Natural Language Processing: System Demonstrations, pages 182–188, Online. Association for Computational Linguistics. 23

[223] Song, G., Long, Q., Luo, Y., Wang, Y., and Jin, Y. (2022). Deep Convolutional Neural Network Based Medical Concept Normalization. *IEEE Transactions on Big Data*, 8:1195–1208. 24

[224] Sorokin, D. and Gurevych, I. (2018). Mixing Context Granularities for Improved Entity Linking on Question Answering Data across Entity Categories. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 65–75, New Orleans, June 5-6, 2018. Association for Computational Linguistics. 2, 19, 66

[225] Sousa, D. and Couto, F. M. (2020). BiOnt: Deep Learning Using Multiple Biomedical Ontologies for Relation Extraction. In *Advances in Information Retrieval*, pages 367–374, Cham. Springer International Publishing. 118

[226] Stearns, M. Q., Price, C., Spackman, K. A., and Wang, A. Y. (2001). SNOMED clinical terms: overview of the development process and project status. *Proceedings AMIA Symposium*, pages 662–666. 39

[227] Sterckx, L., Vandewiele, G., Dehaene, I., Janssens, O., Ongenae, F., De Backere, F., De Turck, F., Roelens, K., Decruyenaere, J., Van Hoecke, S., and Demeester, T. (2020). Clinical information extraction for preterm birth risk prediction. *Journal of Biomedical Informatics*, 110:103544. 3

[228] Stojanov, R., Kocev, I., Gramatikov, S., Popovski, G., Koroušić Seljak, B., and Eftimov, T. (2020). Toward Robust Food Ontology Mapping. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3596–3601. 23

[229] Su, P., Li, G., Wu, C., and Vijay-Shanker, K. (2019). Using distant supervision to augment manually annotated data for relation extraction. *PLoS One*, 14(7):e0216913. 4, 94

[230] Suh, H. S., Tully, J. L., Meineke, M. N., Waterman, R. S., and Gabriel, R. A. (2022). Identification of preanesthetic history elements by a natural language processing engine. *Anesthesia & Analgesia*, 135(6):1162–1171. 22

[231] Sui, X., Song, K., Zhou, B., Zhang, Y., and Yuan, X. (2022). A Multi-Task Learning Framework for Chinese Medical Procedure Entity Normalization. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8337–8341. 23

[232] Sui, X., Zhang, Y., Cai, X., Song, K., Zhou, B., Yuan, X., and Zhang, W. (2023). BioFEG: Generate Latent Features for Biomedical Entity Linking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11584–11593, Singapore. Association for Computational Linguistics. 23

[233] Sullivan, G. M. and Feinn, R. (2012). Using Effect Size—or Why the P Value Is Not Enough. *Journal of Graduate Medical Education*, 4(3):279. 82

[234] Sun, Y. and Loparo, K. (2019). Information Extraction from Free Text in Clinical Trials with Knowledge-Based Distant Supervision. In *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, volume 1, pages 954–955. 22

[235] Sun, Z. and Tao, C. (2023). Named Entity Recognition and Normalization for Alzheimer's Disease Eligibility Criteria. In *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, pages 558–564. IEEE. 3, 24

[236] Sundermann, C. V., Domingues, M. A., Marcacini, R. M., and Rezende, S. O. (2014). Using topic hierarchies with privileged information to improve context-aware recommender systems. *Proceedings - 2014 Brazilian Conference on Intelligent Systems, BRACIS 2014*, pages 61–66. 19

[237] Sung, M., Jeon, H., Lee, J., and Kang, J. (2020a). Biomedical Entity Representations with Synonym Marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650, Online. Association for Computational Linguistics. 23, 66, 70, 81, 82, 83, 97

[238] Sung, M., Jeong, M., Choi, Y., Kim, D., Lee, J., and Kang, J. (2022). BERN2: an advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics*, 38(20):4837–4839. 24, 70, 97

[239] Sung, S. F., Lin, C. Y., and Hu, Y. H. (2020b). EMR-Based Phenotyping of Ischemic Stroke Using Supervised Machine Learning and Text Mining Techniques. *IEEE Journal of Biomedical and Health Informatics*, 24(10):2922–2931. 3

[240] Tan, A.-H. (1999). Text Mining: The state of the art and the challenges. *Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases*, 8:65–70. 17

[241] Tang, H., Sun, X., Jin, B., and Zhang, F. (2021). A Bidirectional Multi-paragraph Reading Model for Zero-shot Entity Linking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13889–13897. 97

[242] Tari, L., Mulwad, V., and von Reden, A. (2016). Interactive online learning for clinical entity recognition. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, HILDA '16, New York, NY, USA. Association for Computing Machinery. 22

[243] Thompson, P., Daikou, S., Ueno, K., Batista-Navarro, R., Tsujii, J., and Ananiadou, S. (2018). Annotation and detection of drug effects in text for pharmacovigilance. *Journal of Cheminformatics*, 10(1):1–33. 78

[244] Tian, L., Zhang, W., Bikakis, A., Wang, H., Yu, Y., Ni, Y., and Cao, F. (2013). Medetect: a lod-based system for collective entity annotation in biomedicine. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pages 233–240. IEEE. 23

[245] Tohti, T., Abdurxit, M., and Hamdulla, A. (2022). Biomedical Entity Linking Based on Global and Local Feature Fusion. In *2022 International Conference on Asian Language Processing (IALP)*, pages 253–258. 23

[246] Topaz, M., Lai, K., Dowding, D., Lei, V. J., Zisberg, A., Bowles, K. H., and Zhou, L. (2016). Automated identification of wound information in clinical notes of patients with heart diseases: Developing and validating a natural language processing application. *International Journal of Nursing Studies*, 64:25–31. 19

[247] Tsujimura, T., Miwa, M., and Sasaki, Y. (2023). Large-scale neural biomedical entity linking with layer overwriting. *J. of Biomedical Informatics*, 143(C). 23

[248] Tutubalina, E., Kadurin, A., and Miftahutdinov, Z. (2021). Fair Evaluation in Concept Normalization: a Large-scale Comparative Analysis for BERT-based Models. In *Proceedings ofthe 28th International Conference on Computational Linguistics*, pages 6710–6716, Barcelona, Spain (Online), December 8-13, 2020. 79, 95, 109

[249] Tutubalina, E., Miftahutdinov, Z., Nikolenko, S., and Malykh, V. (2018). Medical concept normalization in social media posts with recurrent neural networks. *Journal of Biomedical Informatics*, 84:93–102. 23

[250] Varma, M., Orr, L., Wu, S., Leszczynski, M., Ling, X., and Ré, C. (2021). Cross-Domain Data Integration for Named Entity Disambiguation in Biomedical Text. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4566–4575, Punta Cana, Dominican Republic. Association for Computational Linguistics. 23

[251] Vashishth, S., Newman-Griffis, D., Joshi, R., Dutt, R., and Rose, C. (2020). Improving Broad-Coverage Medical Entity Linking with Semantic Type Prediction and Large-Scale Datasets. 23, 79

[252] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *31st Conference on Neural Information Processing Systems (NIPS 2017*, Long Beach, CA, USA. 23, 29, 30, 72

[253] Vretinaris, A., Lei, C., Efthymiou, V., Qin, X., and Özcan, F. (2021). Medical entity disambiguation using graph neural networks. In *Proceedings of the 2021 international conference on management of data*, pages 2310–2318. 23

[254] Wajsbürt, P., Sarfati, A., and Tannier, X. (2021). Medical concept normalization in French using multilingual terminologies and contextual embeddings. *Journal of Biomedical Informatics*, 114:103684. 23

[255] Wang, J., Mathews, W. C., Pham, H. A., Xu, H., and Zhang, Y. (2020). Opioid2FHIR: A system

for extracting FHIR-compatible opioid prescriptions from clinical text. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1748–1751. 3, 22

[256] Wang, M., Zhang, J., Liu, J., Hu, W., Wang, S., Li, X., and Liu, W. (2017). PDD Graph: Bridging Electronic Medical Records and Biomedical Knowledge Graphs via Entity Linking. In *The Semantic Web – ISWC 2017*, pages 219–227, Cham. Springer International Publishing. 3

[257] Wang, Y., Fan, X., Chen, L., Chang, E. I., Ananiadou, S., Tsujii, J., and Xu, Y. (2019). Mapping anatomical related entities to human body parts based on Wikipedia in discharge summaries. *BMC Bioinformatics*, 20(1):430. 3, 22

[258] Wei, C.-H., Allot, A., Lai, P.-T., Leaman, R., Tian, S., Luo, L., Jin, Q., Wang, Z., Chen, Q., and Lu, Z. (2024). PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge. *Nucleic Acids Research*, page gkae235. 9, 100, 101

[259] Wei, C.-H., Allot, A., Leaman, R., and Lu, Z. (2019). PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Research*, 47(W1):W587–W593. 3, 23

[260] Wei, C.-H., Allot, A., Riehle, K., Milosavljevic, A., and Lu, Z. (2022). tmVar 3.0: an improved variant concept recognition and normalization tool. *Bioinformatics*, 38(18):4449–4451. 3

[261] Wei, C.-H., Harris, B. R., Li, D., Berardini, T. Z., Huala, E., Kao, H.-Y., and Lu, Z. (2012). Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts. *Database*, 2012:bas041. 2

[262] Wei, Q., Chen, T., Xu, R., He, Y., and Gui, L. (2016). Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks. *Database*, 2016:baw140. 23

[263] Wiatrak, M. and Iso-Sipila, J. (2020). Simple Hierarchical Multi-Task Neural End-To-End Entity Linking for Biomedical Text. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 12–17, Online. Association for Computational Linguistics. 23

[264] Wu, G., He, Y., and Hu, X. (2018). Entity Linking: An Issue to Extract Corresponding Entity with Knowledge Base. *IEEE Access*, 6:6220–6231. 70

[265] Wu, L., Petroni, F., Josifoski, M., Riedel, S., and Zettlemoyer, L. (2020). Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics. 97

[266] Xu, D., Gopale, M., Zhang, J., Brown, K., Begoli, E., and Bethard, S. (2020). Unified Medical Language System resources improve sieve-based generation and Bidirectional Encoder Representations from Transformers (BERT)–based ranking for concept normalization. *Journal of the American Medical Informatics Association*, 27(10):1510–1519. 23

[267] Xu, J., Gan, L., Cheng, M., Wu, Q., et al. (2018). Unsupervised medical entity recognition and linking in Chinese online medical text. *Journal of healthcare engineering*, 2018. 22

[268] Yamada, I., Ito, T., Takeda, H., and Takefuji, Y. (2018). Linkify: Enhancing Text Reading Experience by Detecting and Linking Helpful Entities to Users. *IEEE Intelligent Systems*, 33(5):37–46. 19

[269] Yamada, I. and Shindo, H. (2019). Pre-training of Deep Contextualized Embeddings of Words and Entities for Named Entity Disambiguation. 49

[270] Yan, C., Zhang, Y., Liu, K., Zhao, J., Shi, Y., and Liu, S. (2021). Enhancing unsupervised medical entity linking with multi-instance learning. *BMC medical informatics and decision making*, 21:1–10. 24

[271] Yang, S., Zhang, P., Che, C., and Zhong, Z. (2023). B-LBConA: a medical entity disambiguation model based on Bio-LinkBERT and context-aware mechanism. *BMC bioinformatics*, 24(1):97. 23

[272] Yao, Z., Cao, L., and Pan, H. (2020). Zero-shot Entity Linking with Efficient Long Range Sequence Modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2517–2522, Online. Association for Computational Linguistics. 97

[273] Yin, X., Huang, Y., Zhou, B., Li, A., Lan, L., and Jia, Y. (2019). Deep Entity Linking via Eliminating Semantic Ambiguity With BERT. *IEEE Access*, 7:169434–169445. 25, 49, 72

[274] Yu, H.-F., Zhong, K., Zhang, J., Chang, W.-C., and Dhillon, I. S. (2022). PECOS: Prediction for Enormous and Correlated Output Spaces. *Journal of Machine Learning Research*. 9, 99, 116

[275] Yuan, H., Yuan, Z., and Yu, S. (2022a). Generative Biomedical Entity Linking via Knowledge Base-Guided Pre-training and Synonyms-Aware Fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4038–4048, Seattle, United States. Association for Computational Linguistics. 31, 96

[276] Yuan, Z., Zhao, Z., Sun, H., Li, J., Wang, F., and Yu, S. (2022b). CODER: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of biomedical informatics*, 126:103983. 23

[277] Zhang, S., Cheng, H., Vashishth, S., Wong, C., Xiao, J., Liu, X., Naumann, T., Gao, J., and Poon, H. (2022). Knowledge-Rich Self-Supervision for Biomedical Entity Linking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 868–880, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. 4, 23, 95, 96, 97

[278] Zhang, Y., Chen, Q., Yang, Z., Lin, H., and Lu, Z. (2019). BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific data*, 6(1):52. 8, 75

[279] Zhang, Y., Ma, X., and Song, G. (2018). Chinese medical concept normalization by using text and comorbidity network embedding. In *2018 IEEE international conference on data mining (ICDM)*, pages 777–786. IEEE. 23

[280] Zhao, S., Su, C., Lu, Z., and Wang, F. (2020). Recent advances in biomedical literature mining. *Briefings in Bioinformatics*, 22(3):bbaa057. 3

[281] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., and Wen, J.-R. (2023). A Survey of Large Language Models. 30

[282] Zheng, J. G., Howsmon, D., Zhang, B., Hahn, J., McGuinness, D., Hendler, J., and Ji, H. (2015).

Entity linking for biomedical literature. *BMC Medical Informatics and Decision Making*, 15(1):1–9. 21, 23, 47

[283] Zhou, H., Ning, S., Liu, Z., Lang, C., Liu, Z., and Lei, B. (2020). Knowledge-enhanced biomedical named entity recognition and normalization: application to proteins and genes. *BMC bioinformatics*, 21:1–15. 23

[284] Zhu, T., Qin, Y., Chen, Q., Mu, X., Yu, C., and Xiang, Y. (2023a). Controllable Contrastive Generation for Multilingual Biomedical Entity Linking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5742–5753, Singapore. Association for Computational Linguistics. 31

[285] Zhu, T., Qin, Y., Feng, M., Chen, Q., Hu, B., and Xiang, Y. (2023b). BioPRO: Context-Infused Prompt Learning for Biomedical Entity Linking. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 32:374–385. 31

[286] Zwicklbauer, S., Seifert, C., and Granitzer, M. (2015). Search-based entity disambiguation with document-centric knowledge bases. In *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business*, pages 1–8. 22

[287] Zwicklbauer, S., Seifert, C., and Granitzer, M. (2016). Robust and Collective Entity Disambiguation through Semantic Embeddings. In *SIGIR '16: Conference on Research and Development in Information Retrieval*, pages 425–434, Pisa, Italy. ACM. 25

Appendix A

Systematic Review of Named Entity Linking and Knowledge Organization Systems in Biomedical and Clinical Domains

Systematic Review of Named Entity Linking and Knowledge Organisation Systems in Biomedical and Clinical Domains

PEDRO RUAS*, SOFIA I. R. CONCEIÇÃO*, and FRANCISCO M. COUTO*, LASIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Portugal

Knowledge and information in the biomedical domain are predominantly represented in text format and available in diverse sources, such as articles, patents, electronic health records, and social media. Automated approaches for organising these resources face challenges understanding natural language. Knowledge organisation systems serve as a crucial standard for enhancing information retrieval and sharing for humans and automated approaches. Named entity linking plays a crucial role in bridging the gap between text and knowledge organisation systems: it maps relevant entities described in text, such as diseases, chemicals, and genes, to unambiguous entries in target knowledge organisation systems that accurately describe their meaning. In this review, we analysed 102 articles published between 2013 and 2024 related to named entity linking, which provides an overview of the landscape of knowledge organisation systems in the biomedical and clinical domains and of the evolution of this task over the past decade. The findings in this study also uncover the limitations of existing approaches and propose future strategies for improvement.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Information systems** → **Digital libraries and archives**; • **Applied computing** → **Health informatics**; **Bioinformatics**.

Additional Key Words and Phrases: Natural Language Processing, Text Mining, Named Entity Linking, Knowledge Organisation System, Biomedical, Clinical, Information Extraction

ACM Reference Format:
Pedro Ruas, Sofia I. R. Conceição, and Francisco M. Couto. 2024. Systematic Review of Named Entity Linking and Knowledge Organisation Systems in Biomedical and Clinical Domains. 1, 1 (September 2024), 31 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Biomedical and clinical information and knowledge are available in multiple formats. One of the most prevalent formats is text, which comes from heterogeneous sources, such as articles, patents, electronic health records, and social media, among others. The high volume of biomedical and clinical text available today raises challenges regarding access, sharing and reuse.

Knowledge organisation systems (KOS), including vocabularies, terminologies, ontologies, knowledge bases and graphs, are thus essential to managing biomedical and clinical information and knowledge [9]. KOS standardise knowledge and information, providing easier retrieval and sharing by scientists and automated approaches. However, to ensure the quality of the KOS content, a continued curation process is necessary for processing the information presented in text and storing it in the structure of the KOS. The curation is either manual, in which a human expert

*Pedro Ruas and Sofia I. R. Conceição contributed equally to this research.

Authors' address: Pedro Ruas, psruas@fc.ul.pt; Sofia I. R. Conceição, sconceicao@lasige.di.fc.ul.pt; Francisco M. Couto, fjcouto@ciencias.ulisboa.pt, LASIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, Lisbon, Portugal, 1749-016 Lisboa.

Manuscript submitted to ACM

reads the relevant text source and maps the present information to the KOS, or automatic, in which a computer program does the process. Text mining approaches based on natural language processing (NLP) techniques play an essential role in automatic curation of KOS content [8, 80, 116]. In particular, the approaches focusing on entities, i.e., parts of text, such as diseases, chemicals, genes or species, that are relevant for a given knowledge domain (for an interesting discussion about what constitutes an entity, see [114]). **Named entity recognition** (NER) and **named entity linking** (NEL) approaches identify entities in text and link them to the most relevant entries in a given target KOS, respectively. For example, in biomedical literature there is abundant variation in terminologies, and diverse obstacles are present: acronyms and abbreviations that lead to ambiguity since, without context or background, some acronyms can be mapped to diverse expanded forms [73]; sub-language in which, within a specific field, there are distinct terms or expressions to that field that are not normally employed on others, additionally evolving through time in the field requiring constant update [73] and homonyms, where different entities have the same label [27].

Due to all these challenges, the NEL task plays a crucial role in organising all these convoluted terms by normalising them to the most relevant entry in a KOS. Due to its significance within the biomedical field and the multitude of approaches proposed in recent years, evaluating the current state of the task and the existing KOS is essential for understanding its constraints and devising potential strategies to address them.

There are several reviews addressing the NEL task, the majority outside the biomedical and clinical domains [4, 39, 101, 119, 120, 151]. [60] surveys biomedical NER and NEL dataset resources, [46] focus on clinical NEL, more concretely on electronic health records. [45] recently provided a comprehensive review of the NEL task in the biomedical domain from its origin until today.

Our goal is to assess the current state of the NEL task in the biomedical and clinical domains, by performing a systematic review, following the *Preferred Reporting Items for Reviews and Meta-Analyses* (PRISMA) guidelines, of the literature published from 2013 to 2024. In particular, we focus on the proposed approaches to the task and on the role of KOS. We address the following research questions:

- RQ1: What is the evolution of the NEL task in the last decade in terms of published approaches and used methods?
- RQ2: What KOS are used in the NEL task and how are they used?
- RQ3: What are the current limitations of the NEL task and its future directions?

The contributions of the present work are:

- Systematic PRISMA review of the evolution of NEL task in the biomedical and clinical domains from 2013 to 2024 with 102 reviewed articles;
- Overview of the KOS landscape in the biomedical and clinical domains;
- Public dataset containing information about the reviewed approaches and the datasets used.

We further provide the relevant background information to enhance the understanding of our study.

2 BACKGROUND

2.1 The named entity linking task

The text mining pipeline encompasses two primary entity-centred tasks: NER and NEL. The processing of a given text piece begins by performing NER, with the identification of each entity mentioned e within it. The collection of identified entities within the text or the corpus is denoted as E, and these identified entities are then categorised accordingly.

The goal of NEL is to associate each $e \in E$ with the respective identifier of a target repository or KOS that accurately represents its meaning.

Given an input text corpus or a document I containing n entity mentions $M = \{m_1, m_2, ..., m_n\}$, and a target KOS K containing l entries $E = \{e_1, e_2, ..., e_l\}$, the goal is to assign each mention m_i to its corresponding entry e_j from the target KOS:

$$EntityLinking(M, E) = \{(m_1, e_{j_1}), (m_2, e_{j_2}), ...(m_n, e_{j_n})\}$$

Where e_{j_i} represents the entry linked to mention m_i.

NER and NEL can be tackled as separate tasks or as joint tasks.

The main challenges of the NEL task are:

- Ambiguity: synonyms, homonyms (the same entity mentioned has different meanings according to the context), abbreviations;
- Insufficient coverage of the target KOS: it is necessary to update the existing KOS with information that is newly published, however, manual curation is slow and costly;
- Scarcity of resources for non-English languages.

An example showcasing the ambiguity associated with the NEL task is shown in Fig. 1.

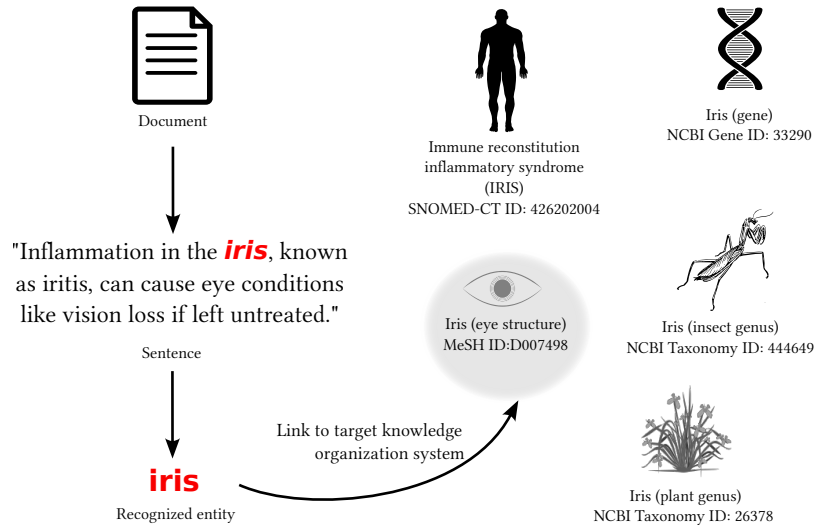


Fig. 1. Example showcasing the linking of the entity mention “iris" to an entry in the target knowledge organisation system *Medical Subject Headings*

.

The typical NEL approach includes two components: candidate generation and candidate disambiguation.

The candidate generation stage builds a list of candidate entities C_{m_i} from the target KOS K for each entity mention $m_i \in M$ (collection of recognised entities) through the function:

$$C_{m_i} = \text{GenerateCandidates}(m, E) = \{c_1, c_2, \ldots, c_n\}$$

This influences the subsequent stages of the NEL approach.

The candidate disambiguation component ranks and selects (or disambiguates) the highest-scoring candidate for each entity mention e through the function:

$$\text{Disambiguate}(m, C_{m_i}) = \{\text{argmax}_{c \in C}\, \text{score}(m, c_i)\}$$

A different axis to classify the approaches consists of the distinction between local and global approaches [111]. Local approaches link each entity mention individually without considering the context of the document where the mention appears. They are usually less complex and thus more efficient. However, they are often unable to deal with ambiguity such as homonymy (when the same entity mention string can have more than one different meaning). On the other hand, global approaches overcome this by performing an interdependent linking of every entity mention. Their goal is to maximize the coherence of the linking decisions at the document or even corpus level. They are usually more precise than local approaches, albeit at the cost of higher complexity.

It is of great importance to ensure the accuracy and reliability of the NEL task because the quality of the outputs influences downstream tasks such as relationship or event extraction. It is essential to consider ethical considerations such as biases in language models, language, and datasets that contribute to skewed distributions in the data. These biases need to be addressed to avoid amplifying the skew [3].

2.2 Knowledge organisation systems

The essential function of a KOS is to organise knowledge and information, which includes their management and representation, by "*organiz[ing] documents, document representations, works and concepts*" [57] and retrieval, to serve "*as a bridge between the user's information need and the material in the collection*" [58].

There are plenty of different typologies of KOS, however, the exact distinction between them is not always clear. One of the most extensive taxonomy of KOS was proposed by Souza [113]. The main division level is based on structure type:

- Unstructured texts: for example, abstracts;
- Concepts, relationship and layout: examples include mind maps, data models, and entity-relationship models;
- Term and/or concept lists: are simple structures typically characterised by alphabetical displays, usually lacking hierarchical arrangements. Examples: dictionaries, gazetteers, glossaries;
- Concept and relationship structures: includes a variety of structures offering varying degrees of relationship expressiveness. Simpler structures consist of hierarchical arrangements with basic hyponym/hyperonym relationships, more complex systems like thesauri may incorporate meronomy (i.e. part-whole relationships) with ontologies providing the highest level of expressiveness. Examples: taxonomies, thesauri, information retrieval indexes, semantic networks, ontologies, and controlled vocabularies.

The secondary division level is related to the application domain and use cases, which encompass a heterogeneous and wide-ranging set of sixty divisions (some are mentioned in the examples above).

One of the main goals of creating and maintaining KOS is enhancing data sharing and reuse across a given domain. The approach to knowledge representation is thus heavily shaped by the goals and language of the community that will use it. In the biomedical and clinical domains, many KOS are categorised and described as **ontologies**. A widely cited definition of ontology in the context of computer and information sciences (i.e. ontology as a knowledge representation artefact and not as a philosophy branch) is the one proposed by Gruber: "*[a]n ontology is an explicit specification of a conceptualization.*" [49].

Guarino expands on this definition, defining that an ontology is an *engineering artefact* that includes a "*specific vocabulary used to describe a certain reality (...)*" and a "*(...) hierarchy of concepts related by subsumption relationships (...)*", which can be complemented by "*axioms (...) to express other relationships between concepts and to constrain their intended interpretation*" [52]. A subsumption relationship, also referred to as a hyponym-hypernym, "is-a" or "SubClassOf" relationship, describes a connection between two classes: a specific class (hyponym) and a generic class (hypernym). In this relationship, instances of the specific class are encompassed within the broader generic class. This analogy draws a parallel to a child-parent relationship. Figure 2 shows an example of the concept "caffeine" represented in the hierarchy of the *Chemical Entities of Biological Interest* (ChEBI) ontology and the different types of relationships with other concepts. A more complete description of the ChEBI ontology is available at Section 4.

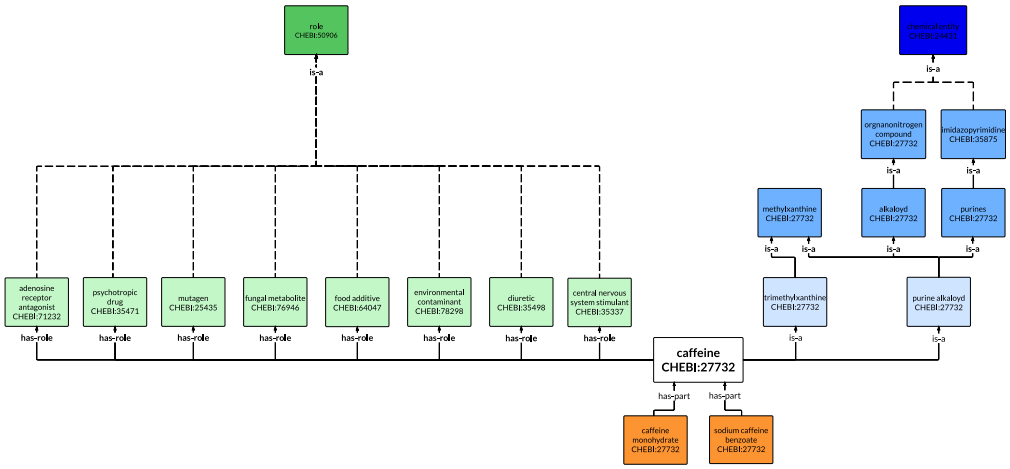


Fig. 2. Context of the concept "caffeine" in the the ChEBI ontology (some relationships and concepts are not shown). Accessed in 24th June 2024. Each box includes the concept name and the respective ChEBI identifier. It is shown three types of relations involving "caffeine": "is-a" (blue boxes), "has-part" (orange boxes), "has-role" (green boxes). Solid lines correspond a to direct connection between two concepts (e.g. "caffeine is-a trimethylxanthine", whereas dashed lines correspond to indirect connections, meaning that one or more ontology levels are omitted (e.g. "organitrogen compound is-a chemical entity.", but "chemical entity is distant ancestor of 'organitrogen compound").

During the last decade, the designation **knowledge graph** has been gaining traction in both academia and industry. Similarly to the confusion in defining different KOS, the definition of what exactly consists of a knowledge graph remains contended. Hogan and coworkers define a knowledge graph as a "*graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent potentially different relations between these entities*" [59], which, overpassing the terminology *nodes* and *edges*, is a definition indistinguishable from that of an ontology. By its turn, the definition by [105] asserts that a knowledge graph is an attempt to model a

certain knowledge domain by providing descriptions of domain entities, designated by *nodes* or *vertices*, and pairwise relations between them, designated by *edges*. A graph can be directed, where an edge between two nodes has direction, or is undirected, in which case the edge is symmetrical.

A common model of knowledge graph consists in the **directed edge-labelled graph** or **multi-relational graph**: set of nodes representing entities, and the directed edges between the nodes represent the relations between the respective entities [59].

To amplify the confusion, in the scientific literature focused on the NEL task the designation **knowledge base** is highly prevalent. However, following the definition of [41], a knowledge base must include a semantic component and a reasoning component able to reason over the "facts" described in the semantic layer, thus, one can argue that designating an ontology as a knowledge base is imprecise.

2.3 The biomedical and clinical landscape of knowledge representation

The current biomedical and clinical knowledge representation landscape is broadly characterised by two differing approaches to ensure data integration and interoperability: the *Unified Medical Language System* (UMLS) [14] and the *Open Biological and Biomedical Ontologies (OBO) Foundry* [61]. Online resources to access biomedical and clinical KOS include: OBO Foundry[1], BioPortal[2], OLS[3] and Ontobee[4].

The UMLS is a collection of resources, including a very large biomedical metathesaurus that provides a standard to connect a wide-ranging set of vocabularies in the biomedical and clinical domains. The metathesaurus includes 3.38 million concepts, from 187 source vocabularies, spanning 27 languages[5]. For some concepts, it provides synonyms and definitions, as well the relations between other concepts. Synonym terms are clustered together to create a concept, which then has relations with other concepts that are either inherited from the respective source vocabulary or are added by the metathesaurus editors. Along with the “Metathesaurus", the UMLS provides two other knowledge sources: “Semantic Network" and the “SPECIALIST Lexicon and Lexical Tools". The "Semantic Network" is a set of broad categories or semantic types and their interrelations. The semantic types categorise every concept present in the UMLS, working in practice as an upper ontology. The latest version (*2023AA*) includes 127 semantic types[6]. The SPECIALIST lexicon is a large syntactic English lexicon that includes both general and biomedical-specific terms[7]. To access the UMLS resources, users must sign a license agreement beforehand and comply with distribution rules. The UMLS includes several vocabularies that can be used separately in NEL.

The OBO Foundry includes 261 ontologies, 184 of them active[8], with a particular focus on basic research. The OBO foundry establishes a set of design patterns and common principles that every ontology must adhere to ensure interoperability. These guidelines focus on ensuring that the ontologies are open, have a common format (OBO format) and are developed collaboratively as a scientific endeavour. Thus, the requirements to access the resources of the OBO foundry are less restrictive compared to the UMLS.

One key difference between the UMLS and the OBO Foundry is the focus of the latter on the development and maintenance of interoperable ontologies, whereas the former focuses on including a wide range of existing terminology

[1] https://obofoundry.org/
[2] https://bioportal.bioontology.org/
[3] https://www.ebi.ac.uk/ols/index
[4] https://ontobee.org/
[5] 2024AA release, statistics accessed at https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html
[6] https://lhncbc.nlm.nih.gov/semanticnetwork/SemanticNetworkArchive.html, accessed in 23th august 2024
[7] https://lhncbc.nlm.nih.gov/LSG/Projects/lexicon/current/web/index.html, accessed in 23th august 2024
[8] As of 23th August 2024. Source: https://obofoundry.org/

and classification systems from various sources. The OBO foundry follows a bottom-up approach, i.e., it intends to guide the development of ontologies according to a set of common rules, whereas the UMLS follows a top-down approach, where the goal is to ensure the interoperability of already existing, heterogeneous KOS.

One must note that the availability of a given KOS in either UMLS or OBO foundry is not mutually exclusive, i.e., several KOS are accessible within both frameworks.

Besides UMLS and OBO foundry, other models exist to ensure data standardisation, particularly in the clinical domain. For instance *The Observational Medical Outcomes Partnership* (OMOP) Common Data Model (CDM)[9] attempts to structure and standardise the storing of data from observational studies, and depends on *The Observational Health Data Sciences and Informatics* (OHDSI) vocabularies and *The Clinical Data Interchange Standards Consortium* (CDISC)[10] focus on standardising clinical data collection.

There is no consensus on classifying semantic resources. For instance, UMLS is described as a metathesaurus, terminological system, or ontology. The classification often depends on the use case and application domain, i.e., an ontology has a different definition whether one is a philosopher or a bioinformatician.

Despite this populated landscape, a major obstacle lies in the conversion of text into structured, useful, shareable data stored within a KOS. A resource is as good as the quality of the data being integrated. Continuously populating the structure with updated and relevant data from the ever-evolving biomedical and clinical fields is crucial to the relevance of KOS. Thus, information extraction pipelines, including NEL approaches, assume a pivotal role in connecting text with KOS [105].

In addition to the variety of KOS employed in the NEL task, it is crucial to consider the diverse approaches used for a comprehensive understanding of the task.

2.4 Categorisation of named entity linking approaches

We expanded the terminology proposed by [44] to categorise the approaches described in the reviewed articles. Note that a given approach can be categorised according to more than one category. This terminology is represented in Fig. 3 and is described below:

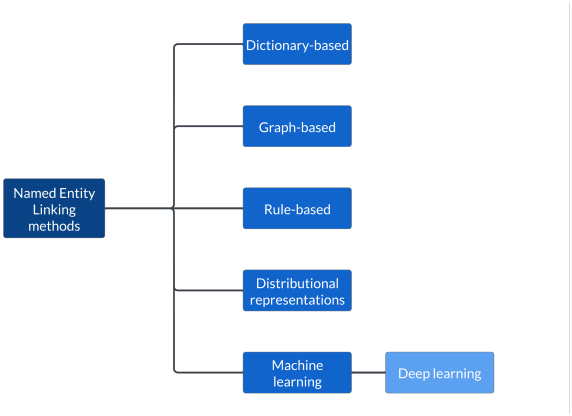


Fig. 3. Categorisation of named entity linking approaches. The categories are not mutually exclusive.

[9] https://ohdsi.org/data-standardization/
[10] https://www.cdisc.org/standards

- **Dictionary-based** (*dict*): approaches based on matching the surface form of the mention (i.e. their text) with the surface forms of the candidates. The matching is based on both mentions and candidates' morphological and syntactic properties (i.e. lexical comparison).
- **Graph-based** (*graph*): approaches involving the building of graphs, usually including every mention in a given document and the respective candidates. A measure of coherence or similarity is usually calculated between mentions/candidates to pick the best candidate.
- **Rule-based** (*rule*): approaches based on symbolic rules crafted by human experts (e.g. "if-then" rules).
- **Distributional representations** (*dist*): mentions and concepts are represented by embeddings instead of their surface forms. The similarity between mention and candidate embeddings can be used directly to pick the best candidate or the embeddings can be included in more complex approaches, such as machine or deep learning models.
- **Machine learning** (*ml*): approaches that learn to do a task without explicit instructions.
  - **Deep learning** (*dl*): approaches based on artificial neural networks including multiple hidden layers.
- other

Assessing the performance of an approach in the NEL task requires common standards to allow for comparison between approaches.

2.5 Evaluation of named entity linking

The performance of NEL is evaluated on annotated datasets or benchmarks, with annotations carried out either automatically (referred to as silver standard) or manually by human annotators (referred to as gold standard). While gold standards allow a more accurate assessment of the performance, human annotations are more expensive, since they require additional time, effort, and domain expertise.

Common dataset formats have been proposed to facilitate interoperability including brat/standoff[11], BioC[12] and Pubtator[13].

The performance of a NEL approach is measured using different metrics according to the context and goal. True positives (TP) refer to the number of entities correctly linked to the appropriate entries in a target KOS, false positives (FP) to the number of entities wrongly linked, and false negatives (FN) to the number of entities that the model should link but fails to do so. **Precision** corresponds to the proportion of true positives out of all the entities linked, measuring how often the predictions of the approach are correct. **Recall** corresponds to the proportion of true positives out of all the entities that should have been linked, measuring how well the approach identifies all relevant entities. The **micro F1-score** evaluates the overall performance by aggregating precision and recall across all entities. To calculate the **macro F-1** score first it is necessary to calculate the F1-score for each entity type or concept in the target KOS and then it is necessary to average all the scores. The macro F1-score gives more weight to less frequent concepts, whereas the micro F1-score treats all concepts equally. **Accuracy@k** checks whether the correct entity is among the top k predictions.

[11] https://brat.nlplab.org/standoff.html
[12] http://bioc.sourceforge.net/
[13] https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/tmTools/Format.html

3 METHODOLOGY

We followed the guidelines of the PRISMA 2020 statement [106]. The review methodology is summarised in Figure 4 and described below.

