Rui Marcos
Brandão Antunes

Extração de informação biomédica usando processamento de linguagem natural e aprendizagem automática

Biomedical information extraction with natural language processing and machine learning methods

Rui Marcos
Brandão Antunes

Extração de informação biomédica usando processamento de linguagem natural e aprendizagem automática

Biomedical information extraction with natural language processing and machine learning methods

Tese apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Engenharia Eletrotécnica, realizada sob a orientação científica do Doutor Sérgio Guilherme Aleixo de Matos, Professor Auxiliar do Departamento de Eletrónica, Telecomunicações, e Informática da Universidade de Aveiro.

Thesis presented to the University of Aveiro for fulfillment of the necessary requirements to obtain the degree of Doctor in Electrical Engineering, developed under scientific supervision of the Doctor Sérgio Guilherme Aleixo de Matos, Assistant Professor in the Department of Electronics, Telecommunications and Informatics at the University of Aveiro.

Trabalho financiado pela Fundação para a Ciência e a Tecnologia (FCT) no contexto do projeto IF/01694/2013/CP1162/CT0018 e da bolsa de doutoramento SFRH/BD/137000/2018. Financiado também pelo Programa Integrado SR&TD SOCA (Ref. CENTRO-01-0145-FEDER-000010), co-financiado pelo programa Centro 2020, Portugal 2020, União Europeia, através do Fundo Europeu de Desenvolvimento Regional.

o júri / the jury

presidente / president

Luís António Ferreira Martins Dias Carlos
Professor Catedrático da Universidade de Aveiro, Portugal
Full Professor at the University of Aveiro, Portugal

vogais / examiners committee

Francisco José Moreira Couto
Professor Associado com Agregação da Universidade de Lisboa, Portugal
Associate Professor with Aggregation at the University of Lisbon, Portugal

Anália Maria Garcia Lourenço
Professora Titular da Universidade de Vigo, Espanha
Associate Professor at the University of Vigo, Spain

Petia Georgieva Georgieva
Professora Associada com Agregação da Universidade de Aveiro, Portugal
Associate Professor with Aggregation at the University of Aveiro, Portugal

Carla Alexandra Teixeira Lopes
Professora Auxiliar da Universidade do Porto, Portugal
Assistant Professor at the University of Porto, Portugal

Sérgio Guilherme Aleixo de Matos
Professor Auxiliar da Universidade de Aveiro (orientador), Portugal
Assistant Professor at the University of Aveiro (supervisor), Portugal

palavras-chave Bioinformática · Extração de informação · Processamento de linguagem natural · Aprendizagem automática · Desambiguação de conceitos · Classificação de textos · Extração de relações.

resumo Assistimos a uma sobrecarga de dados textuais: uma quantidade avassaladora de informação é registada em texto de linguagem natural e armazenada em formato digital. Nas áreas ligadas às ciências da vida, o número crescente de publicações científicas no domínio da biomedicina e de relatórios clínicos retém uma riqueza de conhecimento que deve ser descoberto e associado através de métodos automáticos de extração de informação. Estes são essenciais para auxiliar a curadoria em bases de dados biológicos e desempenham um papel importante na descoberta de medicamentos, medicina de precisão, e investigação clínica.

Esta tese investiga o uso de processamento de linguagem natural, aprendizagem automática, e métodos baseados em conhecimento para extrair informação a partir de textos biomédicos em língua inglesa. Especificamente, estudamos e propomos métodos para desambiguação de entidades, classificação de documentos, e extração de relações. Em suma, este trabalho contribui com um estudo exaustivo de avaliação de várias abordagens para distintas tarefas de extração de informação biomédica, que são um suporte vital para o avanço do conhecimento atual.

keywords Bioinformatics · Information extraction · Natural language processing · Machine learning · Concept disambiguation · Text classification · Relation extraction.

abstract We witness an overload of textual data: a vast amount of information is recorded in natural language text and stored in digital media. In the life sciences fields, the increasing number of biomedical scientific publications and of clinical reports retains a wealth of knowledge that must be unearthed and linked through automatic information extraction methods. These are imperative to assist curation in biological databases and play an important role in drug discovery, precision medicine, and pharmacological and clinical research.

This thesis investigates the use of natural language processing, machine learning, and knowledge-based methods to extract information from biomedical text in English language. Specifically, we study and propose methods for entity disambiguation, document classification, and relation extraction. Overall, this work contributes with an exhaustive evaluation study of several approaches for distinct biomedical information extraction tasks, which are a vital support for the advancement of the current knowledge.

Table of contents

List of figures

List of tables

List of abbreviations

ACE . . . . . Automatic Content Extraction
ADE . . . . . Adverse Drug Event / Adverse Drug Effect
ADR . . . . . Adverse Drug Reaction
AI . . . . . Artificial Intelligence
ANN . . . . . Artificial Neural Network
AUC . . . . . Area Under the Curve

BERT . . . . . Bidirectional Encoder Representations from Transformers
BILOU . . . . . Beginning, Inside, Last, Outside, Unit-length
BiLSTM . . . . . Bidirectional LSTM
BIO . . . . . Beginning, Inside, Outside
BioGRID . . . . . Biological General Repository for Interaction Datasets
BioNLP . . . . . Biomedical NLP
BLLIP . . . . . Brown Laboratory for Linguistic Information Processing
BoW . . . . . Bag-of-Words
BRAT . . . . . Brat Rapid Annotation Tool

CDR . . . . . Chemical–Disease Relation
CNN . . . . . Convolutional Neural Network
CoNLL . . . . . Computational Natural Language Learning
COVID-19 . . . . . Coronavirus Disease 2019
CPI . . . . . Chemical–Protein Interaction
CPR . . . . . Chemical–Protein Relation
CPU . . . . . Central Processing Unit
CRF . . . . . Conditional Random Field
cTAKES . . . . . Clinical Text Analysis and Knowledge Extraction System
CTD . . . . . Comparative Toxicogenomics Database
CUI . . . . . Concept Unique Identifier

DDI . . . . . Drug–Drug Interaction

ELMo . . . . . Embeddings from Language Models

e-mail . . . . . Electronic Mail
EHR . . . . . Electronic Health Record
eMERGE . . . . . Electronic Medical Records and Genomics

GO . . . . . Gene Ontology
GPRO . . . . . Gene and Protein Related Object

HTML . . . . . Hypertext Markup Language

i2b2 . . . . . Informatics for Integrating Biology and the Bedside
ICD . . . . . International Classification of Diseases
IDF . . . . . Inverse Document Frequency
IE . . . . . Information Extraction

KBP . . . . . Knowledge Base Population
k-NN . . . . . K-Nearest Neighbors
KDD . . . . . Knowledge Discovery and Data Mining

LDC . . . . . Linguistic Data Consortium
LLL . . . . . Learning Language in Logic
LOINC . . . . . Logical Observation Identifiers Names and Codes
LR . . . . . Logistic Regression
LSTM . . . . . Long Short-Term Memory

MCN . . . . . Medical Concept Normalization
MEDLARS . . . . . Medical Literature Analysis and Retrieval System
MEDLINE . . . . . MEDLARS Online
MeSH . . . . . Medical Subject Headings
MGI . . . . . Mouse Genome Informatics
MIMIC . . . . . Medical Information Mart for Intensive Care
MLP . . . . . Multi-Layer Perceptron
MRD . . . . . Machine-Readable Dictionary
MUC . . . . . Message Understanding Conference

n2c2 . . . . . National NLP Clinical Challenges
NB . . . . . Naive Bayes
NCBI . . . . . National Center for Biotechnology Information
NER . . . . . Named Entity Recognition
NIST . . . . . National Institute of Standards and Technology
NLM . . . . . National Library of Medicine
NLP . . . . . Natural Language Processing

NLTK . . . . . Natural Language Toolkit
NP . . . . . Noun Phrase
NPMI . . . . . Normalized PMI

OBO . . . . . Open Biological and Biomedical Ontologies
OHNLP . . . . . Open Health NLP
OIE . . . . . Open IE
OMIM . . . . . Online Mendelian Inheritance in Man

PHI . . . . . Protected Health Information
PMC . . . . . PubMed Central
PMI . . . . . Pointwise Mutual Information
PMID . . . . . PubMed Unique Identifier
PoS . . . . . Part-of-Speech
PPI . . . . . Protein–Protein Interaction
PubMed . . . . . Public MEDLINE

RE . . . . . Relation Extraction
ReLU . . . . . Rectified Linear Unit
RNA . . . . . Ribonucleic Acid
RNN . . . . . Recurrent Neural Network

SDP . . . . . Shortest Dependency Path
SemEval . . . . . Semantic Evaluation
SNOMED . . . . . Systematized Nomenclature of Medicine
STS . . . . . Semantic Textual Similarity
SVM . . . . . Support Vector Machine

TAC . . . . . Text Analysis Conference
TDM . . . . . Text Data Mining
TEES . . . . . Turku Event Extraction System
TF . . . . . Term Frequency
TM . . . . . Text Mining
TREC . . . . . Text Retrieval Conference

UMLS . . . . . Unified Medical Language System
UniProt . . . . . Universal Protein Resource

VA . . . . . Veterans Affairs

WLPC . . . . . Wet Lab Protocol Corpus
WSD . . . . . Word Sense Disambiguation

Chapter 1

Introduction

Artificial intelligence (AI) has been widely embraced in our daily lives to help solving diverse tasks. AI is associated with giving machines the intelligence and ability to perform tasks or functions that would require human intelligence. Specific examples of the use of AI include poetry writing, image generation, car driving, medical diagnosis, and playing strategic games (Russell and Norvig, 2009).

Natural language processing (NLP) is a subfield of AI that is related with making machines able to 'understand' and communicate in human natural language, or simply perform processing of text (Jurafsky and Martin, 2008; Indurkhya and Damerau, 2010). It has many applications ranging from more simple tasks, such as sentence boundaries detection and text classification, to more complex tasks including text summarization, question answering, and machine translation. Using computers to automatically process text eases the analysis of large amounts of textual data.

Text data mining (TDM), or simply text mining (TM), involves the use of NLP methods to find new information from raw-text sources (Hearst, 1999; Hotho *et al.*, 2005; Allahyari *et al.*, 2017). Hearst (1999) argues that TDM makes use of text to directly discover heretofore unknown information. The author refers an example where text can be used to form hypotheses for causes of rare diseases. Information extraction (IE) is the process of creating structured information, for example saved in the form of a database, from unstructured data. As stated by Grishman (2015), NLP researchers commonly refer to text as *unstructured data*. However, in fact, natural language text has structure but it is not explicit—it is the goal of IE to make the text's semantic structure explicit. More precisely, in the particular case of textual data, IE encompasses the identification of relationships and their arguments (Grishman, 1997; Sarawagi, 2008; Grishman, 2019). This is related with TDM since new knowledge can be unearthed and inferred from automatically extracted relationships between specific concepts mentioned in text. For instance in the biomedical domain this could mean to find interactions between concepts such as

genes, diseases, chemical compounds, and food items.

Hearst (1999) also highlights the difference between information retrieval (or information access) and TDM. The former aims to help users find relevant documents (full-text or excerpts) according to their information needs (Baeza-Yates and Ribeiro-Neto, 1999), while the goal of the latter is to derive or discover new information from free text (for example, finding previously unnoticed patterns across several datasets).

In this work we investigate the use of NLP and machine learning methods for expediting IE in the biomedical domain. Large amounts of biomedical information, found in the life-sciences scientific literature and clinical narratives from electronic health records, are recorded in natural language text. Therefore, it is imperative to use automatic IE solutions that can create structured information, from this vast data, for further use by TDM approaches to discover new knowledge—it is inconceivable for an individual to read and interpret all this textual data.

We consider that biomedical IE comprises two major tasks: (1) named entity recognition (NER), which is responsible for identifying biomedical entities in free text (such as genes and chemicals); and (2) relation extraction (RE) which aims to determine biomedical relations between the previously recognized entities (such as protein–protein interactions). Nevertheless, biomedical IE can benefit from other NLP tasks. For instance, word sense disambiguation (WSD) aims to identify the proper meaning of an ambiguous term which is relevant to accurately define target entities. Another task is document triage—its goal is to rank or classify documents according to their importance given a pre-defined criterion. For example, it could be used to find relevant documents for extracting specific biomedical interactions.

1.1 Motivation

Much of the biological, medical, and clinical knowledge is recorded in natural language form. This textual data contains hidden relationships that can be exposed by automatic means, helping researchers to investigate new hypotheses that may contribute to a better health and well-being (for example, by finding potential treatments for specific diseases).

In the life sciences field the number of publications is increasingly high and it is hard for the interested audience—medical researchers, physicians, pharmacologists, and others—to keep up with the most recent research. Figure 1.1 shows an exponential growth of the number of publications indexed in MEDLINE, the leading biomedical bibliographic database compiled by the National Library of Medicine (NLM) of the United States. We see that in the last years, more than 800 thousand scientific articles have been

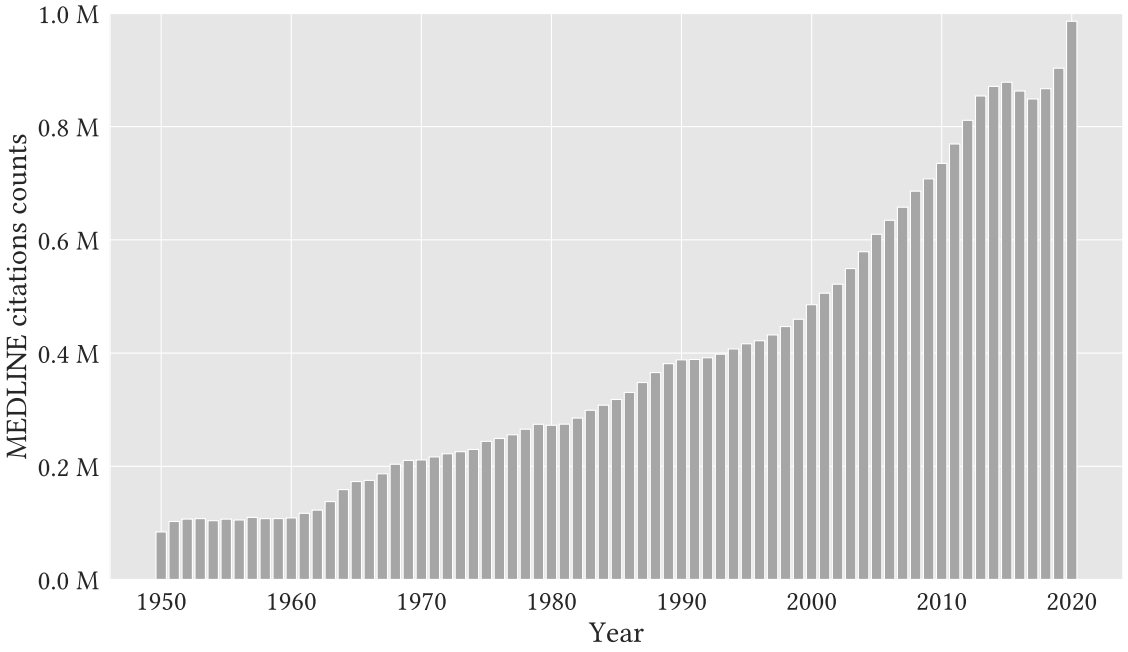

Figure 1.1: Number of MEDLINE indexed publications by year of publication from 1950 to 2020 (as of January 2022). Data retrieved from https://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html.

published per year. The COVID-19 epidemic (Velavan and Meyer, 2020) is a recent example showing that automatic methods are of utmost importance: they help specialists finding more appropriate resolutions for health problems.

Despite the undeniable value of biomedical free text present in scientific literature and clinical reports, the automatic processing of this type of text poses additional challenges when compared to text of the general domain. For example, the biomedical vocabulary is regularly updated with new terms from novel discovered concepts, many terms have distinct meanings within different contexts, and the use of abbreviations further accentuates this ambiguity. In the case of clinical text, this is even more difficult because abbreviations and typographical errors are more frequent.

The summarized aim of this work is to investigate the use of computer-based methods to extract relevant structured information from free text found in biomedical scientific literature and electronic health records.

1.2 Thesis structure

The remainder of this thesis is divided into five chapters. A short description of each chapter follows. Additionally, the rest of this first chapter contains a list of our

publications, a description of our open-source contributions, and clarifications about the document design.

Chapter 2 Preliminaries

We present the fundamental notions of biomedical NLP, some biomedical resources, and the evaluation metrics commonly employed to compare the performance of NLP systems. We explain in detail how NLP is used to process text, and how text is represented to be used by mathematical models. We give emphasis to the biomedical IE task, explaining its components, and what methods and paradigms are usually considered.

Chapter 3 Biomedical concept disambiguation

We enunciate related work describing automatic methods for disambiguation and normalization of biomedical terms in raw text. We describe two different approaches for performing biomedical WSD: supervised machine learning and knowledge-based. We use bag-of-words, word embeddings, and different weighting schemes for representing the texts, showing how these are effective for this task. In addition, we tackle the problem of normalization where we present a model based on word embeddings for linking clinical terms to standard vocabularies.

Chapter 4 Biomedical text classification and similarity measurement

We justify, supported by background work, how text classification is relevant to IE and how measures of text similarity are significant for higher-level biomedical NLP applications. We investigate the use of traditional machine learning classifiers, deep neural networks, and rule-based methods for text classification. Lastly, we explore the application of deep learning, with word embeddings and sentence embeddings, for measuring semantic textual similarity.

Chapter 5 Biomedical relation extraction

We describe related work on biomedical RE from scientific literature and explain the task. We present comprehensive experiments with convolutional and recurrent neural networks using word embeddings for identifying chemical–protein interactions (CPIs) in biomedical scientific text, showing that neural networks methods perform competitively.

Chapter 6 Conclusions

We discuss the overall work, highlight the main contributions, describe some limitations of our methods, and present future work directions.

1.3 Publications

The work described here resulted in several publications. A list of these, with a brief description, is presented chronologically by topic:

1. **Machine learning with word embeddings applied to biomedical concept disambiguation** (Antunes and Matos, 2016).

   We apply traditional machine learning methods in a supervised setting for biomedical word sense disambiguation. We combine bag-of-words and word embeddings to represent the surrounding textual context of ambiguous biomedical terms. We compare word embedding models created from PubMed and Wikipedia, concluding that domain-specific biomedical word embeddings consistently provide better results.
   This publication is relevant to Chapter 3.

2. **Biomedical word sense disambiguation with word embeddings** (Antunes and Matos, 2017a).

   We propose a knowledge-based method for biomedical word sense disambiguation. We use the UMLS (Unified Medical Language System) and the MeSH (Medical Subject Headings) term co-occurrences to extract concept textual definitions and calculate concept associations, respectively. Each concept is represented by a vector weighted by the embeddings of the words in the concept definition. We use word embedding models created from PubMed abstracts. For disambiguation, cosine similarity is used to measure the similarity between the surrounding context of the ambiguous term and each possible concept. We compare this knowledge-based method with machine learning methods using bag-of-words and word embeddings. Despite being outperformed by machine learning methods, our proposed knowledge-based method achieves a comparable performance, does not require training data as in a supervised setting, and can be applied to any biomedical ambiguous term that contains a curated textual definition.
   This publication is relevant to Chapter 3.

3. **Evaluation of word embedding vector averaging functions for biomedical word sense disambiguation** (Antunes and Matos, 2017b).

   We evaluate different word distance weighting schemes using our previously proposed knowledge-based method. This weighting scheme is used to give more importance to the words closest to the ambiguous term. We show that different

weight schemes impact the disambiguation performance, and an adequate weighting can improve it.
This publication is relevant to Chapter 3.

4. **Supervised learning and knowledge-based approaches applied to biomedical word sense disambiguation** (Antunes and Matos, 2017c).

   One limitation of our biomedical WSD approach in past publications (Antunes and Matos, 2017a,b) was that we did not use the whole dataset for testing our knowledge-based method, because we did not have access to textual definitions for every concept in the dataset. The main difference in this work is that we fetched textual definitions from UMLS for all concepts, enabling our results to be directly compared with other works in the literature. This article presents an exhaustive compilation of experiments with different settings using supervised learning and knowledge-based systems. We conclude that our knowledge-based method performs robustly, yet machine learning models provide higher performance but require labeled training data.
   This publication is relevant to Chapter 3.