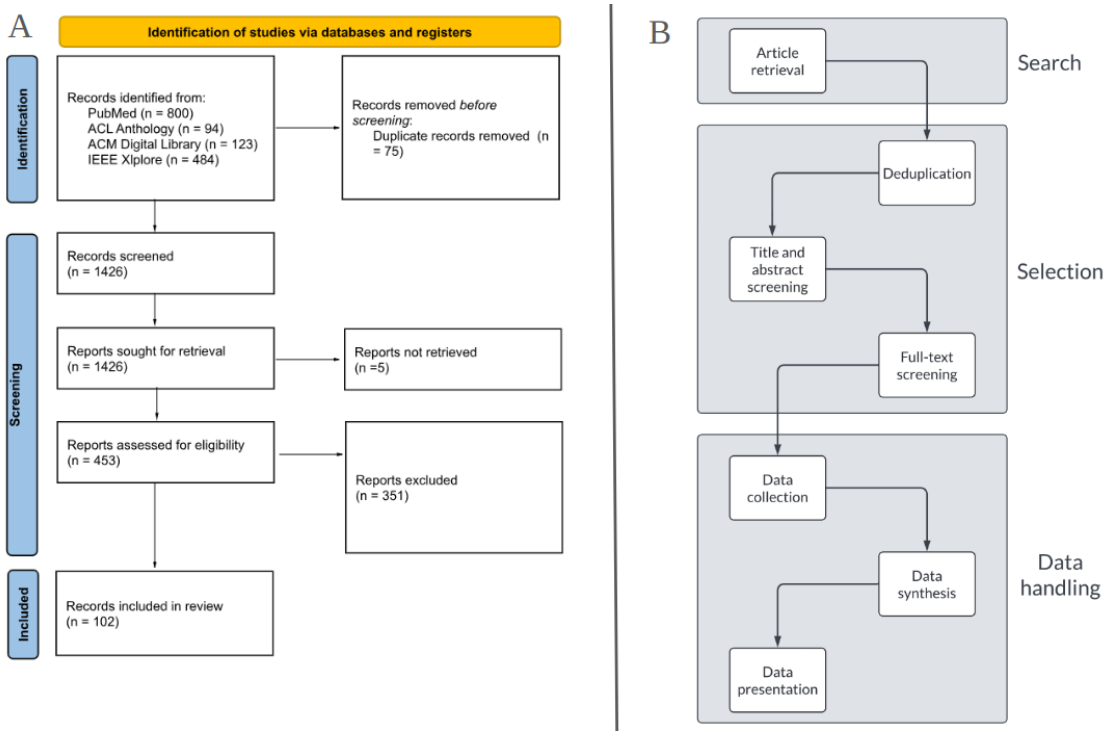


Fig. 4. Review workflows. A- PRISMA flow diagram: results from the databases searches from 2013-2024. B- Review workflow: step by step overview.

3.1 Search

3.1.1 Information sources. We retrieved relevant articles from the following repositories: *IEEE Xplore*[14], a digital library of literature published by the *Institute of Electrical and Electronics Engineers* (IEEE) and partners; *ACM Digital Library*[15], a digital library containing the full-text collection of all literature of the *Association for Computing Machinery* (ACM); *PubMed*[16], a database that comprises citations and abstracts of biomedical literature; *ACL Anthology*[17], provides papers on the field of computational linguistics and NLP.

3.1.2 Search strategy. For each repository, we applied a different search strategy using both manual and automatic techniques according to the available options and characteristics of the repository. Nevertheless, each search query attempted to capture the differing designations for the NEL task. The general query is the following:

[14] https://ieeexplore.ieee.org/Xplore/home.jsp
[15] https://dl.acm.org/
[16] https://pubmed.ncbi.nlm.nih.gov/
[17] https://aclanthology.org/

```
("biomedical" OR "medical" OR "clinical") AND ("entity" OR "term" OR "concept")
AND ("linking" OR "normalization" OR "mapping" OR "disambiguation"
OR "annotation" OR "extraction")
```

We applied additional filters to refine the results further when available, such as topic ("computer science", "natural language processing", "text analysis"), language (English), and publication date. The detailed search strategy for each repository is available in the Appendix A.

Three search rounds were performed:

(1) 13, March 2023 in the repositories *IEEE Xplore*, *ACM Digital Library*, *PubMed*, *ACL anthology*;
(2) 3, April 2024 in the repositories *IEEE Xplore*, *ACM Digital Library*, *PubMed*, *ACL anthology*;
(3) 11, June 2024 in the repositories *IEEE Xplore*, *ACM Digital Library*, *PubMed*, *ACL anthology*.

The search rounds resulted in a large set of articles, which had to be further selected.

3.2 Selection

The selection process was carried out by two NLP experts with biomedical backgrounds and included the following steps:

(1) Deduplication
(2) Title and abstract screening
(3) Full-text screening

We considered the following exclusion criteria:

- Duplicate publication;
- Article is written in non-English languages;
- Article not published in the time interval 2013-2024 (for the year 2024, only articles until 11th of June were considered);
- Articles not addressing NEL (e.g. articles focusing on NER, word sense disambiguation, record linkage, etc.);
- Article that proposes a NEL approach but that is not focused on the biomedical domain;
- Article that does not propose a novel approach (e.g. reviews and surveys, chapters, or articles describing approaches that use existing tools without modifications);
- Article without the methodological description of the approach;
- Article without evaluation of the approach;
- Unavailable full-text.

3.3 Data handling

3.3.1 Data collection. The data was manually collected on the selected articles by reading their text and filling in a pre-defined template based on the research questions.

Appendix B includes an exhaustive description of the data extracted from each article.

4 RESULTS AND DISCUSSION

At the end of the selection and data collection phase and the full-text screening sub-phase, 102 articles were considered for further analysis. Data collection was carried out in these articles. The distribution of the selected articles by source

is the following: 48 from the *PubMed* (47.1 %), 20 from the *ACL Anthology* (19.61 %), 20 from the *IEEE Xplore* (19.61 %) and 14 from the *ACM Digital Library* (13.7 %).

Two datasets with the results are available as supplemental data: reviewed articles and evaluation datasets.

4.1 Evolution of article publication (2013-2024)

This section aims to categorise the selected papers based on the progression of techniques over time. However, given that many of these papers often overlap across multiple categories, it is challenging to classify them without excessive specificity, thus this was done by a sweeping analysis.

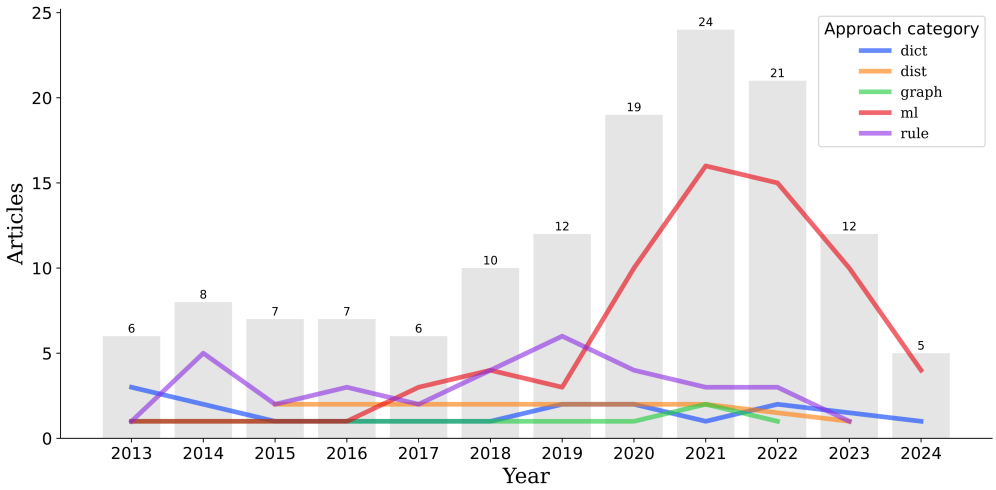


Fig. 5. Distribution of article publication by year (2013-11th June 2024) and the respective categorisation. dict: dictionary-based; ml: machine learning-based; graph: graph-based; rule: rule-based; dist: based on distributional representations.

The distribution of articles and approaches by year is shown in Fig. 5. 2021 is the year with more articles published (N=24), whereas between 2013 and 2017 are the years with fewer articles (N=6). For the year 2024, only articles published until 11th June were considered, which resulted in five selected articles. Until early 2020, the approaches focused more on rules (N=24), followed by machine learning (N=14) and dictionary-based (N=11).

Rule-based models and dictionary-based mainly use techniques like string or semantic matching algorithms based on similarity measure [6, 26, 42, 81, 117, 121, 126, 153], term frequency-inverse document frequency (TF-IDF) [78, 96], mapping systems to a KOS [5, 18, 35, 64, 67, 70, 82, 112, 144, 146], distances such as Levenshtein [69, 85, 89] or probabilities [15, 129].

Although these approaches have some strengths such as flexibility, adaptability and more control due to their transparency they convey many limitations. The limitations that are most encountered in these systems include mismatches due to very similar concepts, wrong string matching due to concept boundaries in the NER phase and the inability to handle synonyms, co-references and syntax-level processing.

For distributional representations, techniques such as vector space [134], k-nearest neighbours [165], semi-Markov models [79] and clustering [7] are employed. Limitations in this category are mainly reported to be confusion between entity types, for example, confusing genes with chemicals or diseases with general biomedical vocabulary.

Graph-based approaches use techniques such as the Personalized Paged Rank algorithm [77, 109] and node importance and inter-node coherence [135]. Reported limitations consist of problems with a lack of edges between candidates and errors dealing with parent-child concepts.

Earlier machine learning approaches consisted in Naive Bayes [47], recurrent neural networks (RNN) [53, 148], convolutional neural networks (CNN) [33, 53, 147], long short-term memory (LSTMs) [93, 97, 138] and n-gram models [108]. These approaches also explored more with unsupervised techniques compared to current approaches [68, 159, 160].

Approaches that fall in the 'Other' category, include more diverse strategies such as in [29] that uses a genetic algorithm based on hypergeometric distribution for candidate generation, having as main drawback the use of incorrect UMLS concepts.

In 2020, the rapid growth of machine learning became more apparent (N=10) following the introduction of the Transformer architecture [141] and the development of models like Bidirectional Encoder Representations from Transformers (BERT) [34]. Research increasingly focused on using pre-trained models such as BERT, BioBERT [83], SciBERT [10], PubMedBERT [50], and ELMo [110]. These models are initially trained on large-scale text datasets, which can include web pages, Wikipedia pages, scientific articles, or other formats, in an unsupervised manner. This means that only unlabelled data is required. This pre-training step generates a language model that has learned the statistical relationships present in the training corpus. The pre-trained model can then be fine-tuned, i.e., further trained on a specific downstream task for which labelled data is available. [65] provides a comprehensive overview of biomedical transformer-based pre-trained language models.

The majority of these approaches are based on learning features to rank the candidates' concepts or better entity disambiguation. Overall, these approaches adopt a more straightforward approach through which the data is fed into the model without any further input or manipulation, holding on the potential to automatically learn the natural features of a dataset [3, 11, 21, 25, 31, 37, 55, 63, 66, 76, 88, 100, 103, 125, 127, 128, 131, 136, 137, 139, 140, 149, 158, 161]. Other approaches attempt to learn representations by integrating certain KOS concepts or semantic types [43, 122, 152, 157], to generate deep contextualized embeddings [12, 19, 28, 30, 38, 86, 143, 155], to include graphs embeddings to better incorporate complex representations [91, 142] and, additionally, other approaches focus on low resource and zero-shot problems [98, 104]. A great deal of issues in the machine learning category are related to ambiguous annotations including entity overlap, hypernyms, hyponyms, failure to process abbreviations and difficulty to deal with complex phrases or expressions outside KOS.

While currently pre-trained models are commonly used, there is still some research regarding more conventional machine learning techniques like LSTMS with Word2Vec and its similar such as BioWordVec, or CNNs [13, 51, 72, 75, 107, 115, 123, 154]. A few models go in more elaborate ways to tackle this task, using a combination of approaches [2, 23, 130, 132].

Recently, Large Language Models (LLMs) and prompt-based learning gained popularity. Given its ability to efficiently incorporate context, prompt-based learning in the context of NEL helps with the issues of ambiguity and name variations [156, 162, 163]. Although these new approaches solve some crucial problems, one drawback is that it requires a large amount of training data, making it computationally expensive.

[74] tries to tackle the problem of specific acronyms and jargon in clinical text by comparing a web mining approach using BING versus the conversational agent ChatGPT[18]. Using discharge summaries from cardiology, dermatology and

[18] https://openai.com/index/chatgpt/

oncology fields as datasets, acronyms were extracted as the context. The results showed that ChatGPT provides a better output than BING since the latest could not resolve acronyms that were scarcely found on the web.

Since only one paper employs LLM in the NEL task in this review, it is not visible swift to heavily employ LLM in the task. It is plausible that the NEL task is no longer being developed independently, but rather refined as part of an end-to-end system using these models. Another explanation is that the models are not yet fit or are an overkill for this task.

4.2 What knowledge organisation systems are used in named entity linking?

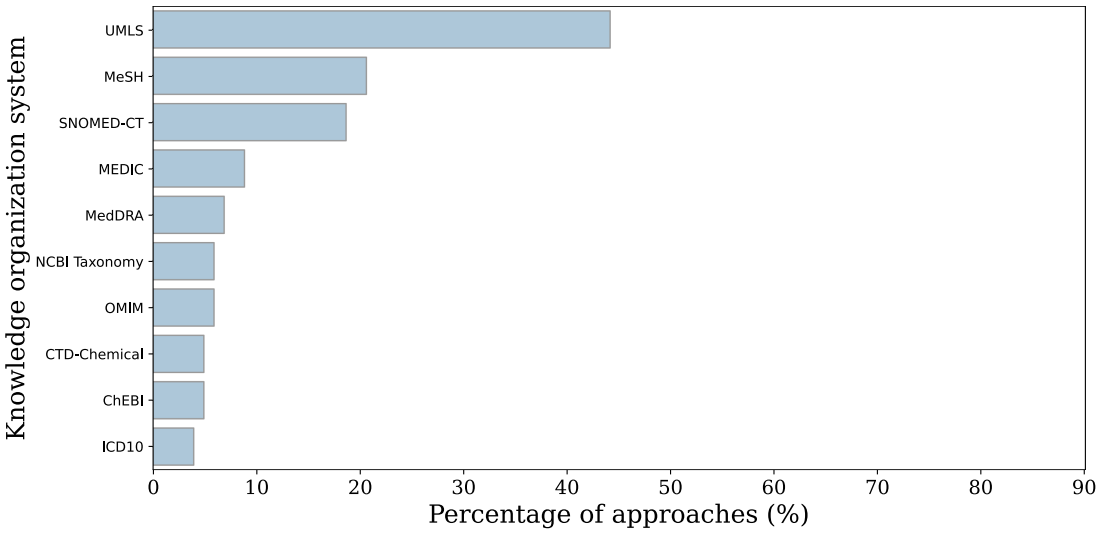


Fig. 6. Top-10 KOS used in the approaches described in the reviewed articles and the respective percentage (e.g. 44.12% of the approaches use UMLS).

Figure 6 shows the target KOS of the NEL approaches described in the reviewed articles. 44.12% (N=45) of the approaches have UMLS as a target knowledge base, followed by MeSH (20.59%, N=21) and *Systematized Nomenclature of Medicine - Clinical Terms* (SNOMED-CT) (18.63 %, N=19).

Below is the description of some of the most commonly used KOS in the NEL task in the biomedical and clinical domains (excluding UMLS, described in Subsection 2.3).

The **MeSH** thesaurus[19] is a controlled and hierarchically organised vocabulary that is used to index journals, cataloguing and search for biomedical articles and health-related information [90]. This thesaurus is actively maintained to modify and update the emerging new concepts while deprecating older ones. In general, these updates are divided into daily updates for supplemental records and annual updates for MeSH Descriptors and MeSH Qualifiers. The main unit of indexing and retrieval is the Descriptors, which are the main headings. Qualifiers are the subheadings and are combined with Descriptors to index citations according to a specific aspect beyond the main headings. There are 16 descriptors categories such as A for anatomic terms, B for organisms, C for diseases, and D for drugs and chemicals,

[19] https://www.nlm.nih.gov/mesh/meshhome.html, accessed in 7th March 2024

which in turn are subdivided into more categories. Each unique descriptor appears at least in one place in the trees and may appear in as many additional locations as necessary.

SNOMED-CT[20] focus on healthcare-related concepts, to facilitate the organisation and storage of clinical data extracted from electronic health records. It was developed by an extensive set of experts, including "*physicians, nurses, physician assistants, pharmacists, informaticians, medical technicians*". It includes four components: Concept Codes, Descriptions, Relations, and Reference Sets (clustering of concepts into sets and cross-references to other KOS). The latest version (version: 2024-05-01, international edition) includes 367,584 concepts. [124]. This KOS is proprietary and maintained by and distributed by SNOMED International, being available in multiple languages.

The **Medical Dictionary for Regulatory Activities** (MedDRA)[21] is an international medical terminology developed by the *International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use* (ICH) [16]. It is focused on pharmacovigilance and clinical research, providing standardisation for "*(...) adverse event information associated with the use of biopharmaceuticals and other medical products (e.g., medical devices and vaccines).*"[22].

The **National Center for Biotechnology Information** (NCBI) includes a comprehensive repertoire of relevant biomedical and clinical information, including several KOS. For instance, the **Gene database**[23], contains information about known and predicted genes from all major taxonomic groups, ranging from viruses to eukaryotes [17, 94]. The gene records provide detailed information such as nomenclature, sequence, structure, function, variations and phenotypes. The **Taxonomy database**[24] [118], as its name suggests, provides a hierarchy of organism names arranged according to phylogenetic criteria. The highest level is the domain, followed by kingdom, phylum, class, order, family, genus, and species.

The **Online Mendelian Inheritance in Man** (OMIM) database[25] contains information about human genes and genetic phenotypes, with an emphasis on the relationship between phenotype and genotype [95]. This database is updated daily and contains information on all known Mendelian disorders.

The **Comparative Toxicogenomics Database** (CTD)[26] provides manually curated information about chemical-gene/protein interactions, chemical-disease and gene-disease relationships. This is a public database that focuses on the progress in the understanding of the influence of environmental exposures on human health [32]. It provides a generalised scope of the entities in diverse human health contexts due to its interconnection with other large vocabularies and KOS. It provides a collection of data about commonly used entity types in biomedical and clinical NEL, such as chemicals, diseases and genes. **CTD-Chemicals** combines a subset of the MeSH thesaurus along with information about their chemical structures, interaction with genes and proteins, disease relationships, enriched pathways and functional annotations. CTD-Disease, best known as **MEDIC disease vocabulary**, maps OMIM diseases to the MeSH disease category, including entries from both resources. **CTD-Genes**, which includes symbols, names and synonyms, is a cross-species vocabulary that arises from the NCBI. It provides information about interaction with chemicals, disease relationships, associated pathways and functional annotations.

The advent of genome sequencing generated vast amounts of molecular data, revealing shared biological functions across all eukaryotic organisms. This led to the creation of the **Gene Ontology** (GO)[27] project, a collaborative effort

[20] https://www.snomed.org/, accessed on 7th March 2024
[21] https://www.meddra.org/
[22] https://www.meddra.org/faq, accessed in 7th March 2024
[23] https://www.ncbi.nlm.nih.gov/gene
[24] https://www.ncbi.nlm.nih.gov/taxonomy
[25] https://www.omim.org/
[26] https://ctdbase.org/
[27] https://geneontology.org/

among several model organism databases, including *FlyBase*, *Mouse Genome Informatics*, and *Saccharomyces Genome Database*. The GO Consortium aimed to develop a unified ontology to describe the functions and roles of genes and gene products across various organisms. GO allows researchers to annotate genes and gene products consistently, enhancing data integration and cross-species comparisons. GO terms are categorised into three distinct ontologies: *Biological process* (describes the processes involving a gene or gene product), *Molecular function* (relates to the biochemical activity of a gene or gene product), *Cellular component* (pertains to cellular or extracellular locations). The latest GO release (2024-01-17) includes 42,442 terms, 7,655,937 annotations, and 1,537,348 annotated gene products from 5,387 species[28]. This ontology is freely available as part of the OBO initiative.

The **ChEBI ontology**[29] [56] focuses on small chemical compounds, including metabolites, drugs, and other bioactive molecules involved in biological processes. ChEBI was built using three data sources: the *Integrated Relational Enzyme* database of the *European Bioinformatics Institute* (IntEnz), *KEGG COMPOUND*, and the *Chemical Ontology*. ChEBI, like similar ontologies, is a graph-theoretic structure with terms as nodes and relationships as edges. However, unlike other ontologies, it is a directed cyclic graph, meaning it contains cycles. The ChEBI ontology includes three high-level categories: *chemical entity*, *role*, and *subatomic particle*. Two of these sub-ontologies classify compounds based on structural chemical features (*chemical entity* and *subatomic particle*), while the *role* category classifies compounds according to their biological and chemical roles. In addition to common relationships such as *is-a* and *is part of*, ChEBI includes specific chemical relationships: *is conjugate acid of*, *is conjugate base of*, *is tautomer of*, *is enantiomer of*, *has functional parent*, *has parent hydride*, and *is substituent group from*. The ChEBI database also stores other relevant information, such as chemical names, synonyms, definitions, chemical information (IUPAC names, SMILES representation), and cross-references with other databases like UniProt and PubChem. The data in ChEBI is freely available as part of the OBO initiative.

The **International Classification of Diseases** (ICD)[30] is a medical classification developed for healthcare, epidemiological and clinical purposes developed by the *World Health Organization* [150]. The purpose of the ICD was to be the worldwide standard for tracking morbidity and mortality statistics, support coding tools in clinical settings and reimbursement systems, and facilitate automated decision-making in healthcare. The ICD structure includes clinical codes ranging from diseases and diagnostics to symptoms and findings. The latest version, ICD-11 replaced ICD-10 after 2022, which was active since 1993[31]. ICD-11 includes about 80,000 entries complemented by 40,000 synonyms, each characterising a disease, syndrome, or health-related phenomenon in a way that not only is descriptive but also specifies its relationships with other entities and provides a way for digital systems to take account of meaning that may be assigned to an entity [54]. The ICD is provided as part of the UMLS.

The choice to utilize UMLS may stem from its extensive coverage, derived from the integration of numerous source KOS. On the other hand, employing individual ontologies from the OBO Foundry may be more suitable for specialised applications where the integration of multiple KOS is unnecessary or less relevant.

A significant obstacle to using large coverage KOS is a lack of specificity and loss of sub-field jargon. Maintaining sub-field vocabulary specificity is critical in domains that are dependent heavily on acronyms, abbreviations, and terminologies, such as signalling and metabolic pathway characterisation.

Barely any papers looked into the use of smaller ontologies (e.g.,[15, 160]). In [160], the authors explored the hierarchical and relational structure, such as the semantic relation between entities (e.g., subClassOf), to enrich domain

[28] https://geneontology.org/stats
[29] https://www.ebi.ac.uk/chebi/
[30] https://www.who.int/standards/classifications/classification-of-diseases
[31] https://www.who.int/standards/classifications/classification-of-diseases

knowledge. The first step was to create a graph representation of the entity mentions, followed by the creation of a knowledge subgraph using the ontology structure, taking into account classes, individuals, and properties. Then, using the knowledge subgraph, each entity is linked to others. The results showed that integrating rich structures into ontologies can help to leverage manual annotation. However, there are problems in the ontologies such as these are also evolving and sometimes lack information, such as vague relations like "related to" that do not provide more specificity.

NEL plays a key role in managing KOS, and, conversely, the functionalities provided by KOS can enhance NEL performance. In particular, KOS features improve in both candidate generation step and candidate ranking.

KOS allow the training of concept embeddings, which can be applied for generating the candidates lists: a similarity measure (e.g. cosine similarity) is calculated between the representation of the input entity mention and the representation of the concepts in the target KOS [25, 53, 68, 87, 125, 131, 136].

The relations defined between concepts can be leveraged to enhance disambiguation graphs, which can include edges based on the relations [77, 115, 160].

Also, The use of semantic types can also be used to filter out irrelevant candidates as described in [140].

4.3 Evaluation in the named entity linking task

Evaluation datasets are critical to access performance. Thus, it is of great importance to understand their quality by assessing their diversity, more concretely, by exploring which entities they encompass and which sources and knowledge bases they use. In this section, we try to make a condensed description of the datasets that were used on the selected papers. The most common datasets used by the approaches described in the reviewed articles are shown in Table 1.

Datasets	Frequency	Percentage (%)
BC5CDR	28	15.47
custom	22	12.15
NCBI Disease	17	9.39
MedMentions	16	8.84
Cadec	6	3.31
MCN	5	2.76
ShARe2013	5	2.76
COMETA	5	2.76
CRAFT	4	2.21
AskAPatient	3	1.66

Table 1. Top-10 evaluation datasets

The most used evaluation datasets in the reviewed articles are BC5CDR with 15% (N=28) followed by 12% (N=22) of custom datasets, 9% (N=17) of the NCBI Disease and 8% (N=16) MedMentions. The top-10 datasets are shown in Table 1.

The BC5CDR [84] was created in 2016 as part of the BioCreative V competition. It is an English gold standard created using lit-abs with sources in *PubMed* having tagged disease and chemical entities with their relation annotations. It is composed of 1,500 articles with 4,409 annotated chemicals, 5,818 annotated diseases and 3,116 chemical-disease interactions linked to MeSH.

NCBI Disease [36] is an English gold standard created in 2014 using 793 lit-abs from PubMed, encompassing 6,891 disease mentions linked to MeSH and OMIM.

MedMentions [99] consists of 350,000 linked mentions of biomedical concept recognition linked to UMLS concepts. It is a gold standard that was built in 2019 using 4,000 lit-abs from *PubMed*. It also has a widely used subset named 'ST21pv" (an acronym for "21 Semantic Types from Preferred Vocabularies") which targets the MedMentions sub-corpus baseline.

In view of existing biases and limitations in current datasets, many authors create a tailored dataset to meet their needs. The disadvantage of this approach is that it is not always reproducible, as some datasets are not publicly available.

The language diversity is low since the majority of the datasets (88.2%, N= 30) include English text, with only a few datasets including text in other languages, such as Chinese (N=2), French (N=1), Hebrew (N= 1) and Spanish (N=1). However, we must note that the search strategies included queries with English terms, which will inherently skew the results towards English results, and we explicitly removed articles expressed in non-English languages in the selection phase. In the clinical domain, particularly, it is critical to develop approaches focusing on the native language of the healthcare workers, since they mostly express themselves in their native language.

UMLS is the most used knowledge base target, accounting for 20% of datasets (N=12), followed by NCBI Taxonomy (10%, N=6), SNOMED-CT (8%, N=5), MeSH (7%, N=4), ChEBI, and OntoBiotope Ontology (5%, N=3). The remaining knowledge bases appear in two or fewer datasets.

Gold standards are the most prevailing standard encompassing 86% of the datasets (N=30) being only 14% (N=5) silver standard ones.

The availability is also one important factor. The greater part of the datasets are publicly available (N=22), some require a license agreement or registration (N=7), and a few are unavailable (N=4).

Competitions are another way to stimulate knowledge advancement in NEL research. Competitions offer targeted problem-solving, which often demands original thinking, as well as method benchmarking that has a closer real-world impact. Some of the existing datasets were developed in the scope of a competition task with a percentage of 38.24% (N=13). Challenges such as the BioCreAtIvE and the BioNLP shared task are some examples of dataset progression. BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) is a challenge for evaluating text mining and information extraction systems in the scope of the biological domain by making use of the community effort. Since its beginning, it has contributed with seventeen datasets[32], five of them used on the surveyed papers, one of them the BC5CDR that is widely used.

4.4 Types of text and entity types

The dominant text type is title and abstracts from articles (*lit-abs*) (52.94%, N=54), followed by electronic health records (20.59%, N=21) and social media text (15.69%, N=16). The top-10 text types are shown in Figure 7.

This highlights the fact that despite the abundance of standards and repositories for storing and disseminating scientific information and data, scientific literature remains the primary means to achieve this goal. Because scientific articles are written in natural language, there is a continuous need to develop and refine text mining and NLP methods. These advancements are crucial for effectively utilising the vast amount of scientific information and data that is constantly being generated.

Predominantly, 47% (N=18) of the text source comes from *PubMed* followed by 18% (N=7) of social media such as blogs, forums, *reddit* or *Twitter*.

[32] https://biocreative.bioinformatics.udel.edu/resources/

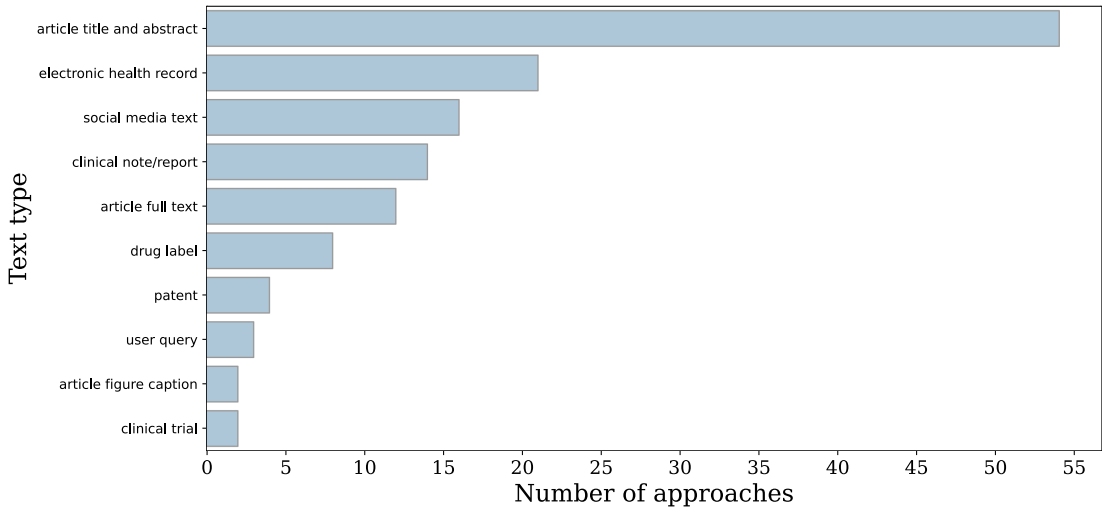


Fig. 7. Number of approaches and the type of text that is addressed

Regarding entity type, most use more diverse entities that fall into the 'bio' type (13%, N=8), followed by 'chemical', 'species' and 'disease' each with 11% (N=8), next with 'symptom' (8%, N=6), 7% cover 'adr' and 'gene' (N=5).

4.5 Named entity linking or normalisation? Disambiguating the task name

In contrast to the central motivation of the NEL task, there is ambiguity in the designation of the task across the literature. In the reviewed papers (N=102), we identified five designation clusters: "normalisation" (N=50), "linking" (N= 37), "disambiguation" (N=8), "mapping" (N=6) and "extraction" (N=1), with differing previous words associated with the object of the task, such as "entity", "named entity", "concept", "term", but also with the domain of the task ("biomedical", "clinical, "medical").

4.6 Limitations of current named entity linking approaches

Several limitations are mentioned by the authors of the reviewed articles. These include dataset-related errors, such as overlapping annotations [131, 138], incorrect delimitation of entity boundaries and incorrect entity typing [64, 78, 89, 148].

Additionally, other limitations are related with the KOS, including the quality [72], ignoring semantic relationships, the existence of NIL entities [140, 149], for example, the correct candidate is not in the generated candidate list [140] or the approaches only identify and cluster NIL entities [155]. The existence of NIL entities arise from a lack of coverage from target KOS: either its structure does not include an existing entity already described in a text source due to the slow curation process, or the entity is completely new and does not fit into the conceptual schema of the KOS.

Evaluation challenges include a lack of benchmarks and many times focusing on a single domain, limiting the ability to generalise across domains [15, 144]. Additionally, the requirement for explainability in clinical settings, as it is essential to understand the reasoning behind the predictions made by deep neural networks[140]. Supervised approaches require labelled training data, which can be difficult to obtain and build [72, 152]. Moreover, there is the

complexity of composite annotations where a single annotation is associated with multiple identifiers simultaneously [79, 138, 148].

Text complexity also presents difficulties, including issues with abbreviations [89, 131, 148], synonyms [112, 117, 131], hypernyms/hyponyms, i.e. parent and child concept with similar names, approach wrongly associate entity with a parent or vice-versa [89, 131, 158]. These challenges can lead to incorrect entity associations and the need to leverage context for disambiguation [33, 72, 115, 152].

4.7 Future directions for named entity linking

The authors of the reviewed papers identified future directions, which we categorised into three groups: domain adaptation and generalisation, knowledge integration and representation, and human evaluation and feedback. We further address these and additional topics that we consider relevant for the future of the NEL task.

Domain adaptation and generalization focus on tailoring approaches to different domains and languages, as well as improving its ability to generalize to new scenarios without requiring extensive training data. This includes evaluating the approach on more domains/datasets [25, 43, 63, 75, 91, 158], addressing differences in dataset characteristics [5], developing and assessing methods in various languages [24, 26, 143, 163], particularly in clinical domains, as well as implementing self-supervision [158] and zero-shot learning to enable models to perform well without supervised training or fine-tuning.

Knowledge integration and representation aim to enhance the approach's ability to incorporate and utilise knowledge from various sources, as well as improve its representation of entities and concepts. This can be achieved by integrating information from the target knowledge base or other external resource (e.g. corpus, structure information from KOS), including concepts, relations, synonyms, and definitions [13, 26, 66, 76, 91, 104, 109, 128, 135, 146, 148, 155, 158]. Further improvements can be made by refining entity and concept representations through the integration of relationships between entities [13, 72, 77, 87, 122, 125, 130], detect terms which can not be represented by a single concept [5, 155], improve inference match on indirect entities [135], acronyms mappings [135] and full entities [121], weight results from different sources to improve accuracy [164], leverage with semantic reasoning and contextual information [64, 68, 115, 137, 160].

Human-in-the-loop processes and feedback emphasize the crucial role of human evaluation in the development and refinement of an approach. This involves expanding traditional evaluation methods by integrating human evaluation [26, 31, 38, 134, 164], with feedback by clinical experts being essential, as they are the primary users of these tools.

Considering the mentioned existing limitations, despite the advancements in new better-performing systems, we find some key areas for future focus.

One of the points is that NEL is understudied as a standalone task, often just being an intermediate step in an end-to-end system or treated as a NER extension [63, 76, 134, 149]. Thus it might be important to explore independently to better understand the phases, candidate generation and entity ranking weaknesses.

Challenges should be promoted to explore more diverse ideas, such as exploring smaller and open access domain ontologies instead of larger KOS (such as UMLS and MeSH). Also, challenges are extremely important in creating more precise datasets and with a bigger variety of entities. Besides, another direction could be the definition of annotation guidelines focusing on more specific entities that could be leveraged to annotate different text sources. For instance, none of the articles we reviewed used a dataset that included rare disease annotations. Rare diseases affect approximately 300 million people globally, placing a significant burden on healthcare systems [40]. Developing annotation guidelines and/or datasets that focus on rare diseases could enhance the extraction and management of information from clinical

documents, for example. To achieve this, resources from the *Orphanet* database, particularly the *Orphanet Rare Disease Ontology* (ORDO)[33], which adheres to the OBO guidelines, could be used.

The advent of ChatGPT and other large language models (LLMs) has tilted NLP research towards adapting these general-purpose models to specific NLP tasks [22], including NEL. Despite their potential, various limitations of these models have been acknowledged, including lack of transparency, data leakage [20], resource intensiveness, and the problem of hallucinations [62]. Nonetheless, they might present benefits in zero or few-shot scenarios, especially in domains where there is a scarcity of human-curated data [22], and they can improve the efficiency of data annotation pipelines.

Although, the recent approaches are very capable and the tendency is of improvement, having a human-in-the-loop remains crucial. Manual validation is essential for datasets and KOS curation, as system performance is connected to data quality. So, efforts should be made to simplify the curation process, possibly through user-friendly tools and open-source collaboration. LLMs can provide more cost-effective and efficient annotations, at the cost of introducing errors and biases, often undetected. To mitigate these issues, several collaborative pipelines between humans and LLM-based approaches have been recently proposed to annotate datasets for NLP tasks [48, 71, 92, 102, 145].

For instance, [48] proposed an approach to annotate clinical notes with medication information that combines human judgement with the LLMs capabilities and compared it with a purely human expert-based pipeline. The pipeline applies the PaLM 2 language model [1] to generate the base annotations (NER and relation extraction) requiring few annotation examples (few-shot setting). Then, medical experts refined the generated annotations. The authors report an average reduction of 58% in the human annotation time, without sacrificing the quality. [102] also explored LLMs to annotate NER datasets, combining both human and LLM annotations to reduce the impact of noisy data.

A more ambitious goal consists in replacing the human feedback component in data annotation pipelines also by a LLM-based approach. [133] provides an overview of the several strategies that have been proposed to automate data annotation in the NLP field using LLMs.

One downside of the proliferation of use cases for LLM is the amount of noise that these models can introduce in the text sources used to develop NEL approaches and to curate KOS resources. Given the current trend in the NEL task towards zero-shot approaches, it will be interesting to assess the real impact of the generated text in the performance of this type of approach. On the other hand, the need for quality data may provide an opportunity for the existing human-curated KOS. In this sense, KOS can be used to assess the quality of the generated output by LLMs.

Another downside is the resource intensiveness associated with the development of LLM-based approaches. Training and fine-tuning LLMs require substantial computational power, including significant memory, specialized hardware such as *graphics processing units* (GPUs) or *tensor processing units* (TPUs), and substantial data storage capabilities. In addition to computational power, knowledge is always evolving and no frozen LLM is capable of handle this issue since developing and maintaining LLM-based pipelines require specialized knowledge and skills. Therefore, there is room to explore more effective ways of deploying these approaches and to investigate alternative, less complex approaches that are not based on LLMs. The decision to work with LLMs should consider the project's objectives, the available data, and the other mentioned resources. It is still a better option, privacy-wise to use an in-house LLM with open-source solutions.

[33] https://www.orphadata.com/ordo/

5 LIMITATIONS OF THE REVIEW

The current work has several limitations. First, there is a source-related limitation, as only four repositories were selected. Additionally, human factors can influence the selection of articles and the data collection process. To address these, we followed the PRISMA guidelines to minimize potential biases. Moreover, it is important to consider the constraints associated with the search strategies employed in the selected repositories.

6 CONCLUSION

We analysed the panorama of NEL literature published between 2013 and 2024, highlighting the evolution of methodology, the application of KOS to the task along with current limitations and future directions. Through the paper, we tried to address three research questions.

For RQ1, we analysed the evolution of article publication on 4.1. It is noticed a focus on machine learning methods in mid-2020, mostly employing pre-trained models.

Regarding RQ2, we analysed which knowledge organisation systems are used in 4.2. It was observed that larger KOS, such as UMLS and MeSH, are generally preferred. Additionally, it was analysed how the evaluation is performed in this task in 4.3 and the incongruity of the task name 4.5.

Finally, for RQ3, limitations of current approaches are addressed on 4.6, being grouped into evaluation, KOS-related and textual challenges. The future directions the task are addressed on 4.7, including those identified in the reviewed articles, and also others identified by us.

Our contributions to this review were three-fold. Firstly, we conducted a systematic PRISMA review of the evolution of NEL tasks in the biomedical and clinical domains from 2013 to 2024, analyzing 102 articles. This review provided insights into the evolution of the publication of these articles, including approach categories, the KOS that were used, as well datasets, text types, entity types, limitations, and future directions. Secondly, we offered an overview of the KOS landscape in the biomedical and clinical domains, highlighting the types of KOS and their implementations. Finally, we contributed with a public dataset containing comprehensive information about the reviewed articles.

DECLARATION OF COMPETING INTEREST

The authors declare no conflict of interests.

ACKNOWLEDGMENTS

This work was supported by FCT (Fundacão para a Ciência e a Tecnologia) through funding of the PhD Scholarships with ref. 2020.05393.BD attributed to PR and ref. UI/BD/153730/2022 attributed to SIRC, and LASIGE Research Unit (UIDB/00408/2020 and UIDP/00408/2020).

REFERENCES

[1] 2023. PaLM 2 Technical Report. arXiv:2305.10403 [cs.CL] https://arxiv.org/abs/2305.10403

[2] Akhila Abdulnazar, Markus Kreuzthaler, Roland Roller, and Stefan Schulz. 2023. SapBERT-based medical concept normalization using SNOMED CT. In *Caring is Sharing–Exploiting the Value in Data for Health and Innovation*. IOS Press, 825–826.

[3] Dhruv Agarwal, Rico Angell, Nicholas Monath, and Andrew McCallum. 2022. Entity Linking via Explicit Mention-Mention Coreference Modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 4644–4658. https://doi.org/10.18653/v1/2022.naacl-main.343

[4] Tareq Al-Moslmi, Marc Gallofre Ocana, Andreas L. Opdahl, and Csaba Veres. 2020. Named Entity Extraction for Knowledge Graphs: A Literature Overview. *IEEE Access* 8 (2020), 32862–32881. https://doi.org/10.1109/ACCESS.2020.2973928

[5] JL Allones, Diego Martinez, and Maria Taboada. 2014. Automated mapping of clinical terms into SNOMED-CT. An application to codify procedures in pathology. *Journal of medical systems* 38 (2014), 1–14.

[6] Noha Alnazzawi, Paul Thompson, and Sophia Ananiadou. 2016. Mapping phenotypic information in heterogeneous textual sources to a domain-specific terminological resource. *PLoS One* 11, 9 (2016), e0162287.

[7] Rico Angell, Nicholas Monath, Sunil Mohan, Nishant Yadav, and Andrew McCallum. 2021. Clustering-based Inference for Biomedical Entity Linking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, Online, 2598–2608. https://doi.org/10.18653/v1/2021.naacl-main.205

[8] Valerio Arnaboldi, Daniela Raciti, Kimberly Van Auken, Juancarlos N Chan, Hans-Michael Müller, and Paul W Sternberg. 2020. Text mining meets community curation: a newly designed curation platform to improve author experience and participation at WormBase. *Database* 2020 (03 2020), baaa006. https://doi.org/10.1093/database/baaa006 arXiv:https://academic.oup.com/database/article-pdf/doi/10.1093/database/baaa006/32923929/baaa006.pdf

[9] Márcia Barros and Francisco Couto. 2016. *Knowledge Representation and Management: a Linked Data Perspective*. Vol. 25. https://doi.org/10.15265/IY-2016-022

[10] Iz Beltagy, Kyle Lo, and Arman Cohan. 2020. SCIBERT: A pretrained language model for scientific text. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics (ACL), Hong Kong, China, 3615–3620. https://doi.org/10.18653/v1/d19-1371 arXiv:1903.10676

[11] Rajarshi Bhowmik, Karl Stratos, and Gerard de Melo. 2021. Fast and Effective Biomedical Entity Linking Using a Dual Encoder. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, Eben Holderness, Antonio Jimeno Yepes, Alberto Lavelli, Anne-Lyse Minard, James Pustejovsky, and Fabio Rinaldi (Eds.). Association for Computational Linguistics, online, 28–37. https://aclanthology.org/2021.louhi-1.4

[12] Yonatan Bitton, Raphael Cohen, Tamar Schifter, Eitan Bachmat, Michael Elhadad, and Noémie Elhadad. 2020. Cross-lingual Unified Medical Language System entity linking in online health communities. *Journal of the American Medical Informatics Association* 27, 10 (2020), 1585–1592.

[13] Mumeng Bo and Meihui Zhang. 2021. Learning Dynamic Coherence with Graph Attention Network for Biomedical Entity Linking. In *2021 International Joint Conference on Neural Networks (IJCNN)*. 1–8. https://doi.org/10.1109/IJCNN52387.2021.9533687

[14] Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 32, Database issue (Jan. 2004), D267–70.

[15] Mayla Boguslav, K. Bretonnel Cohen, William A. Baumgartner Jr., and Lawrence E. Hunter. [n. d.]. *Improving precision in concept normalization*. 566–577. https://doi.org/10.1142/9789813235533_0052

[16] E G Brown, L Wood, and S Wood. 1999. The medical dictionary for regulatory activities (MedDRA). *Drug Saf.* 20, 2 (Feb. 1999), 109–117.

[17] Garth R Brown, Vichet Hem, Kenneth S Katz, Michael Ovetsky, Craig Wallin, Olga Ermolaeva, Igor Tolstoy, Tatiana Tatusova, Kim D Pruitt, Donna R Maglott, et al. 2015. Gene: a gene-centered information resource at NCBI. *Nucleic acids research* 43, D1 (2015), D36–D42.

[18] David Campos, Sérgio Matos, and José Luís Oliveira. 2013. A modular framework for biomedical concept recognition. *BMC bioinformatics* 14 (2013), 1–21.

[19] Yiling Cao, Lu Fang, and Zhongguang Zheng. 2022. Enriching Pre-Trained Language Model with Multi-Task Learning and Context for Medical Concept Normalization. In *Proceedings of the 2022 International Conference on Intelligent Medicine and Health*. 79–83.

[20] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Xiaodong Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting Training Data from Large Language Models. In *USENIX Security Symposium*. https://api.semanticscholar.org/CorpusID:229156229

[21] Gjorgjina Cenikj, Gašper Petelin, Barbara Koroušić Seljak, and Tome Eftimov. 2022. SciFoodNER: Food Named Entity Recognition for Scientific Text. In *2022 IEEE International Conference on Big Data (Big Data)*. 4065–4073. https://doi.org/10.1109/BigData55660.2022.10020459

[22] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.* 15, 3, Article 39 (mar 2024), 45 pages. https://doi.org/10.1145/3641289

[23] Long Chen, Wenbo Fu, Yu Gu, Zhiyong Sun, Haodan Li, Enyu Li, Li Jiang, Yuan Gao, and Yang Huang. 2020. Clinical concept normalization with a hybrid natural language processing system combining multilevel matching and machine learning ranking. *Journal of the American Medical Informatics Association* 27, 10 (2020), 1576–1584.

[24] Luming Chen, Yifan Qi, Aiping Wu, Lizong Deng, and Taijiao Jiang. 2023. TeaBERT: An Efficient Knowledge Infused Cross-Lingual Language Model for Mapping Chinese Medical Entities to the Unified Medical Language System. *IEEE Journal of Biomedical and Health Informatics* (2023).

[25] Hyejin Cho, Dongha Choi, and Hyunju Lee. 2021. Re-Ranking System with BERT for Biomedical Concept Normalization. *IEEE Access* 9 (2021), 121253–121262. https://doi.org/10.1109/ACCESS.2021.3108445

[26] Carlo Combi, Margherita Zorzi, Gabriele Pozzani, Elena Arzenton, and Ugo Moretti. 2019. Normalizing Spontaneous Reports Into MedDRA: Some Experiments With MagiCoder. *IEEE Journal of Biomedical and Health Informatics* 23, 1 (2019), 95–102. https://doi.org/10.1109/JBHI.2018.2861213

[27] Francisco M. Couto and Martin Krallinger. 2020. Proposal of the First International Workshop on Semantic Indexing and Information Retrieval for Health from Heterogeneous Content Types and Languages (SIIRH). In *Advances in Information Retrieval*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer International Publishing, Cham, 654–659.

https://doi.org/10.1007/978-3-030-45442-5_87

[28] Clint Cuffy, Evan French, Sophia Fehrmann, and Bridget T McInnes. 2022. Exploring Representations for Singular and Multi-Concept Relations for Biomedical Named Entity Normalization. In *Companion Proceedings of the Web Conference 2022*. 823–832.

[29] John Cuzzola, Jelena Jovanović, and Ebrahim Bagheri. 2017. RysannMD: A biomedical semantic annotator balancing speed and accuracy. *Journal of Biomedical Informatics* 71 (2017), 91–109. https://doi.org/10.1016/j.jbi.2017.05.016

[30] Jian Dai, Meihui Zhang, Gang Chen, Ju Fan, Kee Yuan Ngiam, and Beng Chin Ooi. 2018. Fine-grained concept linking using neural networks in healthcare. In *Proceedings of the 2018 International Conference on Management of Data*. 51–66.

[31] Ruiyu Dai, Xu Zhang, Feihong Li, and Chenlong Li. 2024. Research on Normalization of Chinese Clinical Terms Based on Keyword Extraction and Data Augmentation Technology. In *Proceedings of the 2023 4th International Symposium on Artificial Intelligence for Medicine Science* (<conf-loc>, <city>Chengdu</city>, <country>China</country>, </conf-loc>) *(ISAIMS '23)*. Association for Computing Machinery, New York, NY, USA, 1291–1298. https://doi.org/10.1145/3644116.3644334

[32] Allan Peter Davis, Thomas C Wiegers, Robin J Johnson, Daniela Sciaky, Jolene Wiegers, and Carolyn J Mattingly. 2022. Comparative Toxicogenomics Database (CTD): update 2023. *Nucleic Acids Research* 51, D1 (09 2022), D1257–D1262. https://doi.org/10.1093/nar/gkac833 arXiv:https://academic.oup.com/nar/article-pdf/51/D1/D1257/48441054/gkac833.pdf

[33] Pan Deng, Haipeng Chen, Mengyao Huang, Xiaowen Ruan, and Liang Xu. 2019. An ensemble CNN method for biomedical entity normalization. In *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*. Association for Computational Linguistics, Hong Kong, China, 143–149. https://doi.org/10.18653/v1/D19-5721

[34] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (oct 2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[35] Guy Divita, Qing T Zeng, Adi V Gundlapalli, Scott Duvall, Jonathan Nebeker, and Matthew H Samore. 2014. Sophia: a expedient UMLS concept extraction annotator. In *AMIA Annual Symposium Proceedings*, Vol. 2014. American Medical Informatics Association, 467.

[36] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics* 47 (2014), 1–10.

[37] Hang Dong, Jiaoyan Chen, Yuan He, Yinan Liu, and Ian Horrocks. 2023. Reveal the Unknown: Out-of-Knowledge-Base Mention Discovery with Entity Linking. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (<conf-loc>, <city>Birmingham</city>, <country>United Kingdom</country>, </conf-loc>) *(CIKM '23)*. Association for Computing Machinery, New York, NY, USA, 452–462. https://doi.org/10.1145/3583780.3615036

[38] Hang Dong, Víctor Suárez-Paniagua, Huayu Zhang, Minhong Wang, Emma Whitfield, and Honghan Wu. 2021. Rare disease identification from clinical notes with ontologies and weak supervision. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2294–2298.

[39] Shuang Duan, Yan Guang, Wenjuan Bu, and Ju Yang. 2023. A Survey of Named Entity Disambiguation in Entity Linking. In *2023 3rd International Conference on Intelligent Communications and Computing (ICC)*. 296–303. https://doi.org/10.1109/ICC59986.2023.10421092

[40] Editorial. 2024. The landscape for rare diseases in 2024. *The Lancet Global Health* 12, 3 (March 2024), e341. https://doi.org/10.1016/S2214-109X(24)00056-1

[41] Lisa Ehrlinger and Wolfram Wöß. 2016. Towards a Definition of Knowledge Graphs. In *International Conference on Semantic Systems*. https://api.semanticscholar.org/CorpusID:8536105

[42] Ehsan Emadzadeh, Abeed Sarker, Azadeh Nikfarjam, and Graciela Gonzalez. 2017. Hybrid semantic analysis for mapping adverse drug reaction mentions in tweets to medical terminology. In *AMIA Annual Symposium Proceedings*, Vol. 2017. American Medical Informatics Association, 679.

[43] Arnaud Ferré, Robert Bossy, Mouhamadou Ba, Louise Deléger, Thomas Lavergne, Pierre Zweigenbaum, and Claire Nédellec. 2020. Handling Entity Normalization with no Annotated Corpus: Weakly Supervised Methods Based on Distributional Representation and Ontological Information. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 1959–1966. https://aclanthology.org/2020.lrec-1.241

[44] Arnaud Ferré and Philippe Langlais. 2023. An analysis of entity normalization evaluation biases in specialized domains. *BMC Bioinformatics* 24 (06 2023). https://doi.org/10.1186/s12859-023-05350-9

[45] Evan French and Bridget T. McInnes. 2023. An overview of biomedical entity linking throughout the years. *Journal of Biomedical Informatics* 137 (2023), 104252. https://doi.org/10.1016/j.jbi.2022.104252

[46] Sunyang Fu, David Chen, Huan He, Sijia Liu, Sungrim Moon, Kevin J. Peterson, Feichen Shen, Liwei Wang, Yanshan Wang, Andrew Wen, Yiqing Zhao, Sunghwan Sohn, and Hongfang Liu. 2020. Clinical concept extraction: A methodology review. *Journal of Biomedical Informatics* 109 (9 2020). https://doi.org/10.1016/j.jbi.2020.103526

[47] Glenn T. Gobbel, Ruth Reeves, Shrimalini Jayaramaraja, Dario Giuse, Theodore Speroff, Steven H. Brown, Peter L. Elkin, and Michael E. Matheny. 2014. Development and evaluation of RapTAT: A machine learning system for concept mapping of phrases from medical narratives. *Journal of Biomedical Informatics* 48 (2014), 54–65. https://doi.org/10.1016/j.jbi.2013.11.008

[48] Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, Jean Steiner, Itay Laish, and Amir Feder. 2023. LLMs Accelerate Annotation for Medical Information Extraction. In *Proceedings of the 3rd Machine Learning for Health Symposium (Proceedings of Machine Learning Research, Vol. 225)*, Stefan Hegselmann, Antonio Parziale,

Divya Shanmugam, Shengpu Tang, Mercy Nyamewaa Asiedu, Serina Chang, Tom Hartvigsen, and Harvineet Singh (Eds.). PMLR, 82–100. https://proceedings.mlr.press/v225/goel23a.html

[49] Thomas R. Gruber. 1993. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition* 5, 2 (1993), 199–220. http://tomgruber.org/writing/ontolingua-kaj-1993.pdf

[50] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare* 3, 1, Article 2 (oct 2021), 23 pages. https://doi.org/10.1145/3458754

[51] Fengming Guan and Taro Tezuka. 2022. A medical Q&A system with entity linking and intent recognition. In *2022 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 820–829.

[52] Nicola Guarino. 1998. Formal Ontologies and Information Systems.

[53] Erin Gustafson, Jennifer Pacheco, Firas Wehbe, Jonathan Silverberg, and William Thompson. 2017. A Machine Learning Algorithm for Identifying Atopic Dermatitis in Adults from Electronic Health Records. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*. 83–90. https://doi.org/10.1109/ICHI.2017.31

[54] James E. Harrison, Stefanie Weber, Robert Jakob, and Christopher G. Chute. 2021. ICD-11: an international classification of diseases for the twenty-first century. *BMC Medical Informatics and Decision Making* 21, 6 (09 Nov 2021), 206. https://doi.org/10.1186/s12911-021-01534-6

[55] Fons Hartendorp, Tom Seinen, Erik van Mulligen, and Suzan Verberne. 2024. Biomedical Entity Linking for Dutch: Fine-tuning a Self-alignment BERT Model on an Automatically Generated Wikipedia Corpus. *arXiv preprint arXiv:2405.11941* (2024).

[56] Janna Hastings, Gareth Owen, Adriano Dekker, Marcus Ennis, Namrata Kale, Venkatesh Muthukrishnan, Steve Turner, Neil Swainston, Pedro Mendes, and Christoph Steinbeck. 2016. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic acids research* 44, D1 (January 2016), D1214—9. https://doi.org/10.1093/nar/gkv1031

[57] Birger Hjørland. 2008. What is Knowledge Organization (KO)? *Knowledge Organization* 35 (07 2008). https://doi.org/10.5771/0943-7444-2008-2-3-86

[58] Gail Hodge. 2000. *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files.* Technical Report. The Digital Library Federation, Council on Library and Information Resources, 1755 Massachusetts Avenue, NW, Suite 500, Washington, DC 20036. https://www.clir.org/pubs/reports/pub91/contents/

[59] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge Graphs. *ACM Comput. Surv.* 54, 4, Article 71 (jul 2021), 37 pages. https://doi.org/10.1145/3447772

[60] Ming Siang Huang, Po Ting Lai, Pei Yen Lin, Yu Ting You, Richard Tzong Han Tsai, and Wen Lian Hsu. 2020. Biomedical named entity recognition and linking datasets: Survey and our recent development. *Briefings in Bioinformatics* 21 (11 2020), 2219–2238. Issue 6. https://doi.org/10.1093/bib/bbaa054

[61] Rebecca Jackson, Nicolas Matentzoglu, James A Overton, Randi Vita, James P Balhoff, Pier Luigi Buttigieg, Seth Carbon, Melanie Courtot, Alexander D Diehl, Damion M Dooley, William D Duncan, Nomi L Harris, Melissa A Haendel, Suzanna E Lewis, Darren A Natale, David Osumi-Sutherland, Alan Ruttenberg, Lynn M Schriml, Barry Smith, Christian J Stoeckert Jr., Nicole A Vasilevsky, Ramona L Walls, Jie Zheng, Christopher J Mungall, and Bjoern Peters. 2021. OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. *Database* 2021 (10 2021), baab069. https://doi.org/10.1093/database/baab069 arXiv:https://academic.oup.com/database/article-pdf/doi/10.1093/database/baab069/40854912/baab069.pdf

[62] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12, Article 248 (mar 2023), 38 pages. https://doi.org/10.1145/3571730

[63] Zongcheng Ji, Tian Xia, Mei Han, and Jing Xiao. 2021. A Neural Transition-based Joint Model for Disease Named Entity Recognition and Normalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing Volume 1: Long Papers*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 2819–2827. https://doi.org/10.18653/v1/2021.acl-long.219

[64] Jitendra Jonnagaddala, Toni Rose Jue, Nai-Wen Chang, and Hong-Jie Dai. 2016. Improving the dictionary lookup approach for disease normalization using enhanced dictionary and query expansion. *Database* 2016 (08 2016), baw112. https://doi.org/10.1093/database/baw112 arXiv:https://academic.oup.com/database/article-pdf/doi/10.1093/database/baw112/8224995/baw112.pdf

[65] Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2022. AMMU: A survey of transformer-based biomedical pretrained language models. *Journal of Biomedical Informatics* 126 (2 2022). https://doi.org/10.1016/j.jbi.2021.103982

[66] Katikapalli Subramanyam Kalyan and Sivanesan Sangeetha. 2021. Bertmcn: Mapping colloquial phrases to standard medical concepts using bert and highway network. *Artificial Intelligence in Medicine* 112 (2021), 102008.

[67] İlknur Karadeniz and Arzucan Özgür. 2015. Detection and categorization of bacteria habitats using shallow linguistic analysis. *BMC bioinformatics* 16 (2015), 1–14.

[68] Ilknur Karadeniz and Arzucan Özgür. 2019. Linking entities through an ontology using word embeddings and syntactic re-ranking. *BMC bioinformatics* 20 (2019), 1–12.

[69] Rohit J Kate. 2016. Normalizing clinical terms using learned edit distance patterns. *Journal of the American Medical Informatics Association* 23, 2 (2016), 380–386.

[70] Donghyeon Kim, Jinhyuk Lee, Chan Ho So, Hwisang Jeon, Minbyul Jeong, Yonghwa Choi, Wonjin Yoon, Mujeen Sung, and Jaewoo Kang. 2019. A Neural Named Entity Recognition and Multi-Type Normalization Tool for Biomedical Text Mining. *IEEE Access* 7 (2019), 73729–73740.

https://doi.org/10.1109/ACCESS.2019.2920708

[71] Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. 2024. MEGAnno+: A Human-LLM Collaborative Annotation System. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, Nikolaos Aletras and Orphee De Clercq (Eds.). Association for Computational Linguistics, St. Julians, Malta, 168–176. https://aclanthology.org/2024.eacl-demo.18

[72] Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A. Folarin, Angus Roberts, Rebecca Bendayan, Mark P. Richardson, Robert Stewart, Anoop D. Shah, Wai Keong Wong, Zina Ibrahim, James T. Teo, and Richard J.B. Dobson. 2021. Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. *Artificial Intelligence in Medicine* 117 (2021), 102083. https://doi.org/10.1016/j.artmed.2021.102083

[73] Martin Krallinger, Alfonso Valencia, and Lynette Hirschman. 2008. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biology* 9, 2 (2008), 1–14. https://doi.org/10.1186/gb-2008-9-s2-s8

[74] Amila Kugic, Markus Kreuzthaler, and Stefan Schulz. 2023. Clinical Acronym Disambiguation via ChatGPT and BING. In *Telehealth Ecosystems in Practice*. IOS Press, 78–82.

[75] Tuan Lai, Heng Ji, and ChengXiang Zhai. 2021. BERT might be Overkill: A Tiny but Effective Biomedical Entity Linker based on Residual Convolutional Neural Networks. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 1631–1639. https://doi.org/10.18653/v1/2021.findings-emnlp.140

[76] Tuan Manh Lai, ChengXiang Zhai, and Heng Ji. 2023. KEBLM: Knowledge-Enhanced Biomedical Language Models. *J. of Biomedical Informatics* 143, C (jul 2023), 10 pages. https://doi.org/10.1016/j.jbi.2023.104392

[77] Andre Lamurias, Pedro Ruas, and Francisco M. Couto. 2019. PPR-SSM: personalized PageRank and semantic similarity measures for entity linking. *BMC Bioinformatics* 20, 1 (29 Oct 2019), 534. https://doi.org/10.1186/s12859-019-3157-y

[78] Robert Leaman, Ritu Khare, and Zhiyong Lu. 2015. Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics* 57 (2015), 28–37. https://doi.org/10.1016/j.jbi.2015.07.010

[79] Robert Leaman and Zhiyong Lu. 2016. TaggerOne: joint named entity recognition and normalization with semi-Markov Models. *Bioinformatics* 32, 18 (06 2016), 2839–2846. https://doi.org/10.1093/bioinformatics/btw343 arXiv:https://academic.oup.com/bioinformatics/article-pdf/32/18/2839/49020913/bioinformatics_32_18_2839.pdf

[80] Heonwoo Lee, Junbeom Jeon, Dawoon Jung, Jung-Im Won, Kiyong Kim, Yun Joong Kim, and Jeehee Yoon. 2023. RelCurator: a text mining-based curation system for extracting gene–phenotype relationships specific to neurodegenerative disorders. *Genes & Genomics* 45, 8 (2023), 1025–1036. https://doi.org/10.1007/s13258-023-01405-6

[81] Hsin-Chun Lee and Hung-Yu Kao. 2017. CDRnN: A high performance chemical-disease recognizer in biomedical literature. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 374–379.

[82] Jeongeun Lee, Hyun-Je Song, Eunsil Yoon, Seong-Bae Park, Sung-Hye Park, Jeong-Wook Seo, Peom Park, and Jinwook Choi. 2018. Automated extraction of Biomarker information from pathology reports. *BMC medical informatics and decision making* 18 (2018), 1–11.

[83] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682 arXiv:1901.08746

[84] Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database* 2016 (2016).

[85] Sheng-Jie Lin, Wen-Chao Yeh, Yu-Wen Chiu, Yung-Chun Chang, Min-Huei Hsu, Yi-Shin Chen, and Wen-Lian Hsu. 2022. A BERT-based ensemble learning approach for the BioCreative VII challenges: full-text chemical identification and multi-label classification in PubMed articles. *Database* 2022 (2022), baac056.

[86] Tzu-Mi Lin, Man-Chen Hung, and Lung-Hao Lee. 2024. Leveraging Dual Gloss Encoders in Chinese Biomedical Entity Linking. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 23, 2, Article 28 (feb 2024), 15 pages. https://doi.org/10.1145/3638555

[87] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-Alignment Pretraining for Biomedical Entity Representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, Online, 4228–4238. https://doi.org/10.18653/v1/2021.naacl-main.334

[88] Guillermo López-García, José M. Jerez, Nuria Ribelles, Emilio Alba, and Francisco J. Veredas. 2023. Explainable clinical coding with in-domain adapted transformers. *J. of Biomedical Informatics* 139, C (mar 2023), 14 pages. https://doi.org/10.1016/j.jbi.2023.104323

[89] Yinxia Lou, Yue Zhang, Tao Qian, Fei Li, Shufeng Xiong, and Donghong Ji. 2017. A transition-based joint model for disease named entity recognition and normalization. *Bioinformatics* 33, 15 (03 2017), 2363–2371. https://doi.org/10.1093/bioinformatics/btx172 arXiv:https://academic.oup.com/bioinformatics/article-pdf/33/15/2363/50756459/bioinformatics_33_15_2363.pdf

[90] Henry J Lowe and G Octo Barnett. 1987. MicroMeSH: a microcomputer system for searching and exploring the National Library of Medicine's Medical Subject Headings (MeSH) Vocabulary. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association, 717.

[91] Wenpeng Lu, Guobiao Zhang, Xueping Peng, Hongjiao Guan, and Shoujin Wang. 2024. Medical Entity Disambiguation with Medical Mention Relation and Fine-grained Entity Knowledge. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language*

Resources and Evaluation (LREC-COLING 2024), Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 11148–11158. https://aclanthology.org/2024.lrec-main.972

[92] Kun Luo, Tong Zhou, Yubo Chen, Jun Zhao, and Kang Liu. 2024. Open Event Causality Extraction by the Assistance of LLM in Task Annotation, Dataset, and Method. In *Proceedings of the Workshop: Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning (NeusymBridge) @ LREC-COLING-2024*, Tiansi Dong, Erhard Hinrichs, Zhen Han, Kang Liu, Yangqiu Song, Yixin Cao, Christian F. Hempelmann, and Rafet Sifa (Eds.). ELRA and ICCL, Torino, Italia, 33–44. https://aclanthology.org/2024.neusymbridge-1.4

[93] Yen-Fu Luo, Weiyi Sun, and Anna Rumshisky. 2018. A Hybrid Method for Normalization of Medical Concepts in Clinical Narrative. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*. 392–393. https://doi.org/10.1109/ICHI.2018.00069

[94] Donna Maglott, Tanya Barrett, Terence Murphy, Michael Feolo, Lukas Wagner, and Richa Agarwala. 2013. Genes and Gene Expression. In *The NCBI Handbook [Internet]. 2nd edition*. National Center for Biotechnology Information (US).

[95] McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). [n. d.]. Online Mendelian Inheritance in Man, OMIM. https://en.wikipedia.org/w/index.php?title=Commonsense_knowledge_(artificial_intelligence)&oldid=1117005558 [accessed 29-November-2023].

[96] Farrokh Mehryary, Kai Hakala, Suwisa Kaewphan, Jari Björne, Tapio Salakoski, and Filip Ginter. 2017. End-to-End System for Bacteria Habitat Extraction. In *Proceedings ofthe BioNLP 2017 workshop*. Association for Computational Linguistics, 80–90. https://doi.org/10.18653/v1/w17-2310

[97] Zulfat Miftahutdinov and Elena Tutubalina. 2017. End-to-end deep framework for disease named entity recognition using social media data. In *2017 IEEE 30th Neumann Colloquium (NC)*. 000047–000052. https://doi.org/10.1109/NC.2017.8263281

[98] Sunil Mohan, Rico Angell, Nicholas Monath, and Andrew McCallum. 2021. Low resource recognition and linking of biomedical concepts from a large ontology. In *Proceedings of the 12th ACM conference on bioinformatics, computational biology, and health informatics*. 1–10.

[99] Sunil Mohan and Donghui Li. 2019. MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. *CoRR* abs/1902.09476 (2019). arXiv:1902.09476 http://arxiv.org/abs/1902.09476

[100] Khalil Mrini, Shaoliang Nie, Jiatao Gu, Sinong Wang, Maziar Sanjabi, and Hamed Firooz. 2022. Detection, Disambiguation, Re-ranking: Autoregressive Entity Linking as a Multi-Task Problem. In *Findings of the Association for Computational Linguistics: ACL 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 1972–1983. https://doi.org/10.18653/v1/2022.findings-acl.156

[101] Cedric Möller, Jens Lehmann, and Ricardo Usbeck. 2021. Survey on English Entity Linking on Wikidata. (12 2021). http://arxiv.org/abs/2112.01989

[102] Yuji Naraki, Ryosuke Yamaki, Yoshikazu Ikeda, Takafumi Horie, and Hiroki Naganuma. 2024. Augmenting NER Datasets with LLMs: Towards Automated and Refined Annotation. arXiv:2404.01334 [cs.CL] https://arxiv.org/abs/2404.01334

[103] Annisa Maulida Ningtyas, Allan Hanbury, Florina Piroi, and Linda Andersson. 2021. Data augmentation for layperson's medical entity linking task. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*. 99–106.

[104] Jiho Noh and Ramakanth Kavuluru. 2021. Joint learning for biomedical NER and entity normalization: encoding schemes, counterfactual examples, and zero-shot evaluation. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. 1–10.

[105] Natasha Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. 2019. Industry-scale knowledge graphs: lessons and challenges. *Commun. ACM* 62, 8 (jul 2019), 36–43. https://doi.org/10.1145/3331166

[106] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372 (2021). https://doi.org/10.1136/bmj.n71 arXiv:https://www.bmj.com/content/372/bmj.n71.full.pdf

[107] L Pape-Haugaard et al. 2020. Clinical concept normalization on medical records using word embeddings and heuristics. *Digital Personalized Health and Medicine: Proceedings of MIE* 2020, 270 (2020), 93.

[108] Alicia Pérez, Aitziber Atutxa, Arantza Casillas, Koldo Gojenola, and Álvaro Sellart. 2018. Inferred joint multigram models for medical term normalization according to ICD. *International journal of medical informatics* 110 (2018), 111–117.

[109] Naiara Perez, Pablo Accuosto, Àlex Bravo, Montse Cuadros, Eva Martínez-Garcia, Horacio Saggion, and German Rigau. 2020. Cross-lingual semantic annotation of biomedical literature: experiments in Spanish and English. *Bioinformatics* 36, 6 (2020), 1872–1880.

[110] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 2227–2237. https://doi.org/10.18653/v1/N18-1202

[111] Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and Global Algorithms for Disambiguation to Wikipedia. In *HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics Stroudsburg, PA, USA ©2011, Portland, Oregon — June 19 - 24, 2011, 1375–1384.

[112] Kaiyu Ren, Albert M. Lai, Aveek Mukhopadhyay, Raghu Machiraju, Kun Huang, and Yang Xiang. 2014. Effectively processing medical term queries on the UMLS Metathesaurus by layered dynamic programming. *BMC Medical Genomics* 7, 1 (08 May 2014), S11. https://doi.org/10.1186/1755-8794-7-S1-S11

[113] Renato Rocha Souza, Douglas Tudhope, and Mauricio Almeida. 2012. Towards a Taxonomy of KOS: Dimensions for Classifying Knowledge Organization Systems. *KNOWLEDGE ORGANIZATION* 39 (01 2012), 179–192. https://doi.org/10.5771/0943-7444-2012-3-179

[114] Henry Rosales-Méndez, Barbara Poblete, and Aidan Hogan. 2018. What should Entity Linking link?

[115] Pedro Ruas and Francisco M. Couto. 2022. NILINKER: Attention-based approach to NIL Entity Linking. *Journal of Biomedical Informatics* 132 (2022), 104137. https://doi.org/10.1016/j.jbi.2022.104137

[116] P. Ruch. 2017. Text Mining to Support Gene Ontology Curation and Vice Versa. In *Methods in Molecular Biology*. Vol. 1446. Humana Press, New York, NY, 69–84. https://doi.org/10.1007/978-1-4939-3743-1_6

[117] Alex Rudniy, Min Song, and James Geller. 2014. Mapping biological entities using the longest approximately common prefix method. *BMC Bioinformatics* 15, 1 (14 Jun 2014), 187. https://doi.org/10.1186/1471-2105-15-187

[118] Conrad L Schoch, Stacy Ciufo, Mikhail Domrachev, Carol L Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, Richard Mcveigh, Kathleen O'Neill, Barbara Robbertse, Shobha Sharma, Vladimir Soussov, John P Sullivan, Lu Sun, Seán Turner, and Ilene Karsch-Mizrachi. 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)* 2020 (Jan. 2020).

[119] Wei Shen, Yuhan Li, Yinan Liu, Jiawei Han, Jianyong Wang, and Xiaojie Yuan. 2021. Entity Linking Meets Deep Learning: Techniques and Solutions. *IEEE Transactions on Knowledge and Data Engineering* (2021). https://doi.org/10.1109/TKDE.2021.3117715

[120] Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering* 27 (2 2015), 443–460. Issue 2. https://doi.org/10.1109/TKDE.2014.2327028

[121] Sunghwan Sohn, Cheryl Clark, Scott R Halgrim, Sean P Murphy, Christopher G Chute, and Hongfang Liu. 2014. MedXN: an open source medication extraction and normalization tool for clinical text. *Journal of the American Medical Informatics Association* 21, 5 (03 2014), 858–865. https://doi.org/10.1136/amiajnl-2013-002190 arXiv:https://academic.oup.com/jamia/article-pdf/21/5/858/17375955/21-5-858.pdf

[122] Mohammad Golam Sohrab, Khoa Duong, Makoto Miwa, Goran Topić, Ikeda Masami, and Takamura Hiroya. 2020. BENNERD: A Neural Named Entity Linking System for COVID-19. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Qun Liu and David Schlangen (Eds.). Association for Computational Linguistics, Online, 182–188. https://doi.org/10.18653/v1/2020.emnlp-demos.24

[123] Guojie Song, Qingqing Long, Yi Luo, Yiming Wang, and Yilun Jin. 2022. Deep Convolutional Neural Network Based Medical Concept Normalization. *IEEE Transactions on Big Data* 8 (2022), 1195–1208. https://api.semanticscholar.org/CorpusID:226487932

[124] M Q Stearns, C Price, K A Spackman, and A Y Wang. 2001. SNOMED clinical terms: overview of the development process and project status. *Proc. AMIA Symp.* (2001), 662–666.

[125] Riste Stojanov, Ilija Kocev, Sasho Gramatikov, Gorjan Popovski, Barbara Koroušić Seljak, and Tome Eftimov. 2020. Toward Robust Food Ontology Mapping. In *2020 IEEE International Conference on Big Data (Big Data)*. 3596–3601. https://doi.org/10.1109/BigData50022.2020.9378066

[126] Harrison S Suh, Jeffrey L Tully, Minhthy N Meineke, Ruth S Waterman, and Rodney A Gabriel. 2022. Identification of preanesthetic history elements by a natural language processing engine. *Anesthesia & Analgesia* 135, 6 (2022), 1162–1171.

[127] Xuhui Sui, Kehui Song, Baohang Zhou, Ying Zhang, and Xiaojie Yuan. 2022. A Multi-Task Learning Framework for Chinese Medical Procedure Entity Normalization. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 8337–8341. https://doi.org/10.1109/ICASSP43922.2022.9747858

[128] Xuhui Sui, Ying Zhang, Xiangrui Cai, Kehui Song, Baohang Zhou, Xiaojie Yuan, and Wensheng Zhang. 2023. BioFEG: Generate Latent Features for Biomedical Entity Linking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 11584–11593. https://doi.org/10.18653/v1/2023.emnlp-main.710

[129] Yingcheng Sun and Kenneth Loparo. 2019. Information Extraction from Free Text in Clinical Trials with Knowledge-Based Distant Supervision. In *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, Vol. 1. 954–955. https://doi.org/10.1109/COMPSAC.2019.00158

[130] Zenan Sun and Cui Tao. 2023. Named Entity Recognition and Normalization for Alzheimer's Disease Eligibility Criteria. In *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*. IEEE, 558–564.

[131] Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical Entity Representations with Synonym Marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 3641–3650. https://doi.org/10.18653/v1/2020.acl-main.335

[132] Mujeen Sung, Minbyul Jeong, Yonghwa Choi, Donghyeon Kim, Jinhyuk Lee, and Jaewoo Kang. 2022. BERN2: an advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics* 38, 20 (2022), 4837–4839.