5. **Clinical concept normalization on medical records using word embeddings and heuristics** (Silva *et al.*, 2020).

   We employ sieve-based models (comprised of several steps)[1], combined with heuristics and word embeddings, in clinical entity normalization. This involves linking clinical named entities—such as drugs, disorders, and procedures—to concepts in established medical terminologies. We show that the sole use of a strategy based on word embeddings presents competitive results.
   This publication is relevant to Chapter 3.

6. **Identifying relevant literature for precision medicine using deep neural networks** (Matos and Antunes, 2017a).

   We evaluate traditional classifiers against deep learning models for document classification, both using word embeddings, and show that deep neural network architectures obtain better results. Our methods performed competitively in a document triage task, which aimed to identify relevant PubMed abstracts that mention protein–protein interactions affected by genetic mutations.
   This publication is relevant to Chapter 4.

[1] Specifically, I was responsible for the first 'sieve' of the model which was based on biomedical word embeddings for representing clinical terms.

7. **Rule-based and machine learning hybrid system for patient cohort selection** (Antunes *et al.*, 2019).

   We propose an automatic system to identify which patients, given their clinical textual reports, meet or not meet certain criteria (for example, does the patient uses drugs or speaks English). The system is composed of rule-based methods with handcrafted text patterns and machine learning classifiers. We show that some criteria are more easily solved with simple heuristics while others are more complex, demand specialized clinical knowledge for designing appropriate text patterns, and benefit from the use of machine learning models.
   This publication is relevant to Chapter 4.

8. **Evaluating semantic textual similarity in clinical sentences using deep learning and sentence embeddings** (Antunes *et al.*, 2020).

   We present a deep neural network model for measuring the semantic similarity between clinical sentences. In this task, a real value is attributed to each pair of sentences to specify the degree of the semantic meaning they share. We assess the impact of using different pre-processing methods and feature representation methods (word embeddings against sentence embeddings).
   This publication is relevant to Chapter 4.

9. **Extraction of chemical–protein interactions from the literature using neural networks and narrow instance representation** (Antunes and Matos, 2019).

   We propose deep neural network models using word embeddings for extracting chemical–protein interactions from PubMed abstracts. Our best model was based on recurrent neural networks and only used information from the shortest dependency path between the target entities. We present an extensive study of experiments and make a detailed error analysis.
   This publication is relevant to Chapter 5.

Apart from the works summarized above, I colaborated in other works that are not discussed in this thesis. These are presented in chronological order:

1. Protein–protein interaction article classification using a convolutional recurrent neural network with pre-trained word embeddings (Matos and Antunes, 2017b).

2. Overview of the BioCreative VI Precision Medicine Track: mining protein interactions and mutations for precision medicine (Doğan *et al.*, 2019).

3. Understanding depression from psycholinguistic patterns in social media texts (Trifan *et al.*, 2020a).

4. Machine learning for depression screening in online communities (Trifan *et al.*, 2020b).

5. Automatic analysis of artistic paintings using information-based measures (Silva *et al.*, 2021b).

6. Chemical–protein relation extraction in PubMed abstracts using BERT and neural networks (Antunes *et al.*, 2021).

7. Chemical detection and indexing in PubMed full text articles using deep learning and rule-based methods (Almeida *et al.*, 2021).

8. Drug mention recognition in Twitter posts using a deep learning approach (Silva *et al.*, 2021a).

9. Chemical identification and indexing in PubMed full-text articles using deep learning and heuristics (Almeida *et al.*, 2022).

10. Chemical identification and indexing in full-text articles: an overview of the NLM-Chem track at BioCreative VII (Leaman *et al.*, 2023).

1.4 Open-source contributions

Some of the code developed for this thesis was made publicly available. For that purpose, the following two open-source repositories were created:

- `https://github.com/ruiantunes/2018-n2c2-track-1`
  This repository contains part of the source code developed from our participation in the 2018 n2c2 Track 1, comprising handcrafted rules and classical machine learning classifiers for automatic patient cohort selection (Antunes *et al.*, 2019).

- `https://github.com/ruiantunes/biocreative-vi-track-5-chemprot`
  This repository contains the source code of our deep learning–based RE extraction system for BioCreative VI Track 5 integrating all post-challenge improvements (Antunes and Matos, 2019). We also share our word embedding models created from PubMed abstracts, and detailed statistics about the task dataset and our system predictions.

1.5 Document writing and design

Preparing a thesis document, that readers find comprehensible and compelling, is a challenging task. As Zobel (2014) clarifies, the writing style must be adequate to communicate science: the text should be rigorous, readable, and based on logical thinking. The author also presents a comprehensive discussion of many aspects of scientific writing, which I frequently consulted for improving the writing of this document. Also relevant is a suitable design of the document by presenting its basic elements in an organized way. It makes the message of the author clearer, helping readers understanding it and keeping them interested (Schriver, 1989; Telg and McLeod-Morin, 2021). It is difficult, if not impossible, to mention all the works that influenced and inspired me for better designing and structuring this document. Still, I would like to point the reader to some major works that, in different ways, helped me to shape this document (Oliveira e Silva, 1994; Campos, 2013; Gal, 2016; Baker, 2017; Amos, 2019; Oleynik, 2020).

The LaTeX typesetting system was used to compose this document, and Inkscape was used for designing the vector images with the exception of Figure 1.1 that was created using Matplotlib. The source code for generating this thesis document is publicly available at:

`https://github.com/ruiantunes/ua-thesis`.

Chapter 2

Preliminaries

In this chapter we present background knowledge that is essential for a clearer understanding of the work presented in this thesis. We introduce the research field of natural language processing (NLP) and detail its most common tasks. Next, we clarify how computational text representation has shifted from simple high-dimensional binary vectors, that carry information if a specific word is within a text, to advanced representations based on real-number vectors that aim to represent the lexical meaning of words in a lower dimensional vector space.

Then, we describe what is information extraction detailing some of its major tasks, methods, and evaluation techniques commonly employed. Finally, we reveal how biomedical text mining involves the application of NLP to extract information from text found in the scientific literature and electronic health records. Related community-wide challenges and several biomedical resources are highlighted.

Some of the content presented in this chapter is based upon well-established literature, to which we point the reader for further consulting (Manning *et al.*, 2008; Bird *et al.*, 2009; Nadkarni *et al.*, 2011; Ingersoll *et al.*, 2013; Jurafsky and Martin, 2018).

2.1 Natural language processing

Natural language processing, a subfield of artificial intelligence, investigates how a machine may 'understand' natural language. It includes a variety of text processing tasks, ranging from simpler exercises, such as finding the boundaries of sentences within a text, to more difficult challenges that can make machines answer questions, summarize or translate documents, or even maintain a human-like conversation. Machines that can successfully address those more difficult NLP tasks are often referred to as 'intelligent' because they show similar language reasoning abilities to those of humans.

To give some context, Turing (1950) defied the computational linguistics community

by reformulating the question 'Can machines think?' so that it could be expressed in a less ambiguous form. The author further presented the *imitation game* (also known as the *Turing Test*) which, in short, is a simple assessment for evaluating if a machine can show intelligence similar to humans by communicating, as good as a human, through natural language (that is, if it is capable of imitating human answers). Since then, a lot of research has been conducted regarding automatic (computational) processing of natural language—a very short historical roadmap about the field of NLP follows.

According to Nadkarni *et al.* (2011), NLP had its beginning around the 1950s with the intersection of the fields of artificial intelligence and linguistics. Rindflesch (1996) presents a solid review of early NLP tasks that were solved using statistical techniques, probabilistic models, and rule-based approaches; and discusses some of the NLP applications such as information retrieval and question-answering systems. According to Marquez and Salgado (2000), the application of machine learning drawn attention in the NLP community around the early 1990s addressing mainly natural language disambiguation.

However, in recent years deep learning has shown great promise in several computational research areas and NLP is no exception. Deep learning–based strategies have established the state-of-the-art in many tasks achieving sometimes performance almost as good or better than a human. Computational linguistics and deep learning have a huge potential that has been explored and there is still plenty room for investigation and future testing, with still much exciting experiments to offer (Manning, 2015).

One of the most important applications of NLP in the biomedicine field is that these technologies help biologists, medical researchers, and physicians with tools that provide automatic annotation of text. These are also fundamental for biocuration and help to keep biomedical databases and ontologies up-to-date (Singhal *et al.*, 2016a).

In this section we detail the many tasks of NLP which deal with the basic processing of text. These are required to construct much more complex tasks such as those of information retrieval or extraction. These tasks are usually addressed using heuristics, rule-based approaches, or machine learning strategies.

2.1.1 Tasks

The field of natural language processing has many applications and it is in constant evolution with new use cases often emerging. However, the main fundamental NLP tasks have remained essentially the same throughout the years. Processing natural language usually consists in employing a variety of methods that solve simpler to more complex tasks, depending on the problem at hand. The simpler, low-level, tasks constitute the main processing blocks, which are then used when addressing more complex, high-level, tasks. We start by enumerating and briefly explaining some of the key *lower-level* NLP

tasks:

- *Sentence boundaries detection*, also referred to as *sentence splitting*, is the task of identifying the boundaries of each sentence, that is, where sentences start and end. It is relevant for further processing steps that take advantage of processing one sentence at a time.

- *Tokenization* consists in splitting an excerpt of text into *words*, though some specific tokenizers have a more granular approach and further split words into *subword units*, also known in the literature as *wordpieces* (Wu *et al.*, 2016). Each individual (sub)word is considered a *token*. Often this method is performed in a sentence (that is, after sentence splitting), though it can be applied in shorter phrases or longer excerpts such as a paragraph or a full-text document without sentence splitting.

- *Stop words removal* discards words that may be considered noisy or irrelevant for a specific language understanding task. Usually, pre-compiled lists of stop words, found in several NLP libraries and databases, are employed; though methods that dynamically identify the stop words are also used, for instance by finding highly frequent non-informative words in a collection of text documents (such as 'the', 'and', 'or'). This technique is often employed for tasks such as document or topic classification, where main terms related to a certain topic contribute for a successful prediction. For instance, Saif *et al.* (2014) studied whether removing stop words is effective for sentiment analysis of Twitter posts.

- *Stemming* and *lemmatization* are two different techniques with the same goal, to convert or normalize related forms of a word to a common base form. For example, *activate*, *activates*, and *activating* could be chopped off to *activat*. This is what stemming performs, it employs a heuristic process that removes the ends of words (even if the resulting word form does not exist in the language). One of the most known stemmers for the English language is the Porter stemmer (Porter, 1980). On the other hand, *lemmatization* finds the root form of a word within a vocabulary corresponding to the base or dictionary form of a word, known as *lemma* (usually a verb or a noun). The resulting *lemma* of the previous example would be *activate* (verb) or *activation* (noun).

- *Part-of-speech tagging* consists in attributing to each word in a text its part-of-speech, that is, identifying whether its *lexical category* is a noun, verb, adjective, or other. These part-of-speech (PoS) categories, also known as *word classes*, are helpful for downstream tasks such as text chunking and named entity tagging.

Probably, the most well-known and established PoS tagset for English is the one from the Penn Treebank Project[1] (1989–1992) which we point the reader for further consultation (Santorini, 1990; Marcus *et al.*, 1993).

- *Text chunking* or *shallow parsing* splits a sentence into contiguous and non-overlapping segments according to the part-of-speech tagged tokens. Consecutive tokens are grouped into major grammatical units such as noun phrases, verb phrases, adjective phrases, and prepositional phrases. Consult Abney (1991) and Sang and Veenstra (1999) for further reading.

- *Dependency parsing* identifies the grammatical relations between words within a sentence, thus exposing its syntactic structure. The relations are directed and labeled from a fixed set of grammatical relations or *dependencies*, forming a structured tree of relationships known as *dependency tree*. The Stanford typed dependencies representation provides a standard description of grammatical relationships (de Marneffe and Manning, 2008, 2016), which was later extended and improved according to the Universal Dependencies representation (Schuster and Manning, 2016).

- *Passage segmentation* is responsible for finding and splitting the constituent sections of a full-text document. Often a complete document contains several parts, and each of these sections may require a different text processing. For example, in clinical reports, one section may have a list of medical prescriptions (with dosages, route of intake, and other relevant information), other section can contain a table with clinical analysis results, and other sections may simply have notes in raw text.

- *Semantic role labeling*, also known as *shallow semantic parsing* or *slot-filling*, identifies semantic relationships, within a sentence, between noun and verb phrases with thematic roles such as *agent*, *instrument*, or *destination* (Gildea and Jurafsky, 2000, 2002; Jurafsky and Martin, 2018). As Jurafsky and Martin (2018) further explain, this task aims to answer how participants relate to events addressing questions like "who did what to whom" and "when and where".

We have identified most of the major *low-level* processing tasks of natural language, which are commonly employed in the implemententation of *higher-level* tasks, that solve specific problems, such as:

- *Word sense disambiguation* aims to find the correct senses of ambiguous terms

[1] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

given their surrounding contexts. Natural language contains many equivocal expressions and frequently it is open to interpretation. Therefore, finding the proper meaning of ambiguous text is relevant for accurate information extraction.

- *Named entity linking* or *concept normalization* attributes unique identifiers—from databases, terminologies, or vocabularies—to terms identified in the text. Often, this task is tackled jointly with the sense disambiguation task because it requires linking every ambiguous expression to a unique meaning using a specific identifier.

- *Document classification* consists in categorizing documents according to predefined criteria. For instance, documents can be classified within several topics or simply as relevant or not (example of binary classification).

- *Named entity recognition* detects the spans of text that refer to specific entities (or concepts). For example in the biomedical domain, this task is used to detect diseases, adverse effects, chemicals, and others.

- *Relation extraction* identifies interactions between named entities in the text. Traditionally, this task started by a *trigger recognition* step where the term expression—usually a verb—involving the two entities would be firstly detected. However, with the recent advance of machine learning, this separate step for trigger recognition was discarded due to more recent NLP models that are able to perform relation extraction in a single step, usually through the use of distributed word representations.

- *Text summarization* is the task of creating a short excerpt of a few sentences or paragraphs, given a larger input text, containing the most relevant information presented in the original document.

- *Question answering* addresses the extraction or generation of textual answers to questions, which can be based on text data alone or by exploiting external information from knowledge sources.

- *Machine translation* is the task of converting text in one language to another. For instance, given a clinical narrative in Spanish, transform it to English.

All the aforementioned *higher-level* tasks are related to the broader concept of measuring the *semantic textual similarity* between two text excerpts. For instance, sense disambiguation and concept normalization compare the surrounding context of an ambiguous term to other contexts of other senses, identifying the *closest* one and therefore

the likely meaning. Similarly, supervised relation extraction models identify interactions between two entities by using and *interpreting* their surrounding context.

2.2 Text representation

Representing text in a numerical form is what allows mathematical models to be used for automatic processing of natural language. Early approaches for text representation were based on binary vectors that indicated the presence or absence of a specific word from a determined vocabulary. These evolved to integer vectors that could tell how many times a word appeared in a text, and then more elaborated formulas appeared for giving a certain importance to each word. One of the most fundamental NLP tasks is the splitting of the text into basic units such as words, also commonly referred to as tokens. This process is known as tokenization and is the basis for representing the text.

2.2.1 One-hot encoding and bag-of-words

A simple way of thinking on how to convert a text document into a numerical representation is to simply attribute an integer number to each word in the vocabulary. In this context, a vocabulary is the set of distinct words found in a collection of documents, or corpus.

One-hot encoding is a common technique for representing words using binary vectors. First, the vocabulary is built, and the length of the vocabulary is considered the length of the binary vector. Then, every word is attributed an integer number, corresponding to its index in the vocabulary, and a document is represented by a vector containing mostly zeros with ones indicating the words that are present in the document.

The bag-of-words (BoW) technique consists in representing a text by specifying only its constituent words and their respective occurrence or frequency—their position in the text are ignored. The most commonly used weighting scheme applied with BoW is the TF–IDF (term frequency–inverse document frequency).

The TF–IDF mechanism has the goal to give higher importance to most common words within a document, but reduce importance to words that are frequent across the corpus. For instance, considering the English language, words such as ‘the’ and ‘of’ are very frequent in a single document, but also in a set of documents, therefore carrying little information about a specific subject.

These techniques allow a simple and fast numerical representation of text that can achieve strong performances in several NLP tasks relevant for information extraction such as document classification. In this case, each document corresponds to a vector

where each dimension corresponds to a specific word, and its value can be the frequency count or the TF–IDF value. Commonly, this provides a solid baseline that can be used as a starting point for further development and improvement. However, one drawback of this method is that does not benefit from the sequential order in which the words appear, that is, the real context and meaning of a structured sentence.

2.2.2 Distributional semantics and word embeddings

Distributed representations of words, also known as word embeddings, are lower-dimensional vector representations of words. These are estimated from large corpora with millions or billions of words. Although distributed word representations were proposed before, the first well-known algorithm for efficiently calculating these word embeddings is word2vec (Mikolov *et al.*, 2013a,b), and then other models such as GloVe appeared (Pennington *et al.*, 2014). These methods calculate fixed word vectors using deep learning models that during training try to predict the surrounding context given the target word, or try to predict the target word given the surrounding context.

These word vector representations are used frequently with deep neural network architectures. An alternative type of representation is *character embeddings* where each character is represented by a single vector. These have been proven to be useful also for a variety of NLP tasks since these can encode and carry further information—for example they can retain information regarding prefixes and suffixes.

2.3 Information extraction

This section gives an overview of the *Information Extraction* task. It presents background work, state-of-the-art methods, and summarizes the description of many subtasks. Within the text data mining field, information extraction is the process of using computerized and automatic methods to *discover knowledge* from digital media sources such as textual data. However, in a more broad sense, information extraction is often related with simply distilling data from document sources, and therefore, in fact, does not always imply *discovering new knowledge*.

Arguably, the most addressed tasks for information extraction from free text are *named entity recognition* and *relation extraction*. The former is responsible for identifying terms or concepts in the text, whereas the latter finds relationships between those concepts. Automatic annotation of entities and their relations is relevant to generate new hypotheses that help to create new knowledge. For instance, particularly within the biomedical domain, adverse drug events (ADEs) can be determined, and candidate drugs or therapies for specific health problems can be detected from clinical reports.

The implementation of these tasks require the use of NLP methods which are able to handle and transform textual data. These are usually employed in a first stage of data preparation for representing the text in a numerical form, which can then be interpreted by mathematical models such as machine learning methods.

Information extraction can be thought as the task of identifying relevant information in the text such as entity mentions and interactions (or relations) between those. Linking (normalizing) entity mentions to standard terminologies or ontologies, or disambiguating them is also considered to be part of information extraction. Classifying documents as pertinent to discover specific knowledge is also considered a relevant task for information extraction.

We also explain the basics of major NLP tasks including word sense disambiguation, document classification, entity recognition, and relation extraction. Frequent methodologies including rule-based approaches, machine learning models, and deep neural networks are briefly introduced. We also clarify different learning paradigms, used in information extraction, such as supervised, semi-supervised, distant supervision, unsupervised or knowledge-based, and explain how they make use of labeled and unlabeled data, and external information sources such as (curated) databases. Additionally, we highlight the differences between pipeline and joint learning methods. Concurrently, we present a brief overview of the biomedical information extraction research area, and detail further background work for different tasks in each chapter separately.

2.3.1 Tasks

There are a plethora of information extraction tasks. In this section we highlight the ones that we considered more relevant according to the work developed for this thesis. These primarily include the retrieval or classification of documents relevant for text mining, and further identification of named entities and their relationships. These are the major steps required to help database curation and potentially find new hypotheses for testing.

Document classification

Document classification is the task of categorizing documents, for example by labeling them relevant or irrelevant regarding predefined criteria. This task can be employed in an IE pipeline where the first step is to identify potentially relevant documents for text mining. Document classification is sometimes interwoven with document retrieval, where a large collection of documents is present, and the most relevant regarding a specific subject need to be selected; in these case, the advantage is that only a smaller

subset of documents needs to be processed with further IE tasks, alleviating the need of computational power.