[133] Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large Language Models for Data Annotation: A Survey. arXiv:2402.13446 [cs.CL] https://arxiv.org/abs/2402.13446

[134] Luis Tari, Varish Mulwad, and Anna von Reden. 2016. Interactive online learning for clinical entity recognition. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics* (San Francisco, California) *(HILDA '16)*. Association for Computing Machinery, New York, NY, USA, Article 8, 6 pages. https://doi.org/10.1145/2939502.2939510

[135] Li Tian, Weinan Zhang, Antonis Bikakis, Haofen Wang, Yong Yu, Yuan Ni, and Feng Cao. 2013. Medetect: a lod-based system for collective entity annotation in biomedicine. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, Vol. 1. IEEE, 233–240.

[136] Turdi Tohti, Mamatjan Abdurxit, and Askar Hamdulla. 2022. Biomedical Entity Linking Based on Global and Local Feature Fusion. In *2022 International Conference on Asian Language Processing (IALP)*. 253–258. https://doi.org/10.1109/IALP57159.2022.9961242

[137] Tomoki Tsujimura, Makoto Miwa, and Yutaka Sasaki. 2023. Large-scale neural biomedical entity linking with layer overwriting. *J. of Biomedical Informatics* 143, C (jul 2023), 11 pages. https://doi.org/10.1016/j.jbi.2023.104433

[138] Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko, and Valentin Malykh. 2018. Medical concept normalization in social media posts with recurrent neural networks. *Journal of Biomedical Informatics* 84 (2018), 93–102. https://doi.org/10.1016/j.jbi.2018.06.006

[139] Maya Varma, Laurel Orr, Sen Wu, Megan Leszczynski, Xiao Ling, and Christopher Ré. 2021. Cross-Domain Data Integration for Named Entity Disambiguation in Biomedical Text. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 4566–4575. https://doi.org/10.18653/v1/2021.findings-emnlp.388

[140] Shikhar Vashishth, Denis Newman-Griffis, Rishabh Joshi, Ritam Dutt, and Carolyn P. Rosé. 2021. Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets. *Journal of Biomedical Informatics* 121 (2021), 103880. https://doi.org/10.1016/j.jbi.2021.103880

[141] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[142] Alina Vretinaris, Chuan Lei, Vasilis Efthymiou, Xiao Qin, and Fatma Özcan. 2021. Medical entity disambiguation using graph neural networks. In *Proceedings of the 2021 international conference on management of data*. 2310–2318.

[143] Perceval Wajsbürt, Arnaud Sarfati, and Xavier Tannier. 2021. Medical concept normalization in French using multilingual terminologies and contextual embeddings. *Journal of Biomedical Informatics* 114 (2021), 103684.

[144] J. Wang, W. Mathews, H. Pham, H. Xu, and Y. Zhang. 2020. Opioid2FHIR: A system for extracting FHIR-compatible opioid prescriptions from clinical text. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE Computer Society, Los Alamitos, CA, USA, 1748–1751. https://doi.org/10.1109/BIBM49941.2020.9313258

[145] Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 303, 21 pages. https://doi.org/10.1145/3613904.3641960

[146] Yipei Wang, Xingyu Fan, Luoxin Chen, Eric I-Chao Chang, Sophia Ananiadou, Junichi Tsujii, and Yan Xu. 2019. Mapping anatomical related entities to human body parts based on wikipedia in discharge summaries. *BMC bioinformatics* 20 (2019), 1–11.

[147] Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. 2019. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic acids research* 47, W1 (2019), W587–W593.

[148] Qikang Wei, Tao Chen, Ruifeng Xu, Yulan He, and Lin Gui. 2016. Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks. *Database* 2016 (10 2016), baw140. https://doi.org/10.1093/database/baw140

[149] Maciej Wiatrak and Juha Iso-Sipila. 2020. Simple Hierarchical Multi-Task Neural End-To-End Entity Linking for Biomedical Text. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*. Association for Computational Linguistics, Online, 12–17. https://doi.org/10.18653/v1/2020.louhi-1.2

[150] World Health Organization (WHO). 2019/2021. International Classification of Diseases, Eleventh Revision (ICD-11). https://icd.who.int/browse11 Licensed under Creative Commons Attribution-NoDerivatives 3.0 IGO licence (CC BY-ND 3.0 IGO).

[151] Gongqing Wu, Ying He, and Xuegang Hu. 2018. Entity Linking: An Issue to Extract Corresponding Entity with Knowledge Base. *IEEE Access* 6 (1 2018), 6220–6231. https://doi.org/10.1109/ACCESS.2017.2787787

[152] Dongfang Xu, Manoj Gopale, Jiacheng Zhang, Kris Brown, Edmon Begoli, and Steven Bethard. 2020. Unified Medical Language System resources improve sieve-based generation and Bidirectional Encoder Representations from Transformers (BERT)–based ranking for concept normalization. *Journal of the American Medical Informatics Association* 27, 10 (07 2020), 1510–1519. https://doi.org/10.1093/jamia/ocaa080 arXiv:https://academic.oup.com/jamia/article-pdf/27/10/1510/34152904/ocaa080.pdf

[153] Jing Xu, Liang Gan, Mian Cheng, Quanyuan Wu, et al. 2018. Unsupervised medical entity recognition and linking in Chinese online medical text. *Journal of healthcare engineering* 2018 (2018).

[154] Cheng Yan, Yuanzhe Zhang, Kang Liu, Jun Zhao, Yafei Shi, and Shengping Liu. 2021. Enhancing unsupervised medical entity linking with multi-instance learning. *BMC medical informatics and decision making* 21 (2021), 1–10.

[155] Siyu Yang, Peiliang Zhang, Chao Che, and Zhaoqian Zhong. 2023. B-LBConA: a medical entity disambiguation model based on Bio-LinkBERT and context-aware mechanism. *BMC bioinformatics* 24, 1 (2023), 97.

[156] Hongyi Yuan, Zheng Yuan, and Sheng Yu. 2022. Generative Biomedical Entity Linking via Knowledge Base-Guided Pre-training and Synonyms-Aware Fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 4038–4048. https://doi.org/10.18653/v1/2022.naacl-main.296

[157] Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. 2022. CODER: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of biomedical informatics* 126 (2022), 103983.

[158] Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Knowledge-Rich Self-Supervision for Biomedical Entity Linking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 868–880. https://doi.org/10.18653/v1/2022.findings-emnlp.61

[159] Yizhou Zhang, Xiaojun Ma, and Guojie Song. 2018. Chinese medical concept normalization by using text and comorbidity network embedding. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 777–786.

[160] Jin G Zheng, Daniel Howsmon, Boliang Zhang, Juergen Hahn, Deborah McGuinness, James Hendler, and Heng Ji. 2015. Entity linking for biomedical literature. *BMC medical informatics and decision making* 15 (2015), 1–9.

[161] Huiwei Zhou, Shixian Ning, Zhe Liu, Chengkun Lang, Zhuang Liu, and Bizun Lei. 2020. Knowledge-enhanced biomedical named entity recognition and normalization: application to proteins and genes. *BMC bioinformatics* 21 (2020), 1–15.

[162] Tiantian Zhu, Yang Qin, Qingcai Chen, Xin Mu, Changlong Yu, and Yang Xiang. 2023. Controllable Contrastive Generation for Multilingual Biomedical Entity Linking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 5742–5753. https://doi.org/10.18653/v1/2023.emnlp-main.350

[163] Tiantian Zhu, Yang Qin, Ming Feng, Qingcai Chen, Baotian Hu, and Yang Xiang. 2023. BioPRO: Context-Infused Prompt Learning for Biomedical Entity Linking. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 32 (nov 2023), 374–385. https://doi.org/10.1109/TASLP.2023.3331149

[164] Stefan Zwicklbauer, Christin Seifert, and Michael Granitzer. 2013. Do We Need Entity-Centric Knowledge Bases for Entity Disambiguation?. In *Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies* (Graz, Austria) *(i-Know '13)*. Association for Computing Machinery, New York, NY, USA, Article 4, 8 pages. https://doi.org/10.1145/2494188.2494198

[165] Stefan Zwicklbauer, Christin Seifert, and Michael Granitzer. 2015. Search-based entity disambiguation with document-centric knowledge bases. In *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business*. 1–8.

A SEARCH STRATEGIES

A.1 ACL anthology

```
1. Download ACL anthology file from \url{} to obtain the file 'anthology+abstracts.bib'
% .bib file downloaded from the ACL site including all references to articles published in ACL
    conferences part of the ACL Anthology.
2. output = [] % the list that will store the information about the articles included in the review
3. Instantiate the variables:
    query_terms_1 = ['biomedical', 'medical', 'clinical']
    query_terms_2 = ['entity', 'term', 'concept']
    query_terms_3 = ['linking', 'normalization', 'mapping', 'disambiguation', 'annotation', '
    extraction']

4. For doc in data % iterate over each article present in data
5.     title = ''
6.     abstract = ''
7.     year = ''
8.     url = ''
9.     doi = ''
10.    keep_doc = False
11. If title is not None and abstract is not None:
12.    title_abs = title + abstract
13.    If 2013 <= year <= 2024:
14.        For term1 in query_terms_1:
15.            If term1 in title_abs:
16.                For term2 in query_terms_2
17.                    If term2 in title_abs:
18.                        For term3 in query_terms_3:
19.                            If term3 in title_abs:
20.                                keep_doc = True
21.    If keep_doc: entry = ['acl', year, title, abstract, url, doi] % include current article in the
    review `output.append(entry)` %

Output: output % list containing the articles included in the review
```

literate

A.2 IEEE Xplore

(1) Start page: https://ieeexplore.ieee.org/Xplore/home.jsp

(2) Search using the query: “(’biomedical’ OR ’medical’ OR ’clinical’) AND (’entity’ OR ’term’ or ’concept’) AND (’linking’ OR ’normalization’ OR ’mapping’ OR ’disambiguation’ OR ’annotation’ OR ’extraction’)"
(3) Set “Year" → “Custom range" to “2013-2024"
(4) Set “Publication topics" to “Natural Language Processing Tasks" and “Text Analysis"
(5) Click on “Export" → “Download"

A.3 ACM Digital Library

(1) Start page: https://dl.acm.org/;
(2) Click on “Advanced search";
(3) Set “Search within" to “Title" and replace set search term to: “((’biomedical’ OR ’medical’ OR ’clinical’) AND (’entity’ OR ’term’ OR ’concept’) AND (’linking’ OR ’normalization’ OR ’mapping’ OR ’disambiguation’ OR ’annotation’ OR ’extraction’))"
(4) Click on “Add search field" and set “Search within" to “Abstract" and replace set search term to: ‘((’biomedical’ OR ’medical’ OR ’clinical’) AND (’entity’ OR ’term’ OR ’concept’) AND (’linking’ OR ’normalization’ OR ’mapping’ OR ’disambiguation’ OR ’annotation’ OR ’extraction’)"
(5) Set “Publication date" → Set “Custom range" From “Jan" of the year “2013" To “Dec" of the year “2024";
(6) Click on “Search"
(7) Apply the filter “Content type" → “Research article"
(8) Click on “Select all" → “Export citations" → Bibtex

A.4 PubMed

Retrevied articles from *PubMed* using Bio.Entrez package [34] with the following query:

```
2024/01/01:2024/12/31[dp]
AND ('biomedical'[tiab] OR 'medical'[tiab] OR 'clinical'[tiab])
AND ('entity'[tiab] OR 'term'[tiab] OR 'concept'[tiab])
AND ('linking'[tiab] OR 'normalization'[tiab] OR 'mapping'[tiab] OR 'disambiguation'[tiab] OR '
    annotation'[tiab] OR 'extraction'[tiab])
AND ('Natural Language Processing'[mh] OR 'Software'[mh] OR 'Biological Ontologies'[mh] OR 'Data
    Mining/methods'[mh] OR 'Knowledge Bases'[mh])
```

literate

B DATA EXTRACTION FROM REVIEWED ARTICLES

Data extracted from the reviewed articles are shown in Table 2.

Received 20 April 2024

[34] https://biopython.org/docs/1.75/api/Bio.Entrez.html

Name	Description	Options
'name'	Designation of the proposed approach	-
'task'	Designation for the named entity linking task	-
'title'	Title of the article	-
'source'	Source repository	'ieee', 'acl', 'acm', 'pubmed'
'journal'	Publication	-
'year'	Year of publication	Range [2013-2024]
'doi'	Digital Object Identifier (DOI) of the article	-
'lang'	Target language according to the ISO language codes	'en' (English), 'zh' (Chinese), 'es' (Spanish), 'fr' (French), 'de' (German), 'nl' (Dutch), 'hb' (Hebrew), 'ko' (Korean)
'text type'	Sub-domain of application or target text type	'lit-abs' (abstract from scientific article), 'lit-full' (full-text from scientific article), 'lit-fig' (figure captions from scientific article), 'ehr' (electronic health record), 'pat' (patents), 'trial' (clinical trial), 'notes' (clinical notes/reports), 'social' (social media or user reviews), 'web' (websites), 'dial' (dialog), 'labels' (drug labels), 'recipes' (food recipes), 'query' (query specified by a user), 'wiki', 'other'
'ent type'	Entity types addressed by the approach	'bio' (general biomedical entity, e.g., UMLS entities), 'disease', 'symptom' (symptom or phenotype), 'treat' (treatment), 'proc' (medical procedure), 'biomark' (biomarker), 'adr' (adverse drug reaction), 'chemical' (chemical or drug), 'gene' (gene or protein), 'taxon' (taxonomic or species), 'mutation', 'anat' (anatomical), 'ima' (imaging modality), 'food', 'cell line', 'cell type', 'cell component' (cellular component), 'bioprocess' (biological process), 'function' (molecular function), 'habitat' (microbial habitat), 'chem reaction' (chemical reaction), 'other'
'kos'	The target KOS to which the approach links entities	custom, ICD9, UMLS, MEDIC, MedDRA, Entrez Gene, MeSH, OMIM, SNOMED-CT, PySearch2, ChEBI, DrugBank, US FDA approved drugs, NCBI Taxonomy (NCBITaxon), dbSNP, ClinVar, ICD10, CTD-Chemical, Hansard, FoodOn, FHIR, ORDO, RadLex, NCI Thesaurus, NCBI Gene, Cellosaurus, Cell Ontology (CL), CTD-Anatomy, HPO, GO-BP, AMT, SIDER, Uniprot, GO, THBP, OntoBiotope, medical Baidu Baike, Sogou medical dictionaries, Cell Type Ontology, Protein Ontology (PR), Sequence Ontology (SO), First Databank, Micromedex, MediSpan, Gold Standard, Multum, NDF-RT, WikiData, Wikipedia, OHDSI, DBpedia, DailyMed, Diseasome, Medicare, MONDO Disease Ontology (MONDO), Molecular Process Ontology (MOP), Tree of Human Body Parts (THBP)
'approach'	Type of approach	'dict' (dictionary-based), 'graph' (graph-based), 'rule' (rule-based), 'dist' (distributional representations, i.e., vectors), 'ml' (machine learning-based)
'approach details'	Details of the approach	-
'ml approach'	Whether a machine learning approach is a deep learning approach	'ml' (machine learning) or 'dl' (deep learning)
'supervision'	Approach is supervised or unsupervised	'sp' (supervised), 'un' (unsupervised), 'self' (self-supervised)
'pt model'	Pre-trained model	- (e.g., BERT, BioBERT, etc)
'datasets'	Datasets used for evaluation	-
'task'	The approach was developed in the context of a community task	'Y' (yes) or 'N' (no)
'NER'	Specifies whether the approach also performs NER along NEL	'Y' (Yes), 'N' (No)
'avail'	Availability of the approach	'code' (the respective code is public), 'web' (available as a web tool), 'package' (available as an installable package), 'NA' (not available)
'link'	URL corresponding to the location of the tool	-

Table 2. Description of the data extracted from the reviewed articles

Appendix B

Deep Semantic Entity Linking

Deep Semantic Entity Linking*

Pedro Ruas[0000−0002−1293−4199]

LASIGE, Faculdade de Ciências, Universidade de Lisboa, Lisbon 1749-016, Portugal
psruas@fc.ul.pt

Abstract. Named entity linking systems are an essential component in text mining pipelines, mapping entity mentions in the text to the appropriate knowledge base identifiers. However, the current systems have several limitations affecting their performance: the lack of context of the entity mentions, the incomplete disambiguation graphs and the lack of approaches to deal with unlinkable entity mentions. The PhD project will focus on solving the aforementioned challenges in order to develop a NEL model which outperforms state-of-the-art performance in Biomedical and Life Sciences domains.

Keywords: Named Entity Linking · Text Mining · NIL entities · Multilingual corpora · Graph-based models

1 Introduction

In text mining pipelines, named entity linking (NEL) systems map the entity mentions recognised by named entity recognition tools to the appropriate knowledge base (KB) concepts. These play an important role in several tasks, such as automatic population and curation of KBs [6], improvement of question answering [22] and search engines [15], and identification of diseases in electronic health records [11]. The simplest approach to the NEL problem consists in choosing the KB concept with the most similar label for each entity mention using string matching techniques, which is very limited since it does not consider the context of the mention. At the contrary, language models pre-trained with large amounts of text, like BERT [4] or ELMo [17], learn contextualised representations of words able to express their meaning according to their local context. Graph-based NEL models consider the global context of the mentions, building a disambiguation graph with mentions and the respective KB candidates as nodes and then attempting to maximise the coherence between the disambiguation candidates. Those based on Personalized PageRank algorithm are one of the state-of-the-art approaches in NEL [7]. However, the main limitation of graph-based approaches is incomplete graphs (e.g., graphs with few edges between nodes) affecting their precision, which is usually caused by a lack of domain knowledge in the KB. Also, when the KB is incomplete, a NEL system is not

* Supported by FCT through the DeST: Deep SemanticTagger project, ref. PTDC/CCI-BIO/28685/2017, PhD Scholarship, ref. 2020.05393.BD, LASIGE ResearchUnit, ref. UIDB/00408/2020 and ref. UIDP/00408/2020

able to associate some entity mentions with the respective KB concepts (unlinkable or NIL entities), which leads to a low recall [21]. Wu et al. [23] divided the approaches to NIL entity clustering in three categories: string matching, hierarchical agglomerative clustering and graph-based. These approaches only attempt to group different mentions of the same NIL entity, but none of them attempts to disambiguate it, even if it is not a perfect disambiguation. The main challenges associated with NEL systems that the PhD proposal intends to address are the lack of mention context to determine local similarity, the incompleteness of disambiguation graphs on graph-based models, and the absence of approaches to link NIL entities to KBs.

The main objectives of the PhD proposal are the development of a NEL system that outperforms state-of-the-art approaches in Biomedical and Life Sciences domains in terms of recall, by creating deep semantic links between NIL entities and concepts of a given KB (NIL entity linking), and in terms of precision, by completing disambiguation graphs with relations extracted from text and with the output of the NIL entity linking model, and by improving improving the local similarity determination with contextualised embedding representations for entity mentions, their context and respective KB candidates.

2 Research Methodology

2.1 Improvement of the disambiguation graph of a graph-based NEL model

Usually, the output of NEL improves RE systems. The hypothesis here is that, at the contrary, RE improves the performance of NEL: RE captures relations between entities that are expressed in text but not in the KB, which complete the disambiguation graph with edges. The goal is to develop a graph-based NEL model that integrates the output of RE systems to improve the disambiguation graph (REEL). The novelty is the use of RE systems to improve NEL systems, and not the other way around as it usually happens. This methodology could originate a feedback cycle in which NEL performance impacts RE output, which then improves NEL performance and so forth. RE tools, like BO-LSTM [10], extract the relations between entities in the text and will add them as edges to the disambiguation graph. This module evaluates the impact of denser disambiguation graphs on the performance of the PPR algorithm. The work relative to this module has been executed and published in a journal paper [19].

2.2 Improvement of a local NEL model

The hypothesis is that pre-trained language models improve the determination of local similarity between entity mention and KB candidates through contextualised word embeddings. The goal is to develop a local NEL model that goes beyond string similarity methods, leveraging contextualised word embeddings. The precision of the model should increase compared with a string matching based

approach. The first step is the generation of a KB candidates list for each mention resorting to abbreviation expansion, string matching and synonyms lookup, and then pre-trained language models are explored, like BERT [4], ClinicalBERT [1], or BioBERT [13] to create contextualised embeddings for mentions, as well their KB candidates. The comparison of mention embeddings with KB candidate embeddings will return a similarity score that will be further used to filter out less relevant candidates from the respective candidates lists.

2.3 Development of a NIL entity linking model

The basis for this module is the framework by Qi et al. [18] to originate embeddings for multi-word expressions, leveraging the sememes associated with the constituent words. The hypothesis here is that, analogously, the meaning of a NIL entity is expressed through the KB concepts associated with its words. The goal is to develop an attention-based model to find the most relevant KB candidate concept for the respective NIL entity in order to disambiguate it. The model converts NIL entities into deep semantic links to the KB, which are added to the disambiguation graph in Subsection 2.1. The first step is to build a word-concept dictionary to allow candidate retrieval for NIL entities: a key is a word present either in a KB concept designation or definition, and its values are the KB concept identifiers where that word appears. For each of the words of a NIL entity, the associated KB concepts are retrieved from the dictionary. Both words and candidates are represented by embeddings, the input to the attention model. The second step is the development of an attention-based model to find the most relevant KB candidate concept, and its schema is represented in Fig.1.

2.4 Evaluation

The evaluation of the improved graph-based model REEL (subsection 2.1) consists of the comparison of its performance with two baseline approaches: string matching and PPR-SSM [9]. The performance of these models (F1-score, precision, recall) is measured in two gold standard datasets: BC5CDR Corpus [14] and CRAFT corpus [2]. The evaluation of the improved local model (subsection 2.2) consists of the comparison of its performance (F1-score, precision, recall) with a string matching baseline approach in the following datasets: NCBI Disease Corpus [5], BC5CDR Corpus, and CRAFT corpus, MedMentions [16]. The evaluation of the NIL entity linking model (subsection 2.3) includes two different steps. The first step involves a specific silver standard built from existing NEL datasets. To build the dataset, existing annotations are converted into NIL annotations, by associating each entity to the label of the direct ancestor of the gold label. The performance of two models (accuracy), string matching (baseline) and NIL entity linking model, is measured in the referred dataset. In the second step, deep semantic links, i.e. the output generated by the NIL entity linking model, are added to the disambiguation graph (nodes and respective edges) of PPR-SSM [9]). The performance of two models, PPR-SSM and improved PPR-SSM,

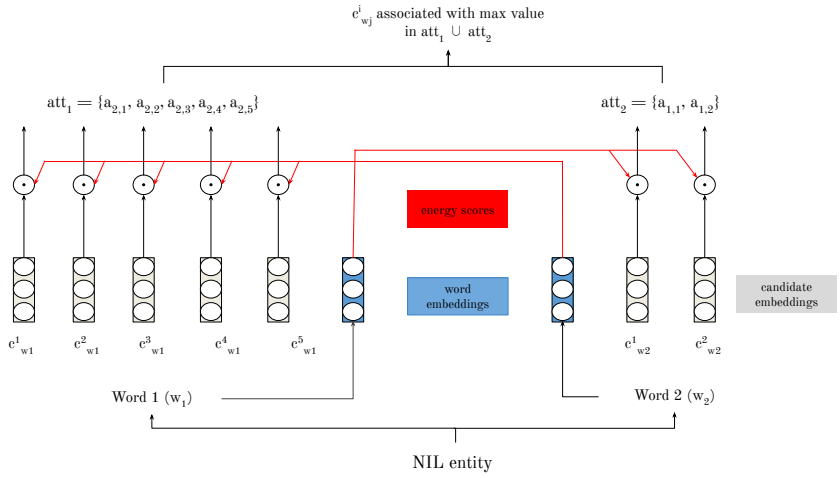


Fig. 1. Attention model schema. A NIL entity with two words is tokenized, word 1 has five KB candidates associated, whereas word 2 has two candidates. Word and candidate embeddings are the input for the attention model. The attention weights of word 1 candidates (att_1) are based on the energy scores computed from the embeddings for word 2, and vice-versa. At the end, the candidate associated with the highest attention weight in both att_1 and att_2 disambiguates the NIL entity.

is measured (F1-score, precision, recall) in NCBI Disease Corpus, BC5CDR Corpus, CRAFT corpus, and MedMentions dataset. The three referred modules are then integrated into a unique hybrid NEL model, which is evaluated in several datasets [5, 14] and, additionally, in a new parallel, multilingual dataset. Since there is no Portuguese biomedical NEL dataset available, the goal is to build one containing biomedical and clinical text in Portuguese, English and Spanish. The documents are retrieved from SciELO[1] and PubMed[2] repositories, automatic NER and NEL tools, like MER [3], recognise medical diagnostic entities present in the documents and link them to terms of the *International Classification of Diseases 10 - Clinical Modification* (ICD10-CM), and then there is a manual validation of the obtained annotations by crowdsourcing and by expert analysis over a selected subset of the documents. The performance of the hybrid model (F1-score, precision, recall), as well of other SOTA approaches (BERT-Based Biomedical Entity Normalization [8] or TaggerOne [12]) is measured on the referred datasets. Preliminary work relative to the development of the parallel, multilingual dataset has been done and published in a workshop paper [20].

[1] https://scielo.org/
[2] https://pubmed.ncbi.nlm.nih.gov/

References

1. Alsentzer, E., Murphy, J., Boag, W., Weng, W.H., Jindi, D., Naumann, T., McDermott, M.: Publicly Available Clinical BERT Embeddings. In: Proceedings ofthe 2nd Clinical Natural Language Processing Workshop. pp. 72–78. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/w19-1909
2. Cohen, K.B., Verspoor, K., Funk, C., Bada, M., Palmer, M., Hunter, L.E.: The Colorado Richly Annotated Full Text (CRAFT) Corpus : Multi-Model Annotation in the Biomedical Domain The Colorado Richly Annotated Full Text (CRAFT) Corpus : Multi-Model Annotation In The Biomedical Domain. In: The Handbook of Linguistic Annotation. No. June (2017). https://doi.org/10.1007/978-94-024-0881-2
3. Couto, F.M., Lamurias, A.: MER: a shell script and annotation server for minimal named entity recognition and linking. Journal of Cheminformatics **10**(1), 58 (dec 2018). https://doi.org/10.1186/s13321-018-0312-9, https://jcheminf.biomedcentral.com/articles/10.1186/s13321-018-0312-9
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (oct 2018), http://arxiv.org/abs/1810.04805
5. Doğan, R.I., Leaman, R., Lu, Z.: NCBI disease corpus: A resource for disease name recognition and concept normalization. Journal of Biomedical Informatics **47**, 1–10 (2014). https://doi.org/10.1016/j.jbi.2013.12.006
6. Dredze, M., Mcnamee, P., Rao, D., Gerber, A., Finin, T.: Entity Disambiguation for Knowledge Base Population. In: 23rd International Conference on Computational Linguistics. pp. 277–285. No. August (2010). https://doi.org/10.3115/1119176.1119181
7. Guo, Z., Barbosa, D.: Robust named entity disambiguation with random walks. Semantic Web **9**(4), 459–479 (2018). https://doi.org/10.3233/SW-170273
8. Ji, Z., Wei, Q., Xu, H.: BERT-based Ranking for Biomedical Entity Normalization (2019), http://arxiv.org/abs/1908.03548
9. Lamurias, A., Ruas, P., Couto, F.M.: PPR-SSM: Personalized PageRank and semantic similarity measures for entity linking. BMC Bioinformatics **20**(1), 1–12 (2019). https://doi.org/10.1186/s12859-019-3157-y
10. Lamurias, A., Sousa, D., Clarke, L.A., Couto, F.M.: BO-LSTM: Classifying relations via long short-term memory networks along biomedical ontologies. BMC Bioinformatics **20**(10) (2019). https://doi.org/10.1186/s12859-018-2584-5
11. Leaman, R., Khare, R., Lu, Z.: Challenges in clinical natural language processing for automated disorder normalization. Journal of Biomedical Informatics **57**, 28–37 (2015). https://doi.org/10.1016/j.jbi.2015.07.010, http://dx.doi.org/10.1016/j.jbi.2015.07.010
12. Leaman, R., Lu, Z.: TaggerOne: Joint named entity recognition and normalization with semi-Markov Models. Bioinformatics **32**(18), 2839–2846 (2016). https://doi.org/10.1093/bioinformatics/btw343
13. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics (September), 1–7 (2019). https://doi.org/10.1093/bioinformatics/btz682
14. Li, J., Sun, Y., Johnson, R.J., Sciaky, D., Wei, C.H., Leaman, R., Davis, A.P., Mattingly, C.J., Wiegers, T.C., Lu, Z.: BioCreative V CDR task corpus: a resource for

chemical disease relation extraction. Database : the journal of biological databases and curation **2016**, 1–10 (2016). https://doi.org/10.1093/database/baw068
15. Meij, E., Balog, K., Odijk, D.: Entity linking and retrieval for semantic search. In: WSDM 2014 - Proceedings of the 7th ACM International Conference on Web Search and Data Mining. p. 683. No. February 2014, New York, New York, USA (2014). https://doi.org/10.1145/2556195.2556201
16. Mohan, S., Li, D.: Medmentions: A large biomedical corpus annotated with umls concepts (2019)
17. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep Contextualized Word Representations pp. 2227–2237 (2018). https://doi.org/10.18653/v1/n18-1202
18. Qi, F., Huang, J., Yang, C., Liu, Z., Chen, X., Liu, Q., Sun, M.: Modeling Semantic Compositionality with Sememe Knowledge. In: Proceedings ofthe 57th Annual Meeting ofthe Association for Computational Linguistics. pp. 5706–5715. Association for Computational Linguistics, Florence, Italy, July 28 - August 2, 2019 (2019). https://doi.org/10.18653/v1/p19-1571
19. Ruas, P., Lamurias, A., Couto, F.M.: Linking chemical and disease entities to ontologies by integrating PageRank with extracted relations from literature. Journal of Cheminformatics **12**(1), 1–11 (2020). https://doi.org/10.1186/s13321-020-00461-4, https://doi.org/10.1186/s13321-020-00461-4
20. Ruas, P., Lamúrias, A., Couto, F.M.: Towards a multilingual corpus for named entity linking evaluation in the clinical domain. In: CEUR Workshop Proceedings. vol. 2619, pp. 2–4 (2020)
21. Shen, W., Wang, J., Han, J.: Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. IEEE Transactions on Knowledge and Data Engineering **27**(2), 443–460 (2015). https://doi.org/10.1109/TKDE.2014.2327028, http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6823700
22. Sorokin, D., Gurevych, I.: Mixing Context Granularities for Improved Entity Linking on Question Answering Data across Entity Categories. In: 7th Joint Conference on Lexical and Computational Semantics (*SEM). pp. 65–75. Association for Computational Linguistics (2018)
23. Wu, G., He, Y., Hu, X.: Entity Linking: An Issue to Extract Corresponding Entity with Knowledge Base. IEEE Access **6**(c), 6220–6231 (2018). https://doi.org/10.1109/ACCESS.2017.2787787, http://ieeexplore.ieee.org/document/8246707/

Appendix C

LASIGE and UNICAGE solution to the NASA LITCOIN NLP competition

LASIGE AND UNICAGE SOLUTION TO THE NASA LITCOIN NLP COMPETITION

Pedro Ruas[†], **Diana F. Sousa**[†]
LASIGE
Lisbon, Portugal
{psruas, dfsousa}@fc.ul.pt

André Neves[†], **Carlos Cruz**
Unicage Europe
Lisbon, Portugal
{andre.neves, carlos.cruz}@unicage.com

Francisco M. Couto
LASIGE
Lisbon, Portugal
fcouto@di.fc.ul.pt

ABSTRACT

Biomedical Natural Language Processing (NLP) tends to become cumbersome for most researchers, frequently due to the amount and heterogeneity of text to be processed. To address this challenge, the industry is continuously developing highly efficient tools and creating more flexible engineering solutions. This work presents the integration between industry data engineering solutions for efficient data processing and academic systems developed for Named Entity Recognition (LasigeUnicage_NER) and Relation Extraction (BiOnt). Our design reflects an integration of those components with external knowledge in the form of additional training data from other datasets and biomedical ontologies. We used this pipeline in the 2022 LitCoin NLP Challenge, where our team LasigeUnicage was awarded the 7th Prize out of approximately 200 participating teams, reflecting a successful collaboration between the academia (LASIGE) and the industry (Unicage). The software supporting this work is available at https://github.com/lasigeBioTM/Litcoin-Lasige_Unicage.

***K*eywords** Data Processing · Named-Entity Recognition · Relation Extraction · Knowledge Graphs

1 Introduction

Biomedical data is presented normally in complex, large, and diverse formats. Whether in free-text form or highly specialized knowledge graphs, the data needs to be processed before one can integrate it into predictive pipelines or derive new research hypotheses to target. The data to be processed frequently falls in more than one format and comes in large volumes, which requires efficient data processing [1].

Due to the elevated data needs in the industry, there is a need to develop efficient, flexible data engineering solutions to accommodate different formats and volumes. For that, Unicage[*] offers a set of commands that allow the user to build efficient programs that can be combined in a modular way to build robust, yet flexible, big data processing pipelines. Unicage Europe is a data engineering startup company with a focus on big data processing through the usage of a shell scripting development methodology alongside a set of command-line tools. These commands are format agnostic, meaning they can work with any type of text data. The tools are written in the C programming language and have been geared towards performance, leveraging the OS's memory and resource management capabilities to deliver high-speed processing. The variety of commands is also able to cover gaps that are present in the toolbox of built-in OS utilities. When compared with other big data processing solutions, Unicage tools and methodology are able to fare reasonably well, with cases where it is able to surpass them in terms of speed and efficiency [2, 3].

[†]Authors contributed equally to this research.
[*]https://unicage.eu/

LASIGE also has vast experience using shell scripting to perform data and text Processing for Health and Life Sciences [4]. Lately, LASIGE's NLP academic research focused on two main common tasks, Named-Entity Recognition (NER) and Relation Extraction (RE), by developing systems such as BiOnt [5]. These tasks correspond to the 2-phase 2022 LitCoin NLP Challenge[†] in which we participated with our team, LasigeUnicage, creating a successful industry-academia collaboration. Our solution allied the data processing efficiency of Unicage with LASIGE's expertise in NLP.

The 2022 LitCoin NLP Challenge was a part of the NASA Tournament Lab, hosted by the National Center for Advancing Translational Sciences (NCATS) and the National Library of Medicine (NLM). The competition aimed to create a data-driven technological solution that leverages the vast volumes of biomedical publications published daily to advance the biomedical field by increasing discoverability and formulating new research hypotheses. Specifically, the goal was to extract scientific concepts from scientific articles (Part 1), connect them by generating knowledge assertions, and label them as novel findings or background information (Part 2).

Our design to target the 2022 LitCoin NLP Challenge relied on two steps: a NER pipeline developed explicitly for the task (Part 1) and the use of the BiOnt system [5] for RE (Part 2). In Part 1, we used the Unicage commands to build a pipeline to gather all the datasets by category and then convert them to the BIO/IBO ("inside-outside-beginning") format required for the NER step. Then, we ensemble six models based on PubMedBERT [6] to recognize the six different types of entities (DiseaseOrPhenotypicFeature, ChemicalEntity, OrganismTaxon, GeneOrGeneProduct, SequenceVariant, CellLine). For Part 2 of the challenge, we used Unicage commands to preprocess the input data and the BiOnt system to perform RE. This system relies not only on the training data itself but can also integrate external knowledge in the form of biomedical ontologies such as ChEBI [7] and GO [8] to further improve the RE process. Finally, we resorted to Unicage commands to post-process the output to identify whether the relations were considered novel.

We state our main contributions described in this paper below:

- Efficient biomedical NLP Pipeline based on industry data processing tools and academically developed systems for NER and RE.
- Application of external data into the biomedical NLP pipeline in the NER and RE stages.
- Integration of LASIGE's NER and RE systems and Unicage industry data processing solutions.

2 Part 1: Named-Entity Recognition

In Part 1, given an abstract text the goal was to find/recognize all biomedical entities of types: DiseaseOrPhenotypicFeature, ChemicalEntity, OrganismTaxon, GeneOrGeneProduct, SequenceVariant, or CellLine. For example, given the sentence *Late-onset metachromatic leukodystrophy: molecular pathology in two siblings.*, the goal is to identify the entity *metachromatic leukodystrophy*, a DiseaseOrPhenotypicFeature.

We started by preprocessing the training datasets containing documents from several Named-Entity Recognition (NER) corpora. Then, we ensemble six trained models (PubMedBERT + linear layer for token classification) to recognize the six different types of entities.

2.1 Data Processing

For data processing, we first converted the corpora to the BIO format, merged the different corpora into a single file for each entity type, and generated the final training datasets for each entity type. Most data was pre-processed using Unicage commands and Shell Scripting. The majority of Unicage commands used were data manipulation commands, such as *self* and *delf*, which are two commands that allow easy manipulation of the data fields on each record of the files, independent of their format, or commands that present an optimization over OS built-in tools, such as *uawk*, which is an optimized version of GNU awk [9]. To complement this pipeline, the usage of specific Python libraries, such as bconv[‡] and the standoff2conll [§], were also used in order to convert the datasets to specific file types so that their manipulation by the Unicage tools could be facilitated.

The datasets used per entity type are the following:

[†] `https://ncats.nih.gov/funding/challenges/litcoin`
[‡] `https://pypi.org/project/bconv/`
[§] `https://github.com/spyysalo/standoff2conll`

- DiseaseOrPhenotypicFeature: BC5CDR [10], PGxCorpus [11], NCBI Disease [12], Disease Names and Adverse Effects [13], MedMentions [14], and PHAEDRA [15].
- ChemicalEntity: Corpora for Chemical Entity Recognition [16], CRAFT [17], BC5CDR [10], CHR [18], and PHAEDRA [15].
- OrganismTaxon: LINNAEUS [19], CRAFT [17], Species-800 [20], and Cell Finder [21].
- GeneOrGeneProduct: BC2GM [22], JNLPBA [23], CRAFT [17], PGxCorpus [11], FSU_PRGE [24], and Cell Finder [21].
- CellLine: JNLPBA [23], GELLUS [25], CLL [¶], and Cell Finder [21].
- SequenceVariant: tmVar [26], PGxCorpus [11], and SNPPhenA [27].