Earlier approaches for classifying documents included rule-based methods built with heuristics. Though, with the increasing of hand-labeled data, traditional supervised machine learning methods started to be applied using bag-of-words features. Most recent works have been using deep learning architectures with word embeddings for text classification (Yang *et al.*, 2016; Fergadis *et al.*, 2018). These methods have been applied for a variety of purposes including finding depressive tendencies in written text (Yates *et al.*, 2017) and performing sentiment analysis (Cambria *et al.*, 2013).

Classification of documents can expedite next information extraction steps since irrelevant, or less relevant, documents can be discarded beforehand. The removal of this unnecessary texts saves valuable processing time that can be employed for more important tasks such as finding relationships between relevant terms. However, the document classification task may not be necessary when it is the aim—and there is enough computational power—to analyze all the documents of a collection, or the number of documents is relatively small.

Named entity recognition

Named entity recognition (NER) is a major task in information extraction from text. Its aim is to identify named entities such as persons, organizations, locations in the case of general-domain text; or chemicals, diseases, and adverse drug effects in the case of biomedical text. This is the most fundamental task for enabling relationship extraction between the pre-detected named entities. For example, considering biomedical text, chemical–protein or drug–drug interactions can be found, and verified or added to external databases by expert curators.

Earlier approaches for NER were based on machine-readable dictionaries (MRDs) that contain lists of entity terms of different types. This strategy constitutes one of the most simple method for entity recognition, since it is only required to match the entity terms within the text—for that case, regular expressions can be used. Regular expression is a method that facilitates the programming of finding text patterns in a concise way, and it has been used for several years in many text processing related areas.

Afterwards, named entity recognition was addressed as a sequence labeling problem, where each token is labeled as part of an entity or not. This is the most common way to solve NER, since it provides a simple way for detecting the entity boundaries (character offsets), and classify the entity type if it is the case of multiple entity classes. Conditional random fields (CRFs) are statistical models that have been largely used in natural language processing. These do not predict the label of each token independently, but

take context (neighboring tokens) into account.

To the best of our knowledge, Ramshaw and Marcus (1995) proposed the {I, O, B} chunk tag set when implementing a rule-based tagging system for text chunking. In their work they used the "I" tag to mark words inside some noun phrase (NP), the words marked with "O" are outside the NP, and the "B" tag is used to mark the first word of a NP which immediately follows another NP.

However, different representations for chunking and named entity recognition (NER) have been proposed and been used as well. For instance, Sang and Veenstra (1999) examined seven different representations for the problem of recognizing NP chunks, where they cite the work of Ratnaparkhi (1998) highlighting that in his work all the chunk initial words receive the same start tag differently from the representation used by Ramshaw and Marcus (1995). Sang and Veenstra (1999) refer to the tagging formats proposed by Ramshaw and Marcus (1995) and Ratnaparkhi (1998) as "IOB1" and "IOB2" respectively. They also present the "IO" partial representation in which words inside a NP receive an "I" tag and others receive an "O" tag. However, this encoding is insufficient since adjacent NPs or named entities cannot be distinguished.

In the NER task across multiple domains, Ratinov and Roth (2009) found that the BILOU (also known as IOBES) representation of text chunks significantly outperforms the widely adopted IOB scheme. In the biomedical domain, Dai *et al.* (2015) studied the effect of different tagging schemes showing that the IOBES scheme obtained better results than the IOB scheme in the recognition of mentions of chemical entities. However, in another work in NER, Lample *et al.* (2016) did not observe a significant improvement of the IOBES tagging scheme over the IOB tagging scheme.

Different names for the same tagging schemes have been used. Table 2.1 presents an incomplete list of these nomenclature variations and their abbreviations. For more details, we point the reader to other works reporting investigation of different tagging schemes (Kudo and Matsumoto, 2001; Cho *et al.*, 2013). An example, adapted from Cho *et al.* (2013), with the use of the "IO", "IOB2" and "IOBES" representations is shown in Table 2.2.

Named entity linking

Named entity linking, or entity normalization, is the task of attributing unique codes from standard terminologies to the previously detected entity mentions. This step is usually performed after the named entity recognition task, where entities are predicted, but are not linked to a curated ontology or database. This task also addresses word sense disambiguation (WSD), since it is guaranteed that every entity mention has a specific meaning because it is connected to a single code.

Table 2.1: Tagging schemes, their nomenclatures and abbreviations.

Nomenclatures	Initials	Meanings
IO	I	Inside
	O	Outside
IOB, BIO	I	Inside
	O	Outside
	B	Beginning
IOBES, BILOU	I	Inside
	O	Outside
	B	Beginning
	E / L	End / Last
	S / U	Single / Unit-length

Table 2.2: A sample text annotated with different tagging schemes. Adapted from Cho *et al.* (2013).

Tokens	IO	IOB2	IOBES
Gamma	I-gene	B-gene	B-gene
glutamyl	I-gene	I-gene	I-gene
transpeptidase	I-gene	I-gene	E-gene
(	O	O	O
GGTP	I-gene	B-gene	S-gene
)	O	O	O
activity	O	O	O
in	O	O	O
the	O	O	O
...	...	...	...

The task is usually solved using sieve-based approaches, where in each sieve (step) it is performed a different strategy. It is common that the first sieves are based on string matching patterns using dictionaries from standard terminologies, followed by more advanced techniques for example including measuring similarities using word embeddings. The resolution of this task is relevant since it helps in the relationship extraction task, and guarantees exactly which concepts are being referred, which is important for accurate information extraction.

Relation extraction

Relation extraction (RE) or relationship extraction aims to identify associations between specific entities in the text. This task commonly follows the entity recognition task where named entities are firstly properly identified. The relations detected and discovered are then relevant to help curators keep their databases well-verified and up-to-date.

Traditionally this task started with the identification of a trigger for a relation. This is known as trigger recognition, where for example a verb could identify the relation between two concepts. Campos *et al.* (2014) present a machine learning model for biomedical event trigger recognition. They use a CRF with a comprehensive feature set achieving an F-score of 0.627 in the BioNLP 2009 shared task corpus.

2.3.2 Pipelined *versus* joint extraction

Traditionally, extraction of entities and relations have been addressed by solving two ordered tasks: NER and relation extraction. This is commonly referred in literature as the *pipeline* approach, since it combines two separate tasks where the relation extraction requires beforehand the named entities.

There are two main paradigms for extracting information:

- Pipelined: concept recognition followed by relation extraction.
- Joint extraction where entities and relations are simultaneously extracted.

Pipelined extraction

Pipelined extraction is built with two separate steps: named entity recognition is followed by relation extraction. In this approach, these steps require training two different models. One weakness of this approach is that the NER step can propagate errors into the last step (Li and Ji, 2014).

Joint extraction

Joint extraction is the task of cooperatively identifying entities and their semantic relations from free text. Traditional joint methods require complex feature engineering or heavily rely on other NLP tools. However, the use of external tools might lead to error propagation where wrongly detected named entities negatively impact the relation extraction task. In recent years, the use of neural networks (deep learning) has been investigated for developing end-to-end models dramatically reducing the manual effort in feature extraction.

In this section we present works that either apply feature-engineered or deep learning methods, discussing what are the conveniences over each other.

Handcrafted feature engineering First works on joint extraction were built with two separate models, where one was responsible for NER and the other performed relation extraction. Afterward, works using only a single joint model were proposed. To the best of our knowledge, Li and Ji (2014) were the first to present a single joint model to predict entities mentions and relations. The entity and relation extraction tasks they addressed were those of the Automatic Content Extraction (ACE) program (Doddington *et al.*, 2004) presenting results on the ACE04 and ACE05 corpora. In contrast to previous research where entity mentions are assumed to be given, this work aimed to investigate an end-to-end model for NER and relation extraction. For comparison, they developed a baseline pipeline system composed of a CRF for entity mention extraction and a maximum entropy model for relation extraction. In the ACE04 corpus their joint model achieved a 0.453 F1-score outperforming previous works and their pipeline baseline model (0.429). In the ACE05 corpus their proposed joint model achieved a 0.495 F1-score, whereas human annotators obtained an F1-score about 0.70 with an inter-annotator agreement of 0.519, showing how end-to-end relation is challenging.

Miwa and Sasaki (2014) introduced a flexible table representation of entities and relations from a single sentence. They employed the BILOU tagging scheme assuming that entities do not overlap. They evaluated their model in the CoNLL04 dataset (Roth and Yih, 2004). Their F1-score metrics showed improved performance in using joint learning (0.610) over a pipeline approach (0.577).

Ren *et al.* (2017) combined joint extraction of entities and relations with distant supervision. They evaluated their model in three datasets from different domains (news articles, Wikipedia articles, biomedical abstracts).

Deep learning Miwa and Bansal (2016) presented the first neural network based model for joint extraction of entities and relations. Their end-to-end model represents word sequence and dependency tree structures by using bidirectional sequential and tree-structured LSTM (long short-term memory) networks. In the ACE04 and ACE05 datasets they achieved F1-scores of 0.484 and 0.556 respectively.

Katiyar and Cardie (2017) employed an attention-based BiLSTM model for joint entity and relation extraction. They made no use of dependency trees neither PoS tags, using only the surface form (sequence of tokens) and achieved competitive results compared to the previous work of Miwa and Bansal (2016).

Zheng *et al.* (2017b) proposed a novel tagging scheme for jointly extracting entities and relations. Their tagging scheme expands the BILOU scheme. Besides each token's

tag having associated information about its position within an entity, it also contains information about the relation and if it is the first or second entity in a triplet. Their end-to-end model, based on LSTM networks, achieved the best results for a dataset created by distant supervision means (Ren *et al.*, 2017). However, their method cannot handle overlapping relations.

Li *et al.* (2017) presented another neural joint model. They evaluated their model in two tasks: the task of extracting ADEs between drug and disease entities, and the task of extracting resident relations between bacteria and location entities. They used parameter sharing to join two BiLSTM networks for extracting entities and relations. They used the BILOU labeling scheme for entity recognition.

Adel and Schütze (2017) applied global normalization of convolutional neural networks. Following Miwa and Sasaki (2014), and Gupta *et al.* (2016) they tackled the problem as a table filling task and did not require to transform it into a token-labeling problem.

Zheng *et al.* (2017a) proposed a hybrid neural network composed of a LSTM for entity extraction and a CNN (convolutional neural network) for relation classification. However, the first layer is a BiLSTM encoding layer which is shared for both tasks. They evaluated their model in the ACE05 dataset surpassing previous works.

Verga *et al.* (2018) proposed a bi-affine relation attention network that simultaneously extract relations at the document-level. They also employed strong distant supervision to create a new dataset, from PubMed abstracts and the Comparative Toxicogenomics Database (CTD), for biological relation mining (chemical-disease, chemical-gene, gene-disease).

Bekoulis *et al.* (2018a) demonstrated that the use of adversarial training, by adding noise to the word representations, improves joint extraction of entities and relations from datasets of different domains. Moreover, their model with adversarial training achieved high performance in the first epochs during the training process.

Bekoulis *et al.* (2018b) employed a BiLSTM encoding layer, a CRF for entity recognition, and a sigmoid function for relation extraction. They conducted a large scale experiment achieving state-of-the-art results in corpora from different domains (news, biomedical, real estate) and languages (English, Dutch).

More recently, Eberts and Ulges (2019) proposed a span-based joint entity and relation model that uses as its core a pre-trained BERT (bidirectional encoder representations from transformers) network (Devlin *et al.*, 2019). Their span-based approach considers that any token subsequence is a potential entity, and that any pair of spans can have a relation. A full search over all span and relation candidates is performed. They claim that one advantage of this approach over the use of BIO or BILOU tagging schemes is

that it can identify overlapping entities (for example, "codeine" within "codeine intoxication"). In their relation classifier they considered the context between the two entities, since it showed to be more profitable than using the whole sentence. They reported state-of-the-art results in three datasets: CoNLL04, SciERC (Luan *et al.*, 2018), and ADE (Gurulingappa *et al.*, 2012b).

Wadden *et al.* (2019) followed the work of Luan *et al.* (2019) additionally performing event extraction, and building span representations on top of multi-sentence BERT encodings. They experimented on four datasets: ACE05, SciERC, GENIA (Kim *et al.*, 2003) and WLPC (Kulkarni *et al.*, 2018).

Luo *et al.* (2020a) proposed a new tagging scheme to represent both entities and relations. Their scheme can represent some overlapped relations, which provides better performance in comparison to other works. They employ a BiLSTM-CRF model with an attention mechanism. They also had to define their own extraction rules according to the proposed scheme.

2.3.3 Methods

There are a variety of methods for extracting information. Earlier methods were built manually implementing handcrafted rules. Then, rule-based and knowledge-based methods followed. And during the past years, due to the the much-higher availability of computational power and also labeled text data, machine learning methods started to thrive, particularly deep learning models that perform better with large amounts of data.

Handcrafted rules are usually constructed using if-else statements or using regular expressions that can find patterns in text. This approach is usually efficient in terms of computational performance, but requires some expertise in the field for constructing these rules, and the accuracy results may not be the best.

Traditional machine learning models also have had great success for document classification, entity recognition, and relation extraction. These models include kernel classifiers such as support vector machine, k-neighbors classifier, decision tree, and probabilistic methods such as the Naive Bayes classifier.

Due to the access to big data, and increasingly gold-standard labeled data, deep learning models have been successfully claiming the state-of-the-art for many language processing tasks. Deep models include convolutional and recurrent neural networks, attention mechanisms, generative adversarial networks, and transformers.

Machine learning

Machine learning models are based on statistical and probabilistic methods that make use of training data to *learn* a specific task. These are used for building robust predictive systems in different scenarios and applications including computer vision and speech recognition, though in here we briefly present only some of the most known or used methods applied in natural language processing.

Naive Bayes This is known to be a straightforward classifier based on a probabilistic method, which has worked well in many different data mining tasks including document classification. These models are much faster compared to more complex methods, and usually do not require large amounts of training data.

k-nearest neighbors This classifier is often used as a solid baseline classifier in many different machine learning tasks. It also provides a robust approach for a variety of NLP problems including document classification. It follows a simple algorithm where a sample is classified as the majority vote of its neighbors. That is, if the majority of the nearest points in the space are from a specific class, then the sample will have this class as prediction. This method is also straightforward and fast to compute, and does not require much training data for reasonably good results.

Decision tree This classifier is also a simple algorithm yet effective in a variety of tasks also providing a solid baseline for machine learning problems. It works by building a flowchart-like, from the root node to low-level nodes, with consecutive binary decisions according to the values of the features. The more depth the tree, the more complex, questions become more refined, which may provide better results or overfit to the training data if there is excess 'memorization' on noisy patterns of the data.

Support vector machine This model is one of the most used in traditional machine learning, and often provides strong results. This classifier tries to separate training examples from different classes by maximizing their distance in the space. New samples are then predicted according to which side they belong to. SVMs are efficient when dealing with high-dimensional feature spaces.

Conditional random field Lafferty *et al.* (2001) proposed the conditional random field model which has been used in computer vision and natural language processing for sequence labeling problems. During the last years, these models combined with LSTMs have been achieving state-of-the-art results (Lample *et al.*, 2016).

Deep learning

Deep learning has been an emerging AI research area in the last years, being considered a subfield of machine learning. It is related with the use of (deep) neural network models for machine learning tasks. Neural network models, built primarily from simple math-operation cells (neurons), were inspired from the human brain structure and have been researched during the last decades (Hinton, 1992). However, in the last years they have shown greater performance due to improved training methodologies, increased computational power, and higher availability of labeled data for training.

The most common and standard neural network, also known as multi-layer perceptron (MLP), is composed of one input layer and one output layer, and can have multiple intermediate (hidden) layers. If there is at least one hidden layer then the model is considered to be a *deep neural network*—hence the term *deep learning*. Each layer of the network can have multiple neurons increasing the model's complexity.

Other type of neural networks were then proposed and are specialized in different tasks. Convolutional neural networks containing convolutional layers are mostly used in image processing, whereas recurrent neural networks that contain a feedback loop are used for text processing tasks. These different networks can combine the input features in different ways and create robust representations of the input data allowing these models to perform well in supervised learning settings.

Knowledge-based

Knowledge-based methods are similar to unsupervised methods because they do not rely on gold-standard labeled training data. However, these methods make use of external knowledge present in curated databases, terminologies, ontologies, or other types of information sources. Strategies based on external knowledge have the objective to infer or predict new information by finding similarities with the currently known associations.

2.3.4 Learning paradigms

There are different learning paradigms when implementing automatic methods for information extraction. Knowledge-based methods rely on external knowledge resources to empower their decision ability. In machine learning it is common to consider three major learning paradigms: supervised, semi-supervised, and unsupervised. Supervised algorithms require labeled training data to train and teach machine learning models what is what.

Information extraction frequently aims to extract specific and pre-specified relations

from homogeneous data, such as for example, extracting adverse drug events from medical records. In contrast, extracting relations from distinct domains is difficult since it requires to create new handcrafted rules or to annotate new training samples for being used in a supervised setting. To alleviate this issue by facilitating domain-independent relation extraction from large and varied corpora as the Web, Banko *et al.* (2007) introduced a new extraction paradigm, open information extraction (OIE), where the proposed system "makes a single data-driven pass over its corpus and extracts a large set of relational tuples without requiring any human input." As stated by Banko *et al.* (2007):

> Standard IE systems can only operate on relations given to it *a priori* by the user, and are only practical for a relatively small number of relations. In contrast, open IE operates without knowing the relations *a priori*, and extracts information from all relations at once.

They proposed a system for automatically extracting possible relations of interest. Their system is composed of three steps: (1) a self-supervised learner that uses a naive Bayes (NB) classifier and heuristics based on the dependency parsing, (2) a single-pass extractor that tags each word with its part-of-speech (PoS) and it finds relations between noun phrases (NPs), and (3) a redundancy-based assessor that creates normalized forms of the relations which are used to count the number of distinct sentences from which each relation was found.

Open IE systems have been used in several natural language processing (NLP) tasks and these are traditionally built using patterns. Some of its applications are question answering (Fader *et al.*, 2014), relation extraction (Soderland *et al.*, 2010; Fader *et al.*, 2011) and information retrieval (Etzioni, 2011).

Despite these open IE systems being useful for extracting many generic relations, this is not the case when finding specific biomedical relations such as ADEs or protein-protein interactions (PPIs). These type of biomedical relations are harder to extract since they contain specific vocabulary, and in the special case of using clinical reports there are usually many abbreviations and misspelling words increasing the level of ambiguity. Commonly, small datasets with hundreds or thousands of documents are automatically collected using precise queries and manually annotated with these narrow type of relations for empowering the evaluation of biomedical text mining systems.

Supervised

Supervised learning is a strategy that uses gold-standard labeled training data to train machine learning models. These models learn from ground-truth examples and, after being trained, they can identify and distinguish previously seen patterns on new different data. This technique is often the one that performs best, but has the limita-

tion that usually requires large amounts of high-quality training data. And the manual process of expert annotation of data is very burdensome and expensive.

Unsupervised

Unsupervised learning algorithms have the objective to find patterns from unlabeled data. The idea is that the model is able to create compact internal representations of the data, and then differentiate distinct features. Approaches used in unsupervised learning include probabilistic methods (algorithms such as clustering) and neural networks.

Semi-supervised

Semi-supervised learning lies between the two extremes of learning paradigms, that use no labeled data (unsupervised) or only use labeled data (supervised). This technique often consists in using a small set of gold-standard labeled training data, and the remaining training data is usually labeled by an heuristic method or with the help of an existing knowledge base. The advantage of these strategies is that they require a lesser amount of curation work for creating ground-truth annotations.

Distant supervision To the best of our knowledge, the first use of learning from weakly labeled data in the biomedical domain was employed by Craven and Kumlien (1999) to construct knowledge bases for molecular biology, where they exploited databases to automatically label training instances. They present a method to represent unstructured natural language text, from the MEDLINE biomedical scientific literature, into a structured form such as a knowledge base or a database. As Craven and Kumlien (1999) state:

> Our approach is motivated by the observation that, for many IE tasks, there are existing information sources (knowledge bases, databases, or even simple lists or tables) that can be coupled with documents to provide what we term "weakly" labeled training examples. We call this form of training data weakly labeled because each instance consists not of a precisely marked document, but instead it consists of a fact to be extracted along with a document that *may* assert the fact.