The integration of each training dataset and the competition dataset was done by assigning the relevant tags to the targeted entity. For example, in the training dataset for entities of the type DiseaseOrPhenotypicFeature, the tokens relative to entities in the competition dataset were tagged with *B* and *I*, whereas tokens relative to entities of other types were assigned the tag *O*.

2.2 Model architecture

Our approach used PubMedBERT embeddings (originally trained on PubMed articles), jointly with a linear classification layer to classify each token that was fine-tuned in each training dataset. We defined a post-processing rule for entities of type CHEMICAL with a length of 1. We checked if the character corresponded to a letter since chemical elements can be represented by a single letter (e.g., *C* represents *carbon*). For the remaining entity types, we excluded entities with a length of 1 in the output.

2.3 Methodology

For Part 1, our methodology consisted of the following:

1. Hyperaparameter optimization: `training epochs`, `learning rate`, `train batch size`, `test batch size`. Small versions of the training sets and competition training dataset (in BIO format) were used as test sets.
2. Fine-tuning: After finding the optimal number of training epochs and learning rates, we combined the competition training dataset with the rest of the training datasets and then trained the model (90% training, 10% validation). Training of a distinct model for each of the six competition entity types.
3. Prediction: Sequential application of each model in a given sentence of an abstract present in the competition's test set.

2.4 Implementation

- 1 Tesla M10 GPU
- Training time (excluding hyperparameter optimization) took approximately 15 hours.
- Prediction time took approximately 8 minutes.

3 Part 2: Relation Extraction

For Part 2, given the abstract text and the identified entities from Part 1, the goal was to identify relations between the biomedical entities. The relation could be classified in a combination of a first label of Association, Positive Correlation, Negative Correlation, Bind, Cotreatment, Comparison, or Drug Interaction, and a second label of Novel or Not Novel. For example, given the sentence *Midline B3 serotonin nerves in rat medulla are involved in hypotensive effect of methyldopa.* with the identified biomedical entities *serotonin* (ChemicalEntity), *rat* (OrganismTaxon), *hypotensive* (DiseaseOrPhenotypicFeature), and *methyldopa* (ChemicalEntity), the goal is to identify the relation between *serotinin* and *hypontensive* and classify as a Positive Correlation Novel.

Identically to Part 1, we started by preprocessing the datasets provided by the challenge organizers and the biomedical ontologies associated with the entities' identifiers. We assembled two BiOnt [5] models. We used the first model

[¶] `https://turkunlp.org/Cell-line-recognition/`

to identify the eight types of relations: Association, Positive Correlation, Negative Correlation, Bind, Cotreatment, Comparison, and Drug Interaction. The second model was to classify the relation even further between Novel and Not Novel.

3.1 Data Processing

Part of the initial data was pre-processed using Unicage commands and Shell Scripting.

Afterwards, we linked the MESH ontology[‖] to the entity types DiseaseOrPhenotypicFeature and ChemicalEntity and the NCBITaxon ontology[**] to the entity types Species and CellLine.

The competition training set was tokenized using BiOnt, and each entity covered by the ontologies considered above was mapped within the hierarchy of that ontology. Finally, the output generated from the model was processed using Unicage commands and Shell Scripting. These scripts used a small set of rules to choose which No/Novel tag to keep for each relation, while at the same time, it generated the final files in the format required by the competition.

3.2 Model architecture

Our approach used the BiOnt system, a biomedical RE system built using bidirectional LSTM networks. The BiOnt system incorporates Word2Vec word embeddings [28] and uses different combinations of input channels to maximize performance, including ontology embeddings. We used the full provided abstract and considered the multiple relations mentioned within each abstract to have more training cases on the same type of relation.

We are aware of the limitations of our approach, given that the BiOnt system architecture is no longer state-of-the-art for biomedical relation extraction. However, the system's unique approach to external knowledge injection allows us to include each representative ontology within the training pipeline furthering the knowledge about each entity in a candidate relation.

3.3 Methodology

For Part 2, our methodology consisted of the following:

1. Hyperaparameter optimization: `training epochs`, `learning rate`, `train batch size`, `test batch size`, `max text length`, `class weights`. Small versions of the training sets and competition training dataset (in BIO format) were used as test sets.
2. Fine-tuning: After finding the optimal number of training epochs and learning rates, we obtained two trained models, one to predict the different types of relations and the other to predict if they were Novel or not.
3. Prediction: Sequential application of each model in a given abstract present in the competition's dataset.

3.4 Implementtation

- 3 Tesla M10 GPU
- Training time (excluding hyperparameter optimization) took approximately 10 hours.
- Prediction time took approximately 5 minutes.

4 Evaluation

For both parts of the challenge (Part 1 and Part 2), we followed the evaluation guidelines provided by the 2022 LitCoin NLP Challenge organizers. The evaluation metric was the average of the Jaccard similarity calculated for each document:

$$J(O, P) = \frac{|P \cap O|}{|P| + |O| - |P \cup O|} \tag{1}$$

In Part 1, P corresponded to the set of predicted mentions and O to the set of correct mentions in a given abstract. A match between two mentions occurred when they had the same type and similar offsets. Our pipeline achieved a score

[‖] urlhttps://www.ncbi.nlm.nih.gov/mesh/
[**] urlhttps://www.ebi.ac.uk/ols/ontologies/ncbitaxon

of 0.8423 in this part (calculated from 50.0% of the test data), which corresponded to 30% of the final score. The highest-scoring team achieved 0.9067.

In Part 2, P corresponded to the set of predicted relations and O to the set of correct relations in a given abstract. A relation is characterized by a pair of entities, its type (Association, Positive Correlation, Negative Correlation, Bind, Cotreatment, Comparison, and Drug Interaction) and its novelty (No, Novel).

For each correct relation in a given abstract, it is calculated the following intersection score with the predictions:

$$intersection\_score = 0.25 \times A + 0.5 \times B + 0.25 \times C \tag{2}$$

Where A, B, and C can either have a value of 1 or 0. If a relation in P includes the same pair of entities present in the correct relation, A has a value of 1. If the relation is also of the same type, B has a value of 1. If the relation also has the same novelty, C has a value of 1. This means that the intersection score for a correct relation and the predictions is a value between 0 and 1. The intersection between O and P in a given abstract is calculated by the averaged intersection scores for each correct relation in O.

Our pipeline achieved a score of 0.2124 in this part (calculated from 50.0% of the test data), which corresponded to 70% of the final score. The highest-scoring team achieved 0.6279.

5 Conclusion

This work presented the pipeline elaborated to participate in the 2022 LitCoin NLP Challenge by our team, Lasige-Unicage, highlighting a successful collaboration between the academia (LASIGE) and the industry (Unicage). Our biomedical NLP pipeline used data engineering pre-processing tools and two systems to perform NER and RE that could incorporate external knowledge. The NER system was explicitly designed to tackle the challenge, whereas, for RE, we used the BiOnt system [5] with minimal modifications. We were awarded the 7th Prize ($ 5000) in the LitCoin competition out of approximately 200 participating teams.

In the future, the goal is to improve the NER module by expanding and refining training datasets and exploring different classification layers. As for RE, we could link more external ontological data to boost performance. Unicage commands offered advantages, namely in the ease of use and the versatility of the commands, which allowed us to convert and merge the different corpora files to the BIO/IBO format efficiently, along with the post-processing of the results from the Relation Extraction model. We intend to explore how we can further integrate Unicage approaches in NLP tasks and pipelines, with a particular focus on data processing aspects.

Acknowledgments

This work has been supported by FCT through Deep Semantic Tagger (DeST) Project under Grant PTDC/CCIBIO/28685/2017 (`http://dest.rd.ciencias.ulisboa.pt/`), in part by LASIGE Research Unit under Grants UIDB/00408/2020 and UIDP/00408/2020, and in part by FCT and FSE through PhD Scholarship under Grant SFRH/BD/145221/2019 and PhD scholarship ref. 2020.05393.BD.

References

[1] Reihaneh H Hariri, Erik M Fredericks, and Kate M Bowers. Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big Data*, 6(1):1–16, 2019.

[2] João MP Moreira, Helena Galhardas, and Miguel L. Pardal. Leanbench: comparing software stacks for batch and query processing of IoT data. *Procedia computer science*, 130:448–455, 2018.

[3] Duarte M. Nascimento, Miguel Ferreira, and Miguel L. Pardal. Does big data require complex systems? a performance comparison between spark and unicage shell scripts, 2022.

[4] F. Couto. *Data and Text Processing for Health and Life Sciences*. Number 1137 in Advances in Experimental Medicine and Biology. Springer, 2019.

[5] Diana Sousa and Francisco M Couto. BiOnt: deep learning using multiple biomedical ontologies for relation extraction. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, pages 367–374. Springer, 2020.

[6] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1), oct 2021.

[7] Kirill Degtyarenko, Paula De Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36(suppl_1):D344–D350, 2007.

[8] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[9] Alfred V Aho, Brian W Kernighan, and Peter J Weinberger. *The AWK programming language*. Addison-Wesley Longman Publishing Co., Inc., 1987.

[10] Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016, 05 2016. baw068.

[11] Joël Legrand, Romain Gogdemir, Cédric Bousquet, Kevin Dalleau, Marie-Dominique Devignes, William Digan, Chia-Ju Lee, Ndeye-Coumba Ndiaye, Nadine Petitpain, Patrice Ringot, et al. Pgxcorpus, a manually annotated corpus for pharmacogenomics. *Scientific data*, 7(1):3, 2020.

[12] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10, 2014.

[13] Harsha Gurulingappa, Roman Klinger, Martin Hofmann-Apitius, and Juliane Fluck. An empirical evaluation of resources for the identification of diseases and adverse effects in biomedical literature. In *International Conference on Language Resources and Evaluation*, 2010.

[14] Sunil Mohan and Donghui Li. MedMentions: A large biomedical corpus annotated with umls concepts. *ArXiv*, abs/1902.09476, 2019.

[15] Paul Thompson, Sophia Daikou, Kenju Ueno, Riza Theresa Batista-Navarro, Junichi Tsujii, and Sophia Ananiadou. Annotation and detection of drug effects in text for pharmacovigilance. *Journal of Cheminformatics*, 10, 2018.

[16] Corinna Kolárik, Roman Klinger, C. Friedrich, Martin Hofmann-Apitius, and Juliane Fluck. Chemical names: Terminological resources and corpora annotation. In *International Conference on Language Resources and Evaluation*, 2008.

[17] Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A. Baumgartner, K. B. Cohen, Karin Verspoor, Judith A. Blake, and Lawrence E. Hunter. Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, 13, 7 2012.

[18] Sunil Kumar Sahu, Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. Inter-sentence relation extraction with document-level graph convolutional neural network. *ArXiv*, abs/1906.04684, 2019.

[19] Martin Gerner, G. Nenadic, and Casey M. Bergman. LINNAEUS: A species name identification system for biomedical literature. *BMC Bioinformatics*, 11:85 – 85, 2010.

[20] Evangelos Pafilis, Sune P. Frankild, Lucia Fanini, Sarah Faulwetter, Christina Pavloudi, Aikaterini Vasileiadou, Christos Arvanitidis, and Lars Juhl Jensen. The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text. *PLoS ONE*, 8, 2013.

[21] Mariana Neves, Alexander Damaschun, Andreas Kurtz, and Ulf Leser. Annotating and evaluating text for stem cell research. 2012.

[22] Lawrence H. Smith, Lorraine K. Tanabe, Rie Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, C. Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig A. Struble, Richard J. Povinelli, Andreas Vlachos, William A. Baumgartner, Lawrence E. Hunter, Bob Carpenter, Richard Tzong-Han Tsai, Hong-Jie Dai, Feng Bin Liu, Yifei Chen, Chengjie Sun, Sophia Katrenko, Pieter W. Adriaans, Christian Blaschke, Rafael Torres, Mariana L. Neves, Preslav Nakov, Anna Divoli, Manuel Maña-López, Jacinto Mata, and W. John Wilbur. UvA-DARE (digital academic repository) overview of biocreative II gene mention recognition. 2008.

[23] Nigel Collier and Jin-Dong Kim. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78, 2004.

[24] Udo Hahn, Katrin Tomanek, Elena Beisswanger, and Erik Faessler. A proposal for a configurable silver standard. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 235–242, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[25] Suwisa Kaewphan, Sofie Van Landeghem, Tomoko Ohta, Yves Van de Peer, Filip Ginter, and Sampo Pyysalo. Cell line name recognition in support of the identification of synthetic lethality in cancer from text. *Bioinformatics*, 32(2):276–282, 2016.

[26] Chih-Hsuan Wei, Bethany R. Harris, Hung-Yu Kao, and Zhiyong Lu. tmVar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics*, 29 11:1433–9, 2013.

[27] Behrouz Bokharaeian, Alberto Díaz, Nasrin Taghizadeh, Hamidreza Chitsaz, and Ramyar Chavoshinejad. SNPPhenA: a corpus for extracting ranked associations of single-nucleotide polymorphisms and phenotypes from literature. *Journal of Biomedical Semantics*, 8, 2017.

[28] Xin Rong. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*, 2014.

Appendix D

LasigeBioTM team at CLEF2020 ChEMU evaluation lab: Named Entity Recognition and Event extraction from chemical reactions described in patents using BioBERT NER and RE

LasigeBioTM team at CLEF2020 ChEMU evaluation lab: Named Entity Recognition and Event extraction from chemical reactions described in patents using BioBERT NER and RE

Pedro Ruas, Andre Lamurias, and Francisco M Couto

LASIGE, Faculdade de Ciências, Universidade de Lisboa, Lisbon 1749-016, Portugal
psruas@fc.ul.pt
alamurias@lasige.di.fc.ul.pt
fcouto@di.fc.ul.pt

Abstract. This manuscript describes the participation of the LasigeBioTM team in the NER and EE tasks of the ChEMU evaluation lab. We have fine-tuned the BioBERT NER model to locate and tag named entities and the BioBERT RE model to detect relations between trigger words and named entities. For the NER task, we obtained a F1-score of 0.9392 (exact matching) and 0.9630 (relaxed matching), which was an improvement over the baseline approach and achieving the 3rd best team result. For the EE task, we were not able to produce all the required annotation files due to the dimension of the test set and, consequently, we did not obtain results in time to submit to the competition. However, we obtained an accuracy of 0.9849 when we applied the BioBERT RE model on the development set.

1 Introduction

Chemical patents are a valuable source of information for chemical research. Every year, thousands of patents are registered, increasing the already large wealth of information available. Considering only the European Patent Office (EPO) and the year 2019, there were 7697 new patents filled in the "Pharmaceuticals" category and 6197 in the "Organic fine chemistry" category[1]. The manual analysis of these documents is costly, both in terms of time and effort, so it is necessary to develop text mining approaches to extract information in a more efficient way.

There are some challenges associated with patent text, like the presence of longer sentences, the use of specific terminology, the complexity of the syntactic

[1] http://documents.epo.org/projects/babylon/eponet.nsf/0/BC45C92E5C077B10C1258527004E95C0/

Category	Entity type
Chemical entity	"REACTION_PRODUCT"
	"STARTING_MATERIAL"
	"REAGENT_CATALYST"
	"SOLVENT"
	"OTHER_COMPOUND"
Reaction property	"TEMPERATURE"
	"TIME"
	"YIELD_PERCENT"
	"YIELD_OTHER"
Reaction label	"EXAMPLE_LABEL"

Table 1. Entity types and the respective broad category in which they are included.

structure, which limits the performance of general text mining models, usually developed for news text [3]. In addition, chemical patents typically contain a large number of chemical compounds with a complex structure, which originates ambiguity if, for example, we aim to link those entities to a reference knowledge base [1].

The ChEMU evaluation lab [7] proposed the chemical named entity recognition (NER) task and the chemical reaction event extraction (EE) task on a large corpus of chemical patents.

Language representation models, like Bidirectional Encoder Representations from Transformers (BERT) [2], have shown state-of-the-art performance across several natural language processing tasks, including Named Entity Recognition (NER) and Relation Extraction (RE). The goal of NER is to find in a given text the location of named entities and to classify them according to pre-defined types. Several works have described the application of BERT to the NER task [6, 9]. In turn, the goal of the RE task is the detection and classification of semantic relations between two given entities in a text, and BERT has also been applied to this task [8].

We describe here our approach to Task 1 (NER) and Task 2 (EE) of ChEMU, which consisted of the application of the BioBERT NER model for Task 1 and modelling Task 2 as a joint NER + RE task, in which we applied the BioBERT NER and the BioBERT RE models.

2 Methodology

2.1 Task 1 - NER

Data preparation The objective of the task was to recognise named entities related to chemical reactions in patents and to tag them according to ten different types:

We used a rule-based tokenizer proposed by one of the BioBERT authors[2] which uses regular expressions. We started by tokenizing the text of the docu-

[2] `https://github.com/dmis-lab/biobert/issues/107#issuecomment-615558492`

ments belonging to the released train (900 documents) and development (225 documents) sets. Each token was then tagged according to the IOB2 notation. The separator between each sentence was an empty line. For each set, there was a file that included all tokens, one per line, and the respective tags. These files were the input for fine-tuning the model we have used, which is detailed in the next section. When the test set was released, we applied the same process to the respective documents.

Model We used BioBERT [5], which consists of the BERT model pre-trained on several general corpora (English Wikipedia and BooksCorpus) and additionally in biomedical-specific corpora (PubMed abstracts and PMC full-text articles). The results reported by the authors show that BioBERT achieves better performance in the NER of biomedical entities in comparison with the classical implementation of BERT [5].

We fine-tuned the BioBERT NER model using the files corresponding to the competition datasets converted to the IOB2 notation, more concretely, 900 documents in the train set and 113 randomly chosen documents in the development set. We changed the training batch size from 32 to 24 to lower the memory requirements. To lower the required time for training, we used the pre-trained weights "BioBERT-Base v1.0 (+ PubMed 200K)", which corresponds to the smallest vocabulary available for BioBERT. The number of training epochs, the learning rate and the maximum sequence length were set to the default values, respectively, 10.0, 1^{-5} and 128. We did not have time to explore different values for the referred hyper-parameters, so we opted for the default values as it was the safest approach (with exception of the training batch size).

Then, we applied the fine-tuned model to recognise the entities and to predict the respective type in the remaining 112 documents of the development set that had not been previously used (inference mode). We used the same values for the hyper-parameters. We developed a module to process the BioBERT NER output and to generate the respective annotation files according to the BRAT standoff format[3]. We submitted the resulting annotation files to the competition page and obtained a F1-score of 0.9524 using the exact matching criterion and a F1-score of 0.9904 using the relaxed matching criterion. Given these results, we fine-tuned again the model, but this time using all documents belonging to the train (900) and the development (225) sets.

With the release of the test set, we applied the fine-tuned model to the converted file containing the tokenized text. We maintained the previous values used in inference mode for all hyper-parameters, with the exception of the maximum sequence length, which we changed from 128 to 384, due to the presence of longer sentences in the test set when comparing with the train and the development sets.

[3] `https://brat.nlplab.org/standoff.html`

2.2 Task 2 - Event extraction

Data preparation Our approach was to model the Task 2 as a joint NER and RE task. The goal was to detect trigger words and to recognise arguments involving trigger words and the entities described in Task 1. We followed a similar approach for the Task 1 and converted both the documents of the train and the development sets to the IOB2 format. But in this case, we only considered the annotations relative to event trigger words, i.e., words associated with individual steps in the context of the reaction. These event trigger words belonged to two additional entity types not present in Task 1: "REACTION_STEP" and "WORKUP". Like in the Task 1, for each set there was a file that included all tokens and the respective tags.

For the RE part of the task, there were two labels for a relation between an event trigger words and a chemical entity: "ARG1", to label a relation between a trigger word and a chemical entity, and "ARGM", to label a relation between a trigger word an adjunct entity, like temperature, yield or time. First, we performed sentence segmentation of the text present in the documents of the train and the development sets and, for each sentence containing at least a trigger word and an entity, we assumed that it could potentially contain a relation. For sentence segmentation, we used the same script referred in the Task 1, which consists of a rule-based model proposed by one of the BioBERT authors. In each sentence, the trigger word and the entity were replaced, respectively, by the tags "@TRIGGER$" and "@LABEL$". If the trigger word and the entity were effectively part of an argument in the gold standard annotations, the sentence would be assigned the label "1", otherwise the label would be "0". Besides, if in a given sentence, for example, there were present a trigger word and two different entities, the sentence would appear in two different lines of the final file, each one associated with a trigger word - entity pair. At the end, for each set we obtained a file containing a sentence per line, with the respective relation label and the tags @TRIGGER$ and @LABEL$ in the correct position.

Model For the NER step, we fine-tuned the BioBERT NER model using the files corresponding to the documents of the train (900) and development (225) sets converted to the IOB2 notation. We used the same hyper-parameter values for fine-tuning as described in Task 1. The documents of the test set were also converted into a single file according to the IOB2 notation and the approach was similar to that of Task 1.

For the RE step, we considered the files containing the sentences of the train in the format described above to fine-tune the BioBERT RE model. We used the pre-trained weights "BioBERT-Base v1.0 (+ PubMed 200K) and changed the batch size from 32 to 24. The number of training epochs, the learning rate and the maximum sequence length were set to the default values, respectively, 3.0, 2^{-5} and 128. We evaluated the fine-tuned model on the development set and obtained an accuracy of 0.9849. We fine-tuned again the model using all the sentences in the train and development sets. We applied the fine-tuned model to predict the relation labels in the converted documents of the test set only

	Exact evaluation			Relaxed evaluation		
Model	Precision	Recall	F1-score	Precision	Recall	F1-score
Baseline	0.9071	0.9071	0.8893	0.9219	0.8723	0.9053
LasigeBioTM	**0.9327**	**0.9457**	**0.9392**	**0.9590**	**0.9671**	**0.9630**

Table 2. Task 1 (NER) evaluation using the exact and the relaxed matching criteria considering all entity types. The performance of the baseline approach and our approach are shown in terms of Precision, Recall and F1-score. The highest values for each metric in the exact and relaxed evaluations are highlighted.

	Exact evaluation			Relaxed evaluation		
Entity type	Precision	Recall	F1-score	Precision	Recall	F1-score
EXAMPLE_LABEL	0.9577	0.9742	0.9659	0.9718	0.9885	0.9801
OTHER_COMPOUND	0.9529	0.9534	0.9531	0.9648	0.9658	0.9653
REACTION_PRODUCT	0.8387	0.8877	0.8625	0.9204	0.9498	0.9349
REAGENT_CATALYST	0.8887	0.8869	0.8878	0.9145	0.9091	0.9118
SOLVENT	0.9410	0.9696	0.9551	0.9433	0.9720	0.9574
STARTING_MATERIAL	0.8920	0.9058	0.8988	0.9460	0.9421	0.9440
TEMPERATURE	0.9674	0.9706	0.9690	0.9870	0.9934	0.9902
TIME	0.9846	0.9889	0.9868	0.9799	0.9955	0.9876
YIELD_OTHER	0.9732	0.9909	0.9820	0.9799	0.9955	0.9876
YIELD_PERCENT	**0.9897**	**0.9923**	**0.9910**	**0.9974**	**1.0000**	**0.9987**

Table 3. Task 1 (NER) evaluation using the exact and relaxed matching criteria according to each entity type. The performance of our approach is shown in terms of Precision, Recall and F1-score. The highest values for each metric in the exact and relaxed evaluations are highlighted.

changing the maximum sequence length from 128 to 384. At last, we developed a module to import the output of the RE model of BioBERT and to determine the type of the argument: "ARG1" if the relation was between a trigger word and a chemical entity, "ARGM" if the relation was between a trigger word and an adjunct entity.

3 Results

The evaluation results for Task 1 (NER) are available in Table 2 and the Table 3 shows the results for the task according to each entity type.

For Task 2 (EE) we did not obtain any results for the test set, but we obtained an accuracy of 0.9849 when we applied the BioBERT RE model on the dev set.

4 Discussion

4.1 Task 1 (NER)

Overall, the results obtained for this task, both using exact (F1-score of 0.9392) and relaxed matching (F1-score of 0.9630) criteria were positive. Comparing with the baseline approach, we obtained a higher F1-score, both considering the exact matching (+0.0499) and the relaxed matching (+0.0577) criteria. These results represent the 3rd best position in terms of team results and the 5th best position in the overall submission rank.

The good performance of the BioBERT model is due to the fact that it produces contextualised word representations that consider both the left and the right contexts of the words. The meaning of the words is usually related with the context where they appear, i.e., a given word can have two (or more) different meanings in different contexts. This is particularly relevant for chemical compounds, which can have different roles according to the reaction (i.e. the context) in which they participate. The BioBERT NER model obtained higher F1-score when recognising the entities of the type "YIELD_PERCENT", both using the exact (0.9910) and the relaxed matching (0.9987) criteria. This is related with the fact that this type of entities always included the character "%" in their surface form after a numerical value (for example, "53%"), which was an immutable pattern easily recognisable. In the other hand, the model obtained the lowest results when recognising the entities of the type "REACTION_PRODUCT" using the exact matching criteria (F1-score of 0.8625) and the entities of the type "REACTION_CATALYST" using the relaxed matching (F1-score of 0.9118). There was more ambiguity associated with these two types of entities (and also with the entities of the type "STARTING_MATERIAL") since they were related with chemical compounds, which can have different roles. This means that, for example, the chemical compound "4-nitropyridin-2-amine" can be the reaction product of a given reaction but in other reaction can participate as the reaction catalyst or as the starting chemical compound. The BioBERT NER model was able to correctly recognise almost all entities belonging to these types, however, the fact that it was originally pre-trained on scientific articles and general corpora and not on patents including chemical compounds, prevented an even higher performance.

4.2 Task 2 (EE)

The test set was too large (10000 documents) when comparing with the other sets, and the text of the documents was significantly larger too: the average number of characters present in the documents of the test set was 61002, whereas in the documents of the train and development sets was 835 and 772, respectively. This difference increased the required time to fine-tune and apply our model. Due to the limited time participants had to submit the results, we were only able to produce annotations for 1126 documents out of the necessary 10000 comprising the test set.

The entities and events to extract were located in the description of chemical reactions within the patents text. So the inclusion of a module able to detect the chemical reactions within the text would filter out irrelevant text and, consequently, would allow a faster application of our approach, which would be specially relevant for Task 2.

As it was previously referred, instead of using BioBERT, a model pre-trained directly over chemical patents text would possibly obtain higher performance in both tasks.

5 Conclusion

In Task 1 (NER), we obtained a F1-score of 0.9392 (exact matching) and 0.9630 (relaxed matching), which corresponds, respectively, to an increase of 0.0499 and 0.0577 over the baseline performance and to the 3rd best performing system. For Task 2 we were not able to produce results due to the lack of time to apply our approach over the entire test set. Consequently, our future work will focus mainly on the resolution of the problems associated with the poor performance in this task. First, we will explore other RE systems, like for example "BO-LSTM" [4]. Second, we will apply a module like Yoshikawa et al. [10] to extract the specific snippets describing chemical reactions within the patents text. The machine learning model (BiLSTM-CRF) proposed by the authors obtained a significantly higher performance in the extraction of chemical reactions comparing with simpler baseline approaches, including a rule-based model, which we expect will enhance our approach.

Acknowledgements

This project was supported by FCT through funding of the DeST: Deep SemanticTagger project, ref. PTDC/CCI-BIO/28685/2017, and the LASIGE ResearchUnit, ref. UIDB/00408/2020

References

1. Akhondi, S.A., Klenner, A.G., Tyrchan, C., Manchala, A.K., Boppana, K., Lowe, D., Zimmermann, M., Jagarlapudi, S.A., Sayle, R., Kors, J.A., Muresan, S.: Annotated chemical patent corpus: A gold standard for text mining. PLoS ONE **9**(9), 1–8 (2014). https://doi.org/10.1371/journal.pone.0107477
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (oct 2018), `http://arxiv.org/abs/1810.04805`
3. Hu, M., Cinciruk, D., Walsh, J.M.: Improving Automated Patent Claim Parsing: Dataset, System, and Experiments (2016), `http://arxiv.org/abs/1605.01744`
4. Lamurias, A., Sousa, D., Clarke, L.A., Couto, F.M.: BO-LSTM: Classifying relations via long short-term memory networks along biomedical ontologies. BMC Bioinformatics **20**(10), 1–12 (2019). https://doi.org/10.1186/s12859-018-2584-5

5. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020). https://doi.org/10.1093/bioinformatics/btz682
6. Moon, T., Awasthy, P., Ni, J., Florian, R.: Towards Lingua Franca Named Entity Recognition with BERT. Tech. rep. (2019)
7. Nguyen, D.Q., Zhai, Z., Yoshikawa, H., Fang, B., Druckenbrodt, C., Thorne, C., Hoessel, R., Akhondi, S.A., Cohn, T., Baldwin, T., Verspoor, K.: ChEMU: Named entity recognition and event extraction of chemical reactions from patents. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **12036 LNCS**, 572–579 (2020). https://doi.org/10.1007/978-3-030-45442-5_74
8. Papanikolaou, Y., Roberts, I., Pierleoni, A.: Deep Bidirectional Transformers for Relation Extraction without Supervision pp. 67–75 (2019). https://doi.org/10.18653/v1/d19-6108
9. Peng, Y., Yan, S., Lu, Z.: Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets (2019). https://doi.org/10.18653/v1/w19-5006, `https://www.mendeley.com/catalogue/2347f426-b409-3772-9174-688480ed2a76/`
10. Yoshikawa, H., Nguyen, D.Q., Zhai, Z., Druckenbrodt, C., Thorne, C., Akhondi, S.A., Baldwin, T., Verspoor, K.: Detecting Chemical Reactions in Patents. Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association (3), 100–110 (2019), `https://www.aclweb.org/anthology/U19-1014`

Appendix E

LasigeBioTM at CANTEMIST: Named Entity Recognition and Normalization of Tumour Morphology Entities and Clinical Coding of Spanish Health-related Documents

LasigeBioTM at CANTEMIST: Named Entity Recognition and Normalization of Tumour Morphology Entities and Clinical Coding of Spanish Health-related Documents

Pedro Ruas[a], Andre Neves[a], Vitor D.T. Andrade[a] and Francisco M. Couto[a]

[a]*LASIGE, Faculdade de Ciências da Universidade de Lisboa, Portugal*

Abstract

CANTEMIST included three subtasks for the automatic assignment of codes related with tumour morphology entities to Spanish health-related documents: CANTEMIST-NER, CANTEMIST-NORM and CANTEMIST-CODING. For CANTEMIST-NER, we trained Spanish biomedical Flair embeddings on PubMed abstracts and then trained a BiLSTM+CRF Named Entity Recognition tagger on the CANTEMIST corpus using the trained embeddings. For CANTEMIST-NORM, we adapted a graph-based model that uses the Personalized PageRank algorithm to rank the eCIE-O-3.1 candidates for each entity mention. As for CANTEMIST-CODING, we adapted X-Transformer, a state-of-the-art deep learning Extreme Multi-Label Classification algorithm, to the multilingual and biomedical panorama to classify the clinical cases with a ranked list of eCIE-O-3.1 terms. The results obtained were a F1-score of 0.749 and 0.069 for the CANTEMIST-NER and the CANTEMIST-NORM subtasks, respectively, and our best scoring submission achieved a MAP score of 0.506 in the CANTEMIST-CODING subtask.

Keywords

CANTEMIST, Named Entity Recognition, Normalization, Coding, Text Mining, Natural Language Processing, Clinical Text, Extreme Multi-Label Classification

1. Introduction

There are several benefits arising from the application of Natural Language Processing (NLP)/Text Mining approaches to clinical text, like for example, the improvement of the decision-making process in clinical context. The use of electronic health records is associated with less doctor-patient interaction [1], so tools that are able to automatically extract relevant information from clinical notes can free up the doctors to contact directly with patients. Besides, these tools have the potential to improve biomedical [2, 3] and pharmaceutical research [4] and to democratise the access to clinical information for the layman user [5].

In the present work, we describe the participation of the LasigeBioTM team in CANTEMIST ("CANcer TExt Mining Shared Task – tumor named entity recognition") competition [6], which

CANTEMIST 2020

EMAIL: psruas@fc.ul.pt (P. Ruas); aneves@lasige.di.fc.ul.pt (A. Neves); fc49005@alunos.fc.ul.pt (V.D.T. Andrade); fcouto@di.fc.ul.pt (F.M. Couto)

ORCID: 0000-0002-1293-4199 (P. Ruas); 0000-0003-0627-1496 (F.M. Couto)

included a corpus of Spanish health-related documents and three different subtasks with the following goals:

- CANTEMIST-NER: Automatically recognise and locate tumour morphology mentions.
- CANTEMIST-NORM: Returning and normalising all tumour morphology mentions along with their respective codes from the eCIE-O-3.1 ("Clasificación Internacional de Enfermedades para Oncología - 3ª edición, 1ª revisión"[1]) terminology.
- CANTEMIST-CODING: Classification of clinical cases by returning a list of ranked eCIE-O-3.1 codes for each document.

For the CANTEMIST-NER substask we used the Flair framework [7] to train new Flair embeddings over Spanish translated PubMed abstracts and to train a NER tagger with BiLSTM+CRF architecture on the CANTEMIST Corpus leveraging the trained embeddings. For the CANTEMIST-NORM substask, we used the PPR-SSM model [8] to normalise the entities recognised by the NER tagger. This model builds a disambiguation graph for each document, where the nodes are the retrieved candidate codes from the eCIE-O-3.1 terminology for the present entities and the relations are based on the hierarchy of eCIE-O-3.1 terminology. A variation of this model additionally retrieves candidates from other terminologies, like CIE-10-ES and DeCS, and extracts relations between concepts from these terminologies and the codes from the eCIE-O-3.1 terminology to improve the edge structure in the graph. The Personalised PageRank algorithm (PPR) assign weights to each candidate and the highest scored one is the selected code for the respective entity. As for the CANTEMIST-CODING subtask, we adapted and built a pipeline using X-Transformer, a deep learning Extreme Multi-Label Classification (XMLC) algorithm, to the multilingual biomedical panorama, so that it could successfully process and classify each clinical case with the eCIE-O-3.1 terms more related with each document.

1.1. Related work

In the Named Entity Recognition (NER) task, state-of-the-art approaches usually have a BiLSTM-CRF architecture, which was initially proposed by Huang et al. [9]. LSTM (Long-Short Term Memory) networks are Recurrent Neural Networks (RNN), which means that these networks have a recurrent layer connecting different features at different time frames. In fact, BiLSTM networks can leverage the past features and the future features for a given time frame. An input layer represents a given set of features, in this case text tokens, at a given time and the output layer represents the probability distribution for each label at that time. The CRF (Conditional Random Fields) models, in turn, focus on sentence level tag information. More recently, pre-trained language models, like BERT [10] or ELMo [11], have been fine-tuned to the NER task. BERT has a multilayer bidirectional transformer encoder architecture and, contrarily to RNNs, employs an attention mechanism to establish the dependency between the input features and the output. The original BERT implementation has been trained in general corpora, such as the BookCorpus and the English Wikipedia, but since then a plethora of domain-specific versions have been proposed, including BioBERT [12] and ClinicalBERT [13].

[1] https://eciemaps.mscbs.gob.es/ecieMaps/browser/index_o_3.html

The state-of-the-art approaches in Entity Normalization (also called Disambiguation or Linking) include graph-based models [14, 15], neural networks-based models [16, 17] and, similarly to what happens for the NER task, more recently the fine-tuned pre-trained language models [18, 19]. The graph-based models usually focus on building a graph containing the candidates for the entity mentions and then on ranking the candidates according to the relevance or coherence of each candidate in the graph. These are global models, since the disambiguation decision of a given entity mention is dependant of the other disambiguations in the same graph, but there is also sometimes a module responsible for the determination of the local similarity between candidates and mentions or the candidate retrieval. The neural network-based models usually take into account both the global coherence of the candidates and the local similarity, however, the candidates and the entities are typically represented by word embeddings, which are then integrated in the neural network. On the other hand, BERT-based models like the one proposed by Ji et al. [18] focus on the generation of contextualised word embeddings for candidates and entities and then on candidate ranking, which is considered a sequence-pair classification task. In this case, the model performs disambiguation of each entity independently based on word representations and other local features.

As to XMLC, several machine learning solutions have been developed in the last decade [20], but only more recently there have been deep learning solutions applied to XMLC. One of the first attempts to was the XML-CNN [21], a convolutional neural network that was adapted from a state-of-the-art approach to a multi-class classification task [22], with some changes on the neural network layers that allowed it to capture features more precisely from different regions of text. There was also HAXMLNet [23] which used a BiLSTM RNN with a multi-label attention layer to capture the most relevant parts of the text, along with a hierarchical clustering algorithm to divide labels through clusters, which proved efficient on larger datasets. Lastly, there is X-Transformer [24] the first deep learning approach to scale pre-trained Transformer models, such as BERT [10], RoBERTa[25] or XLNet[26] to XMLC. The algorithm uses a three-stage framework that firstly, semantically indexes all the possible labels in clusters using ELMo [11]. Then, using a deep learning Transformer model, it indexes each text instance to the most relevant cluster and, finally, ranks the labels retrieved from the previous cluster indices. X-Transformer surpassed other state-of-the-art methods in XMLC in four benchmark datasets and it was also applied to a query recommendation dataset from Amazon, where it showed improvements of more than 10% over Parabel [27], one of the most commonly used and competitive XMLC algorithms.

Our team has already made an adaptation of X-Transformer for the biomedical panorama in the BioASQ MESINESP competition that occurred earlier this year. In MESINESP, the goal was indexing a large dataset of biomedical articles written in Spanish using DeCS terms. In the final scoreboard, our approach using X-Transformer has achieved high scores in the precision measures, surpassing most competing systems in those measures.

2. Methodology

2.1. CANTEMIST-NER

The goal of this task was to recognise and locate tumour morphology entities in Spanish health-related documents. We used the Flair framework [7] to develop a Spanish biomedical NER tagger.