2.3.5 Evaluation metrics

In order to measure the performance of an IE system, specific metrics are usually employed to measure the quality of the predicted annotations (Chinchor and Sundheim, 1993; Fawcett, 2006; Dalianis, 2018). In this calculation, the predicted annotations are

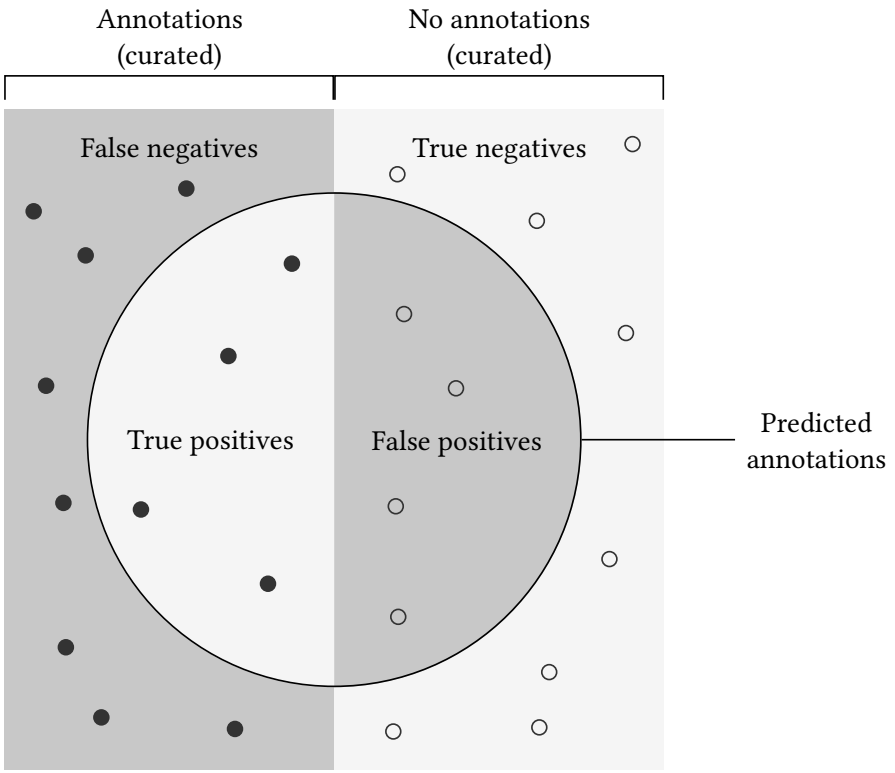

Figure 2.1: Spatial visualization of true (false) positives and true (false) negatives. Image adapted from https://en.wikipedia.org/wiki/Precision_and_recall.

compared with the gold-standard annotations commonly made by expert curators. Predicted annotations are considered true if they strictly match the gold-standard annotations, and false otherwise. Furthermore, system predictions can be positive if the system provides an annotation (for example, it identifies a chemical entity mention in a text), or negative if the system does not provide any annotation (for example, no chemical entity mention is found). Therefore, predictions can take one of distinct four classes (Figure 2.1):

- True Positive (TP): correct prediction, the annotation exists in the curated corpus;
- False Positive (FP): incorrect prediction, the annotation does not exist in the curated corpus but the system predicted it;
- True Negative (TN): correct prediction, the annotation does not exist in the curated corpus;
- False Negative (FN): incorrect prediction, the annotation exists in the curated corpus but the system did not predict it.

In this scenario, the used metrics to evaluate the classification problem are Precision, Recall, Accuracy, Specificity, and F1-score. These metrics assume values between zero, in the worst case, and one, in the best case.

Precision is the ratio between the number of true positives and the number of positive predictions. In contrast, recall or sensitivity is the ratio between the number of true positives and the number of positive curated annotations. Accuracy is the ratio between the correct (positive or negative) predictions and the total number of (positive or negative) predictions. Specificity is the ratio between the number of true negatives and the number of negative curated annotations. F1-score, or F1-measure, is the harmonic mean of precision and recall. These metrics are shown in Equations (2.1) to (2.5).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.2}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2.3}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2.4}$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.5}$$

In diverse information extraction and NLP tasks such as document classification, named entity recognition, and relation extraction, the most used metrics are precision, recall, and F1-score. Particularly, in multi-class classification problems that deal with the prediction of multiple classes—for example, multiple document topics, entity or relation types—it is common to consider specific different averaging formulas for taking into account the individual performance for every class. These are known as *macro-average* and *micro-average* (Yang, 1999; Tsoumakas *et al.*, 2009). Macro-averaging consists in first calculating the score for every class and then averaging the per-class scores to obtain the final macro-averaged score. On the other hand, micro-averaging consists in first summing up the number of true (false) positives (negatives) of every class, and then calculate the micro-averaged score using these global counts. Curiously, Opitz and Burst (2019) present two slightly different ways of calculating the macro-averaged F1-score, but conclude that the commonly employed formula—average of F1-scores per each class—is the more robust.

2.4 Biomedical text mining

In this section we present background material about NLP applied in the biomedical domain. In the biomedicine research field, new concepts are discovered regularly and their term expressions are added to the current biomedical vocabulary. Therefore, it is of utmost importance to keep standard terminologies, vocabularies, and databases updated with biological and medical curated information. We present (1) several resources, such as databases and thesauri, that are commonly used in biomedical information extraction pipelines, and (2) worldwide shared-tasks and challenges aiming to improve biomedical text mining systems.

Biomedical information extraction is concerned with the understanding of natural language texts in the biomedical domain, and has the objective of mining information from these unstructured textual data and represent them in a structured and unambiguous way. Due to the enormous quantity of biomedical information registered in textual form, such as scientific documents or clinical health records, it is impractical to manually acquire and link all the existing knowledge (Cases *et al.*, 2013). Therefore, automatic IE and text mining solutions are helpful for integrating current biomedical information (Krallinger *et al.*, 2005; Ananiadou *et al.*, 2006).

2.4.1 Resources

Biomedical text mining is achievable due to the large number of resources that are available (Simpson and Demner-Fushman, 2012; Rebholz-Schuhmann *et al.*, 2012; Zhu *et al.*, 2013; Przybyła *et al.*, 2016; Rosário-Ferreira *et al.*, 2021). Artificial intelligence–based techniques, such as machine learning and knowledge-based methods, rely on training or external data for extracting information from free text. It is due to the manual curation efforts of biomedical experts that nowadays there is a large amount of standard datasets for biomedical IE tasks and other well-established sources of information such as databases and terminologies. Biomedical curated data store biomedical knowledge that is essential for building text mining systems, and these data resources can be mainly split into two distinct groups:

- *Corpora* are collections of annotated documents that are used for assessing and comparing different IE solutions. For instance, these documents may contain annotations about concepts and their relationships that are previously annotated by expert curators.

- *Knowledge bases* are repositories that collect information with the goal of modeling the behavior and reality about some field. Databases and ontologies are two types

of knowledge bases. While databases are restrictively a structured collection of data (tables, schemes), ontologies are more rich since they also interconnect the data (graphs, tree structures). The type of knowledge base to be used should be chosen accordingly with the user needs, since ontologies are considered better than databases for representing domain concepts with higher level of detail, but databases may be more efficient (Martinez-Cruz *et al.*, 2012).

Aside from these data resources, other materials relevant to the development of biomedical text mining systems include:

- *Toolkits and frameworks* enable a fast and easier prototyping of more complex NLP architectures. Examples include NLTK (Loper and Bird, 2002), cTAKES (Savova *et al.*, 2010), Gensim (Řehůřek and Sojka, 2010), and Flair (Akbik *et al.*, 2019).

- *Annotation tools and applications* provide a medium to aid domain experts manually annotate documents (Neves and Ševa, 2021). Examples include BRAT (Stenetorp *et al.*, 2012), MyMiner (Salgado *et al.*, 2012), PubTator (Wei *et al.*, 2013), and LitSuggest (Allot *et al.*, 2021).

- *Standard formats for interoperability*, such as BioC (Comeau *et al.*, 2013), easen the processing of distinct annotated datasets by different NLP systems.

- *Pre-trained models* are usually made publicly available to allow reuse, evaluation, and reproducibility by other researchers. These include, for example, word and sentence embeddings such as BioWordVec (Zhang *et al.*, 2019b) and BioSentVec (Chen *et al.*, 2019b), and contextualized word representation models such as BioBERT (Lee *et al.*, 2020) and PubMedBERT (Gu *et al.*, 2021).

Table 2.3 presents some publicly available databases, ontologies, and terminologies that are relevant in the biomedical domain and useful for supporting the construction of biomedical IE systems. Nowadays there are a considerable number of resources for biomedical text mining and new databases are frequently published. Therefore, choosing the most adequate databases for specific biomedical text mining challenges, in spite of being an arduous task requiring the knowledge and experience of experts, is important because different external knowledge sources lead to fluctuations in the performance of IE systems.

Table 2.3: A list of biomedical databases, ontologies, and terminologies.

Resource	Description
BioGRID (Chatr-aryamontri *et al.*, 2017)	Biological General Repository for Interaction Datasets (BioGRID) is a database containing protein, genetic, and chemical interactions. https://thebiogrid.org
CTD (Davis *et al.*, 2017)	Comparative Toxicogenomics Database (CTD) contains information about interactions between chemicals, genes, and diseases. https://ctdbase.org
DrugBank (Wishart *et al.*, 2018)	DrugBank is a database containing molecular information about drugs, their mechanisms, interactions, and targets. https://www.drugbank.com
MeSH (Lipscomb, 2000)	Medical Subject Headings (MeSH) is the terminology used for indexing articles for MEDLINE literature. https://www.ncbi.nlm.nih.gov/mesh
MIMIC-III (Johnson *et al.*, 2016)	Medical Information Mart for Intensive Care (MIMIC) is a large database comprising information, such as medications and clinical notes, about patients admitted to critical care units. https://mimic.mit.edu
OBO Foundry (Smith *et al.*, 2007)	The Open Biological and Biomedical Ontologies (OBO) Foundry is an initiative for maintaining a family of interoperable ontologies in the biomedical domain. https://obofoundry.org
PubMed (Sayers *et al.*, 2021)	PubMed (Public MEDLINE) is a database comprising scientific abstracts and citations published in life science journals. https://pubmed.ncbi.nlm.nih.gov
SNOMED (Cornet and de Keizer, 2008)	Systematized Nomenclature of Medicine (SNOMED) is a global standard for clinical health terminology. https://www.snomed.org
UMLS (McCray, 1989)	Unified Medical Language System (UMLS) integrates biomedical terminology, coding standards, and other resources such as a semantic network for improving interoperability between biomedical information systems. https://www.nlm.nih.gov/research/umls/index.html
UniProt (The UniProt Consortium, 2017)	Universal Protein Resource (UniProt) is a repository containing information about protein sequences and associated information. https://www.uniprot.org

2.4.2 Community-wide efforts and shared tasks

Several community-wide efforts have been set up for evaluating the performance of biomedical text mining and its applicability to real IE problems (Huang and Lu, 2016). These are relevant to foster biomedical text mining research since several teams around the world present their state-of-the-art solutions. These efforts and challenges are often prepared by researchers and organizations. Each of these competitive events is usually composed by a variety of tasks where participating teams can choose in which task they want to compete. Some of the most known international challenges for biomedical and clinical text mining, along with some of their past NLP tasks are presented below.

- BioCreative[2] is a community-wide effort organized by several researchers from different universities and investigation centers, approximately every two or three years. This challenge aims to evaluate text mining systems applied to the life science literature. Some of their previous shared tasks are:
  - Gene mention recognition—its goal was to identify genes and gene products mentions in MEDLINE sentences (Smith *et al.*, 2008);
  - Protein–protein interactions (PPIs) extraction—the aim was to detect PubMed abstracts containing PPIs, and then recognize the interations in the relevant documents (Krallinger *et al.*, 2011);
  - Chemical compound and drug name recognition—the objective was to identify chemical and drug names in PubMed abstracts (Krallinger *et al.*, 2015).
- n2c2[3] (National NLP Clinical Challenges) continues the legacy of the i2b2 (Informatics for Integrating Biology and the Bedside) NLP shared tasks. Some of their tasks include:
  - De-identification—it consisted in removing automatically protected health information (PHI) from medical records (Uzuner *et al.*, 2007);
  - Obesity disease classification—the goal was to classify obesity and its comorbidities from patient discharge summaries (Uzuner, 2009);
  - Medication identification—it focused on identifying medication information such as their dosages, durations, and reasons for administration (Uzuner *et al.*, 2010).

[2] http://www.biocreative.org/
[3] https://n2c2.dbmi.hms.harvard.edu/

- BioASQ[4] organizes challenges on large-scale biomedical semantic indexing and question answering:
  - Semantic indexing—the goal was to label biomedical-related documents using the MeSH (Medical Subject Headings) terminology used for indexing MEDLINE articles (Tsatsaronis *et al.*, 2015);
  - Question answering—the objective was to compose correct answers to given biomedical questions (Tsatsaronis *et al.*, 2015).

2.5 Summary

In this chapter we explained what is natural language processing. We discussed how representing text evolved in the last years due to deep learning advances, how NLP is crucial to information extraction, and detail that these tasks can follow a pipeline or joint extraction paradigm. We detailed the most common methods, evaluation metrics, and learning paradigms. Finally, we explained how this has been useful for biomedical text mining and presented some commonly used resources for assessing biomedical information extraction and enumerated international competitions that have been boosting the development of this research field.

In the following chapters we will address different NLP tasks in the context of biomedical text mining. We will present our methodologies, results, and compare with related work. Given the preliminaries we detailed here that consisted in explaing the most fundamental concepts in natural language processing we believe the reader can now understand in more detail the remaining content of this thesis.

[4] http://www.bioasq.org/

Chapter 3

Biomedical concept disambiguation

In natural language text it is frequent that words or phrases[1] are ambiguous, that is, they can convey different meanings depending on the surrounding context. As stated by Navigli (2009), identifying the correct sense of an ambiguous word in a specific context is only apparently simple—while humans generally do not even notice the ambiguities of language, machines need to process *unstructured text* and extract structured information to determine the underlying meaning.

In detail, as Vicente and Falkum (2021) explain, a *monosemous* term has only one meaning, and terms with multiple senses can be considered *polysemous* or *homonymous*. *Polysemous* terms are associated with two or more related senses, whereas in contrast *homonymous* terms are associated with two or more unrelated meanings. These phenomena are denominated as *monosemy*, *polysemy* and *homonymy*. In this work, to simplify the task at hand, we make no distinction between *polysemous* and *homonymous* terms, henceforward referring to them as ambiguous terms. However, we stress that tackling *polysemy* and *homonymy* separately has the potential to improve downstream NLP (natural language processing) tasks. For instance, Krovetz (1997) has demonstrated its value in information retrieval.

The computational identification of the correct meaning of an ambiguous word (or term) given a specific context is known as word sense disambiguation (WSD), and it is considered an AI-complete problem relevant for natural language understanding (Navigli, 2009; Ide and Véronis, 1998). This is an important task for extracting accurate information from text. Generally, the first major step in information extraction is *concept recognition* which is responsible for identifying concepts of specific classes, such as chemicals or diseases, in the text. This task can be articulated as a pipeline of two subtasks: (1) named entity recognition (NER) followed by (2) disambiguation and nor-

[1] In this context, a phrase is considered a group of words that forms a grammatical unit, playing a specific role within the syntactic structure of a sentence. https://en.wikipedia.org/wiki/Phrase

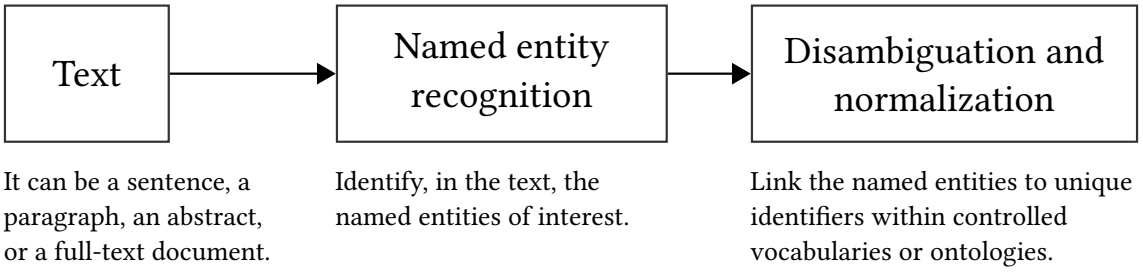

Figure 3.1: Named entity recognition and normalization pipeline.

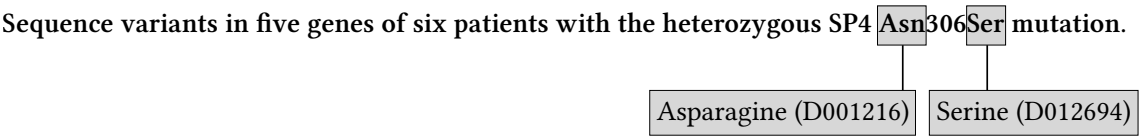

Figure 3.2: Example text with chemical entity annotations. Annotations from the PMID 17356515 document in the NLM-Chem dataset (Islamaj *et al.*, 2021b). The outer boxes contain the associated MeSH heading and unique identifier.

malization (Figure 3.1).

The aim of NER is to identify the text spans mentioning concepts of specific types. However, the sole utility of NER is limited because detected named entities are not linked to controlled vocabularies or ontologies, which is required for many end-user tasks (Leaman and Lu, 2016). Named entity normalization, or named entity linking, is the process of associating detected named entities with unique identifiers from standard knowledge bases. Undoubtedly, this is entwined with disambiguation: often, named entities convey multiple meanings that are associated with several unique identifiers. In this case, disambiguation methods are applied to each ambiguous named entity for selecting the correct unique identifier, amongst a set of candidate unique identifiers, according to its surrounding context. The *concept recognition* task is considered to be completed after the entity mentions are identified and linked to established databases. Figure 3.2 provides an example text with chemical concepts annotated and linked within the MeSH (Medical Subject Headings) vocabulary.

This chapter is mainly focused on disambiguation, but work on entity normalization is also discussed. We summarize related work, detail common resources for assessing these tasks, and describe our solutions. Based on distributed representations of words, or simply *word embeddings*, we propose supervised learning and knowledge-based approaches for biomedical WSD, and a method for normalization of clinical terms.

3.1 Background

For a long time, word sense disambiguation has been a challenging problem in computational linguistics, and even its definition has been a topic of debate. Kilgarriff (1997) dicusses the concept of *word senses* arguing that word senses only exist relative to a specific task. The author further extends this discussion gathering evidence from lexicographers and philosophers, stating that word senses are not easy to inventorize (Kilgarriff, 2007).

Ide and Véronis (1998) present an exhaustive review about WSD putting into perspective the past work on the topic since around the 1950s. The authors make a survey of several methods (knowledge-based and corpus-based), discuss several aspects and issues including the role of context, the disagreement on sense divisions, and the difficulty of evaluation and comparison of systems. Some of the first approaches for sense disambiguation relied on machine-readable dictionaries (Lesk, 1986; Veronis and Ide, 1990) with particular emphasis for improving information retrieval systems (Krovetz and Croft, 1989). Sanderson (1994, 1996) also extensively explored the impact of lexical ambiguity and disambiguation on information retrieval, concluding that very small queries benefit most from disambiguation than other queries that contain a sufficient number of words and provide enough context to implicity resolve ambiguities. Yarowsky (1995) proposed an unsupervised learning algorithm for sense disambiguation that matched the performance of supervised techniques, arguing that the cost of annotating a large training corpus may not be necessary to achieve a good WSD performance. Stevenson and Wilks (2001) showed that combining different knowledge sources can further improve WSD, and proposed a sense tagger that surpassed an accuracy of 94% on their evaluation corpus.

WordNet (Fellbaum, 1998; Miller, 1995) is a large lexical database of the English language, created at Princeton University, that has been extensively used for tackling WSD in the general domain (Banerjee and Pedersen, 2002; Loureiro and Jorge, 2019). It comprises synonyms that are grouped into unordered sets (synsets), each containing a brief definition and, in most cases, sample sentences showing its use. Despite its large coverage for general-domain concepts, other specific-domains—such as the life sciences field—require specialized thesauri and vocabularies. Presumably, the most used lexical database in the biomedical sciences is the Unified Medical Language System (UMLS), comprising many controlled vocabularies, and a comprehensive thesaurus and semantic network of biomedical concepts (Bodenreider, 2004). For instance, it includes clinical terminologies such as SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms), LOINC (Logical Observation Identifiers Names and Codes), and RxNorm (Bodenreider *et al.*, 2018). Besides these, many other biomedical ontologies and controlled

vocabularies exist, but some are specific for a certain class of biomedical concepts such as drugs or diseases. For example, the MeSH vocabulary is used to index PubMed articles which facilitates searching for topics of interest (Lipscomb, 2000).

Much of the research work in WSD and entity linking has been developed within the general-domain, but in this thesis we focus on the extraction of knowledge from biomedical text and are interested in the application of these tasks in free text from life-sciences scientific literature and clinical records. Thus, henceforward, we will focus here on reviewing prior work on biomedical entity disambiguation and normalization, since they have long been considered fundamental tasks for biomedical text mining (Schuemie *et al.*, 2005; Krauthammer and Nenadic, 2004).

Traditionally, external sources of information such as UMLS serve to fuel WSD knowledge-based methods that are employed in an unsupervised fashion and do not require labeled training data (Jimeno-Yepes and Aronson, 2010; McInnes *et al.*, 2011; El-Rab *et al.*, 2013; Garla and Brandt, 2013; McInnes and Pedersen, 2013; Duque *et al.*, 2018). Nonetheless, supervised or semi-supervised methods, based on machine learning, may not use external knowledge and still perform better due to having access to labeled training data (McInnes and Stevenson, 2014; Jimeno-Yepes, 2017). The difference between supervised and semi-supervised methods is in the quality of labeled training data. Supervised learning makes use of *ground-truth* or *gold-standard* training data that is annotated by domain experts. Such models produce less realistic results since the creation of gold-standard training data is very expensive, limited, and may be considered insufficient for obtaining a well-trained supervised model able to generalize to unseen data. Due to the problem of limited annotation human power, semi-supervised, and *weak* or *distant* supervision strategies have been investigated (Li *et al.*, 2019). These make use of *silver standard* training data that are created (1) through heuristics or handcrafted rules, or (2) by using existing knowledge from common databases. Recent approaches have been using external resources to build concept embeddings (Sabbir *et al.*, 2016; Newman-Griffis *et al.*, 2018). And, as shown by Tsai and Roth (2016) and Siu *et al.* (2016), the use of multiple knowledge databases also brings benefits to the problem of biomedical concept disambiguation.

3.2 Available corpora

In this section, for conciseness, we only present available research corpora—datasets used for common and standard evaluation—employed for the specific tasks regarding biomedical word sense disambiguation and entity normalization. However, for more biomedical NLP resources that are frequently adopted for solving these tasks, we point

Table 3.1: Datasets for biomedical word sense disambiguation, presented in chronological order. MSH: Medical Subject Headings. NLM: National Library of Medicine. UMLS: Unified Medical Language System. UMN: University of Minnesota. VUH: Vanderbilt University Hospital. WSD: word sense disambiguation.

Resource	Description
NLM WSD test collection (Weeber *et al.*, 2001)	A collection comprising 50 ambiguous terms with a total of 5000 disambiguated instances. Abstracts from the 1998 MEDLINE citations were used. The terms were mapped to the 1999 version of the UMLS. https://lhncbc.nlm.nih.gov/ii/areas/WSD/original.html
MSH WSD dataset (Jimeno-Yepes *et al.*, 2011)	A dataset containing 203 ambiguous terms with a total of 37 888 ambiguity instances retrieved from the 2010 MEDLINE. The 2009AB UMLS version was used to map the terms. https://lhncbc.nlm.nih.gov/ii/areas/WSD/collaboration.html
VUH admission notes (Wu *et al.*, 2013, 2015) (Wang *et al.*, 2016b, 2018c)	It contains 25 ambiguous abbreviations, from clinical admission notes, with up to 200 sentences containing each abbreviation. These were randomly selected and manually annotated by experts.
UMN clinical notes (Moon *et al.*, 2014) (Wu *et al.*, 2015) (Wang *et al.*, 2016b)	A dataset containing 75 acronyms and abbreviations (short forms) with their possible senses (long forms). Senses of each ambiguous term were manually annotated from 500 random instances and matched with the 2011AB UMLS version. https://hdl.handle.net/11299/137703

the reader to Section 2.4 where we present some of the most well-known external knowledge sources, and additionally enumerate NLP competitions or shared tasks that addressed, and have been fostering, the development of solutions for biomedical concept disambiguation.

Annotated datasets are required to evaluate automatic systems, but manual curation by domain experts is an expensive and cumbersome task that is often not enough (Baumgartner *et al.*, 2007; Howe *et al.*, 2008; Karp, 2016). It is therefore of utmost relevance to publicly share such resources for research purposes to further the development and improvement of these text mining methods.

Table 3.1 presents some of the most well-known datasets for evaluating biomedical WSD systems. In our work, we use the MSH WSD dataset since it is one of the most used, and largest, datasets for evaluating WSD in biomedical scientific literature

Table 3.2: Datasets for biomedical named entity normalization, presented in chronological order. CDR: chemical disease relation. CUI: Concept Unique Identifier. MCN: Medical Concept Normalization. MeSH: Medical Subject Headings. NCBI: National Center for Biotechnology Information. NLM: National Library of Medicine. OMIM: Online Mendelian Inheritance in Man. PMC: PubMed Central. ShARe: Shared Annotated Resources. UMLS: Unified Medical Language System.

Resource	Description
NCBI disease corpus (Doğan and Lu, 2012) (Leaman *et al.*, 2013) (Doğan *et al.*, 2014)	A collection of 793 PubMed abstracts containing disease entity mentions and their corresponding MeSH or OMIM identifiers, where each abstract was manually annotated by two experts. A total of 6982 disease mentions are mapped to 790 unique identifiers. https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE
ShARe corpus (Elhadad *et al.*, 2015) (Pradhan *et al.*, 2015)	The SemEval-2015 ShARe corpus is comprised of 531 de-identified clinical notes (discharge summaries and radiology reports) annotated with disorder mentions along with their normalization to the UMLS terminology using CUIs. http://share.healthnlp.org
BC5CDR corpus (Li *et al.*, 2015, 2016)	It contains 1500 PubMed abstracts annotated with biomedical entity mentions linked with MeSH identifiers, with a total of 4409 chemicals and 5818 diseases. https://biocreative.bioinformatics.udel.edu/tasks/biocreative-v/track-3-cdr
MCN corpus (Luo *et al.*, 2019, 2020b)	A wide-coverage corpus for clinical concept normalization including annotated medical problems, treatments, and tests. It consists of 100 discharge summaries with a total of 10 919 concept mentions mapped to 3792 unique identifiers from two controlled vocabularies—RxNorm (Liu *et al.*, 2005; Nelson *et al.*, 2011) and SNOMED CT (Cote, 1986; Stearns *et al.*, 2001; Cornet and de Keizer, 2008; Bodenreider *et al.*, 2018). https://n2c2.dbmi.hms.harvard.edu/2019-track-3
NLM-Chem corpus (Islamaj *et al.*, 2021b,a)	It consists of 150 full-text PMC articles, doubly annotated by ten NLM experts, with a total of 38 942 chemical entity mentions, which correspond to 4867 unique chemical names and 2064 MeSH identifiers. https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-2

(Jimeno-Yepes *et al.*, 2011). Similarly, Table 3.2 presents some of the standard datasets for biomedical named entity normalization. For this task, we use the MCN (Medical Concept Normalization) corpus (Luo *et al.*, 2019) for assessing our knowledge-based normalization method in the clinical domain.

3.3 Biomedical word sense disambiguation

In this section we present our studies about biomedical WSD methods. At first, we describe our research about the impact of using general-domain *versus* domain-specific embedding models for disambiguation of ambiguous terms in biomedical scientific abstracts (Antunes and Matos, 2016). Then, we present an exhaustive study of different methods for tackling biomedical WSD (Antunes and Matos, 2017a,b,c). We compare the use of traditional text feature engineering (bag-of-words) against the use of word embeddings using a plethora of machine learning classifiers. We propose a knowledge-based method based on word embeddings, concept textual definitions, and concept associations—derived from MeSH term co-occurrences—that achieves competitive results. We also investigate the impact of using different word embeddings averaging functions according to the distance between the ambiguous term and the remaining words of the surrounding context. In all these works we used the MSH WSD dataset for evaluation of our methods.

3.3.1 MSH WSD dataset

The MSH WSD dataset is likely the most used resource for evaluating biomedical WSD systems. It was generated by an automatic method proposed by Jimeno-Yepes *et al.* (2011), using the UMLS Metathesaurus and the manual MeSH indexing in MEDLINE citations. The data consists of PubMed scientific abstracts, each with one ambiguous term identified and mapped to the correct sense using UMLS Concept Unique Identifiers (CUIs). It contains 203 ambiguous terms (88 are regular terms, 106 are abbreviations, and 9 are a mix of both) with a total of 423 distinct senses. Most of the terms (189) have only two different meanings, 12 terms have three different meanings, and the remaining 2 terms have four and five different meanings. For each possible sense there is a maximum of 100 instances, that is, abstracts where the ambiguous term occurs. There are a total of 37 888 examples of ambiguity, and each term has on average 187 ambiguity cases.

3.3.2 General-domain *versus* domain-specific word embeddings

In this section we present our experiments with traditional machine learning classifiers for evaluating the impact of using general-domain and domain-specific (biomedical) word embeddings models for biomedical WSD (Antunes and Matos, 2016).

Methodology

In order to get a greater judgment of the performance of the two word embeddings models, we employed several machine learning classifiers using the scikit-learn framework[2] (Pedregosa *et al.*, 2011): decision tree, k-nearest neighbors, passive aggressive linear model, ridge regression, and two different implementations of support vector machines. For a fair evaluation, we applied 5-fold cross-validation to split the abstracts set, of each ambiguous term, in training and test subsets. A list of 313 stop words obtained from the MEDLINE repository[3] was used to filter out less relevant words in the corpus.

Word embeddings Two word embeddings models were created: one for the general-domain and another specific to the biomedical domain using textual data from Wikipedia and PubMed, respectively. Wikipedia is range-wide having no specific domain. We used the full Wikipedia dump, obtained in September 2015, amounting to approximately four million articles and containing about two million distinct words. On the other hand, PubMed is specific to the biomedical domain. Around six million abstracts corresponding to the years 2010 to 2015 were used, containing around 400 thousand distinct words. Both models were trained with the Gensim framework (Řehůřek and Sojka, 2010) using a window of five words and for a feature vector of size 100. Each abstract was represented by the weighted average of the vector embeddings of the respective words, with the TF–IDF value of each word used as weight.

Results and discussion

Table 3.3 presents the macro-accuracy results using several machine learning classifiers and two different word embeddings models (Wikipedia and PubMed). The model specific to the biomedical domain—created with PubMed abstracts—consistently outperformed the general model created from Wikipedia articles. Nevertheless, the results obtained with the latter indicate that even features extracted from general-domain corpora may contribute to these methods. The highest accuracy result using the Wikipedia model (support vector classification) exceeds the lowest accuracy result using

[2] For details about these classifiers we point the reader to the scikit-learn web page: https://scikit-learn.org.

[3] https://data.lhncbc.nlm.nih.gov/public/ii/information/MBR/WordCounts/2009/wrd_stop

Table 3.3: Accuracy disambiguation results, in the MSH WSD dataset, using different machine learning classifiers and word embeddings from the general and biomedical domains (Wikipedia and PubMed). Results shown are the average across five folds. DT: decision tree; kNN: k-nearest neighbor (k=5); PA: passive aggressive linear model; RR: ridge regression; SGD: linear support vector machine with stochastic gradient descent; SVC: support vector classification.

Classifier	Model of word embeddings from	
	Wikipedia	PubMed
DT	0.817	0.849
kNN	0.896	0.918
PA	0.893	**0.928**
RR	0.905	0.910
SGD	0.874	0.916
SVC	**0.912**	0.924

the PubMed model (decision tree) by around 6 percentage points (0.912 *vs* 0.849), which demonstrates that the selection of an appropriate classifier is also important. The highest accuracy result obtained was 0.928 corresponding to the use of the word embeddings model from PubMed and the application of the passive aggressive classifier.

3.3.3 Supervised learning and knowledge-based methods

In this section, we present an exhaustive study of supervised learning and knowledge-based methods for biomedical WSD (Antunes and Matos, 2017c). We describe our knowledge-based method and evaluate the impact of using different word vector embeddings averaging functions for representing the surrounding context of ambiguous terms (Antunes and Matos, 2017a,b).

Implementation

Supervised learning and knowledge-based methods are applied to the MSH WSD dataset in order to measure and compare the accuracy results of disambiguation. Bag-of-words features are used only in the supervised setting whereas word embeddings—created with unlabeled MEDLINE abstracts—are used in both approaches. From the UMLS Metathesaurus we extracted CUI textual definitions to be used in the knowledge-

based approach. The word embeddings are used to calculate vector embeddings for: (1) the surrounding contexts of the ambiguous terms, henceforward denoted as context embeddings or context vectors; and (2) CUI textual definitions extracted from UMLS, henceforward denoted as concept embeddings or concept vectors. The knowledge-based method finds the most likely sense for a specific ambiguous term by evaluating the similarity between its context vector and all the concept vectors weighted by CUI–CUI association values. Each step is described in detail in the following paragraphs.

Word embeddings The word embeddings models were generated using PubMed articles which are specific to the biomedical domain. MEDLINE abstracts corresponding to the years 1900 to 2015 were used, containing around 15 million documents with a total of around 800 thousand unique words. We trained six word embeddings models using context windows of 5, 20, and 50 words, and vector sizes of 100 and 300. For generating the word embeddings vectors we used the continuous bag-of-words model proposed by Mikolov *et al.* (2013a), implemented in the Gensim framework (Řehůřek and Sojka, 2010). The word embeddings are used to calculate the context embeddings and the concept embeddings as explained next.

Context embeddings The context embeddings are vectors that represent the surrounding contexts of the ambiguous terms. We consider the surrounding context of an ambiguous term to be all the words of the respective document excluding the ambiguous term occurrences. Each context vector is obtained by weighting the vector embeddings of the respective words using different weighting schemes. In the end, all the context vectors are normalized (L2 norm equal to 1).

In the supervised learning setting we considered the TF–IDF weighting scheme, whereas in the knowledge-based method we additionally tested four averaging functions which consisted of word distance decay functions multiplied by the IDF value. The objective of using decay functions, instead of the term frequency component, was to give greater importance to closest words of the ambiguous term, since we hypothesized that the nearest words to the ambiguous term may be more relevant for disambiguation. The absolute word distance d between some specific word and the closest occurrence of an ambiguous term was defined as being the input variable of the decay function. The five weighting schemes, with the IDF value being used as a multiplicative factor, were:

- Term frequency;
- No decay (constant): $\mathrm{f}(d) = 1$;
- Fractional decay: $\mathrm{f}(d) = 1/d$;

- Exponential decay: $f(d) = \exp(-d)$;
- Logarithmic decay: $f(d) = 1/\ln(1+d)$.

Concept embeddings The concept embeddings are only relevant to the knowledge-based method. We extracted, from UMLS knowledge sources[4], textual definitions for every CUI which were used to create the concepts vector embeddings—the TF–IDF value was used to weight the word vectors. Alike the context embeddings, all the vectors were normalized.

CUI–CUI association values We calculated CUI–CUI association values as normalized pointwise mutual information (NPMI)[5] from the MeSH co-occurrence counts in MEDLINE citations[6,7]. A NPMI value is between -1 and 1, with -1 for never occurring together, 0 representing independence, and 1 a complete association (a concept in relation to itself has a value of 1). Since there is a large number of CUIs, and consequently the number of possible CUI–CUI associations is much higher, we kept only the association values with NPMI values greater or equal than 0.3. These are only used in the knowledge-based method.

Supervised learning classification We tested five machine learning classifiers from the scikit-learn framework (Pedregosa *et al.*, 2011): decision tree, k-nearest neighbors, logistic regression, multi-layer perceptron, and support vector machine. Bag-of-words, containing unigrams and bigrams weighted by TF–IDF, and context embeddings were used as input features to train the classifiers. For each ambiguous term, 5-fold cross-validation was used to split the corresponding abstracts set for training and testing the classification models.

Knowledge-based method The knowledge-based method does not require training from the target dataset, being only dependent on knowledge from external sources. Our method relies on the idea of comparing the surrounding contexts of the ambiguous terms with the concepts' textual definitions—the aim is to find the most similar concept (likely meaning) given a specific context.

As described earlier, each CUI was represented by a concept vector, and the surrounding context of an ambiguous term was represented by a context vector. Therefore,

[4] https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html

[5] https://en.wikipedia.org/wiki/Pointwise_mutual_information

[6] Since the MSH WSD dataset uses CUIs to identify the distinct term senses, we used the MeSH to CUI mapping in UMLS to translate the MeSH term associations to UMLS concept–concept associations.

[7] https://ii.nlm.nih.gov/MRCOC.shtml

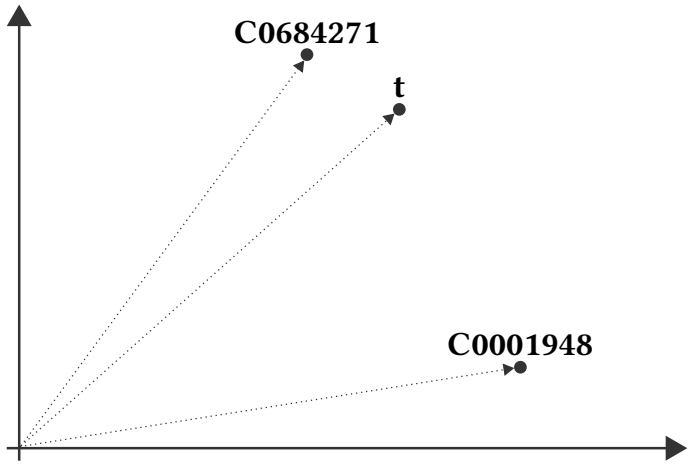

Figure 3.3: Exemplificative spatial representation of the context vector of an ambiguous term and the vectors of two candidate CUIs. In this example, one can visualize that the closest concept vector to the context **t** is relative to the C0684271 identifier which would be the one selected as the correct meaning.

it is straightforward to infer the most related sense of an ambiguous term by calculating the cosine similarity between the context vector and each CUI vector, and selecting the most similar one. Figure 3.3 illustrates a visual example.

We extended this baseline approach by calculating a score for each candidate CUI of an ambiguous term, and the CUI with highest score is selected as the correct meaning. The score is obtained using the cosine similarities between the context vector and every concept vector weighted by CUI–CUI association values (one concept is the candidate sense, and the other is the one whose textual definition is being compared with the context). The intuition behind this idea is that if a distinct concept has a strong association with the candidate concept, and its textual definition is similar to the context in which the ambiguous term is inserted, then the likelihood of the respective candidate CUI to be the correct sense must be increased (likewise, if the association is weak then the likelihood must be decreased). The score function is defined in Equation (3.1).

$$\text{score}(\text{CUI}) = \frac{1}{N} \sum_{j} \text{NPMI}(\text{CUI}, \text{CUI}_j) \cdot \text{CS}(\mathbf{t}, \mathbf{CUI}_j) \tag{3.1}$$

According to Equation (3.1): the CUI variable represents a candidate meaning; the CUI_j variable represents any other related concept; the **t** variable corresponds to the context vector; and $\mathbf{CUI}_j$ is the concept vector of the related concept CUI_j. The context **t** of a specific candidate is compared to every concept textual definition $\mathbf{CUI}_j$ by its cosine similarity $\text{CS}(\mathbf{t}, \mathbf{CUI}_j)$, which is then weighted by the $\text{NPMI}(\text{CUI}, \text{CUI}_j)$ association value. The N variable is the total number of associations considered, corresponding to

the number of NPMI values, and it is used to normalize the final score. For each candidate CUI—from a set of possible concepts, given a specific ambiguous term—a score is calculated, and the one that obtains the highest score is considered as the correct sense.