2.1.1. Training of Flair embeddings

We trained new Flair contextualised embeddings [28] in Spanish biomedical text, more concretely, in translated abstracts of PubMed articles available at https://temu.bsc.es/mesinesp/index.php/download/translated-pubmed-articles/. We considered 4 subsets of articles, each one with 80%/10%/10% of the articles included in the train, validation and test files, respectively:

1. 32,500 articles, 40,987,614 tokens
2. 32,500 articles, 35,352,727 tokens
3. 32,500 articles, 39,021,229 tokens
4. 32,500 articles, 40,005,075 tokens

The 4 splits contained a total of 130,000 articles and 155,366,645 tokens, of which 143,387,385 corresponded to training tokens.

We generated a language model for each split with hidden_size = 1024, nlayers = 1, dropout = 0.1, using the following training parameters:

- sequence_lenght = 250
- mini_batch_size = 100
- max_epochs = 2000
- patience = 25

We trained forward and backward embeddings in each split using a single NVIDIA Tesla P4 GPU, since Flair does not have multi-GPU support in the current version, and interrupted the training after a variable number of epochs that ranged from 71 in the backward embeddings training in split 1 and 99 in the backward embeddings training in split 2.

2.1.2. Pre-processing of the CANTEMIST corpus

We converted the train, development 1 (dev1) and development 2 (dev2) sets of the CANTEMIST corpus[2] to the IOB2 format[3] using the Flair sentence segmenter jointly with the Flair tokenizer. Each token was tagged with the label “B-MOR_NEO" if corresponded to the beginning of an annotation, the label “I-MOR_NEO" if corresponded to the inside of an annotation, and the label “O" if it was outside of any annotation. The content present in the train, dev1 and dev2 sets originated, respectively, the files “train.txt", “dev.txt" and “test.txt". The corpus was then loaded into a Flair “ColumnCorpus" object to allow the further training of the NER tagger.

[2] https://temu.bsc.es/cantemist/?p=4338

[3] https://en.wikipedia.org/wiki/Inside%E2%80%93outside%E2%80%93beginning_(tagging)

2.1.3. Training of the Spanish biomedical NER tagger

The following models were considered:

1. “base": uses Flair embeddings (es-forward and es-backward) trained on Spanish Wikipedia + Spanish FastText embeddings.
2. “medium": uses Flair embeddings (es-forward and es-backward) trained on Spanish Wikipedia + Spanish FastText embeddings + PubMed Flair embeddings trained in 1 PubMed split.
3. “large": uses Flair embeddings (es-forward and es-backward) trained on Spanish Wikipedia + Spanish FastText embeddings + PubMed Flair embeddings trained in 2 PubMed splits.
4. “pubmed": uses PubMed Flair embeddings trained in 4 PubMed splits.

We considered the default architecture for the sequence tagger: BiLSTM with a CRF decoding layer, hidden_size = 256. The training parameters were set to:

- learning_rate = 0.1
- mini_batch_size = 32
- max_epochs = 55
- patience = 3

Due to the lack of time, we only selected the “medium" model for training, as we considered it was the safest approach: to leverage trained biomedical embeddings jointly with general available embeddings. After the training, we applied the model to predict the labels in the 5232 documents belonging to the background + test set of the CANTEMIST corpus and to create an annotation file in the BRAT format for each text document.

2.2. CANTEMIST-NORM

The goal of this task was to perform NER and the normalization or disambiguation of the recognised entities to the eCIE-O-3.1 terminology. We applied the model previously developed for the NER task to recognise the entities and adapted the PPR-SSM model [8] to assign the entities a eCIE-O-3.1 code.

2.2.1. Pre-processing of the NER output

In each document, the first step was to apply the NER tagger to generate the NER output files and then to retrieve the ten best eCIE-O-3.1 candidates for each recognised entity through string matching, more concretely, according to the edit distance. The model then built a disambiguation graph with the eCIE-O-3.1 candidates for all present entities in the document. Two candidates/nodes were considered linked in the graph if they were linked in the eCIE-O-3.1 hierarchy. For each candidate was calculated the extrinsic information content (IC), which is a measure of rareness: the IC of a given entity is high if that entity has few entries in an external dataset [29]. In this case, we considered the external dataset the train, dev1 and dev2 sets of the CANTEMIST corpus.

2.2.2. Entity disambiguation

The model applied the Personalized PageRank (PPR) algorithm over each disambiguation graph. PPR is a variation of PageRank [30], which was originally proposed as an algorithm to rank the relative importance of web pages. It considers the web a graph where each node is a page and has links to other pages (forward links) and links from other pages (backlinks). The PageRank algorithm simulates the behaviour of a "random surfer" in the web: from a given page the surfer can either follow one of the forward links in that page or jump to a random page belonging to the graph. In the personalised variation, this jump is not random and instead always occur to a chosen page. After successive iterations, the algorithm returns the probability distribution of reaching each node in the graph. Nodes containing more links will be reached more times, so they will have more relevance in the context of the graph. PPR have also been applied in the normalization of entities, but in this case the web graph is replaced by the disambiguation graph containing the candidates for all entities in a given document. PPR traverses the graph and then assigns weights to each candidate according to its coherence or relevance to the graph: more connected nodes will have higher weight. Additionally, in our model, the IC of each candidate was also considered in the candidate ranking. The model selected the highest scored candidate for each entity and added the eCIE-O-3.1 codes to the annotation files outputted by the NER tagger.

2.2.3. Multiple terminologies

We also explored the use of more than one terminology in the candidate retrieval and graph building. We considered the CIE-10-ES ("Classificación Internacional de Enfermedades 10.ª Revisión, Modificación Clínica"[4]) and the Spanish DeCS ("Descriptores en Ciencias de la Salud")[5] terminologies. For each entity, besides the ten best eCIE-O-3.1 candidates, we also retrieved the five best CIE-10-ES and the five best DeCS candidates through string matching and built the disambiguation graph accordingly. We considered that a eCIE-O-3.1 candidate and a CIE-10-ES or a DeCS candidate were linked if they were present in the same sentence of any document belonging to the CANTEMIST corpus. We applied the python implementation[6] of MER [31] for the fast recognition of the entities in the sentences. With these additional candidates, the disambiguation graph was denser and contained more semantic information, which we expected that would improve the precision of the disambiguation process. After the application of the PPR algorithm, the model only selected eCIE-O-3.1 candidates to normalise the mentions.

2.2.4. Models

The following models were considered:

1. "single-ont": this model only used candidates retrieved from the eCIE-O-3.1 terminology.
2. "multi-ont": besides eCIE-O-3.1, this model additionally retrieved candidates from CIE-10-ES and DeCS terminologies to improve the disambiguation graph.

[4] https://eciemaps.mscbs.gob.es/ecieMaps/browser/index_10_mc.html
[5] http://decs.bvs.br/E/homepagee.htm
[6] https://pypi.org/project/merpy/

2.3. CANTEMIST-CODING

The goal of this task was the classification of clinical cases in Spanish by returning a list of ranked eCIE-O codes related to the content of each clinical case. To tackle this challenge, we decided to use X-Transformer, a state-of-the-art deep learning XMLC solution and apply it to multilingual biomedical panorama.

2.3.1. X-Transformer modifications

Some modifications in the algorithm code were required. The first one was made in the vectorization of the labels of the training and test sets. We have chosen to use all possible labels, including the labels that were not present in the train or test sets. This change was needed since the algorithm would fail to work correctly if the number of labels between sets did not match.

Another modification was the inclusion of BETO [32][7] in the choices of models to train X-Transformer, since we considered that using a Transformer model specifically designed for the Spanish language could lead to improved results over the Multilingual version of BERT. Finally, we have also adapted the algorithm so that it could process input data containing diacritical marks, such as accents, that are common in the Spanish language.

2.3.2. Pipeline

A pipeline was developed for this task as it can be seen in Figure 1. After retrieving the data from the competition organisers, the first step of this pipeline was merging each separate text file that composed the training and development datasets into a single file for each dataset, so that in the end we could have only two files, one for train and another for test. Then, using the '.ann' files given for the other two tasks of the CANTEMIST competition and that were associated with each clinical case, we extracted all labels that were attributed to each document and appended them to the beginning of the corresponding clinical case separated by a tab character ('\t'). This way, X-Transformer could distinguish between the labels and the text. The text was then stemmed using a Snowball stemmer[8].

The next step was creating a vocabulary file containing all labels used in the datasets, which was required as input by X-Transformer. Each line of this file had an eCIE-O code and its corresponding internal identifier, which corresponds to a number from 0 to N, where N is the total number of eCIE-O codes minus 1. This internal identifier is a label standardization method that allows X-Transformer to classify the text using labels from any kind or domain. For the creation of this file, we used a file containing a list of valid eCIE-O codes that was provided by the competition in their evaluation scripts folder, from which we retrieved all codes and corresponding descriptions. Then, we included the eCIE-O codes descriptions that were present in the '.ann' files and that were not present in the list of codes retrieved, adding them to existing descriptions if they had differences. Then, the codes without any description were removed, since they would be of no use to classify the clinical cases using X-Transformer. In the end, our vocabulary file consisted of a total of 4360 eCIE-O codes.

[7] https://github.com/dccuchile/beto

[8] https://www.nltk.org/_modules/nltk/stem/snowball.html

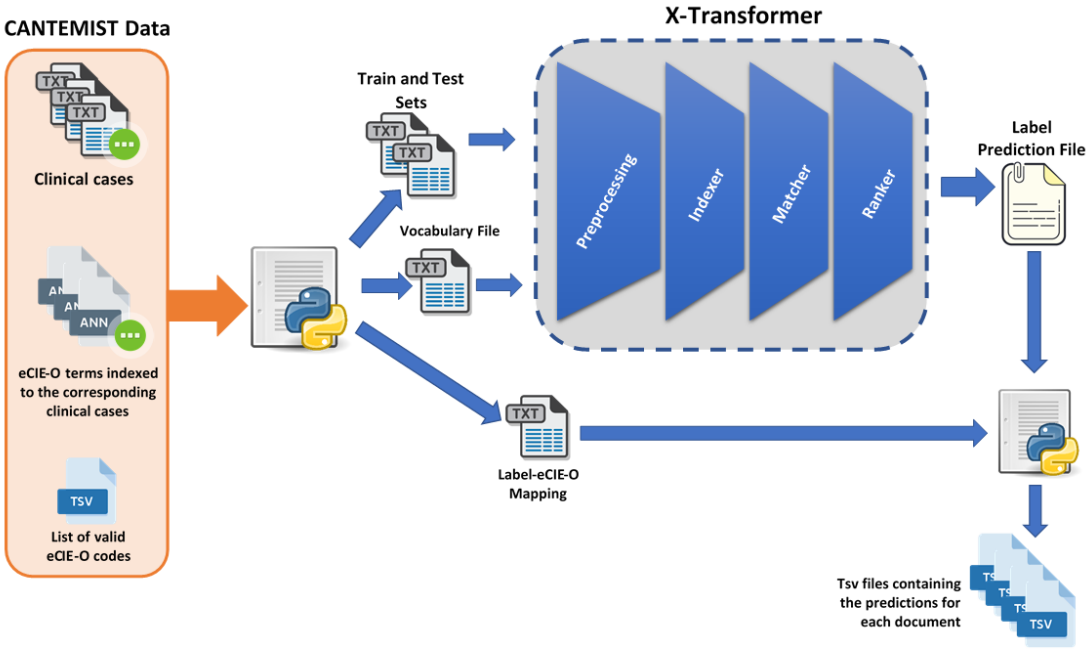


Figure 1: Developed pipeline for CANTEMIST-CODING task. It processes the CANTEMIST data before running it through X-Transformer. In the end, the predictions are converted to the format required by the competition.

In addition, we have also created a label mapping file that contained the correspondence between the eCIE-O code and its numeric identifier in the vocabulary file. For example, the term 'Células tumorales benignas', which has the corresponding eCIE-O code '8001/0', is the eleventh element in the vocabulary file, thus it's numeric identifier will be '10'. This label mapping file will later be used to map the predictions from their numeric identifiers to the corresponding eCIE-O codes required for the task.

The results of X-Transformer are given in the form of sparse matrices, with a number of rows equal to the number of clinical cases that compose the test set, and the number of columns corresponding to the possible labels. The prediction for each clinical case was retrieved, comprising a top K of most relevant labels and their confidence values ranging from -0.99 to 1. We used K=20 labels per clinical case. Then, using a Python script, we converted the predicted labels from their numeric identifiers to their corresponding eCIE-O codes using the label mapping file previously created. The script also discarded each label with a confidence score under a threshold chosen to achieve the highest Precision, Recall or F1 scores. For each of these measures, a '.tsv' file was created with the predictions for each clinical case in the format required by the competition. A fourth '.tsv' file was also created for the score threshold equal to 0, which was used as a baseline score. In the end, the files were used as input for the evaluation script given by the CANTEMIST competition so that we could retrieve the Mean Average Precision (MAP) scores for the predictions of our models.

In the test and background sets given by competition organizers, there were no eCIE-O codes

indexing the clinical cases, so we had to put a placeholder label on each document, because X-Transformer was not prepared to run on unlabelled data. We also had to artificially adapt the size of the given test and background sets by splitting the sets into a total of 48 smaller sets of 250 clinical cases each. This procedure was necessary so that the files could have the same size as the test sets used on the trained X-Transformer models, which had a total of 250 articles. The first 109 lines of each of those files was composed by 109 clinical cases from the test and background sets to classify, and the remaining 141 came from the dev1 set, which was already classified with eCIE-O codes. These 141 clinical cases were used as an additional validation set to define the confidence threshold values of our submissions.

2.3.3. Developed Models

In a first iteration, we trained 4 models using the 501 indexed clinical cases that composed the train set, and the dev2 set that comprised a total of 250 indexed clinical cases, was used as test set. One of our models was trained using BERT Base Multilingual Cased and another was trained using BETO, the Spanish version of BERT.

The other two models were trained with two X-Transformer models that were previously developed by us for the Spanish biomedical domain using biomedical articles in Spanish retrieved from the IBECS, LILACS and PubMed databases, along with a list of keywords identified for each article using MER[31], a NER software. The major difference between the two models was that one of them was developed using 318,658 articles, while the other one used 50% more data, with a total of 637,316 articles. We shall call this two models Spanish Biomedical X-Transformer and Spanish Biomedical X-Transformer large, correspondingly. Summarizing, the 4 models had the following characteristics:

- Model 1: BERT base Multilingual Cased finetuned with the clinical records.
- Model 2: BETO finetuned with the clinical records.
- Model 3: Spanish Biomedical X-Transformer finetuned with the clinical records.
- Model 4: Spanish Biomedical X-Transformer large finetuned with the clinical records.

In a second iteration, we trained 4 additional models following the same characteristics of the previous ones, but using a larger train set composed by the 501 clinical cases that composed the CANTEMIST-CODING train set, plus 249 additional clinical cases from the dev1 set that were already classified. This way, we expected to achieved better results, since the models were trained with additional clinical cases. Summarizing, these four models had the following characteristics:

- Model 5: BERT base Multilingual Cased finetuned with 750 clinical records.
- Model 6: BETO finetuned with 750 clinical records.
- Model 7: Spanish Biomedical X-Transformer finetuned with 750 clinical records.
- Model 8: Spanish Biomedical X-Transformer large finetuned with 750 clinical records.

All models were trained using the default parameters of X-Transformer, except for the eval and train batch sizes which were both changed from their original values of 64 and 32 to 4 due to hardware constrains. We have also set the number of gradient accumulation steps to 2 to compensate for the small batch size. Each model was trained for 12 epochs, on a single NVIDIA Tesla P4 GPU.

2.3.4. Preliminary Results

In order to choose which model predictions to submit, we decided to evaluate the predictions made by each model on the dev2 set using the evaluation script given by the competition organizers. As was explained before, each model had 4 '.tsv' files as output, with each file containing the predictions with a confidence score superior to the confidence score threshold defined to best precision, recall or F1-score, and the baseline score, which corresponds to the confidence score threshold set to 0, which was the middle of the X-Transformer confidence score scale. Then, each file was used as input for the evaluation script given by the competition organizers.

The results for each model can be seen in Table 1. As it can be seen, the highest MAP scores are achieved when the predictions are focused on achieving the highest recall scores especially when using the models trained with the Spanish Biomedical X-Transformer models. We can notice that the models that use BETO seem to achieve higher scores when compared with the ones that used BERT Multilingual. In addition, we notice that there is not a clear difference between the usage of more clinical cases to train the models, with some models achieving slightly higher scores in MAP if the evaluation was focused on precision or in the F1-score, while in other models the score was inferior when compared with the models that used lesser articles.

Taking this into consideration, we decided to choose the predictions focused on recall of 5 distinct models to submit for the CANTEMIST-CODING task so we could also compare their performance in the competition. The chosen models were Models 2, 3, 4, 5 and 7. We then retrieved the predictions made by the models for each of the 48 text files that contained the test and background sets data. Then, for each of the resulting prediction files, the first 109 lines corresponding to the predictions made for the test and background sets were stored in the '.tsv' files, while the other 141 lines which corresponded to the predictions of the labelled cases from the first development set, were used to find the confidence score threshold that achieved the best recall score, and that would be used to choose which predictions would be stored in the final '.tsv' file.

3. Results and discussion

The results for the CANTEMIS-NER subtask are available in Table 2.

Our model obtained a F1-score of 0.749 in this subtask. Some errors prevented a higher performance, such as those related with the span of some detected entities. For example, in the document "cc_onco94.ann", the mention "linfoma" is correctly recognised by the NER tagger, but the processing script attributed the span "1690 1697", whereas the correct span would be "1691 1698". Besides, in some cases the NER tagger only recognised incomplete entity mentions.

Model	Focus	MAP	Threshold
Model 1 BERT Base Multilingual Cased	Baseline	0.222	0
	F1	0.366	-0,66
	Precision	0.18	0,22
	Recall	0.384	-0,81
Model 2 BETO	Baseline	0.267	0
	F1	0.385	-0,48
	Precision	0.188	0,48
	Recall	0.438	-0,83
Model 3 Spanish Biomedical X-Transformer	Baseline	0.293	0
	F1	0.378	-0,34
	Precision	0.191	0,52
	Recall	**0.446**	-0,82
Model 4 Spanish Biomedical X-Transformer large	Baseline	0.281	0
	F1	0.396	-0,45
	Precision	0.18	0,6
	Recall	**0.448**	-0,82
Model 5 BERT Base Multilingual Cased (750 CC)	Baseline	0.203	0
	F1	0.372	-0,46
	Precision	0.186	0,12
	Recall	0.407	-0,83
Model 6 BETO (750 CC)	Baseline	0.224	0
	F1	0.371	-0,46
	Precision	0.18	0,39
	Recall	0.417	-0,82
Model 7 Spanish Biomedical X-Transformer (750 CC)	Baseline	0.25	0
	F1	0.392	-0,4
	Precision	0.2	0,32
	Recall	**0.442**	-0,82
Model 8 Spanish Biomedical X-Transformer large (750 CC)	Baseline	0.268	0
	F1	0.379	-0,43
	Precision	0.194	0,29
	Recall	0.427	-0,81

Table 1
Preliminary results of the trained models for the CODING task using the second development set.
Bold values correspond to the three highest values achieved in the Mean Average Precision (MAP) measure using the evaluation script given by the competition organizers.
Green lines correspond to the models and corresponding evaluation focus whose predictions were chosen to submit for the CANTEMIST-CODING task.

For example, in the document “cc_onco89.ann", the NER tagger recognised the entity “implantes mediastínicos", whereas the full entity mention would be “implantes mediastínicos pleurales".

For future work, it would be interesting to train the other models beside the “medium" (“base", “large" and “pubmed") in the CANTEMIST corpus and to apply them to the test set to verify if they obtain a higher performance. Besides, we only trained the Flair embeddings during less than 100 epochs, so with more epochs, probably the performance of the NER tagger would be higher. The NER tagger itself could also be trained for more epochs, up until 150, according to a

Model	P	R	F1
medium	0.787	0.714	0.749

Table 2
Results for the CANTEMIST-NER subtask. P, R and F1 refer, respectively, to Precision, Recall and F1-score.

Model	P	R	F1	P-No-M	R-No-M	F1-No-M
1.single-ont	0.063	0.057	0.060	**0.059**	**0.082**	**0.069**
2.multi-ont	**0.064**	**0.058**	**0.061**	**0.059**	0.080	0.068

Table 3
Results for the CANTEMIST-NORM subtask. P, R and F1 refer, respectively, to Precision, Recall and F1-score, and P-No-M, R-No-M and F1-No-M refer, respectively, to the Precision, Recall and F1-score calculated without considering the mentions to metastasis (8000/6 code). Bold values correspond to the highest achieved scores in our submissions

suggestion by the authors of Flair. We will also address the errors associated with the span of the recognised entities.

The results for the CANTEMIS-NORM subtask are available in Table 3.

The model “multi-ont" obtained a F1-score of 0.061, which represents a slight improvement of +0.001 comparing to the “single-ont" model. Still, the performance of the model was too low, which is mainly related with the candidate retrieval step. Since we used string matching for candidate retrieval and selected the top candidates according to the edit distance between candidate/entity mention, most of the candidates lists did not contain the correct codes for the respective entity mentions. For example, the correct code for the entity mention “cáncer" would be the eCIE-O-3.1 code 8000/6, corresponding to the concept “Neoplasia metastásica". However, neither the “single-ont" model nor the “multi-ont" model were able to retrieve this candidate code for the entity mention, because the edit distance between “cancer" and “Neoplasia metastásica" is too high. Besides, the lack of a synonyms list in the eCIE-O-3.1 terminology further exacerbates the problem. Another aspect that is worth mentioning is the fact that the performance of the normalization model is always dependant on the performance of the NER tagger, so an incorrect output returned by the latter will hinder the results outputted by the former.

In order to improve the normalization model we will explore alternative methods for candidate generation, like for example, the use of word embeddings, both for entity mentions and for the terminology concepts. Instead of just considering that two entities in the same sentence are related, we will also use a proper relation extraction tool to get more semantically meaningful relations between concepts, either belonging to the same terminology, or belonging to different terminologies (for example, a relation between a eCIE-O-3 and a DeCS concept), which will densify the disambiguation graph and improve the disambiguation precision.

The results for the CANTEMIS-CODING subtask are available in Table 4.

Looking at our results, we can observe that, contrarily to what we expected, the best scoring model was Model 5 which, in our preliminary evaluation, had achieved the lowest MAP score of the five models submitted for this task. The lowest MAP scores were achieved by Models 3, 4 and 7, which were trained using the Spanish Biomedical X-Transformer models and that

Model	MAP	P	R	F1	MAP-No-M	P-No-M	R-No-M	F1-No-M
Model 2	0.463	0.157	0.549	0.244	0.350	0.119	0.466	0.189
Model 3	0.449	0.159	0.517	0.243	0.333	0.118	0.427	0.184
Model 4	0.455	0.151	0.532	0.235	0.344	0.113	0.445	0.180
Model 5	**0.506**	**0.211**	**0.601**	**0.312**	**0.399**	**0.167**	**0.527**	**0.254**
Model 7	0.459	0.197	0.541	0.289	0.346	0.151	0.456	0.226

Table 4
Results for the CANTEMIST-CODING subtask. MAP, P, R and F1 refer, respectively, to Mean Average Precision, Precision, Recall and F1-score, and MAP-No-M, P-No M, R-No-M and F1-No-M refer, respectively, to the Mean Average Precision, Precision, Recall and F1-score calculated without considering the mentions to metastasis (8000/6 code).
Bold values correspond to the highest scores for each metric in our submissions

had achieved the highest MAP scores in the preliminary evaluation. We can also notice that the precision and F1 scores were lower than the recall scores, but that was expected, since our submissions contained the predictions that used a confidence score threshold focused on achieving the highest recall scores.

As a proposal for a future work, we could try to develop additional models with a Transformer architecture, like the Biomedical X-Transformer models, and use them to train other X-Transformer models expecting to improve the number of correctly identified eCIE-O codes. These models could be trained using scientific biomedical articles in Spanish or even the clinical cases given by the CANTEMIST competition.

Another possible solution could pass by preprocessing the clinical case files before running them through X-Transformer by reducing the amount of text from each clinical case. This could be achieved by using automatic text summarization tools to leave only the essential information about each clinical case. Then, using NER tools, we could retrieve key entities and/or related terms and include them in the summarized text. This way, by reducing the original clinical case to a smaller and more objective text, along with identified key terms and entities given by the NER tools, it is expected that the X-Transformer model will be able to achieve better results since it has a smaller and more concise string of words, than a larger amount of text with the key topics more diluted.

4. Conclusion

We obtained a F1-score of 0.749 and 0.069 for the CANTEMIST-NER and the CANTEMIST-NORM subtasks, respectively, and a MAP of 0.506 for the CANTEMIST-CODING subtask. The code to run the developed models is available in our GitHub page: https://github.com/lasigeBioTM/CANTEMIST-Participation.

To improve the NER tagger, we intend to resume the training of the Flair embeddings up until 2000 epochs and to generate a larger language model, with 2048 hidden layers instead of 1024, and to train the tagger over more text, by including the development 2 set in the training process. For the normalization model, we intend to explore the use of word embeddings in the candidate generation process and the use of a relation extraction tool to build better disambiguation graphs. As to improvements in the classification model, we intend to explore

additional X-Transformer models trained with new models using biomedical data in Spanish and through the combination with other NLP techniques, such as automatic text summarization and NER, in order to further improve the results achieved by the algorithm.

Acknowledgements

This project was supported by FCT through funding of the DeST: Deep SemanticTagger project, ref. PTDC/CCI-BIO/28685/2017, and the LASIGE ResearchUnit, ref. UIDB/00408/2020

References

[1] O. Asan, P. Smith, E. Montague, More Screen Time, Less Face time – Implications for EHR Design, Journal of Evaluation in Clinical Practice 20 (2014) 896–901. doi:10.1111/jep.12182.

[2] P. Sfakianaki, L. Koumakis, S. Sfakianakis, G. Iatraki, G. Zacharioudakis, N. Graf, K. Marias, M. Tsiknakis, Semantic biomedical resource discovery : a Natural Language Processing framework, BMC Medical Informatics and Decision Making 15 (2015). URL: http://dx.doi.org/10.1186/s12911-015-0200-4. doi:10.1186/s12911-015-0200-4.

[3] P. Ernst, A. Siu, D. Milchevski, J. Hoffart, G. Weikum, DeepLife: An Entity-aware Search, Analytics and Exploration Platform for Health and Life Sciences, in: Proceedings ofthe 54th Annual Meeting ofthe Association for Computational Linguistics—System Demonstrations, Association for Computational Linguistics, Berlin, Germany, August 7-12, 2016, 2016, pp. 19–24. URL: http://www.aclweb.org/anthology/P16-4004{%}0Ahttp://aclweb.org/anthology/P16-4004. doi:10.1111/j.1348-0421.2010.00272.x.

[4] M. Vazquez, M. Krallinger, F. Leitner, A. Valencia, Text mining for drugs and chemical compounds: Methods, tools and applications, Molecular Informatics 30 (2011) 506–519. doi:10.1002/minf.201100005.

[5] J. He, M. de Rijke, M. Sevenster, R. van Ommering, Y. Qian, Generating Links to Background Knowledge: A Case Study Using Narrative Radiology Reports, in: CIKM'11, ACM, October 24–28, 2011, Glasgow, Scotland, UK., 2011, p. 1867. doi:10.1145/2063576.2063845.

[6] A. Miranda-Escalada, E. Farré, M. Krallinger, Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.

[7] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, R. Vollgraf, FLAIR: An easy-to-use framework for state-of-the-art NLP, in: NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Demonstrations Session, 2019, pp. 54–59.

[8] A. Lamurias, P. Ruas, F. M. Couto, PPR-SSM: Personalized PageRank and semantic similarity measures for entity linking, BMC Bioinformatics 20 (2019) 1–12. doi:10.1186/s12859-019-3157-y.

[9] Z. Huang, W. Xu, K. Yu, Bidirectional LSTM-CRF Models for Sequence Tagging (2015). URL: http://arxiv.org/abs/1508.01991. arXiv:1508.01991.

[10] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[11] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2227–2237. URL: https://www.aclweb.org/anthology/N18-1202. doi:10.18653/v1/N18-1202.

[12] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics (2019) 1–7. doi:10.1093/bioinformatics/btz682. arXiv:1901.08746.

[13] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly Available Clinical, in: Proceedings ofthe 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 72–78. doi:10.18653/v1/w19-1909.

[14] M. Pershina, Y. He, R. Grishman, Personalized Page Rank for Named Entity Disambiguation, in: Human Language Technologies: The 2015 Annual Conference ofthe North American Chapter ofthe ACL, Section 4, Association for Computational Linguistics, Denver, Colorado, May 31 – June 5, 2015, 2015, pp. 238–243.

[15] Z. Guo, D. Barbosa, Robust named entity disambiguation with random walks, Semantic Web 9 (2018) 459–479. doi:10.3233/SW-170273.

[16] Y. Cao, L. Hou, J. Li, Z. Liu, Neural Collective Entity Linking (2018). URL: http://arxiv.org/abs/1811.08603. arXiv:1811.08603.

[17] O.-E. Ganea, T. Hofmann, Deep Joint Entity Disambiguation with Local Neural Attention, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, September 7–11, 2017, 2017, pp. 2619–2629. URL: http://arxiv.org/abs/1704.04920. doi:10.18653/v1/d17-1277. arXiv:1704.04920.

[18] Z. Ji, Q. Wei, H. Xu, BERT-based Ranking for Biomedical Entity Normalization (2019). URL: http://arxiv.org/abs/1908.03548. arXiv:1908.03548.

[19] X. Yin, Y. Huang, B. Zhou, A. Li, L. Lan, Y. Jia, Deep Entity Linking via Eliminating Semantic Ambiguity With BERT, IEEE Access 7 (2019) 169434–169445. doi:10.1109/ACCESS.2019.2955498.

[20] K. Bhatia, H. Jain, P. Kar, M. Varma, P. Jain, Sparse local embeddings for extreme multi-label classification, in: Advances in Neural Information Processing Systems, 2015.

[21] J. Liu, W. C. Chang, Y. Wu, Y. Yang, Deep learning for extreme multi-label text classification, in: SIGIR 2017 - Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017. doi:10.1145/3077136.3080834.

[22] Y. Kim, Convolutional neural networks for sentence classification, in: EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 2014. doi:10.3115/v1/d14-1181. arXiv:1408.5882.

[23] R. You, Z. Zhang, S. Dai, S. Zhu, Haxmlnet: Hierarchical attention network for extreme multi-label text classification, CoRR abs/1904.12578 (2019). URL: http://arxiv.org/abs/1904.

12578. arXiv:1904.12578.

[24] W.-C. Chang, H.-F. Yu, K. Zhong, Y. Yang, I. Dhillon, Taming Pretrained Transformers for Extreme Multi-label Text Classification, 2020. URL: http://arxiv.org/abs/1905.02331v4. arXiv:1905.02331.

[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[26] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, CoRR abs/1906.08237 (2019). URL: http://arxiv.org/abs/1906.08237. arXiv:1906.08237.

[27] Y. Prabhu, A. Kag, S. Harsola, R. Agrawal, M. Varma, Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising, in: 2018 World Wide Web Conference, International World Wide Web Conferences Steering Committee, International World Wide Web Conferences Steering Committee, 2018, pp. 993–1002.

[28] A. Akbik, D. Blythe, R. Vollgraf, Contextual String Embeddings for Sequence Labeling, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 1638–1649. URL: https://github.com/zalandoresearch/flair.

[29] F. M. Couto, A. Lamurias, Semantic Similarity Definition, Reference Module in Life Sciences (2018) 0–16. URL: http://linkinghub.elsevier.com/retrieve/pii/B9780128096338204019. doi:10.1016/B978-0-12-809633-8.20401-9.

[30] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web, Technical Report, Stanford InfoLab, 1998. URL: http://ilpubs.stanford.edu:8090/422/.

[31] F. M. Couto, A. Lamurias, MER: a shell script and annotation server for minimal named entity recognition and linking, Journal of Cheminformatics 10 (2018) 58. URL: https://jcheminf.biomedcentral.com/articles/10.1186/s13321-018-0312-9. doi:10.1186/s13321-018-0312-9.

[32] J. Cañete, G. Chaperon, R. Fuentes, J. Pérez, Spanish pre-trained bert model and evaluation data, in: to appear in PML4DC at ICLR 2020, 2020.

Appendix F

COVID-19 recommender system based on an annotated multilingual corpus

G&I
Genomics & Informatics

BLAH7
Application note

eISSN 2234-0742
Genomics Inform 2021;19(3):e24
https://doi.org/10.5808/gi.21008

Received: February 26, 2021
Revised: August 4, 2021
Accepted: August 12, 2021

*Corresponding author:
E-mail: mcbarros@fc.ul.pt

**Corresponding author:
E-mail: psruas@fc.ul.pt

***Corresponding author:
E-mail: dfsousa@lasige.di.fc.ul.pt

#These authors contributed equally to this work.

COVID-19 recommender system based on an annotated multilingual corpus

Márcia Barros[1,2#*], Pedro Ruas[1#**], Diana Sousa[1#***], Ali Haider Bangash[3,4], Francisco M. Couto[1]

[1]Large-Scale Informatics Systems Laboratory, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal
[2]Center for Astrophysics and Gravitation, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal
[3]Shifa College of Medicine, Shifa Tameer-e-Millat University, Islamabad 46000, Pakistan
[4]Working Group 3, COST Action EVidence-Based RESearch (EVBRES), Western Norway University of Applied Sciences, Inndalsveien 28, 5063 Bergen, Norway

Tracking the most recent advances in Coronavirus disease 2019 (COVID-19)–related research is essential, given the disease's novelty and its impact on society. However, with the publication pace speeding up, researchers and clinicians require automatic approaches to keep up with the incoming information regarding this disease. A solution to this problem requires the development of text mining pipelines; the efficiency of which strongly depends on the availability of curated corpora. However, there is a lack of COVID-19–related corpora, even more, if considering other languages besides English. This project's main contribution was the annotation of a multilingual parallel corpus and the generation of a recommendation dataset (EN-PT and EN-ES) regarding relevant entities, their relations, and recommendation, providing this resource to the community to improve the text mining research on COVID-19–related literature. This work was developed during the 7th Biomedical Linked Annotation Hackathon (BLAH7).

Keywords: COVID-19, entity extraction, recommendation, relation extraction, text mining
Availability: The code supporting our work and the resulting datasets are publicly available at https://github.com/lasigeBioTM/blah7.

Introduction

Coronavirus disease 2019 (COVID-19) pandemic took the world by surprise due to its impact on global public health. The scientific community, sensing the danger posed by this global health emergency, was quick to join hands in a bid to mitigate its effects. Natural language processing and machine learning (ML) research also focused in the quest of curbing the morbidity and mortality associated with the pandemic, being LitCovid [1] and CORD-19 [2] good examples of this effort. These resources are massive databases of scientific literature generated throughout the world pertinent to the COVID-19 pandemic—the characteristics of its causative organism (severe acute respiratory syndrome coronavirus 2), pathophysiology of the ailment as well as preventive measures that are suggested to be employed.

Recommender systems (RS) are tools for predicting the best items of interest for the users of a system, being mostly based on the past interests of the users. The interests of the users are usually collected through explicit or implicit feedback, for example, using a 5 stars system or the products opened by the users, respectively. The feedback is then used to create recommendation datasets of < user,item,rating > , useful for developing and evaluating

recommendation algorithms. The main approaches used in RS are collaborative-filtering, which uses the similarity between the ratings of the users and its only dependent on the feedback of the users, and content-based, which uses the similarity/relation between the items. RS have been widely used for recommending movies, books, or e-commerce, achieving excellent results. In scientific fields, such as Health and Life Sciences, RS began to be used with the goal of helping health staff and researchers, for example, by recommending drugs to a researcher based on the drugs that she/he already had interest in. The major challenge for RS in scientific fields is the lack of open source recommendation datasets. Some alternatives have been developed, one in particular called LIBRETTI, which uses the scientific literature for creating such datasets [3].

Earlier on, it was realized that a massive resource of literature surely would come in handy while developing management protocols and RS by training ML models. Therefore, efforts have been made to create such pipelines and to fit them onto ML models that allow recommendation [4]. However, since medical literature has its own specific linguistic characteristics and that is fairly more complex than generic text, it was observed that semi-automatic annotation is critical in creating a richer constellation of medical data that can be used for superior training of recommender ML models. Moreover, a large portion of health related text is normally generated in the native language, so text mining tools should also be able to process multilingual corpora.

Therefore, the goals of the present project were to retrieve COVID-19 related documents, to automatically annotate them with entities and relations, generate recommendation datasets of scientific entities, and to manually validate a sample of the obtained annotations. The recommendation datasets are then used to develop new recommendation algorithms in the field of COVID-19.

The contributions of the present work are an automatic pipeline for document retrieval, entity and relation extraction, and recommendation, as well as a set of multilingual parallel datasets (English/Portuguese/Spanish) related with COVID-19 that allows the evaluation of Named Entity Recognition/Linking, Relation Extraction (RE), and Recommendation Systems. We also developed a new recommendation algorithm, called Relation Recommendation Algorithm (RelRA), and conducted preliminary tests with it.

Methodology

Fig. 1 presents the general workflow and the tools used throughout our work.

Document retrieval

The first step was to retrieve COVID-19 related abstracts from PubMed repository using the Bio.Entrez package (https:// biopython.org/docs/1.75/api/Bio.Entrez.html), which is part of Biopython. We used PubMed since it allows the abstract retrieval in more than one language, in our case, we needed English, Spanish, and Portuguese abstracts. Two versions of the dataset were created using different queries: *abstracts_covid_19*, which includes abstracts directly related with COVID-19 and *abstracts_large*, which includes abstracts directly and indirectly related with COVID-19. The queries used are present in Table 1.

Entity extraction

The second step was to extract named entities present in the retrieved documents, more concretely, by performing Named Entity Recognition and Named Entity Linking. This step was accomplished by the Python implementation of MER [5], a dictionary matching system that, given a lexicon with the terms of an ontology or any knowledge base, recognizes entities in text and links them to the respective identifiers. The MER tool is light and efficient, since it does not require neither labelled data for training, as the SOTA supervised approaches usually requires (e.g., BERT) nor extensive time-consuming training. Besides, it works with any given lexicon, even if it is non-English. Consequently, we considered the tool as adequate to achieve our goals in this short-term project. Biomedical entities present in Portuguese, Spanish, and English abstracts were recognized by MER and then linked to the respective DeCS ("Descritores em Ciências da Saúde", https:// decs.bvsalud.org/) term (September 2020 edition). DeCS is multilingual biomedical vocabulary built upon MeSH terminology. It includes versions in several languages, such as Portuguese, Spanish, and English, so it is suitable for our goal of creating a multilingual dataset. Almost all DeCS terms are MeSH terms, however, there exist some specific DeCS terms that do not correspond to any MeSH term (4,378 out of 34,294 terms). Additionally, recognized biomedical entities in English abstracts were linked to the following ontologies (latest edition available at January 2021): Human Disease Ontology (DO, https://disease-ontology.org/), Gene Ontology (GO, http://geneontology.org/), Human Phenotype Ontology (HPO, https://hpo.jax.org/app/), Chemical Elements of Biological Interest (ChEBI) Ontology (https://www.ebi.ac.uk/chebi/), and Coronavirus Infectious Disease Ontology (CIDO, https://github.com/CIDO-ontology/cido).

Relation extraction

In the third step, the relation extraction module performs RE by applying the BiOnt [6] system, built to allow the extraction of relations between biomedical entities supported by ontologies (e.g., HPO and GO). We opted for BiOnt due to its unique use of added

https://doi.org/10.5808/gi.21008

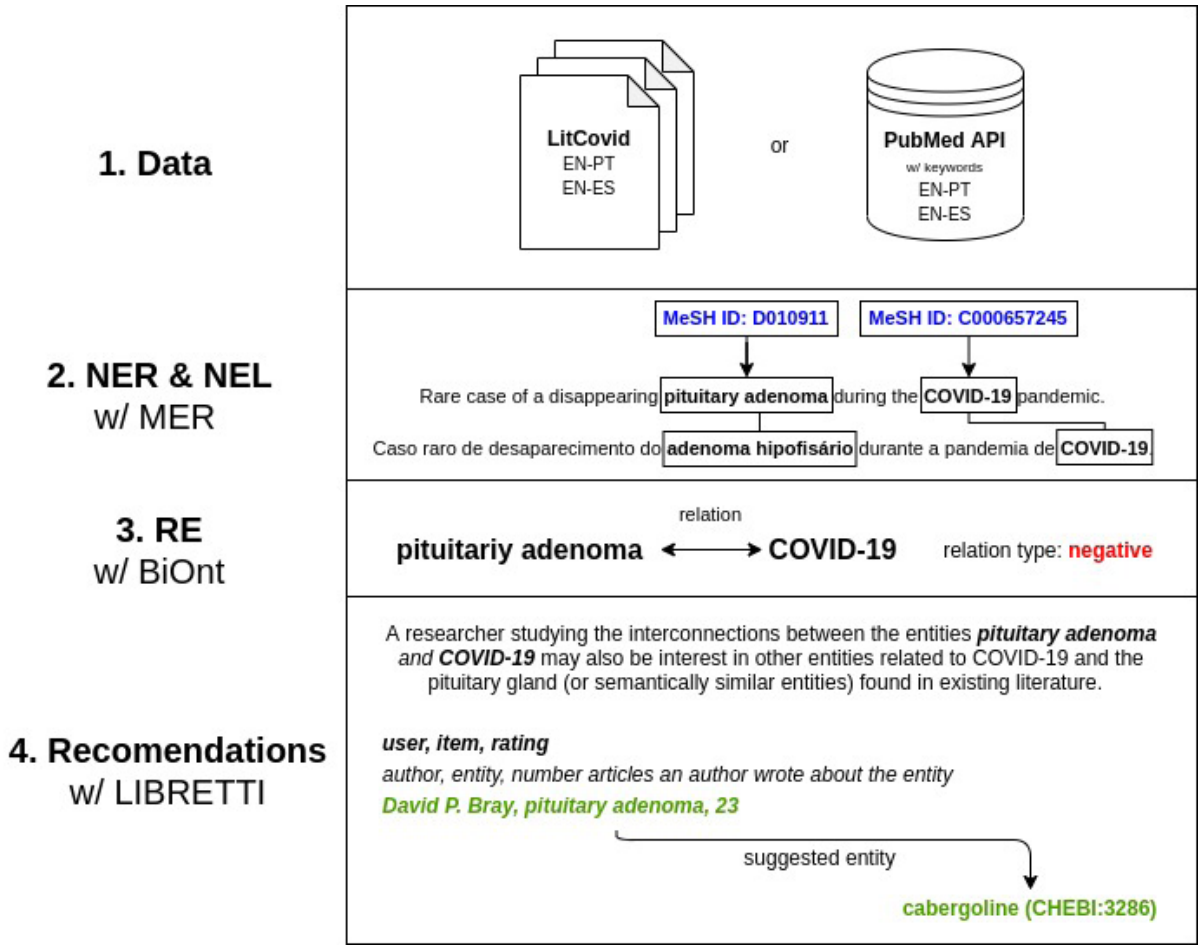


Fig. 1. Pipeline with the tools used at each stage with an example retrieved from article PMID:33220478. NER, Named Entity Recognition; NEL, Named Entity Linking; RE, Relation Extraction; LIBRETTI, LIterature Based RecommEndaTion of scienTific Items; COVID-19, coronavirus disease 2019.

Table 1. Queries used for document retrieval

Set	English query	Portuguese query	Spanish query
abstracts_covid_19	*covid-19 AND English [LANG]*	*AND English [LANG] AND Portuguese [LANG]*	*AND English [LANG] AND Spanish [LANG]*
abstracts_large	*2019 Novel Coronavirus Disease OR 2019 Novel Coronavirus Infection OR 2019-nCoV Disease OR 2019-nCoV Infection OR COVID-19 Pandemic OR COVID-19 Pandemics OR COVID-19 Virus Disease OR COVID-19 Virus OR Infection OR COVID19 OR Coronavirus Disease 2019 OR Coronavirus Disease-19 OR SARS Coronavirus 2 Infection OR SARS-CoV-2 Infection AND English [LANG]*	*2019 Novel Coronavirus Disease OR 2019 Novel Coronavirus Infection OR 2019-nCoV Disease OR 2019-nCoV Infection OR COVID-19 Pandemic OR COVID-19 Pandemics OR COVID-19 Virus Disease OR COVID-19 Virus OR Infection OR COVID19 OR Coronavirus Disease 2019 OR Coronavirus Disease-19 OR SARS Coronavirus 2 Infection OR SARS-CoV-2 Infection AND English [LANG] AND Portuguese [LANG]*	*2019 Novel Coronavirus Disease OR 2019 Novel Coronavirus Infection OR 2019-nCoV Disease OR 2019-nCoV Infection OR COVID-19 Pandemic OR COVID-19 Pandemics OR COVID-19 Virus Disease OR COVID-19 Virus OR Infection OR COVID19 OR Coronavirus Disease 2019 OR Coronavirus Disease-19 OR SARS Coronavirus 2 Infection OR SARS-CoV-2 Infection AND English [LANG] AND Spanish [LANG]*

external knowledge in the form of biomedical ontologies to potentiate RE. Thus, instead of relying on just the training data for the learning process as most RE systems, BiOnt also adds the ancestry information to each entity in a candidate pair by matching it to an ontology term.

In new data, the BiOnt system can identify relations between different and the same type of biomedical entities, such as diseases and human phenotypes, provided we use the pre-trained models trained on available training data. Using the pre-trained models available, we extracted relations between ChEBI and DO entities

and between GO and HPO entities for this project. BiOnt does not make available pre-trained models for all other combinations that we could apply to our entities. We only considered relations in English abstracts since both Portuguese and Spanish abstracts only had annotated MeSH terms, for which we did not have pre-trained models. We did not restrain the relations to sentence-level and instead considered relations within the same abstract. However, since our Portuguese and Spanish abstracts can be linked to their respective English versions, it is also possible to map the extracted relations from English to the corresponding translated abstracts.

Recommendation

Dataset creation

In the fourth step, the goal was to create multilingual recommendation datasets. The datasets were created using a methodology called LIterature Based RecommEndaTion of scienTific Items (LIBRETTI), which consists in developing a standard <user,item,rating> dataset from research articles. The users are the authors from the scientific articles, the items are biomedical entities mentioned in the articles, and the ratings are the number of articles an author wrote about an item [3]. For this work, the input research corpus is the one retrieved in phase 1: Document retrieval, more specifically the abstracts_large collection. The items are biomedical entities recognized in phase 2, i.e., diseases from the DO, gene terms from GO, phenotypes from HPO, and chemical compounds from ChEBI.

RelRA

The primary goal of the RS developed in this work is to recommend entities related to the COVID-19 disease to the researchers. To that end, we developed a new recommender content-based algorithm, based on the relations between the items - RelRA. We developed RelRA during 7th Biomedical Linked Annotation Hackathon (BLAH7), and conducted the first experiments. RelRA is based on the relations between the entities extracted in phase 2. It integrates phase 3: Relation extraction.Consider a user who has already rated some items, we want to know which items are suitable recommendations for this user. The goal of the algorithm is to provide a score to each unrated item in order to rank them. For that, we use the relations between the items, and the score is calculated considering how many relations an unrated item has with the items in the rated list. For this work we used a list of relations created using the method described in phase 3: Relation Extraction.

The RelRa algorithm was evaluated in the datasets EN_PT and EN_ES, created in phase 4. To avoid biases, the list of relations used for evaluating the RelRA algorithm were extracted from a sample of the CORD-19 corpus (nine thousand documents, version from 2020-03-13) with research articles completely different from those used to create the recommendation datasets.

Evaluation

For the evaluation, we tested the RelRa algorithm against a random algorithm. We used a cross-validation strategy, with 80% for the training set and 20% for the test set. For the evaluation the datasets were filtered, thus each user had at least 20 items rated. The evaluation metrics are Precision, Recall, and Mean Reciprocal Ranking (MRR) [7].

Manual validation at BLAH7

For manual validation of the obtained annotations, we randomly selected a sample of 40 English (20) and Portuguese (20) abstracts belonging to *abstracts_covid_19* set with entity annotations and, in the case of the English abstracts, also with relation annotations. During BLAH7, annotations were uploaded to PubAnnotation (http://pubannotation.org/) and 4 participants were responsible for the correction of the existing entity and relation annotations, but also for the addition of new annotations, if deemed necessary.

Results and Discussion

Statistics about the retrieved documents and the dimensions of the recommendations datasets created from each corpus are available in (Table 2). As expected, the abstracts_large datasets have a much higher number of documents, both in PT and ES, than the limited

Table 2. Number of documents in each version of the dataset, and respective dimensions of the recommendation datasets

Set	Languages	Abstracts	nUsers	nItems	nRatings
abstracts_covid_19	EN_PT	80	1,750	1,507	36,614
	EN_ES	53	1,744	669	14,920
abstracts_large	EN_PT	346	1,869	2,403	49,839
	EN_ES	390	1,855	1,036	20,417

EN_PT, created from abstracts in English ("EN") and Portuguese ("PT"); PT, created from abstracts in Portuguese ("PT"); EN_ES, created from abstracts in English ("EN") and Spanish ("ES"); ES, created from abstracts in Spanish ("ES"). nUsers, number of users; nItems, number of items; nRatings, number of ratings.

https://doi.org/10.5808/gi.21008

abstracts_COVID_19. Therefore, the recommendations datasets created from the abstracts_large corpus have as well a higher number of users, items and ratings.

The results of the algorithms RelRA and Random (RAND) for the datasets EN_PT and EN_ES are presented in Fig. 2, respectively, for the evaluation metrics of Precision, Recall and MRR, for the top@5 ranked results.

Fig. 2 shows that RelRa achieved better results for all the evaluation metrics when compared with a random recommendation of the items, for both EN_PT and EN_ES datasets. RelRa seems to have great potential for the recommendation of scientific entities based on their relations, however, the number of documents and relations needs to be improved for further testing.

Fig. 3 shows an example of recommendation. The user represented had interest in six different entities. We have a list of three entities that we wish to know if are suitable for recommendation to this user. Using RelRA, we find the relations between the list of liked items and these unknown items. Severe acute respiratory syndrome (DOID_2945) is related to two items in the liked list, thus it has a score of 2/6. Propyzamide (CHEBI_34935) is related to three entities in the liked list, achieving a score of 3/6. Inflammatory response (GO_0006954) does not have any relation with the items liked by

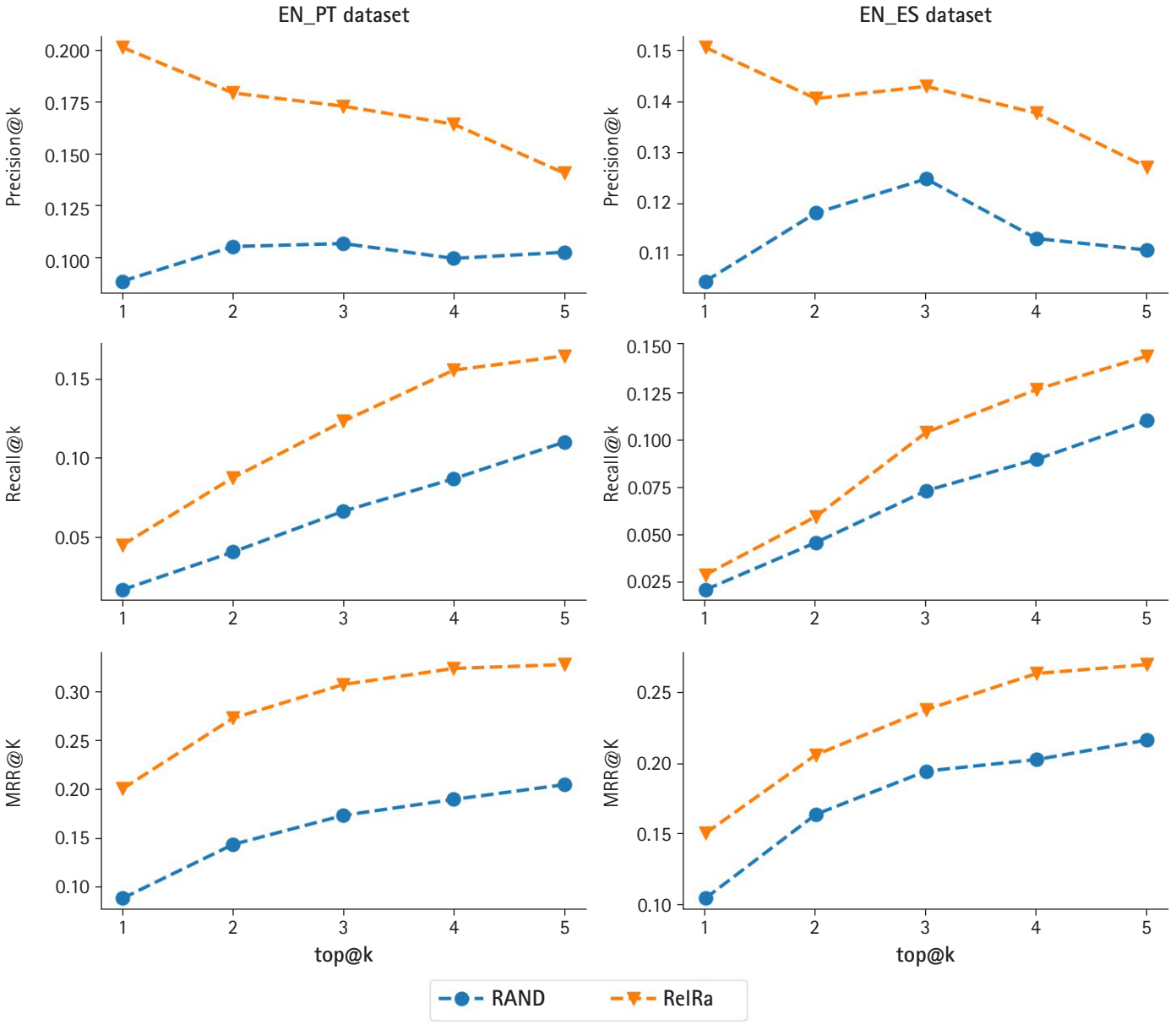


Fig. 2. Results for RelRa vs. RAND, for precision, recall and MRR, for EN_PT and EN_ES datasets. RelRA, Relation Recommendation Algorithm; RAND, Random; MRR, Mean Reciprocal Ranking; EN_PT, created from abstracts in English ("EN") and Portuguese ("PT"); EN_ES, created from abstracts in English ("EN") and Spanish ("ES").

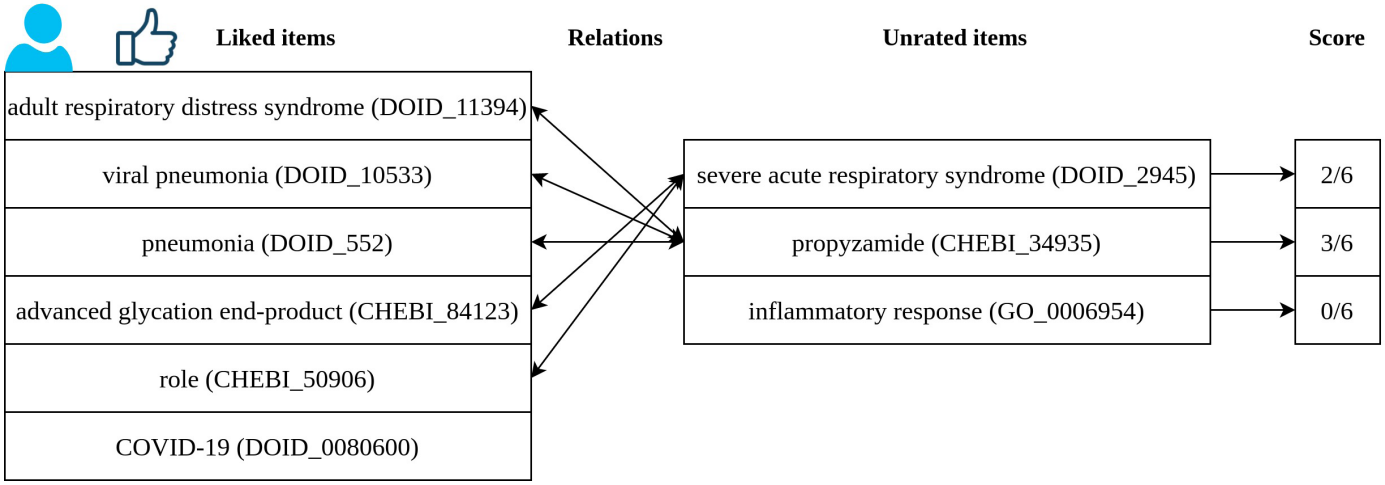


Fig. 3. Example of recommendation using the RelRA algorithm. RelRA, Relation Recommendation Algorithm; COVID-19, coronavirus disease 2019.

this user. Severe acute respiratory syndrome (DOID_2945) and Propyzamide (CHEBI_34935) would be recommended to this user.

These are the first tests with RelRA, which was specially developed during BLAH7.

Manual validation at BLAH7

Table 3 presents the results for the manual validation stage at BLAH7. After the validation process, for the 40 documents, we retrieved more entities, more relations, and with better quality by discarding illy annotated entities and relations. Further, by reaching a consensus between the four annotators, we increased our datasets' quality by adding even more annotations. These datasets are available in the PubAnnotation (http://pubannotation.org/collections/LASIGE:%20Annotating%20a%20multilingual%20COVID-19-related%20corpus%20for%20BLAH7) platform (in their original and consensus format).

Table 3. Final counts for the 40 abstracts sample (20 English and 20 Portuguese), the mean number of each subset for the annotators/curators task, and the final consensus numbers of manual validation

Dataset		Original	Annotated/Curated	Consensus
Portuguese	Entities	245	322	354
English	Entities	493	511	607
	Relations	224	238	250

Conclusion

Our goals for the present project were to retrieve COVID-19 related documents, to automatically annotate them with entities and relations, generate recommendation datasets of scientific entities, and to manually validate a sample of the obtained annotations.

We were able to create an automatic pipeline for document retrieval, entity and relation extraction, and recommendation, as well as a set of multilingual parallel datasets (English/Portuguese/Spanish) related with COVID-19 that allows the evaluation of Named Entity Recognition/Linking, RE, and Recommendation Systems. Further, we partially manually validated our datasets using the PubAnnotation platform.

For future work, the manual validation of the annotations could be improved, more concretely, by leveraging crowdsourcing platforms to recruit a large number of annotators [8]. Besides, due to time constraints, we were not able to manually validate the annotations present in EN_ES datasets during BLAH7, so future work could accomplish this.

ORCID

Márcia Barros: https://orcid.org/0000-0002-9728-9618
Pedro Ruas: https://orcid.org/0000-0002-1293-4199
Diana Sousa: https://orcid.org/0000-0003-0597-9273
Ali Haider Bangash: https://orcid.org/0000-0002-8256-3194
Francisco M. Couto: https://orcid.org/0000-0003-0627-1496

Authors' Contribution

Conceptualization: MB, PR, DS, FMC. Data curation: MB, PR, DS, AHB. Formal analysis: MB, PR, DS, AHB. Funding acquisition: FMC. Methodology: MB, PR, DS, FMC. Writing - original

https://doi.org/10.5808/gi.21008

draft: MB, PR, DS, AHB. Writing - review & editing: MB, PR, DS, AHB, FMC.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This work was supported by FCT through funding of Deep Semantic Tagger (DeST) project (ref. PTDC/CCI-BIO/28685/2017) and LASIGE Research Unit (ref. UIDB/00408/2020 and ref. UIDP/00408/2020); and FCT and FSE through funding of PhD Scholarship, ref. 2020.05393.BD, PhD Scholarship, ref. SFRH/BD/128840/2017, and PhD Scholarship, ref. SFRH/BD/145221/2019.

References

1. Chen Q, Allot A, Lu Z. LitCovid: an open database of COVID-19 literature. Nucleic Acids Res 2021;49:D1534-D1540.
2. Lu Wang L, Lo K, Chandrasekhar Y, Reas R, Yang J, Eide D, et al. CORD-19: The Covid-19 Open Research Dataset. Preprint at: https://arxiv.org/abs/2004.10706 (2020).
3. Barros M, Moitinho A, Couto FM. Using research literature to generate datasets of implicit feedback for recommending scientific items. IEEE Access 2019;7:176668-176680.
4. Tworowski D, Gorohovski A, Mukherjee S, Carmi G, Levy E, Detroja R, et al. COVID19 Drug Repository: text-mining the literature in search of putative COVID19 therapeutics. Nucleic Acids Res 2021;49:D1113-D1121.
5. Couto FM, Lamurias A. MER: a shell script and annotation server for minimal named entity recognition and linking. J Cheminform 2018;10:58.
6. Sousa D, Couto FM. BiOnt: deep learning using multiple biomedical ontologies for relation extraction. In: Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science, Vol. 12036 (Jose JM, Yilmaz E, Magalhaes J, Castells P, Ferro N, Silva MJ, et al., eds.). Cham: Springer, 2020. pp. 367-374.
7. Shani G, Gunawardana A. Evaluating recommendation systems. In: Recommender Systems Handbook (Ricci F, Rokach L, Shapira B, Kantor P, eds.). Boston: Springer, 2011. pp. 257-297.
8. Sousa D, Lamurias A, Couto FM. A hybrid approach toward biomedical relation extraction training corpora: combining distant supervision with crowdsourcing. Database (Oxford) 2020; 2020:baaa104.

Appendix G

LASIGE-BioTM at MESINESP2: entity linking with semantic similarity and extreme multi-label classification on Spanish biomedical documents

LASIGE-BioTM at MESINESP2: entity linking with semantic similarity and extreme multi-label classification on Spanish biomedical documents

Pedro Ruas[1], Vitor D. T. Andrade[1] and Francisco M. Couto[1]

[1]LASIGE, Faculdade de Ciências, Universidade de Lisboa, Lisbon 1749-016, Portugal

Abstract

Our team, LASIGE_BioTM, participated in the three sub-tracks of MESINESP2: (1) scientific literature, (2) clinical trials, and (3) patents. Our system comprises two modules: entity linking and extreme multi-label classification. The first module uses the entities recognized in text and then applies a graph-based entity linking model to link them to the DeCS vocabulary. In the end, it applies a semantic similarity-based filter to determine the most relevant entities in each document, which are then fed to the second module. The second module consists of an adapted version of the X-Transformer algorithm, and is responsible for associating each document with the top-20 relevant DeCS codes, which can be viewed as an extreme multi-label classification algorithm. The obtained results (micro F1-scores) were 0.2007, 0.0686, and 0.0314 for sub-tracks 1, 2, and 3, respectively. These represent low values when compared to other participants, mainly because of the lack of time our team had available to train the models. All of the used software is available in an open access repository.

Keywords

Named Entity Recognition, Named Entity Linking, Extreme Multi-Label Classification, Multilingual, Text Mining

1. Introduction

Automatic semantic indexing is essential to organise the growing text data that is available, which is particularly critical in scientific domains, including the biomedical one, where most of the findings are available in the text format. We can view this task as an extreme multi-label classification (XMC) problem, in which the goal is to tag a given data point with a subset of relevant labels from an extremely large label list. Therefore, the data points are the text documents to classify, and the label list provided by a knowledge base, such as an ontology. Most of the proposed XMC approaches focus on datasets including Wikipedia articles or on datasets with commercial application (e.g. dynamic search advertising) and less attention is devoted to the biomedical domain. Additionally, multilingual approaches focusing on other languages besides English are also scarce, such is the case of Spanish.

In this sense, initiatives such as BioASQ [1] are necessary to stimulate the development of biomedical, multilingual-focused approaches. In particular, the Medical Semantic Indexing

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
psruas@fc.ul.pt (P. Ruas); fc49005@alunos.fc.ul.pt (V. D. T. Andrade); fcouto@di.fc.ul.pt (F. M. Couto)
0000-0002-1293-4199 (P. Ruas); 0000-0003-0627-1496 (F. M. Couto)

In Spanish (MESINESP) task was first introduced in the BioASQ 2020 challenge and the goal was to perform semantic indexing of Spanish health-related documents, like scientific articles, clinical trials, and healthcare project summaries, with terms from the Spanish version of the *Descriptores en Ciencias de la Salud* (DeCS). The second edition, the MESINESP2 shared-task [2] was extended and included the following sub-tracks: **MESINESP-L** – Scientific Literature: Automatic indexing with DeCS terms of Spanish abstracts from two databases, IBECS and LILACS; **MESINESP-T** - Clinical Trials: Automatic indexing with DeCS terms of Spanish clinical trials from REEC (Registro Español de Estudios Clínicos); **MESINESP-P** – Patents: Automatic indexing with DeCS terms Spanish patents extracted from Google Patents.

In the past, named entities have been considered important features that aid the classification of texts. For instance, Gui et al [3] proposed a hierarchical text classification method that leverages named entities as features, and, according to the conclusions of the referred study, the features are responsible for the improvement of the method's performance. More recently, Anelic and co-workers [4] have argued that named entities do not improve the performance of text classification, and can even decrease it. However, none of these works attempted to normalise the recognised entities to concepts belonging to structured vocabularies, the approaches only used the surface form of the entities instead of the designations for the associated concepts. Besides, not every entity recognised in a given document has the same importance, i.e., some entities may not be related with the main topic of the document, which can be particularly true in documents containing a large number of different entities. Therefore, we explored the hypothesis that linking the recognised entities to concepts of a structured vocabulary and selecting only the most relevant entities to feed the text classification algorithm improve its performance.

After participating in the first edition [5], this paper describes the participation of our team, LASIGE_BioTM, in the sub-tracks of MESINESP2. We developed a pipeline based on two modules: the first one performs entity linking, by mapping the recognised entities in text to terms of the DeCS vocabulary and then applying a semantic similarity-based filter to obtain the most relevant entities in each document; the second module is based on the X-Transformer algorithm [6], and is responsible to classify each document with the most relevant DeCS terms. The software used in the experiments is available on: https://github.com/lasigeBioTM/MESINESP2.

1.1. Related work

1.1.1. Entity Linking

The extraction of entities is carried out through the text mining process. This process can be executed by different approaches such as: rule-based methods, machine learning and deep learning.

Rule-based methods include a set of terms, regular expressions or sentence constructions defined by experts [7]. Rule-based methods also include dictionary approaches, in which a given text is matched against a lexicon using string matching [8].

Machine learning methods in text mining are trained on training and validation datasets to make predictions on a test dataset [7]. Deep learning is a subset of machine learning that consists of artificial neural networks that include multiple hidden layers between input and

output. An artificial neural network is composed of nodes, processing units with a similar function to the neurons in the brain. The input for the nodes in text mining applications are word embeddings, which are vector representations of words. According to the way the nodes are organised, deep neural networks can be classified as Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), among others.

Usually text mining approaches include the tasks of Named Entity Recognition (NER) and Named Entity Linking (NEL). NER corresponds to the recognition of entities mentioned in the text and NEL to the linking of the recognised entities to concepts of a given knowledge base.

For the NER task, state-of-art approaches usually have a bidirectional long short-term memory - conditional random fields (BiLSTM-CRF) architecture. However, approaches that use pre-trained language models have recently emerged and showed promising results. One of the pre-trained language models that has been highlighted in the tasks of text mining is BERT [9], which is organized in a multilayer bidirectional transformer encoder. This architecture is based on an attention mechanism and allows the finding of dependencies between input and output [10]. Several variations of the original BERT model are trained in different scientific corpora, such as BioBERT [11] which was trained in PubMed and PMC articles and SciBERT [12], that was trained in Semantic Scholar articles. After the pre-training, these variations and the original BERT model can also be fine-tuned for NEL tasks[13].

In addition to the pre-trained language models, NEL state-of-the-art approaches in the biomedical domain also include graph-based models. Usually, these build a disambiguation graph composed by candidates for entity mentions and then ranked according to their relevance and coherence in the graph. Models that use the Personalized PageRank algorithm to determine the relevance of the candidates in the graph have been proposed, such as Pershina et al. [14].

1.1.2. Semantic similarity

The calculation of the relevance of the candidates in a graph normally requires a similarity measure to compare its nodes, as was proposed by Lamurias et al. [15]. A semantic similarity measure is a metric to compare the similarity between sets of text based on their implicit and explicit semantics. In the present work, we measured the semantic similarity between each entity and the remaining entities of a given document through Resnik's metric [16]. This metric is based on the extrinsic information content (IC) of the most informative common ancestor (MICA) of two given concepts [17] and is defined as:

$$SSM_{resnik}(e_1, e_2) = IC_{shared}(e_1, e_2)$$

Being e_1 and e_2 the entity 1 and the entity 2, respectively.

1.1.3. Extreme multi-label classification and biomedical semantic indexing

Chang et al. [18] divided the approaches to the XMC task in four categories: one-vs-all, partitioning methods, embedding-based, and deep-learning-based.

The Parabel algorithm [19] follows a one-vs-all approach because it learns a separate classifier for each label in the label list. It also applies a tree-based method, since it learns a balanced hierarchy over labels, which helps identifying the most similar labels with respect to a given

Table 1
Number of documents in each corpus.

Corpus	Train	Dev	Test
Scientific literature (L)	249,474	1,065	10,179 (500 gold standard)
Clinical Trials (T)	3,560	147	8,919 (250 gold standard)
Patents (P)	—	115	68,404 (150 gold standard)

label, i.e. those that are present in the same leaves. It performs sub-sampling of data points by restricting a given label's negative training examples to those examples that are annotated with similar or confusing labels, which decreases training and prediction times from linear to logarithmic. The approach then applies a hierarchical multi-label model, which is a generalisation of the multi-class hierarchical softmax model. Each classifier learns a joint probability distribution over the possible labels that is based on data point features and on the label hierarchy. Parabel was applied to Dynamic Search Advertising, which aims to predict the subset of search engine queries that will lead to a click on a given ad page.

The current state-of-the-art in XMC consists of approaches that leverage pre-trained deep language models. The first approach of this type was X-BERT (*BERT for eXtreme Multi-label Text Classification*) [18], later renamed to X-Transformer [6], which fine-tunes BERT, RoBERTa, and XLNet for the XMC task. The main challenges of applying Transformer to the XMC problem are the extremely large set of possible labels and the label sparsity, which arises from the fact that too few labels are associated with a large number of training instances. The model includes three components: a semantic label indexer, a deep neural matcher, and a ranker. The authors applied the developed algorithm to four datasets, Eurlex-4K, Wiki10-28K, AmazonCat-13K and Wiki-500K, obtaining the following precision@1 values: 86.00%, 85.75 %, 95.17 %, 67.87 %.

1.1.4. MESINESP1

With respect to the MESINESP task, six teams have participated in the first edition, including our team, which have developed a pipeline [5] based on the X-Transformer algorithm [6] and the MER tool [20] for the named entity recognition and linking step. The approach with best performance was based on AttentionXML with multilingual-BERT [21], which achieved a micro F-measure value of 0.4254, whereas our approach achieved a micro F1-score of 0.2507.

Besides DeCS and MeSH vocabularies, there are also related works that focus on the classification or coding of clinical content with codes belonging to other vocabularies, in particular the International Classification of Diseases (ICD) terminology [22, 23, 24, 25].

2. Methodology

2.1. Data description

The target label list consisted of 34,046 codes belonging to the DeCS vocabulary[1] (2020 edition), complemented with additional COVID-related descriptors added by the organisation. Both corpora (JSON files) and the DeCS vocabulary (TSV file) were provided by the organisation and downloaded from the following link: https://zenodo.org/record/4634129#.YHcShxIo9an.

2.2. Entity Linking

Our approach consisted in using the recognised entities from the documents of each subtrack that were provided in the folder "Additional Data". The entities of these files were then further linked to the respective DeCS codes through an entity linking model. This model searches for the ten best candidates of DeCS through string matching and then develops a disambiguation graph with those candidates. The Personalized PageRank algorithm is applied to the disambiguation graph and estimates the coherence of each node, i.e. candidate, to the graph. The coherence is associated with the node degree, meaning that nodes linked to a high number of other candidate nodes are probable candidates for their respective entities compared with more isolated nodes. Besides coherence, the IC of the DeCS code associated with the nodes is used for ranking: nodes associated with DeCS codes with higher IC receive higher ranking scores. IC corresponds to the presence of an entity in a corpus, if an entity is not common in a corpus its IC will be high. The higher the IC of a candidate is, the better ranking that candidate will have in the graphic. After ranking all the candidates, the PPR selects the candidate with better ranking to map each entity. At the end, all entities in a given document are linked to their respective DeCS concepts.

To explore the guiding hypothesis of this work, we filtered the number of entities to include each document by applying a semantic similarity-based filter, more concretely, by selecting the entities for which there were other similar entities recognised in the same document.

After this step, the average of the several semantic similarity values obtained for an entity corresponded to the final score of that entity. The entities were then sorted by their score. At the end, we explored two values for the semantic similarity-based filter: 1.0 and 0.25. Considering the filter 0.25, we only included the top 25% entities according to their score, and for the filter 1.0, we included all the entities in the document. This way, we could determine the impact of choosing the most relevant entities in the performance of the classifier algorithm.

2.3. Extreme Multi-Label Classification

We approached the sub-tracks as an Extreme Multi-Label Classification (XMC) problem. Our starting point was a pipeline based on the X-Transformer algorithm [6] that was adapted to the biomedical domain by our group in the context of past competitions, such as BioASQ [5] and CANTEMIST [26]. The pipeline was further adapted to the present competition, and includes the following modules: entity linking (subsection 2.2), preprocessing, semantic label indexer, deep neural matcher, and ranker. The main modifications were made in the entity linking and

[1] http://red.bvsalud.org/decs/en/

preprocessing modules. The complete description of the entity linking component is available in the previous subsection 2.2.

The preprocessing module imports the retrieved dataset JSON files (train, dev, and text subsets) and the DeCS TSV file, the JSON files with the output from the entity linking (subsection 2.2), and, for each dataset, it generates several files:

1. vocabulary file ("label_vocab.txt"): it includes the internal numerical identifier for each DeCS term. For example, the term "calcimicina" has the internal numerical identifier "0".
2. label correspondence file ("label_correspondence.txt"): it includes the correspondence between the internal numerical identifiers, and the respective DeCS labels and terms. For example, "0" corresponds to "D000001", which corresponds to "calcimicina".
3. subset files ("*subset*.txt", "*subset*_raw_text.txt", "*subset*_raw_labels.txt"): for each subset (train, dev, and test) it is generated the three aforementioned files. The file "*subset*.txt" includes the DeCS labels that are associated with the respective documents, separated by commas, the stemmed texts of documents' titles, and the DeCS terms that were extracted in the documents appended to the end of the stemmed titles. The file "*subset*_raw_text.txt" includes only the stemmed titles, and the file "*subset*_raw_labels.txt" only the DeCS terms relative to the labels associated with the documents.

We only considered the titles of the documents based on the results described by Neves et al. [5]: the performance of the models using titles is similar to that of models using abstracts, so it is more efficient to use titles since they have less text. The limited time that we had to train models also influenced our decision to only use the titles, since the required time is lower. The titles were stemmed using the Snowball Stemmer implementation for Spanish text provided by the NLTK package[2]. As the documents belonging to the test sets were unlabeled, we added the placeholder "0" to each document in the "*subset*.txt" files. The module was also modified in order to integrate extracted entities independently of the tool employed.

The X-Transformer algorithm includes three modules: semantic label indexer, deep neural matcher, and ranker. The semantic label indexer first obtain meaningful representations for labels that are based on embeddings of the text descriptions associated with the labels, and on Positive Instance Feature Aggregation (PIFA), which is a type of label embeddings based on the TF-IDF features that are relevant instances for the labels. Then, it applies k-means clustering in order to generate label clusters according to the semantic representations described before. The deep neural matcher performs fine-tuning of BERT to encode an instance embedding, which is then used to find the most relevant clusters for the instance. At the end of this step, only a small subset of clusters are considered for the next step, which is performed by the ranker. The ranker determines the relevance of the labels in the chosen clusters to the instance, which is substantially more efficient than performing the ranking of all the initial labels. For a more complete description of the X-Transformer algorithm please refer to the original publication by Chang et al. [6].