Results and discussion

In both approaches, supervised and knowledge-based, the dataset was split into five folds, and the final results were obtained by averaging the results of each fold. Table 3.4 and Table 3.5 present the disambiguation accuracy results obtained by the supervised machine learning classifiers and the knowledge-based method respectively.

Regarding the supervised learning setting, the highest accuracy (0.9557) was obtained combining unigrams and word embeddings features using a multi-layer perceptron. However, the best result using only bag-of-words features—unigrams and bigrams—is very close (0.9552) and was achieved by a support vector machine, showing that the state-of-the-art results for this problem can be reproduced using simple word-based features. Similarly, the sole use of word embeddings features with the multi-layer perceptron attained a close performance (0.9514). We observe that the performance variations using distinct word embeddings models trained with different vector sizes (100 and 300) and context windows (5, 20, and 50) are not significant[8], concluding that any of these models perform reasonably well. It is also noticeable that the use of bigrams contribute only slightly to the results, and unigram features alone achieve almost as good if not better results than the combination of unigrams and bigrams. Finally, we note that the decision tree model obtained the lowest results, below around 2–3 percentage points overall, in comparison to the other four classifiers. Remarkably, the k-nearest neighbors, a simple baseline model, achieved consistent and competitive performances with different combinations of features—attaining a top accuracy of 0.9475—which indicates that the features in use provide effective text representations.

Table 3.5 presents the disambiguation results obtained by the knowledge-based method. Different thresholds for the NPMI values (0.3, 0.5, 0.8, and 1.0) were imposed to select only a subset of CUI–CUI associations. The threshold 1.0 is the particular case of the baseline scenario where only the cosine similarity between the context vector and each candidate CUI vector is computed. We observe that, in all weighting schemes, the threshold 0.3 (rows 19–24) performed the best showing that the use of additional CUI–CUI associations, to an extent, is beneficial. Overall, the use of associations allowed to improve the accuracy by around 2 percentage points when compared to not using any related concepts (rows 1–6), that is, the case when only the similarity between the textual

[8] Using different dataset splits or random seeds for initializing model parameters would likely produce small variations in the results.

Table 3.4: Supervised learning disambiguation results in the MSH WSD dataset using bag-of-words and word embeddings features. The evaluation metric is accuracy and the results were obtained using 5-fold cross-validation. Rows 1–3 present the results using only bag-of-words features (unigrams, bigrams, and both), rows 4–9 present the results using only word embeddings (different vector sizes and windows), and rows 10–15 present the results from combining bag-of-words features (unigrams) with word embeddings. The highest accuracy, in each of these groups, is highlighted in bold. The overall highest accuracy is also underlined.

	BoW*	**WE†**		**Classifier‡**				
Row	Features	Size	Window	DT	kNN	LR	MLP	SVM
1	U	-	-	0.9067	0.9324	0.9205	0.9401	0.9511
2	B	-	-	0.8335	0.8850	0.8704	0.9224	0.9253
3	U+B	-	-	0.9019	0.9354	0.9101	0.9445	**0.9552**
4	-	100	5	0.9219	0.9452	0.9500	0.9503	0.9449
5	-		20	0.9185	0.9452	0.9495	0.9498	0.9452
6	-		50	0.9194	0.9447	0.9495	0.9501	0.9431
7	-	300	5	0.9186	0.9449	0.9505	0.9503	0.9452
8	-		20	0.9186	0.9444	0.9508	**0.9514**	0.9446
9	-		50	0.9166	0.9441	0.9509	0.9508	0.9444
10	U	100	5	0.9244	0.9464	0.9515	**0.9557**	0.9490
11			20	0.9215	0.9468	0.9514	0.9556	0.9486
12			50	0.9229	0.9467	0.9515	0.9555	0.9481
13		300	5	0.9218	0.9475	0.9519	0.9544	0.9499
14			20	0.9194	0.9473	0.9524	0.9550	0.9496
15			50	0.9191	0.9468	0.9520	0.9545	0.9482

* Bag-of-words features. U: unigrams. B: bigrams.

† Word embeddings model trained with a specific vector size and context window.

‡ Machine learning classifier evaluated. DT: decision tree. kNN: k-nearest neighbors (k=5). LR: logistic regression. MLP: multi-layer perceptron. SVM: support vector machine.

definition of the candidate CUI and the context of the ambiguous concept is considered. However, using the NPMI thresholds 0.5 and 0.8 obtained inferior results when compared to the simplest case (threshold 1.0) demonstrating that a lower NPMI threshold for selecting more concept associations is required to improve performance—we conclude that associations with a lower value play a key role. Regarding the weighting schemes, the fractional decay averaging function consistently obtained the highest results for every NPMI threshold considered. The word embeddings models trained with higher context

Table 3.5: Knowledge-based disambiguation results in the MSH WSD dataset using word embeddings, CUI textual definitions, and CUI–CUI association values. Different weighting schemes for averaging the word embeddings in the creation of the context vectors are compared. The evaluation metric is accuracy and the results are the average across five folds. Rows 1–6 present the results when only the cosine similarity between the ambiguous term' context vector and each candidate concept vector is considered. Rows 7–12, 13–18, and 19–24 additionally consider the cosine similarities of related concepts that have a NPMI value greater or equal than 0.8, 0.5, and 0.3 respectively. The highest accuracy, in each of these groups, is highlighted in bold. The overall highest accuracy is also underlined.

	CUI–CUI	**WE**[†]		**Weighting scheme**[‡]				
Row	**associations**[*]	S	W	TF	None	Frac.	Exp.	Log.
1	CS	100	5	0.8144	0.8164	0.8415	0.8259	0.8318
2			20	0.8254	0.8286	0.8473	0.8270	0.8407
3			50	0.8321	0.8341	0.8502	0.8278	0.8468
4		300	5	0.8181	0.8203	0.8457	0.8278	0.8355
5			20	0.8319	0.8352	**0.8533**	0.8302	0.8477
6			50	0.8337	0.8365	**0.8533**	0.8276	0.8501
7	NPMI $\geq$ 0.8	100	5	0.8132	0.8154	0.8395	0.8236	0.8304
8			20	0.8243	0.8277	0.8459	0.8255	0.8395
9			50	0.8314	0.8334	0.8493	0.8264	0.8461
10		300	5	0.8168	0.8193	0.8438	0.8255	0.8340
11			20	0.8312	0.8343	0.8515	0.8283	0.8466
12			50	0.8332	0.8357	**0.8518**	0.8264	0.8491
13	NPMI $\geq$ 0.5	100	5	0.8005	0.8019	0.8234	0.8057	0.8155
14			20	0.8152	0.8178	0.8348	0.8137	0.8290
15			50	0.8197	0.8236	0.8376	0.8150	0.8343
16		300	5	0.8030	0.8057	0.8267	0.8092	0.8190
17			20	0.8174	0.8203	0.8377	0.8162	0.8323
18			50	0.8209	0.8245	**0.8396**	0.8168	0.8352
19	NPMI $\geq$ 0.3	100	5	0.8430	0.8458	0.8617	0.8378	0.8560
20			20	0.8573	0.8600	0.8720	0.8458	0.8704
21			50	0.8600	0.8635	**<u>0.8744</u>**	0.8459	0.8730
22		300	5	0.8446	0.8471	0.8622	0.8404	0.8573
23			20	0.8566	0.8598	0.8730	0.8469	0.8705
24			50	0.8582	0.8611	0.8736	0.8478	0.8719

[*] CS: cosine similarity between the term context vector and each candidate concept vector only. NPMI $\geq$ *threshold*: concepts with a NPMI value, with respect to the candidate concept, greater or equal than the threshold are considered.

[†] Word embeddings model trained with a specific vector size (S) and context window (W).

[‡] Different weighting schemes are used to create the context embeddings. The IDF value is implicitly considered in all cases. TF: term frequency. None: no decay. Frac.: fractional decay. Exp.: exponential decay. Log.: logarithmic decay. These word distance decay functions are described in detail in this section.

Table 3.6: Performance comparison of WSD systems using supervised and knowledge-based methods in the MSH WSD dataset. Macro-accuracy is the evaluation metric.

Work*	Approach	S†	KB†
Zhang *et al.* (2019a)	Long short-term memory networks	0.9600	-
Pesaranghader *et al.* (2019)	Long short-term memory networks	0.9682	0.9267
Duque *et al.* (2018)	Co-occurrence graph	-	0.7152
Ours (Antunes and Matos, 2017c)	Word embeddings, cosine similarity	0.9557	0.8744
Jimeno-Yepes (2017)	Support vector machine	0.9597	-
Sabbir *et al.* (2016)	Word embeddings, k-nearest neighbors	-	0.9434
Tulkens *et al.* (2016)	Word embeddings, cosine similarity	-	0.84
Jimeno-Yepes and Berlanga (2015)	Word–concept statistical model	0.930	0.891
McInnes and Stevenson (2014)	Semantic similarity measures	0.97	0.78
McInnes and Pedersen (2013)	Semantic similarity measures	-	0.75
Garla and Brandt (2013)	Semantic similarity measures	-	0.8071
Jimeno-Yepes *et al.* (2011)	Naive Bayes classifier	0.9386	0.8383

* Works sorted in reverse chronological order.

† S: supervised. KB: knowledge-based. Distinct authors report results with different decimal places. Also, some results are not directly comparable because different strategies and dataset splits—for example, different number of folds in cross-validation—have been used for evaluation.

windows performed slightly better, but we did not find the impact of the vector size (100 *vs* 300) to be notorious. The best accuracy obtained by the knowledge-based method, 0.8744, was achieved using the fractional decay averaging function, the NPMI threshold set to 0.3, and the word embeddings model trained with a vector size of 100 and a context window of 50 words.

From our results, we confirm that the supervised learning approach performs rather better than the knowledge-based method (0.9557 *vs* 0.8744), but the latter does not require a training stage using annotated labels.

Comparison with other works Table 3.6 presents a performance comparison of our approaches with other works employing supervised and knowledge-based approaches in the same dataset. However, our results are not absolutely comparable with the ones from other works since different evaluation strategies have been used. For example, we use the average across five folds and other authors report results from 10-fold cross-validation. Nevertheless, we consider that this comparison allows us to assess the efficacy of our methods and perceive how these compare to the state-of-the-art.

Jimeno-Yepes *et al.* (2011) generated the MSH WSD dataset by automatic means and tested a supervised naive Bayes classifier and four knowledge-based methods. The supervised approach, using only the words occurring in the text, achieved an accuracy only about 2 percentage points below our best supervised result (0.9386 *vs* 0.9557). Their

best-performing knowledge-based method (Automatic Extracted Corpus) consists in automatically creating training data using documents from MEDLINE and queries using English *monosemous* relatives. A naive Bayes classifier is then trained using the automatically generated data. Our best knowledge-based result is around 3 percentage points higher than the one they obtained (0.8744 *vs* 0.8383).

Garla and Brandt (2013), and McInnes and Pedersen (2013) present knowledge-based methods, that use semantic similarity measures derived from the UMLS Metathesaurus, achieving accuracies of 0.8071 and 0.75 respectively. Garla and Brandt (2013) processed the abstracts with biomedical NER systems to capture concepts from UMLS that were used as feature vectors. Similarly to our knowledge-based method, the system proposed by McInnes and Pedersen (2013), UMLS::SenseRelate, assigns a score to each possible concept of an ambiguous term according to a similarity metric between the concept and the surrounding context. McInnes and Stevenson (2014) continued to explore the use of semantic similarity measures and proposed supervised and unsupervised (knowledge-based) methods. Their supervised method combines linguistic and biomedical specific features (including unigrams, bigrams, part-of-speech tags, and MeSH terms) in binary feature vectors. The MeSH terms were assigned manually by expert annotators, to each abstract, for the purpose of indexing. We suspect the inclusion of this information has a solid contribution in their final performance. However, considering a scenario where this *ground-truth* information is not available—for example, in recent publications that have not yet been annotated with MeSH terms—their supervised final performance (0.97) would likely decrease.

Jimeno-Yepes and Berlanga (2015) used a word–concept statistical model estimated from knowledge sources surpassing our method by about 2 percentage points (0.891 *vs* 0.8744). Our knowledge-based method is similar to the one proposed by Tulkens *et al.* (2016), which also compared concept representations with the representations of the context of ambiguous terms, and obtained an accuracy of 0.84. To the best of our knowledge, the highest accuracy achieved without supervised means (0.9434) was obtained by Sabbir *et al.* (2016). They used neural word and concept embeddings, and employed weak supervision—they did not use hand-labeled examples—to build their prediction model (k-nearest neighbors).

Similarly to our work, Jimeno-Yepes (2017) achieved an accuracy of 0.9597 in a supervised fashion combining unigrams and word embeddings using a support vector machine. More recent works have been using LSTM networks. Pesaranghader *et al.* (2019) use concept textual definitions from UMLS to compute concept embeddings, and employ a BiLSTM model achieving a state-of-the-art accuracy of 0.9682 in a supervised setting. Zhang *et al.* (2019a) also use a BiLSTM model achieving a similar accuracy (0.9600). Li

et al. (2019) followed a semi-supervised approach, based on label propagation, and used a BiLSTM attaining an accuracy of 0.9671.

3.4 Clinical named entity normalization

Electronic health records (EHRs) contain medical narratives, such as discharge and admission reports, that contain valuable information about the clinical history of patients in the form of free text. However, it is unfeasible to manually analyze large-scale medical texts, making the process of automatic annotation important to summarize or extract relevant data from clinical reports. This requires recognizing medical entities in the text including drugs, disorders, medical procedures, and laboratory measures. For that, the normalization of the entities is an essential step which consists in linking the entities to established terminologies (grounding). These annotations mitigate the problem of ambiguous and unspecific terms, helping physicians to more quickly get an overview of a patient clinical history.

In this section we present a method based on word embeddings for entity normalization in clinical texts (Silva *et al.*, 2020). We evaluate our approach in the MCN corpus that was developed by Luo *et al.* (2019) and was employed in the 2019 n2c2/UMass Lowell Track 3 challenge (Luo *et al.*, 2020b).

3.4.1 A knowledge-based approach based on word embeddings

The aim of this task is to link detected entities to unique codes within standard medical vocabularies. It is only focused on the normalization step—named entity recognition is dispensed—since mention spans are assumed to be given. We present a knowledge-based method based on word embeddings for representing medical concepts.

Dataset

We used the MCN corpus proposed by Luo *et al.* (2019), consisting of clinical texts annotated with entities linked to unique codes from standard databases. It comprises a wider set of clinical concepts—medical problems, treatments, and tests—in comparison to other datasets that only considered the normalization of disease mentions (Pradhan *et al.*, 2013, 2014; Elhadad *et al.*, 2015).

Each annotated clinical entity is associated with a single CUI from the UMLS 2017AB version. For example, “hypertension” and “HTN”, or “blood pressure” and “BP” are two examples of expressions that refer to the same concepts and are therefore identified by the same CUIs (C0020538 and C0005824, respectively). Although UMLS encompasses

Table 3.7: Detailed MCN dataset statistics. MCN: Medical Concept Normalization. CUI: Concept Unique Identifier.

	Training	Test	Total
Number of clinical records	50	50	100
Number of annotated entities	6684	6925	13 609
Number of unique CUIs	2331	2579	3792

Table 3.8: Examples of text rewrite rules handcrafted according to the MCN training set. MCN: Medical Concept Normalization. The left and right columns, in each of the three groups, contain the original and final text respectively—for example, abbreviations were replaced by their full-form expressions.

b/l	bilateral	po	oral	10%	partial
co2	carbon dioxide	trop	troponin	1/4	fourth
e.	escherichia	u/s	ultrasound scan	x2	double
iv	intravenous	vit	vitamin	2L	two liters
mso4	morphine sulphate	w/u	workup	3 of 6	iii/vi

several vocabularies, only two were used for annotation. RxNorm (Nelson *et al.*, 2011) was used to annotate clinical drugs and medications, whereas SNOMED CT (Stearns *et al.*, 2001), an extensive vocabulary of clinical terminology, was used for normalizing the remaining concepts (disorders, procedures, body structures, and others).

The dataset contains a total of 100 annotated discharge summaries and is split into two subsets: training and test (Table 3.7). We used the training subset to develop our model and the test subset for final evaluation.

Method

Our knowledge-based method involves two sequential steps: text pre-processing, and similarity computation. In the first step, specific text rewrite rules were handcrafted by inspecting the clinical named entities in the training set with the aim of cleansing the surface representation of these mentions (Table 3.8). In addition to the text replacements made, HTML (Hypertext Markup Language) entities and other superfluous symbols were also discarded to reduce the noise in the text.

In the second step, we represented (1) the clinical named entities from the training and test subsets, and (2) the UMLS concept names by using pre-trained biomedical

word embeddings. We employed the publicly available BioWordVec[9] model (Chen *et al.*, 2019b) that was created using the fastText library (Bojanowski *et al.*, 2017) and was generated from over 30 million documents from PubMed articles and clinical notes from the MIMIC-III database (Johnson *et al.*, 2016). Each term (named entity or concept name) was represented by the vector embeddings average of its constituent words. We created a mapping between CUIs and *term embeddings* using (1) the entity mentions and respective identifiers annotated in the training subset and (2) the concept names and identifiers from the UMLS within the RxNorm and SNOMED CT vocabularies.

To predict the most likely CUI for a new entity mention, our knowledge-based method calculates the cosine similarity between the entity vector embedding and every pre-calculated *term embedding*, and the CUI associated with the most similar *term embedding* is the predicted identifier for the entity.

Results and discussion

We evaluated our knowledge-based method in the training and test subsets. The evaluation in the training subset was made by averaging the results from 10 repetitions of 5-fold cross-validation[10], whereas in the test subset only a final one-time prediction was made. Accuracies of 0.812 and 0.801 were obtained in the training and test subsets, respectively, demonstrating that our system generalized well to new data since the accuracy performace on the test subset only decreased about 1 percentage point. However, a posterior evaluation without using handcrafted replacements proved that the created text patterns were biased toward the training set and led to overfitting, since simpler text pre-processing resulted in a lower training accuracy (0.807) and similar test accuracy (0.800).

Table 3.9 shows the official results obtained in the 2019 n2c2/UMass Lowell Track 3 challenge. Our method ranked amongst the top-10 best performing systems[11]. Our system obtained about 4 percentage points improvement compared to the baseline sieve-based model proposed by Luo *et al.* (2019) that obtained an accuracy of 0.764 in the test subset.

[9] https://github.com/ncbi-nlp/BioSentVec

[10] In this scenario, and for each data split, we only used the respective training folds for creating the mapping between CUIs and *term embeddings*.

[11] In total, 33 teams participated in the 2019 n2c2/UMass Lowell Track 3, consisting of 108 total submissions (Luo *et al.*, 2020b).

Table 3.9: Performance of the top-performing teams in the 2019 n2c2/UMass Lowell Track 3. Accuracy is the evaluation metric. This table is adapted from Luo *et al.* (2020b).

Rank	Team name	System brief description	Accuracy
1	TTI	Cascading dictionary matching, deep learning	0.8526
2	KP	Cascading dictionary matching	0.8194
3	UAZ	Cascading dictionary matching	0.8166
4	Ali	Retrieval, machine learning	0.8105
5	MDQ	Retrieval, machine learning	0.8101
6	UWM	Cascading dictionary matching	0.8079
7	UAv (ours)	Cosine similarity	0.8013
8	ezDI	Cascading dictionary matching	0.8006
9	MIT	Cascading dictionary matching	0.7961
10	NaCT	Cascading dictionary matching	0.7957

Ali: Alibaba; ezDI: ezDI, Inc; KP: Kaiser Permanente; MDQ: Med Data Quest, Inc; MIT: Massachusetts Institute of Technology; NaCT: National Centre for Text Mining; TTI: Toyota Technological Institute; UAv: University of Aveiro; UAZ: University of Arizona; UWM: University of Wisconsin-Milwaukee.

3.5 Summary

Automatic identification of entities in *unstructured text* is an imperative task for extracting knowledge from biomedical scientific literature and clinical narratives. It consists not only in detecting the entities' mention spans, but also in attributing them unique codes from standard vocabularies. This process, known as entity normalization or linking, faces the problem of ambiguity in the language—it is frequent that terms can have multiple senses depending on the context in which they are inserted—where word sense disambiguation mechanisms must be employed.

In this chapter, we proposed supervised learning and knowledge-based methods, based on distributed representations of words, to disambiguate multiple-meaning terms from PubMed abstracts. We conclude that simple word-based features provide a strong baseline for sense disambiguation, and the inclusion of word embeddings is advantageous. We also verify that supervised learning models surpass knowledge-based methods because they rely on annotated training data. Finally, we present an automatic method, based on word embeddings and external knowledge from the UMLS Metathesaurus, to normalize entities in patient clinical reports using standardized vocabularies, which performed competitively well.

Chapter 4

Biomedical text classification and similarity measurement

Text classification or categorization can represent a different number of tasks. The most primitive one is to simply identify if a given text is relevant or not according to some criteria—that is, for example, if a news article is about a specific subject or a clinical record contains valuable information about a specific disease. This is known as binary text classification since there are only two possible outcomes—*yes* or *no*. On the other hand, in multi-class classification there are several, three or more, possible classes and only a single one must be chosen. Another problem is multi-label classification where none, one, or more topics (classes) can be associated with a document.

The NLP task of sentiment analysis is an example of text categorization where different emotion states can be associated with a fragment of text often collected from social media (Feldman, 2013; Liu and Chen, 2015; Bouazizi and Ohtsuki, 2016; Tao and Fang, 2020). Another common application of text classification is spam filtering in eletronic mail (e-mail) systems where unsolicited e-mail messages should be detected and discarded (Diao *et al.*, 2000; Zhang *et al.*, 2004; Bhowmick and Hazarika, 2018).

Document triage, document filtering, or document selection is a text classification task that can be employed to select the most relevant documents for a specific information extraction task. Although Hearst (1999) argues that categorizing documents does not lead to discovering new information, Sarawagi (2008) explains that document classification is relevant (for information extraction) as a first step to filter out less-relevant documents especially if the corpus is very large. Similarly, Balog (2018) clarifies how the task of scoring documents according to how relevant they are to a given target entity is helpful for information extraction.

Text classification can be thought as grouping and clustering texts that share some degree of similarity concerning some aspect. For instance, if a text is considered to be

relevant regarding a specific criterion then it is likely that similar texts may also be of interest. For that reason, we consider that measuring the semantic textual similarity is a intertwined task with text classification. Therefore, in this chapter, we also review and present work on text similarity measurement despite being applicable in distinct NLP tasks such as word sense disambiguation and relation extraction.

Measuring the semantic similarity between texts is the basis of many text processing tasks. For example, finding excerpts of text with redundant information can be used for summarization (Aliguliyev, 2009). Another use case is plagiarism detection where texts with equivalent semantics are identified (Lukashenko *et al.*, 2007). Semantic similarity measurement can also be used to find similar textual contexts in which a relationship holds between specific entities (Panchenko and Morozova, 2012)—if two concepts have a certain interaction and two other different concepts appear surrounded by a similar textual context then it is likely they have a similar or the same interaction.

This chapter addresses two analogous NLP tasks: (1) text classification and (2) measurement of semantic textual similarity. We start by giving a brief overview of these problems applied in the biomedical domain and then present our solutions. First, a study with supervised machine learning classifiers is conducted for scientific literature triage. Then, a hybrid system based on rules and machine learning for categorizing clinical text is presented. Finally, a neural network that quantifies the textual similarity between clinical sentences is presented.

4.1 Background

The problem of *text classification* or *text categorization* is likely one of the oldest and most addressed tasks of natural language processing and has been relevant for many years in information retrieval (Lewis, 1995; Manning *et al.*, 2008). Early work on text categorization was focused on selecting different features from syntactic analysis (Lewis, 1992), where the simplest approach would be to treat each word as a feature. Later, Joachims (1998) proposed the use of support vector machines for text categorization and argued these are robust models because they eliminate the need for feature selection, show good performance, and do not require manual parameter tuning.

In the biomedical domain, to the best of our knowledge, Craven and Kumlien (1999) was one of the pioneer works employing text classification methods to identify relevant textual information, from 2889 abstracts in the MEDLINE database, for constructing biomedical knowledge bases.

Feldman *et al.* (2003) review past work on text mining of the biomedical literature. Particularly, the authors emphasize the importance of text categorization as an early step

of pre-processing since reducing the set of documents simplifies the follow-up mining tasks such as entity and relation extraction. Furthermore, they explain that there are two main approaches for text categorization: (1) the *knowledge engineering* approach where expert knowledge is manually encoded using a set of rules, and (2) the *machine learning* approach where a text classifier is built automatically by learning from a set of previously classified documents using a pre-defined set of categories. They also distinguish two main methods in the machine learning approach: in one method, for each pair of document and category, a boolean value is attributed—true if the document belongs to the category or false otherwise; in the other method, a value between 0 and 1 representing the confidence that the document belongs to the category is assigned, and afterwards document are ranked according to their confidence value.

Cohen and Hersh (2005) also present an extensive review of past work on biomedical text mining and text classification. The authors argue that text classification systems are valuable to biomedical database curators because they may have to review some documents until finding a particular piece of information for updating the database. Yeh *et al.* (2003) conducted a text mining competition in the KDD (Knowledge Discovery and Data Mining) Challenge Cup[1] aimed at identifying which biomedical articles were relevant for curating *Drosophila* gene expression information. The organizers provided a training set of 862 journal articles curated in FlyBase (The FlyBase Consortium, 2002) with genes and gene products, and a test set with 213 new articles. Participating teams had to develop a system that indicated which articles contained experimental evidence for gene expression products. The best performing team (Regev *et al.*, 2002)—that obtained 78% F-measure in the document curation sub-task—manually built text rules for matching common patterns in the title, abstract, and figure captions, and performed part-of-speech tagging and text chunking. In a different work, Donaldson *et al.* (2003) used a support vector machine to find protein–protein interaction data on the PubMed literature database and estimated that their system spared curators several days of work because they had to scan a significantly lower number of abstracts. The TREC (Text Retrieval Conference) 2004 Genomics Track (Voorhees, 2004; Hersh *et al.*, 2004; Hersh, 2005) addressed a triage task of full-text documents, simulating the work of curators in the Mouse Genome Informatics (MGI) system, where participants had to develop automatic solutions for detecting articles containing experimental evidence requiring the assignment of Gene Ontology (GO) codes.

The Computational Medicine Center in Cincinnati, Ohio organized a shared task to promote the development of automatic systems for assigning ICD-9-CM[2] codes to ra-

[1] https://kdd.org/kdd-cup/view/kdd-cup-2002/Tasks

[2] International Classification of Diseases, Ninth Revision, Clinical Modification. https://www.cdc.gov/nchs/icd/icd9cm.htm

diology reports (Pestian *et al.*, 2007). The challenge corpus was annotated by coding staff of CCHMC (Cincinnati Children's Hospital Medical Center) and two independent coding companies in order to reduce variation due to human judgment. It was split into training and testing sets with 978 and 976 documents respectively, and tagged with forty-five ICD-9-CM distinct labels. This multi-label classification task allowed each record to be assigned more than one code. The participating team from the University of Pennsylvania (Crammer *et al.*, 2007) employed a cascaded approach combining three coding systems: (1) a specialized policy that searched for specific keywords complying with some heuristics, (2) a rule-based system that checked if ICD-9-CM code descriptions appear in the reports, and (3) a learning system that used several natural language features. They achieved a micro-averaged F1-score of 0.8760 on the test set ranking 4th out of 44 systems that entered the challenge. Another team, from the University of Manchester (Sasaki *et al.*, 2007), employed different machine learning algorithms and tested different features such as n-grams of words weighted by TF–IDF values. They ranked 5th in the challenge and their best result, a micro-averaged F1-score of 0.8594 on the test set, was achieved by a support vector machine. In another work, Farkas and Szarvas (2008) investigated the feasibility of automatically constructing rule sets—in contrast to purely handcrafted rule-based systems—by replacing several laborious steps with machine learning models with the aim to alleviate the manual work from experts. Their system achieved competitive results with a micro-averaged F1-score of 0.8893 on the test set, and the authors concluded that hybrid systems preserve the good performance of rule-based classifiers and that, with the help of machine learning methods, their construction can be accelerated and require less human effort. More recent works explore the use of deep neural networks for automatic ICD-9 coding in clinical reports (Pereira *et al.*, 2018; Zeng *et al.*, 2019).

Other international challenges on biomedical text classification have been conducted in recent years to encourage the creation of text mining solutions and assess the state-of-the-art. We close this brief roadmap by presenting three competitions offered by the BioCreative organizers but we alert the reader that additional research work, not mentioned here, has been carried out (Huang and Lu, 2016). In the BioCreative III PPI ACT (article classification task) participants had to implement systems for detecting PubMed abstracts describing protein–protein interactions (Krallinger *et al.*, 2010, 2011). The organizers prepared a dataset split into training, development, and test partitions with 2280, 4000, and 6000 abstracts respectively which were manually labeled by domain experts. BioCreative VI launched a document triage task for precision medicine where the aim, for participating systems, was to identify PubMed abstracts describing genetic mutations affecting protein-protein interactions—our contribution on this problem is

presented in Section 4.2. In BioCreative VII, Chen *et al.* (2021a, 2022) promoted a challenge on COVID-19 literature curation since the number of COVID-19 related articles was growing at a rate about 10 000 articles per month. Their effort tackled a multi-label topic classification for COVID-19 literature to alleviate the burden of manual topic annotation in the LitCovid database (Chen *et al.*, 2021b) which contained tens of thousands of PubMed articles relevant to COVID-19 that needed to be assigned with up to eight distinct topics (such as *case report*, *diagnosis*, and *treatment*).

To the best of our knowledge, a great part of the research work on the task of measuring the semantic textual similarity (STS) has been performed for the general-domain and originated from the SemEval STS task series (Agirre *et al.*, 2012, 2013, 2014, 2015, 2016). On the other hand, regarding STS between biomedical text snippets there has been less investigation and only a few manually annotated datasets have been made available (Soğancıoğlu *et al.*, 2017; Wang *et al.*, 2018a, 2020). Wang *et al.* (2018b, 2020) organized the first shared tasks on clinical STS and captured the attention of many research teams around the world that proposed their systems—our contribution on this problem is presented in Section 4.4. For example, Chen *et al.* (2021c) benchmarked several top-ranked deep learning models for measuring the relatedness between sentence pairs in the clinical domain. They evaluated word embedding–based models such as convolutional neural networks, sentence embedding–based models such as the BioSentVec pre-trained model (Chen *et al.*, 2019b), and transformer-based models such as BioBERT (Lee *et al.*, 2020), BlueBERT (Peng *et al.*, 2019), and ClinicalBERT (Alsentzer *et al.*, 2019). The authors concluded that BioSentVec and BioBERT achieved the highest results but emphasize that BERT models are much slower than the convolutional neural network and BioSentVec models.

Lastly, for further reading, we point the reader to other survey works on text classification (Aggarwal and Zhai, 2012; Mirończuk and Protasiewicz, 2018; Altınel and Ganiz, 2018; Kowsari *et al.*, 2019; Minaee *et al.*, 2022) and similarity measurement (Wang and Dong, 2020; Chandrasekaran and Mago, 2022).

4.2 Literature triage for precision medicine

Identifying relevant literature for harvesting particular biomedical knowledge is a paramount task, but expert curators invest a significant amount of time manually performing this annotation (Fang *et al.*, 2012; Karp, 2016). There are scientific articles that are more pertinent for extracting specific biomedical information such as protein–protein interactions or adverse drugs effects; and thus, the development of automatic solutions for biomedical document triage is essential and helpful to alleviate the manual

annotation work by professional curators. Also, a reduced number of more appropriate candidate documents allows automatic IE systems to take more time for making their predictions—regarding biomedical named entities and their relations—with a superior performance, and therefore a document classification step is crucial for eliminating noisy or less-relevant documents.

In this section we present supervised machine learning models for selecting biomedical documents relevant for precision medicine. The aim of precision medicine is to select the best treatments for different patient groups, considering individual variability in genes, environment, and lifestyle. Regarding genetic variability, valuable information about variants and how they affect protein–protein interactions is available in the scientific literature. Extracting and curating this information in an efficient manner requires the application of text mining algorithms. In this work, we experimented with classical machine learning and neural network classifiers for predicting which PubMed abstracts contained relevant information for extracting protein–protein interactions (PPIs) affected by genetic mutations. We also evaluated the impact of including additional training data from a similar dataset, containing general PPIs, as a semi-supervised or self-training approach.

The Precision Medicine task, part of the BioCreative VI community challenge in biomedical text mining, aimed to evaluate text mining approaches and tools for identifying and extracting information regarding the impact of genetic mutations on protein–protein interactions (Doğan *et al.*, 2017, 2019). The challenge consisted of two subtasks, namely document triage and relation extraction.

The application of text mining and automatic classification tools for document triage was evaluated in the BioCreative III PPI article classification task where the aim was to classify and rank articles relevant for curating protein–protein interactions (Krallinger *et al.*, 2010, 2011). The best system was based on a large margin classifier with features derived from gene named entity recognition, MeSH terms, and dependency parsing, and reached an area under the interpolated Precision/Recall curve (AUC iP/R) of 0.6798 and an F1-score of 0.6142 (Kim and Wilbur, 2010, 2011).

4.2.1 Materials and methods

We followed a supervised machine learning approach, and evaluated classical classifiers against deep learning models. In both cases, we used word embeddings to represent the words in the documents.

Table 4.1: Statistics of the Precision Medicine track dataset.

Partition	Abstracts	Positive	Negative
Training	4082	1729	2353
Test	1427	704	723
Total	5509	2433	3076

Data

The Precision Medicine task organizers provided a dataset split into training and test subsets consisting of 4082 and 1427 PubMed abstracts respectively, which were manually classified as relevant, that is, containing information regarding the impact of gene mutations on protein–protein interactions, or not relevant. Table 4.1 presents the statistics of the Precision Medicine track dataset in detail.

Apart from this annotated dataset, we exploited the use of the BioCreative III PPI ACT corpus as additional data (Krallinger *et al.*, 2010, 2011). This corpus consists of 12 280 MEDLINE abstracts, 2732 of which were annotated as containing PPI information. Although this annotation does not consider the impact of genetic mutations, as is the case of the task considered here, we tried to incorporate this data in a self-training approach.

Word embeddings

We used the word2vec implementation in the Gensim framework (Řehůřek and Sojka, 2010) and generated word embeddings from the complete MEDLINE database, corresponding to 15 million abstracts in English language. We created six models with vector sizes of 100 and 300 features and windows of 5, 20, and 50. The models contain around 775 thousand distinct words. We conducted several preliminary experiments using cross-validation on training data with these six variants of word embedding models, and our results demonstrated that the word embeddings model with 300 features and window size of 50 provided the best result for almost every configuration. Therefore, in this work, all the classifiers tested and presented here used the model with 300 features and window size of 50.

Classical classifiers

We compared three classifiers from the scikit-learn library (Pedregosa *et al.*, 2011): k-nearest neighbors (k-NN), logistic regression (LR), and multi-layer perceptron (MLP).

For the sake of readability, we denote these as *classical* classifiers because they are commonly employed in text classification and their off-the-shelf implementations, offered by scikit-learn, can be applied straightforward to this task with little effort. We used the default hyperparameters defined by scikit-learn with the exception of the following values: the k-NN used a number of neighbors of 99, and the MLP used a maximum of 2000 iterations for convergence. These classifiers were favorably chosen because their implementation allows to predict probability estimates which was relevant for this task since the system had to return a confidence value for the prediction. For example, if the system predicts that an article is relevant then it is also valuable to know how much confidence the system has in its prediction. Also, these classifiers were selected because they provided solid results in previous research on document classification (Kamath *et al.*, 2018; Kadhim, 2019; Shah *et al.*, 2020).

To obtain the document representation for the classifier, we tokenized the document and obtained the sequence of word vectors by simple look-up in the pre-calculated word2vec embeddings model. However, these classifiers are not directly applicable to sequences of distinct length and some form of aggregating these sequences is required. This is commonly addressed by summing or averaging the word vectors, resulting in a single vector representation of the document. We followed a similar approach, a weighted average of the word vectors, where each word vector was weighted by its TF–IDF value pre-calculated from the training data—note that for cross-validation evaluation the IDF values were calculated for every subgroup containing only the training partitions.

Deep learning classifiers

We applied different deep learning strategies based on (1) convolutional and (2) long short term memory (LSTM) layers. Convolutional neural networks (CNNs) have been extensively applied in image recognition and classification problems with very good performance (Rawat and Wang, 2017), and various works also demonstrate their application in text classification tasks (Rios and Kavuluru, 2015). On the other hand, LSTM networks which are a special type of recurrent neural networks (RNNs) can be more adequate for text-based tasks due to the sequential structure of natural language text, since these models contain feedback connections and can learn long-term dependencies in the input sequences (Hochreiter and Schmidhuber, 1997; Graves, 2012).

Overfitting in the training data is a common problem when using deep neural networks since they have a strong pattern-memorization ability. In general, the higher the number of layers and neurons they have the stronger their tendency to memorize and overfit. For that reason, in our experiments we included different strategies—early stop-

ping, dropout, and regularization—to avoid overfitting.

Our early stopping technique consisted in being vigilant of the loss value in a validation subset (for every training epoch) and terminating the training process if the loss did not decrease after five consecutive epochs. We used 10% of the training data, selected randomly, as validation set. We applied dropout to the output of the embedding and hidden layers so that a random selection of the output tensors is not used for updating the model weights, with the aim of forcing the model to learn a less biased representation of the data. Finally, L2 regularization is applied to the final layer to penalize large weights that could otherwise be assigned to biased input dimensions.

Differently from the classical classifiers that used a weighted average of the word vectors to represent each document, in the case of these deep neural network classifiers each document was represented by the concatenated sequence of its word vectors where a maximum sequence length of 1000 words was considered. This document representation was then forwarded to a convolutional recurrent neural network.

We empirically tested various network architectures and built three different systems based on convolutional and LSTM layers for the official evaluation in the task since, in comparison to the classical classifiers, these deep learning models provided superior results according to preliminary cross-validation experiments on training data (Table 4.2). All models were trained using the binary cross-entropy loss function and the RMSProp algorithm as optimizer (Tieleman and Hinton, 2012). Models were implemented in the Keras framework (Chollet *et al.*, 2015) with the TensorFlow backend (Abadi *et al.*, 2016) and executed on a machine with 12 CPU cores and 192 GB of memory.

System 1 The first system consists of a network architecture starting with an embedding layer, that represents each word in a document by its respective word vector, followed by three convolutional layers with average pooling. Each convolutional layer uses 128 filters with ReLU (rectified linear unit) activation and a kernel size of 3. The output is then connected to a bidirectional LSTM layer with 128 units, and to a final densely connected layer with sigmoid activation and L2 regularization with a penalty factor of 0.01. A dropout of 0.1 was included after the embedding layer and of 0.2 within the LSTM units.

This model was trained using 90% of the training data from the Precision Medicine dataset (10% left for validation), with a batch size of 32 samples and for a maximum of 100 epochs.

System 2 In the second system we followed a self-training approach to incorporate the BioCreative III PPI corpus as additional data. Since this corpus is annotated following different guidelines, we first applied the trained model from System 1 to infer the

Table 4.2: Five-fold cross-validation results on the Precision Medicine training set with classical and deep learning classifiers. The highest F1-score is highlighted in bold. k-NN: k-nearest neighbors. LR: logistic regression. MLP: multi-layer perceptron.

	Precision	Recall	F1-score
Classical classifiers			
k-NN (k=99)	0.618	0.553	0.582
LR	0.674	0.546	0.603
MLP	0.606	0.578	0.592
Deep learning classifiers			
System 1	0.637	0.681	0.651
System 2	0.640	0.692	0.664
System 3	0.698	0.735	**0.715**

relevance of these documents and selected the ones that were classified with a confidence value higher than 0.90. This equated to adding 9673 documents, all *pseudo-labeled* as not relevant, to the training data. The same network was then re-trained from scratch with the combined dataset, and a validation subset containing 10% of the training data from the Precision Medicine dataset was used to monitor the model performance during training.