The models developed for the different sub-tracks are shown in Table 2. We explored the fine-tuning of different deep neural matchers. The BERT Base Multilingual Cased model was trained on the Wikipedia dumps of the top 104 largest languages in Wikipedia and has the following

[2] https://www.nltk.org/

Table 2
Models used for the three sub-tracks, with the respective target datasets, thresholds (top entities to consider according to their relevance), and deep neural matcher.

Model	Target dataset	Threshold	Deep neural matcher
LASIGE_BioTM-1	L	1.0	CANTEMIST
LASIGE_BioTM-2		0.25	
LASIGE_BioTM-3	T	1.0	BERT Multilingual Base Cased
LASIGE_BioTM-4		0.25	
LASIGE_BioTM-5	P	1.0	BERT Multilingual Base Cased

characteristics: 12-layer, 768-hidden, 12-heads, 110M parameters. The X-Transformer algorithm uses the Pytorch implementation from HuggingFace Transformers [27]. The CANTEMIST model corresponds to the Model 7 described by Ruas et al. [26]. It is also based on the the BERT Base Multilingual Cased model and was first fine-tuned on 318,658 Spanish biomedical articles from the IBECS, LILACS and PubMed databases, jointly with extracted entities in the context of the participation in the first edition of MESINESP [5].

2.4. Training approach

We explored several training approaches according to the target corpus:

- L corpus: Fine-tuning of the model CANTEMIST using the provided training dataset of 249,474 documents and the provided test set with 10,179 documents.
- T corpus: Training of the model BERT Multilingual Base Cased using the provided training dataset of 249,474 documents from the L corpus and a generated test set built from the 3560 clinical trials of the training set, the 147 clinical trials of the development set, and the 8919 clinical trials of the test set (total of 12,627 documents).
- P corpus: Training of the model BERT Multilingual Base Cased using the provided training dataset of 249,474 documents from the L corpus and a generated test set built from the 115 patents of the development set and the 68,404 patents from the test set.

The training of the deep neural matcher is the limiting step of the algorithm in terms of time. Each model was trained during a single epoch then evaluated on the respective test set. The training and evaluation time was approx. 2 days for each model using a single NVIDIA Tesla P4 GPU. The values for the hyper-parameters are the following: `depth` = 6, `train_batch_size`=4, `eval_batch_size`=4, `learning_rate`=0.00005, `warmup_rate`=0.1.

3. Results and discussion

The results obtained for each sub-track are shown on Table 3. The official evaluation metric of the competition was the micro F1-score (MiF). Our best models achieved a MiF of 0.2007, 0.0686, and 0.0314 in the sub-tracks L, T, and P, respectively. These results are low when compared to

Table 3
Results on test sets for the three sub-tracks. Performance for the baseline models, the best models, and our models are shown according to the metrics: MiF-micro F1-score, MiP-micro precision, MiR-micro recall, MaF-macro F1-score, MaP-macro precision, MaR-macro recall.

Sub-track	Model	MiF	MiP	MiR	MaF	MaP	MaR
L	Baseline	0.2876	0.2335	0.3746	0.3438	0.2335	0.3746
	BERTDeCS version 4	**0.4837**	**0.5077**	**0.4618**	**0.3926**	**0.5237**	**0.3990**
	LASIGE_BioTM-1	0.2007	0.1584	0.2738	0.0941	0.1016	0.1232
	LASIGE_BioTM-2	0.1886	0.1489	0.2573	0.0920	0.0950	0.1219
T	Baseline	0.1288	0.0781	**0.3678**	0.2403	0.0977	**0.3619**
	BERTDeCS version 2	**0.3640**	0.3666	0.3614	**0.3102**	**0.4177**	0.3391
	LASIGE_BioTM-3	0.0679	0.0575	0.0828	0.0056	0.0050	0.0136
	LASIGE_BioTM-4	0.0686	0.0581	0.0838	0.0061	0.0054	0.0133
P	Baseline	0.2992	0.4293	0.2296	0.2518	**0.5290**	0.2497
	BERTDeCS version 2	**0.4514**	**0.4487**	**0.4541**	**0.4138**	0.5041	**0.4271**
	LASIGE_BioTM-5	0.0314	0.0239	0.0459	0.0071	0.0060	0.0135

the top results in each sub-track, more concretely, there is a difference of 0.2830, 0.2961, and 0.4200 in terms of MiF in the sub-tracks L, T, and P, respectively.

With respect to the initial hypothesis, the obtained results were mixed. In the sub-track L, the LASIGE_BioTM-1 model, which included all the entities recognised in the documents, obtained slightly better results (0.2007 MiF) compared with LASIGE_BioTM-2 model (0.1886 MiF), which only included 25% of the top relevant entities. However, in the sub-track T, the opposite happened, since LASIGE_BioTM-4 (top 25% entities) obtained marginally better results (0.0686 MiF) than LASIGE_BioTM-3 (0.0679 MiF). Consequently, we cannot confirm the initial hypothesis that feeding only the most relevant entities to the classifier algorithm improves its performance.

Assuming that there were no coding errors that may have undermined the results, there are several possible reasons behind the relatively low performance that our models achieved in the three sub-tracks.

Arguably, the main one is related with the impossibility of carrying out an optimisation of the hyper-parameters of the classifier algorithm, in particular the number of training epochs. Each model was only trained or fine-tuned during one epoch in the respective training dataset, which is not enough to accurately learn relevant features. The limited time we had available made it impossible to extend the training process during more epochs. Additionally, we were not able to train the models in a multi-gpu setting due to unresolved errors, so the duration of each training epoch was approximately two days using a single gpu. Beyond the number of training epochs, the optimization of other hyperparameters such as `train_batch_size`, `eval_batch_size`, and `learning_rate`, would probably lead to a better performance.

With respect to the sub-track 2 and sub-track 3, the developed models were trained on documents belonging to the L corpus (sub-track 1), and not on documents of the respective sub-tracks corpora. The text present in scientific literature has different characteristics compared with the text associated with clinical trials and patents, so the models fine-tuned in a certain

type of text will necessarily have a worse performance when their evaluation occurs over a different type of text. For sub-track 3, there was no training dataset available, but for sub-track 2 probably it would have been better if we had trained models 3 and 4 over the training dataset of the task and not over the training dataset for sub-track 1.

4. Conclusion

Our approach including an entity linking model and the X-Transformer algorithm obtained a micro F1-score of 0.2007, 0.0686, and 0.0314 in sub-tracks 1, 2, and 3, respectively, which is a low performance compared with the top participants, and even with the baseline approaches. In order to improve the performance, we need to perform a careful error-analysis to identify any coding errors that may have undermined the results. Next, we need to spend more time in the training process, more concretely, by training the models during more epochs, to perform hyper-parameter optimisation, to solve the problems associated with multi-gpu training, to explore the use of summarisation tools to feed only the relevant content to the classifier, and to explore less resource-demanding pre-trained models, such as DistilBERT. Besides, we only used the titles of the articles based on previous studies, but in the future we will explore the impact of using more text in the performance of the classification algorithm.

Acknowledgments

This work was supported by FCT through funding of Deep Semantic Tagger (DeST) project (ref. PTDC/CCI-BIO/28685/2017) and LASIGE Research Unit (ref. UIDB/00408/2020 and ref. UIDP/00408/2020); and FCT through funding of PhD Scholarship, ref. 2020.05393.BD.

References

[1] A. Nentidis, G. Katsimpras, E. Vandorou, A. Krithara, L. Gasco, M. Krallinger, G. . Paliouras, Overview of BioASQ 2021: The ninth BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. (2021).

[2] L. Gasco, A. Nentidis, A. Krithara, D. Estrada-Zavala, , R.-T. Murasaki, E. Primo-Peña, C. Bojo-Canales, G. Paliouras, M. Krallinger, Overview of BioASQ 2021-MESINESP track. Evaluation of advance hierarchical classification techniques for scientific literature, patents and clinical trials. (2021).

[3] Y. Gui, Z. Gao, R. Li, X. Yang, Hierarchical text classification for news articles based-on named entities, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 7713 LNAI (2012) 318–329. doi:10.1007/978-3-642-35527-1_27.

[4] S. Andelic, M. Kondic, I. Peric, M. Jocic, A. Kovacevic, Text Classification Based on Named Entities, in: 7th International Conference on Information Society and Technology ICIST 2017, 2017.

[5] A. Neves, A. Lamurias, F. M. Couto, Extreme Multi-Label Classification applied to the Biomedical and Multilingual Panorama, in: CLEF 2020 Working Notes, 2020. URL: http://ceur-ws.org/Vol-2696/paper_67.pdf.

[6] W.-C. Chang, H.-F. Yu, K. Zhong, Y. Yang, I. Dhillon, Taming Pretrained Transformers for Extreme Multi-label Text Classification (2020). URL: https://doi.org/10.1145/3394486.3403368. doi:10.1145/3394486.3403368. arXiv:1905.02331v4.

[7] A. Lamurias, F. Couto, Text Mining for Bioinformatics Using Biomedical Literature, 2019, p. 602–611. doi:10.1016/B978-0-12-809633-8.20409-3.

[8] F. M. Couto, A. Lamurias, Mer: a shell script and annotation server for minimal named entity recognition and linking, Journal of Cheminformatics 10 (2018) 58.

[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.

[11] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.

[12] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3615–3620. URL: https://www.aclweb.org/anthology/D19-1371. doi:10.18653/v1/D19-1371.

[13] Z. Ji, Q. Wei, H. Xu, Bert-based ranking for biomedical entity normalization, AMIA Summits on Translational Science Proceedings 2020 (2020) 269.

[14] M. Pershina, Y. He, R. Grishman, Personalized page rank for named entity disambiguation, in: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Denver, Colorado, 2015, pp. 238–243. URL: https://www.aclweb.org/anthology/N15-1026. doi:10.3115/v1/N15-1026.

[15] A. Lamurias, P. Ruas, F. M. Couto, PPR-SSM: Personalized PageRank and semantic similarity measures for entity linking, BMC Bioinformatics 20 (2019) 1–12. doi:10.1186/s12859-019-3157-y.

[16] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, arXiv preprint cmp-lg/9511007 (1995).

[17] F. Couto, A. Lamurias, Semantic similarity definition, Encyclopedia of bioinformatics and computational biology 1 (2019).

[18] W.-C. Chang, H.-F. Yu, K. Zhong, Y. Yang, I. Dhillon, X-BERT: eXtreme Multi-label Text Classification with using Bidirectional Encoder Representations from Transformers (2019) 1–12. URL: http://arxiv.org/abs/1905.02331. arXiv:1905.02331.

[19] Y. Prabhu, A. Kag, S. Harsola, R. Agrawal, M. Varma, Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising, in: Proceedings of the World Wide Web Conference, WWW 2018, ACM, New York, NY, USA, April 23-27, 2018, Lyon, France, 2018, pp. 993–1002. doi:10.1145/3178876.3185998.

[20] F. M. Couto, A. Lamurias, MER: a shell script and annotation server for minimal named entity recognition and linking, Journal of Cheminformatics 10 (2018) 58. URL: https://jcheminf.biomedcentral.com/articles/10.1186/s13321-018-0312-9. doi:10.1186/s13321-018-0312-9.

[21] R. You, Z. Zhang, Z. Wang, S. Dai, H. Mamitsuka, S. Zhu, AttentionXML: Label Tree-based Attention-Aware Deep Model for High-Performance Extreme Multi-Label Text Classification, in: 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada, 2019, pp. 1–11. arXiv:1811.01727.

[22] P. Xie, H. Shi, M. Zhang, E. P. Xing, A Neural Architecture for Automated ICD Coding, in: Proceedings ofthe 56th Annual Meeting ofthe Association for Computational Linguistics (Long Papers), Association for Computational Linguistics, 2018, pp. 1066–1076.

[23] H. Shi, P. Xie, Z. Hu, M. Zhang, E. P. Xing, Towards Automated ICD Coding Using Deep Learning, Technical Report, 2017. arXiv:1711.04075v3.

[24] S. Silvestri, F. Gargiulo, M. Ciampi, G. De Pietro, Exploit Multilingual Language Model at Scale for ICD-10 Clinical Text Classification, Proceedings - IEEE Symposium on Computers and Communications 2020-July (2020). doi:10.1109/ISCC50000.2020.9219640.

[25] C. Sen, B. Ye, J. Aslam, A. Tahmasebi, From Extreme Multi-label to Multi-class: A Hierarchical Approach for Automated ICD-10 Coding Using Phrase-level Attention (2021). URL: http://arxiv.org/abs/2102.09136. arXiv:2102.09136.

[26] P. Ruas, A. Neves, V. D. Andrade, F. M. Couto, Lasigebiotm at cantemist: Named entity recognition and normalization of tumour morphology entities and clinical coding of Spanish health-related documents, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), 2020, pp. 422–437. URL: http://ceur-ws.org/Vol-2664/cantemist_paper11.pdf.

[27] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Appendix H

Lasige-BioTM at ProfNER: BiLSTM-CRF and contextual Spanish embeddings for Named Entity Recognition and Tweet Binary Classification

Lasige-BioTM at ProfNER: BiLSTM-CRF and contextual Spanish embeddings for Named Entity Recognition and Tweet Binary Classification

Pedro Ruas and **Vitor D. T. Andrade** and **Francisco M. Couto**
LASIGE, Faculdade de Ciências da Universidade de Lisboa
1749-016 Lisboa, Portugal
psruas@fc.ul.pt, fc49005@alunos.fc.ul.pt, fcouto@di.fc.ul.pt

Abstract

The paper describes the participation of the Lasige-BioTM team at sub-tracks A and B of ProfNER, which was based on: i) a BiLSTM-CRF model that leverages contextual and classical word embeddings to recognize and classify the mentions, and ii) on a rule-based module to classify tweets. In the Evaluation phase, our model achieved a F1-score of 0.917 (0,031 more than the median) in sub-track A and a F1-score of 0.727 (0,034 less than the median) in sub-track B.

1 Introduction

The track "ProfNER-ST: Identification of professions & occupations in Health-related Social Media" (Miranda-Escalada et al., 2021b) occurred in the context of the "Social Media Mining for Health Applications (#SMM4H) Shared Task 2021" (Magge et al., 2021), and included two different sub-tracks that focused on Spanish Twitter data:

- Track A – Tweet binary classification: to determine if a given tweet has a mention of occupation or not.
- Track B – Named Entity Recognition (NER) offset detection and classification: to recognise the span of mentions of occupations and classify them in the respective category.

This paper describes the participation of the Lasige-BioTM team in the aforementioned sub-tracks. We applied 8 different models NER models (4 supervised models based on BiLSTM-CRF architecture, 3 rule-based models) to predict entities for sub-track B and explored the impact of performing data augmentation in the training set. For sub-track A, we developed a rule-based model for tweet classification that was based on the NER output for sub-track B.

1.1 Related Work

According to Goyal et al. (2018), NER approaches can be divided in two categories: rule-based and machine learning-based, being the latter further subdivided into supervised, semi-supervised, unsupervised; other approaches combine aspects from the two categories and are thus designated by hybrid. The models with an architecture consisting of a bidirectional Long Short-Term Memory (BiLSTM) network and a Conditional Random Field (CRF) decoding layer are among the state-of-the-art approaches for the NER task. (Huang et al., 2015). For a comprehensive overview of the existing NER approaches please refer to Goyal et al. (2018) and, specifically for the biomedical domain, to Lamurias and Couto (2019).

2 Methodology

2.1 Corpus description

The ProfNER corpus (Miranda-Escalada et al., 2020) contains 10,000 health-related tweets in Spanish that were annotated by linguist experts with entities relative to professions, employment statuses, and other work-related activities and includes four categories: "PROFESION", "SITUACION_LABORAL", "ACTIVIDAD", and "FIGURATIVA". For sub-track A, a given tweet was assigned the label "1" if it included at least one entity belonging to any category, but for sub-track B only entities belonging to categories "PROFESION" and "SITUACION_LABORAL" were considered for evaluation.

2.2 Pre-processing

We performed data augmentation on the training set of the corpus using the Python library `nlpaug` (Ma, 2019). For example, considering the mentioned entity "médico" present in the training set, data augmentation consisted of substituting a random character by a keyboard character (i.e. replac-

Proceedings of the Sixth Social Media Mining for Health Workshop 2021, pages 108–111
June 10, 2021.

ing the character by a neighbour character in the keyboard in order to simulate a typing error character, since Twitter data is usually noisy: "médico" → "médLco"), by a random distance character ("médico" → "médicB"), and by a synonym (i.e. replacing the character by a synonym in the Spanish WordNet: "médico" → "dr."). The output of this step consisted of three additional training files besides the original training file, each one associated with the result of a type of augmentation.

2.3 MER

The first approach was based on `MER` (Couto and Lamurias, 2018), a minimal NER tagger that recognizes entities and the respective span in text according to a given lexicon. It is based on the text processing command-line tools `grep` and `awk`, and on an inverted recognition technique that uses the words in input text as patterns to match the lexicon words. Several lexicons were created and processed including: 1) mentions in "PROFESION" category in training set and its WordNet synonyms, 2) mentions in "PROFESION" category in training set and its WordNet synonyms, jointly with entities present in the Occupations gazetteer provided by the organisation (Asensio et al., 2021), 3) mentions in "SITUACION_LABORAL" category in training set and its WordNet synonyms, 4) entities in "ACTIVIDAD" category in train set and its WordNet synonyms, 5) entities in "FIGURATIVA" category in train set and its WordNet synonyms. The first model ("MER 1") included the lexicons 1, 3, 4, and 5, the second model ("MER 2") included the lexicons 2, 3, 4, and 5, the third model ("MER 3") was similar to the first one but the mention "sin" was filtered out. During Practice phase, we built the lexicons from the training set and used the validation set as the test set. For sub-task A, we developed a rule-based module to classify each tweet with the label "1" if at least one mention was recognized in the respective text, and with label "0" otherwise.

2.4 BiLSTM-CRF

To implement the second approach, we resorted to the `FLAIR` framework (Akbik et al., 2019), and created an object of the class `SequenceTagger`, which instantiates a NER model with an architecture consisting of a BiLSTM network and a CRF decoding layer. LSTM are recurrent neural networks (RNNs), which include an input layer x representing features at time t, one or more hidden layers h, and an output layer y, which in the case of the NER task, represents a probability distribution over labels or tags at time t. A CRF network focus on the sentence level and also uses past and future tags/labels to predict the current one. The combination of a BiLSTM network with a CRF network has shown performance improvements over alternative architectures (Huang et al., 2015).

In the NER task, text needs to be tokenized and vectorized before being inputed to the neural network, which can be done leveraging pre-trained embeddings. FastText embeddings (Bojanowski et al., 2017) are an improvement over classic word embeddings, more concretely the skipgram model, by capturing sub-word information. `FLAIR` embeddings (Akbik et al., 2018) are contextual string embeddings that capture syntactic-semantic word features. We have explored the integration of different types of embeddings in the BiLSTM-CRF model through the `StackedEmbeddings` class:

- "Base" : `FLAIR` embeddings ("es-forward" and "es-backward") trained on Spanish Wikipedia (Akbik et al., 2018) + Spanish FastText embeddings
- "Twitter" : FastText Spanish COVID-19 Twitter Embeddings, provided by the organization (Miranda-Escalada et al., 2021a) (uncased version of the cbow model).
- "Medium" : `FLAIR` embeddings ("es-forward" and "es-backward") + Spanish FastText embeddings + FastText Spanish COVID-19 Twitter Embeddings

For the sub-track A, we applied a similar rule-based module as described in Section 2.3. If a model recognizes at least one entity in a given tweet in the context of sub-track B, the module assigns the label "1" to the respective tweet. If no entity is recognized in a given tweet, this receives the label "0". All the tweet IDs and respective label are then outputted in the predictions file for sub-track A.

2.4.1 Training

During Practice phase, we trained the models "Base" and "Twitter" on the original training file ("Base" and "Twitter"), and additionally, on the three files that resulted from the data augmentation step ("Base-aug" and "Twitter-aug"). During Evaluation phase, we merged the training and validation annotations, resulting in a file composed by 14,674 sentences for training and 1,630 sentences

for validation. The training parameters were set to: `hidden size` = 256, `Mini batch size` = 32, `Max epochs` = 55, `Patience` = 3.

3 Results and discussion

3.1 Practice phase

The performance of the referred models in the validation set for sub-tracks A and B are available in Table 1. The "Base" model trained on the original training file achieved the best performance in sub-tracks A and B: F1-scores (strict) of 0.908 and 0.716, respectively. Consequently, we selected this model for further training and application in the test set. The models trained on files resulting from data augmentation achieved lower performances compared with the respective versions trained exclusively on the original training file.

3.2 Evaluation phase

The results achieved by our model in the Evaluation phase and the median results for all competing teams are shown in Table 2. In sub-track A, our model achieved a F1-score of 0.917 (0.031 more than the median) and in sub-track our model achieved a F1-score of 0.727 (0.034 less than the median).

3.3 Error analysis

The model "Base", that uses contextual embeddings trained on a general corpora, obtained higher performance when comparing to the model "Twitter", although this latter model uses Twitter-specific embeddings, more concretely, FastText embeddings that were trained on Twitter data. For instance, consider the following tweet of the validation set: *"Ya que están sesionando la importante pero NO prioritaria #LeyDeAmnistia,será que también vean la cuestión de #Economia y #SaludParaTodos? Digo!Recuerden que su prioridad somos los millones que estamos indefensos ante el #COVID-19 y sin trabajo @MorenaSenadores #LeyDeAmnistiaNo https://t.co/DCiuqiBjEs"*. The model "Twitter" recognizes the mention "@MorenaSenadores" and assigns the "PROFESION" category to it, whereas the model "Base" does not recognize any mention, since is been able to assume in this context that the mention do not correspond to a profession, but instead to a Twitter handle. There is a mention with the string "senadores" classified as "PROFESION" in a tweet of the training set, which maybe leads the model "Twitter" to assume that the words "@MorenaSenadores" must also correspond to a mention, since the string is similar.

4 Conclusion

During the Practice Phase, we explored different approaches to participate in sub-tracks A e B of ProfNER: data augmentation on training set, and application of MER and a BiLSTM-CRF model for NER and further tweet classification. For the Evaluation phase we applied the BiLSTM-CRF model on the test set of ProfNER corpus and achieved F1-scores of 0.917 (0,031 more than the median) and in sub-track our model achieved a F1-score of 0.727 (0,034 less than the median). The code to run the experiments is available in our GitHub page[1]. For future work, we intend to perform hyper-parameter optimisation for the BiLSTM-CRF model, such as learning rate, hidden size, and specially the number of training epochs, since we had limited available time to perform the training of the model. We will also explore the use of different contextualised embeddings, since the models using this type of embeddings seem to achieve better performance compared to those using classical word embeddings. Besides, to improve tweet classification we will explore the application of Named Entity Linking tools (Lamurias et al., 2019) to link the recognized entities in sub-track B to structured vocabularies that contain hierarchical relationships between concepts, such as MeSH or DBpedia. This way, it will be possible to know the ancestors for a given entity, which will provide the context to effectively determine if the entity is associated with an occupation or not.

Acknowledgements

This work was supported by FCT through funding of Deep Semantic Tagger (DeST) project (ref. PTDC/CCI-BIO/28685/2017) and LASIGE Research Unit (ref. UIDB/00408/2020 and ref. UIDP/00408/2020); and FCT through funding of PhD Scholarship, ref. 2020.05393.BD.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for*

[1] https://github.com/lasigeBioTM/LASIGE-participation-in-ProfNER

	Sub-track 7A			Sub-track 7B					
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**Rel-P**	**Rel-R**	**Rel-F1**
MER 1	0.621	0.767	0.687	0.399	0.535	0.457	0.565	0.668	0.612
MER 2	0.498	0.839	0.625	0.290	0.578	0.386	0.418	0.721	0.529
MER 3	0.621	0.767	0.687	0.472	0.535	0.501	0.668	0.667	0.667
Base	**0.941**	0.876	**0.908**	**0.795**	**0.651**	**0.716**	**0.901**	**0.738**	**0.811**
Base-aug	0.848	0.830	0.839	0.705	0.616	0.657	0.826	0.721	0.770
Twitter	0.895	0.874	0.884	0.721	0.616	0.664	0.856	0.730	0.788
Twitter-aug	0.786	0.904	0.841	0.597	0.611	0.604	0.737	0.755	0.746
Medium-aug	0.780	0.887	0.830	0.618	0.601	0.609	0.753	0.733	0.743

Table 1: Practice results for sub-track 7A (left) and sub-track 7B (right). P, R, and F1 refer to precision, recall, and F1-score (strict), respectively and Rel-P, Rel-R, and Rel-F1 refer to relaxed precision, relaxed recall, and relaxed F1-score, respectively

	Sub-track 7A			Sub-track 7B		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**
Lasige-BioTM	**0.951**	**0.886**	**0.917**	0.814	0.657	0.727
Median	0.919	0.855	0.886	**0.842**	**0.727**	**0.761**

Table 2: Evaluation phase results for sub-tracks 7A and 7B. P, R, F1 refer to precision, recall, and F1-score (strict), respectively.

Computational Linguistics (Demonstrations), pages 54–59.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Alejandro Asensio, Antonio Miranda-Escalada, Marvin Aguero, and Martin Krallinger. 2021. Occupations gazetteer - ProfNER & MEDDOPROF - occupations, professions and working status terms with their associated codes. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information.

Francisco M. Couto and Andre Lamurias. 2018. MER: a shell script and annotation server for minimal named entity recognition and linking. *Journal of Cheminformatics*, 10(1):58.

Archana Goyal, Vishal Gupta, and Manish Kumar. 2018. Recent Named Entity Recognition and Classification techniques: A systematic review. *Computer Science Review*, 29:21–43.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging.

Andre Lamurias and Francisco M Couto. 2019. Text Mining for Bioinformatics Using Biomedical Literature. In K. and Ranganathan, S., Gribskov, M., Nakai and C Schoonbach, editors, *Encyclopedia of Bioinformatics and Computational Biology, vol. 1*, January, pages pp. 602–61. Oxford: Elsevier.

Andre Lamurias, Pedro Ruas, and Francisco M. Couto. 2019. PPR-SSM: Personalized PageRank and semantic similarity measures for entity linking. *BMC Bioinformatics*, 20(1):1–12.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Arjun Magge, Ari Z. Klein, Ivan Flores, Ilseyar Alimova, Mohammed Ali Al-garadi, Antonio Miranda-Escalada, Zulfat Miftahutdinov, Eulalia Farre, Salvador Lima, Juan Banda, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Martin Krallinger, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. Overview of the sixth social media mining for health applications (#smm4h) shared tasks at naacl 2021.

Antonio Miranda-Escalada, Marvin Aguero, and Martin Krallinger. 2021a. Spanish covid-19 twitter embeddings in fasttext. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

Antonio Miranda-Escalada, Vicent Briva-Iglesias, Eulàlia Farré, Salvador Lima López, Marvin Aguero, and Martin Krallinger. 2020. ProfNER corpus: gold standard annotations for profession detection in Spanish COVID-19 tweets. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

Antonio Miranda-Escalada, Eulàlia Farré-Maduell, Salvador Lima López, Luis Gascó-Sánchez, Vicent Briva-Iglesias, Marvin Agüero-Torales, and Martin Krallinger. 2021b. The profner shared task on automatic recognition of occupation mentions in social media: systems, evaluation, guidelines, embeddings and corpora. In *Proceedings of the Sixth Social Media Mining for Health Applications Workshop & Shared Task*.

Appendix I

Creating Recommender Systems Datasets in Scientific Fields

Creating Recommender Systems Datasets in Scientific Fields

Marcia Barros
Francisco M. Couto
Matilde Pato
Pedro Ruas
marciabarros@edu.ulisboa.pt
fjcouto@edu.ulisboa.pt
mppato@fc.ul.pt
psruas@fc.ul.pt
LASIGE, Faculdade de Ciências, Universidade de Lisboa
Lisboa, Portugal

ABSTRACT

Recommender systems (RS) have been successfully explored in a vast number of domains, e.g. movies and tv shows, music, or e-commerce. In these domains we have a large number of datasets freely available for testing and evaluating new recommender algorithms. For example, Movielens and Netflix datasets for movies, Spotify for music, and Amazon for e-commerce, which translates into a large number of algorithms applied to these fields. In scientific fields, such as Health and Chemistry, standard and open access datasets with the information about the preferences of the users are scarce. First, it is important to understand the application domain, i.e. "what the recommended item is". Second, who are the end users: researchers, pharmacists, clinicians or policy makers. Third, the availability of data. Thus, if we wish to develop an algorithm for recommending scientific items, we do not have access to datasets with information about the past preferences of a group of users. Given this limitation, we developed a methodology, called LIBRETTI - LIterature Based RecommEndaTion of scienTific Items, whose goal is the creation of <user, item, rating> datasets, related with scientific fields. These datasets are created based on the major resource of knowledge that Science has: scientific literature. We consider the users as the authors of the publications, the items as the scientific entities (for example chemical compounds or diseases), and the ratings as the number of publications an author wrote about an entity. In this tutorial we will approach state-of-the-art recommender systems in scientific fields, explain what is Named Entity Recognition/Linking (NER/NEL) in research literature, and to demonstrate how to create a dataset for recommending drugs and diseases through research literature related to COVID-19. Our goal is to spread the use of LIBRETTI methodology in order to help in the development of recommender algorithms in scientific fields. More info about the tutorial at https://lasigebiotm.github.io/RecSys.Scifi/.

KDD '21, August 14–18, 2021, Virtual Event, Singapore

ACM ISBN 978-1-4503-8332-5/21/08.
https://doi.org/10.1145/3447548.3470805

CCS CONCEPTS

• **Information systems** → **Recommender systems**.

KEYWORDS

datasets, recommender systems, named-entity recognition, named-entity linking, COVID-19

ACM Reference Format:
Marcia Barros, Francisco M. Couto, Matilde Pato, and Pedro Ruas. 2021. Creating Recommender Systems Datasets in Scientific Fields . In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3447548.3470805

1 TUTORIAL OUTLINE

1.1 Recommender Systems

Te first part of the tutorial provides the theoretical knowledge about recommender systems in general, Named Entity Recognition (NER), Named Entity Linking (NEL) and their uses in recommender systems, creation and use of recommender systems in scientific fields (LIBRETTI).

- Introduction to recommender systems [18]
  - Types of recommender systems (collaborative-filtering, content-based, hybrid)
  - Types of users' feedback/ratings (implicit, explicit)
  - Challenges in Recommender Systems
- Scientific recommender systems [1, 3–5, 7, 10, 12, 20–24]
  - State-of-the-art scientific recommender systems, including fields (chemistry, astronomy, health, ...), type of datasets (what is being recommended to whom is being recommended), availability of the datasets.
- Introduction to Named Entity Recognition (NER) and Named Entity Linking (NEL) [6, 8, 9, 11, 13–17, 19]
  - Challenges of natural language
  - Text mining: typical pipeline, use cases in biomedical text
  - Named entity recognition: definition, applications and challenges in the biomedical domain, NER in recommender systems, types of systems (rule-based, machine-learning based, transfer learning)
  - Named entity linking: definition, applications, local vs global named entity linking, state-of-the-art approaches

- LIBRETTI pipeline [2–4]
  - Presentation of the LIBRETTI pipeline
  - Presentation of works which used LIBRETTI datasets in the fields of Astronomy, Chemistry, and Health.

1.2 Creating a recommendation dataset through scientific literature

The second part of the tutorial is hands on, where the participants learn how to use the approaches mentioned in Section 1.1.

- Retrieve the research articles related to COVID-19 [25]
  - The input used for creating the recommendation dataset will be a sample of the COVID-19 Open Research Dataset (CORD-19).
- Named Entity Recognition (NER) and Linking (NEL)
  - This step will show how to extract and normalize entities from the CORD-19 dataset. It includes the opening of each JSON file and the application of a NER tool on the title, abstract, body text and captions, according to the segmentation provided on the original corpus;
  - This step uses the tool MER [6] to recognize the entities in the text. The entities that we want to recognize are diseases and drugs, thus, we will use the Disease Ontology (DO), and the Chemical Entities of Biological Interest (ChEBI) ontology as input lexicons for MER.
- Creating the recommendation dataset
  - In this step we will use the methodology LIBRETTI for creating a standard dataset of <user, item, rating>. The users are the authors from the research articles, and the items are the entities recognized in the NER phase in each article. A similar approach is described in [4];
  - The output of this tutorial will be an open source recommendation dataset for diseases and drugs related to COVID-19, which may be used for training new recommendation algorithms.

1.3 Final discussion

In the third part of the tutorial, we will have an open discussion about the results, such as statistics and uses for the dataset, as well the final conclusions. More info about the tutorial at https://lasigebiotm.github.io/RecSys.Scifi/.

ACKNOWLEDGMENTS

This work was supported by FCT through funding of Deep Semantic Tagger (DeST) project (ref. PTDC/CCI-BIO/28685/2017), LASIGE Research Unit (ref. UIDB/00408/2020 and ref. UIDP/00408/2020), PhD Scholarship ref. SFRH/BD/128840/2017, PhD Scholarship ref. 2020.05393.BD

REFERENCES

[1] Marleen Balvert, Georgios Patoulidis, Andrew Patti, Timo M Deist, Christine Eyler, Bas E Dutilh, Alexander Schönhuth, and David Craft. 2019. A Drug Recommendation System (Dr. S) for cancer cell lines. *arXiv preprint arXiv:1912.11548* (2019).

[2] Márcia Barros, André Moitinho, and Francisco M Couto. 2019. Using research literature to generate datasets of implicit feedback for recommending scientific items. *IEEE Access* 7 (2019), 176668–176680.

[3] Marcia Barros, Andre Moitinho, and Francisco M Couto. 2021. Hybrid semantic recommender system for chemical compounds in large-scale datasets. *Journal of cheminformatics* 13, 1 (2021), 1–18.

[4] Marcia Afonso Barros, Andre Lamurias, Diana F Sousa, Pedro Ruas, and Francisco M Couto. 2020. COVID-19: A Semantic-Based Pipeline for Recommending Biomedical Entities. (2020).

[5] Jonas Boström, Niklas Falk, and Christian Tyrchan. 2011. Exploiting personalized information for reagent selection in drug design. *Drug discovery today* 16, 5-6 (2011), 181–187.

[6] Francisco M Couto and Andre Lamurias. 2018. MER: a shell script and annotation server for minimal named entity recognition and linking. *Journal of cheminformatics* 10, 1 (2018), 1–10.

[7] Charalampos Doulaverakis, George Nikolaidis, Athanasios Kleontas, and Ioannis Kompatsiaris. 2012. GalenOWL: Ontology-based drug recommendations discovery. *Journal of biomedical semantics* 3, 1 (2012), 1–9.

[8] John M Giorgi and Gary D Bader. 2020. Towards reliable named entity recognition in the biomedical domain. *Bioinformatics* 36, 1 (2020), 280–286.

[9] Philip John Gorinski, Honghan Wu, Claire Grover, Richard Tobin, Conn Talbot, Heather Whalley, Cathie Sudlow, William Whiteley, and Beatrice Alex. 2019. Named entity recognition for electronic health records: a comparison of rule-based and machine learning approaches. *arXiv preprint arXiv:1903.03985* (2019).

[10] Ming Hao, Stephen H Bryant, and Yanli Wang. 2018. A new chemoinformatics approach with improved strategies for effective predictions of potential drugs. *Journal of cheminformatics* 10, 1 (2018), 1–9.

[11] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991* (2015).

[12] Tsukasa Ishihara, Yuji Koga, Yoshiyuki Iwatsuki, and Fukushi Hirayama. 2015. Identification of potent orally active factor Xa inhibitors based on conjugation strategy and application of predictable fragment recommender system. *Bioorganic & medicinal chemistry* 23, 2 (2015), 277–289.

[13] Donghyeon Kim, Jinhyuk Lee, Chan Ho So, Hwisang Jeon, Minbyul Jeong, Yonghwa Choi, Wonjin Yoon, Mujeen Sung, and Jaewoo Kang. 2019. A neural named entity recognition and multi-type normalization tool for biomedical text mining. *IEEE Access* 7 (2019), 73729–73740.

[14] Andre Lamurias, Pedro Ruas, and Francisco M Couto. 2019. PPR-SSM: personalized PageRank and semantic similarity measures for entity linking. *BMC bioinformatics* 20, 1 (2019), 1–12.

[15] Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2013. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics* 29, 22 (2013), 2909–2917.

[16] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.

[17] Anand Rajaraman and Jeffrey David Ullman. 2011. *Mining of massive datasets*. Cambridge University Press.

[18] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2015. Recommender systems: introduction and challenges. In *Recommender systems handbook*. Springer, 1–34.

[19] Pedro Ruas, Andre Lamurias, and Francisco M Couto. 2020. Linking chemical and disease entities to ontologies by integrating PageRank with extracted relations from literature. *Journal of Cheminformatics* 12, 1 (2020), 1–11.

[20] Atsuto Seko, Hiroyuki Hayashi, and Isao Tanaka. 2018. Compositional descriptor-based recommender system for the materials discovery. *The Journal of chemical physics* 148, 24 (2018), 241719.

[21] Emre Sezgin and Sevgi Özkan. 2013. A systematic literature review on health recommender systems. In *2013 E-Health and Bioengineering Conference (EHB)*. IEEE, 1–4.

[22] Ekaterina A Sosnina, Sergey Sosnin, Anastasia A Nikitina, Ivan Nazarov, Dmitry I Osolodkin, and Maxim V Fedorov. 2020. Recommender systems in antiviral drug discovery. *ACS omega* 5, 25 (2020), 15039–15051.

[23] Benjamin Stark, Constanze Knahl, Mert Aydin, and Karim Elish. 2019. A literature review on medicine recommender systems. *(IJACSA) International Journal of Advanced Computer Science and Applications* 10, 8 (2019).

[24] Chayaporn Suphavilai, Denis Bertrand, and Niranjan Nagarajan. 2018. Predicting cancer drug response using a recommender system. *Bioinformatics* 34, 22 (2018), 3907–3914.

[25] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. 2020. Cord-19: The covid-19 open research dataset. *ArXiv* (2020).