System 3 The third system is similar to the System 1 but consists of a deeper network composed of three convolutional layers and three LSTM layers. The first layer of the network is the embedding layer as in Systems 1 and 2, but a dropout of 0.2 is applied to the embedding vectors. This layer is followed by three convolutional layers with average pooling, and each convolutional layer uses 64 filters with ReLU activation and a kernel size of 5. A dropout of 0.4 is applied after each pooling stage. This is then followed by a bidirectional LSTM layer and two unidirectional LSTM layers. All LSTM layers are composed of 128 units and use a dropout of 0.2. Finally, a dense layer is applied with sigmoid activation and L2 regularization with a penalty factor of 0.01.

4.2.2 Results and discussion

Table 4.2 shows the cross-validation results obtained on the Precision Medicine training set with the different classifiers. The best cross-validation result was obtained by the deeper network (System 3) which outperformed all classical classifiers by more than 11 percentage points in F1-score. Comparing to the classical classifiers tested, the deep learning systems obtained considerably better results in terms of recall and F1-score.

Table 4.3: Official results of the Precision Medicine track retrieved from the overview paper prepared by the challenge organizers (Doğan *et al.*, 2019). Results are evaluated on the Precision Medicine test set.

R*	Work		Average precision	Precision	Recall	F1-score
1	Fergadis *et al.* (2017)		0.7158	0.6289	0.7656	0.6906
2	Luo *et al.* (2017)		0.7253	0.6073	0.7997	0.6904
3	Matos and Antunes (2017a)	System 2	0.6677	0.5700	0.8736	0.6898
	(ours)	System 1	0.6616	0.5864	0.8338	0.6886
		System 3	0.6929	0.6070	0.7898	0.6864
4	Chen *et al.* (2017a)		0.5797	0.5713	0.8253	0.6752
5	Qu *et al.* (2017)		0.6632	0.5413	0.8835	0.6713
6	Tran and Kavuluru (2017)		0.6439	0.5438	0.8736	0.6703
7	Chen *et al.* (2017b)		0.6744	0.5361	0.8849	0.6677
8	Altınel *et al.* (2017)		0.5077	0.5022	0.9801	0.6641
9	Team 405		0.5871	0.5484	0.5710	0.5595
10	Wang *et al.* (2017c)		0.4904	0.4649	0.3480	0.3981

* R: rank. Teams ranked according to the F1-score evaluation.

The use of additional training data (System 2) helped to improve the results of the first deep network, although only by a small margin. Nevertheless, this result indicates that careful inclusion of related datasets, when available, can lead to better classification performance.

Table 4.3 presents our results obtained in the Precision Medicine test set, which demonstrates the suitability of our proposed systems based on convolutional recurrent neural networks. Our best F1-score result (0.6898) was obtained by System 2 evidencing that external training data was in fact beneficial, though not by a significant margin. We also emphasize that our three systems achieved competitive performance being close to the top teams—our highest F1-score result differs less than 1 percentage point from the best result. Finally, when comparing these results with the ones from cross-validation (Table 4.2) we observe that System 3 suffered from overfitting since its F1-score performance was deteriorated, in contrast to Systems 1 and 2 that improved their performance about 3 percentage points in F1-score when applied to unseen test data. However, we notice that System 3 obtained the highest average precision (0.6929) amongst our systems which indicates its superior adequacy for sorting documents according to their relevance.

4.3 Patient cohort selection for clinical trials

Clinical trials play a critical role in medical studies. However, identifying and selecting cohorts for such trials can be a troublesome task since patients must match a set of complex pre-determined criteria. Patient selection requires a manual analysis of clinical narratives in patients' records, which is a time-consuming task for medical researchers.

To simplify this selection process, attempts have been sought to automate cohort selection by performing patient phenotyping with informatics techniques, and this has in fact been demonstrated to be possible for some studies by the eMERGE (Electronic Medical Records and Genomics) consortium, which showed that algorithms can be used with effectiveness for phenotyping purposes (Pathak *et al.*, 2013).

While automating cohort selection is certainly of great interest, it faces major challenges namely how to define inclusion and exclusion criteria such that an algorithm can automatically and efficiently select patients in a dataset, or even how to integrate data from various sources (Pathak *et al.*, 2013), such as omics and EHR (electronic health record) data. EHR data is of particular interest as it can contain textual information stored in a structured form (data inserted in strict form fields), or in clinical narratives where text data is stored in an unstructured format (for example, free text report in a discharge record). Unstructured data has been getting increased attention since fusing information extracted from structured and unstructured data, instead of only resorting to the structured variant, can lead to significant performance improvements in a system (Ludvigsson *et al.*, 2013).

Extracting proper information from unstructured data such that it can be represented in a structured counterpart is a very difficult task. However, the capability to efficiently perform such extraction is of paramount importance, as automatic patient cohort selection systems can greatly benefit from it (Shivade *et al.*, 2014). It is due to this widely recognized potential that much research has focused on levering unstructured data from EHRs, using for that purpose natural language processing techniques to process unstructured text and extract meaningful content (Pathak *et al.*, 2013).

In this section we present an automatic classification system, based on handcrafted rules and machine learning models, that analyzes clinical reports and identifies which documents meet or do not meet specific medical criteria. The approach herein presented was developed and tested on the 2018 n2c2 (National NLP Clinical Challenges) Track 1 shared task[3] dataset where each patient record is annotated with 13 selection criteria (Stubbs *et al.*, 2019). The resulting hybrid approach attained a micro-average and macro-average F1-score of 0.8844 and 0.7271, respectively, in the n2c2 test set. In the remaining

[3] https://portal.dbmi.hms.harvard.edu/projects/n2c2-2018-t1/

of this section we describe the data resources used, explain the methodology developed, and present and discuss the obtained results. Part of the source code resultant from this work is available at:

https://github.com/ruiantunes/2018-n2c2-track-1.

4.3.1 Materials and methods

The objective of this work was to explore NLP techniques to solve the problem of automatic patient cohort selection. The problem consists in classifying 13 binary criteria for each patient given their clinical textual records. Classifying each criterion as 'met' or 'not met' was considered a single binary problem, where machine learning models were tested separately and rule-based methods were developed individually for each criterion. Our final system was a combination of both, where some criteria were better solved using heuristics and others using machine learning algorithms.

In this work, we used five classical machine learning classifiers from the scikit-learn and XGBoost libraries (Pedregosa *et al.*, 2011; Chen and Guestrin, 2016), and built two deep learning models using the Keras library (Chollet *et al.*, 2015). These are presented in detail in the next sections.

Data

The dataset used for this work was provided by the 2018 n2c2 (Track 1 shared task) organization, and is split into training and test sets containing 202 and 86 samples, respectively. Each sample comprises between 2 to 5 dated records of a single patient where the records are de-identified and the dates are modified to protect the identities of the participants. Nevertheless, the relative time intervals between patient records are kept to allow a timeline interpretation of these.

Each sample of the dataset has a list of 13 binary selection criteria that were manually annotated by medical professionals with a value of 'met' or 'not met' indicating whether or not a patient meets the pre-defined requirements of the criterion. Table 4.4 is based on the guidelines provided by the n2c2 organizers and shows a summary of the 13 selection criteria where each criterion was attributed a unique tag for identification purposes. From here on, for simplicity, we refer to the selection criteria as tags where each tag corresponds to a criterion representing a single binary classification problem.

Table 4.5 shows the dataset distribution where one can see that certain tags are highly imbalanced. There are tags, such as ASP-FOR-MI or MAKES-DECISIONS, where the 'met' class is much more frequent, but the opposite is also verified with the 'not met' class prevailing in tags such as DRUG-ABUSE or MI-6MOS. It is also relevant to note that

Table 4.4: Patient selection criteria of the 2018 n2c2 Track 1 dataset. Based on the annotation guidelines provided by the task organizers (Stubbs *et al.*, 2019).

Tag	Criteria
ABDOMINAL	Intra abdominal surgery, intestine resection, bowel obstruction
ADVANCED-CAD	Having at least two conditions about cardiovascular diseases (taking medications, myocardial infarction, angina, ischemia)
ALCOHOL-ABUSE	Current alcohol abuse
ASP-FOR-MI	Use of aspirin to prevent myocardial infarction
CREATININE	Serum creatinine larger than the limit of normal
DIETSUPP-2MOS	Taken a dietary supplement in the past 2 months
DRUG-ABUSE	Drug abuse
ENGLISH	Patient must speak English
HBA1C	HbA1c value between 6.5% and 9.5%
KETO-1YR	Diagnosis of ketoacidosis in the past year
MAJOR-DIABETES	Major diabetes-related complication
MAKES-DECISIONS	Patient must make their own medical decisions
MI-6MOS	Myocardial infarction in the past 6 months

the tag KETO-1YR only contains 'not met' labels, making supervised machine learning models unable to learn this criterion.

External resources

In order to expand the training data for some criteria, we used as external resource, the MIMIC-III critical care database (Johnson *et al.*, 2016), which is a large and freely-available database containing medications, laboratory measurements, imaging reports, and other clinical data from around 40 thousand adult patients. In this work, we used around 2 million clinical reports (1) to create word embeddings to be used in deep learning algorithms, (2) to be selected beforehand, pseudo-labeled, and used as additional training data in a semi-supervised setting, and (3) to find text patterns to help in the development of handcrafted rules.

Since the clinical reports in the MIMIC-III database possess ICD-9 diagnosis and procedure codes[4], we decided to explore those ICD-9 codes for the selection of relevant clinical reports from the MIMIC-III database. To do that, we manually mapped seven tags

[4] The ICD-9 codes are generated during patient admission for billing purposes. http://www.icd9data.com

Table 4.5: Class distribution of the 2018 n2c2 Track 1 dataset.

	Training set		Test set	
Tag	Met	Not met	Met	Not met
ABDOMINAL	76	126	30	56
ADVANCED-CAD	125	77	45	41
ALCOHOL-ABUSE	7	195	3	83
ASP-FOR-MI	163	39	68	18
CREATININE	82	120	24	62
DIETSUPP-2MOS	106	96	44	42
DRUG-ABUSE	10	192	3	83
ENGLISH	192	10	73	13
HBA1C	67	135	35	51
KETO-1YR	0	202	0	86
MAJOR-DIABETES	113	89	43	43
MAKES-DECISIONS	194	8	83	3
MI-6MOS	18	184	8	78

into a list of possible ICD-9 codes—the resulting mapping is presented in Table 4.6—and used the mapped codes to select relevant records from the database. The filtered list of clinical reports was then classified following a machine learning approach and reports with higher confidence were selected to be used as additional positive ('met') training samples.

Timeline restrictions

For the majority of tags, all the clinical records of each patient were concatenated resulting in a unique textual document per patient and, for simplicity, we ignored date information in clinical records. However, for tags KETO-1YR and MI-6MOS only the records from the past year and past six months, respectively, were considered since these criteria have time restrictions. Despite the criterion DIETSUPP-2MOS restricting intake of dietary supplements in the past two months, older records were also considered since these could indicate past supplements still being ingested.

Rule-based methods

From inspecting the training dataset, its statistics and understanding the selection criteria, we perceived that developing handcrafted rules to find text patterns would be

Table 4.6: ICD-9 medical codes related with some of the selection criteria of the 2018 n2c2 Track 1 dataset. These codes were manually selected. ICD: International Classification of Diseases.

Tag	ICD-9 diagnosis and procedure codes
ABDOMINAL	536.3, 536.4, 537.2, 537.3, 537.5, 539, 555.0, 555.2, 560, 564.4, 569.6, 751.1, 863, 864, 865, 866, 868, 996.81, 996.82, 996.86, 996.87, E879.5, 42, 43, 44, 45, 45.4, 45.7, 47, 50, 51, 52
ALCOHOL-ABUSE	303, 305.0, 980, V11.3
ASP-FOR-MI	E935.3
DIETSUPP-2MOS	V65.3, 280, 264, 265, 266, 267, 269
DRUG-ABUSE	304, 305.2, 305.3, 305.4, 305.5, 305.6, 305.7, 305.8, 305.9
MAJOR-DIABETES	249, 249.4, 249.5, 249.6, 249.7, 249.8, 250, 250.4, 250.5, 250.6, 250.7, 337.1, 357.2, 362.0, 588.1, 997.6, E878.5, 84.0, 84.1, 84.3, 84.91
MI-6MOS	410, 412

the most effective solution for certain tags. For instance, these applied to tags CREATININE and HBA1C where float values had to be found in the text near "creatinine" and "HbA1c" mentions, being an information that is not considered in the supervised learning approach (only in heuristics). Moreover, certain tags had one of the classes with very small support, and in those cases we expected that machine learning classifiers could not correctly learn due to the lack of training samples, whereas rule-based methods were expected to have better prediction capability. With this in mind, rules were implemented for every tag with the exception of the ABDOMINAL and MAJOR-DIABETES tags.

We developed two rule-based classifiers: one for submitting the results to the n2c2 shared task, and a second one after the challenge by improving some of the first rules by doing a more exhaustive error analysis on the training set (referred to as *modified rule-based classifier*). However, we were aware that this manual modification of the rules being evaluated in the training set could lead to overfitting. The rules were altered for the following nine tags: ADVANCED-CAD, ALCOHOL-ABUSE, ASP-FOR-MI, CREATININE, DRUG-ABUSE, ENGLISH, HBA1C, MAKES-DECISIONS, and MI-6MOS.

Both of the developed rule-based classifiers receive as input the raw text of the concatenated dated records. The rules implemented in both classifiers not only try to identify keywords specific to the criterion of interest using regular expressions, but also make complex decisions using if-else conditions. Rules for catching negation cases were also taken into account. Reports from the MIMIC-III database were also consulted to expand the rules, namely for the criteria ALCOHOL-ABUSE, DRUG-ABUSE, ENGLISH,

Table 4.7: The architecture of the deep learning models used in the patient cohort selection task. ReLU: rectified linear unit.

Model	Structure (top–bottom)
Fully connected neural network (FCNN)	Embedding layer Flatten layer Dense layer with 128 units ReLU activation Dense layer with 128 units ReLU activation Dropout with rate 0.2 Single unit with sigmoid activation
Convolutional neural network (CNN)	Embedding layer Convolutional layer with 128 filters ReLU activation Global max pooling operation Dense layer with 128 units ReLU activation Dropout with rate 0.2 Single unit with sigmoid activation

and MAKES-DECISIONS. Additionally, the DrugBank database (Wishart *et al.*, 2018) was used for compiling a list of supplements for the criteria DIETSUPP-2MOS.

Classical machine learning

To feed the classical machine learning classifiers, documents were firstly vectorized using a bag-of-words (BoW) approach. In the tokenization step, words were converted to lowercase, except for those with all uppercase letters as they could represent acronyms, and stop words were discarded. Preliminary results showed that feeding the classifiers with bigrams and trigrams in addition to unigrams did not result in significant improvements, thus in this work we only considered the use of unigrams.

The scikit-learn and XGBoost libraries were used to explore the following classical machine learning classifiers: AdaBoostClassifier, BaggingClassifier, DecisionTreeClassifier, GradientBoostingClassifier, and XGBClassifier. All classifiers were used with their respective default hyperparameter settings.

Deep learning

In this work we tested two deep learning classifiers: a fully connected neural network and a convolutional neural network. Both models were implemented with the Keras

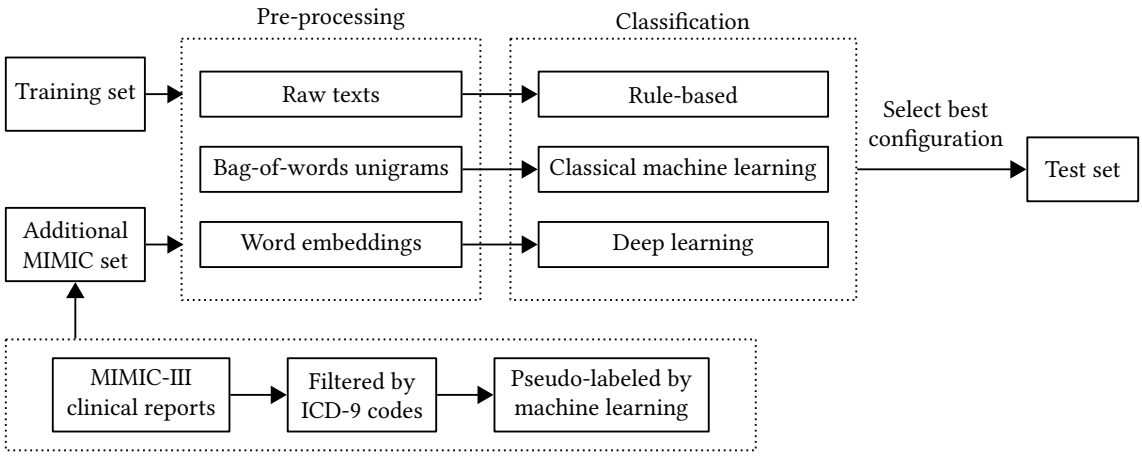

Figure 4.1: Overall system architecture used in the patient cohort selection task.

library and Table 4.7 presents the structure of each model.

Each document was represented by the concatenation of its words using word embeddings, with a fixed length of 5000 words. The word embeddings were created from around 2 million MIMIC-III clinical reports using the word2vec architecture (Mikolov *et al.*, 2013a) from the Gensim library (Řehůřek and Sojka, 2010). The final word embeddings model contained around 100 thousand distinct words.

From preliminary experiments we decided to use word embeddings generated with the skip-gram architecture, a feature size of 50, a window of 5, and all the words converted to lowercase. Furthermore, the models were trained with a batch size of 256 samples for a period of 30 epochs.

Overall system

The system herein described is composed of heuristics and machine learning models. Our approach consisted in selecting the methods which achieved the best results in the training set and applying them to the test set. The rule-based methods take as input raw text, while the classical machine learning classifiers use BoW unigrams, and the deep learning models use word embeddings.

In the supervised learning approaches, a few MIMIC-III clinical reports were first selected using the ICD-9 codes and then classified considering the probability output by an ensemble classifier pre-trained with the training set in a self-training setting. The ensemble calculates the average of the probabilities obtained from the five different classical machine learning classifiers. We tested this setup in the seven tags that were mapped to ICD-9 codes (Table 4.6).

Additionally, an optional pre-processing step was developed for the removal of tab-

Table 4.8: Detailed results with a baseline classifier applied to the 2018 n2c2 Track 1 test set. TP: true positive. TN: true negative. FP: false positive. FN: false negative. P: precision. R: recall. F1: F1-score. MI: micro-averaged. MA: macro-averaged.

Tag*	**Met**							**Not met**							**O. F1†**
	TP	TN	FP	FN	P	R	F1	TP	TN	FP	FN	P	R	F1	
ABD	0	56	0	30	0.0000	0.0000	0.0000	56	0	30	0	0.6512	1.0000	0.7887	0.3944
ADV	45	0	41	0	0.5233	1.0000	0.6870	0	45	0	41	0.0000	0.0000	0.0000	0.3435
ALC	0	83	0	3	0.0000	0.0000	0.0000	83	0	3	0	0.9651	1.0000	0.9822	0.4911
ASP	68	0	18	0	0.7907	1.0000	0.8831	0	68	0	18	0.0000	0.0000	0.0000	0.4416
CRE	0	62	0	24	0.0000	0.0000	0.0000	62	0	24	0	0.7209	1.0000	0.8378	0.4189
DIE	0	42	0	44	0.0000	0.0000	0.0000	42	0	44	0	0.4884	1.0000	0.6562	0.3281
DRU	0	83	0	3	0.0000	0.0000	0.0000	83	0	3	0	0.9651	1.0000	0.9822	0.4911
ENG	73	0	13	0	0.8488	1.0000	0.9182	0	73	0	13	0.0000	0.0000	0.0000	0.4591
HBA	0	51	0	35	0.0000	0.0000	0.0000	51	0	35	0	0.5930	1.0000	0.7445	0.3723
KET	0	86	0	0	0.0000	0.0000	0.0000	86	0	0	0	1.0000	1.0000	1.0000	0.5000
MAJ	43	0	43	0	0.5000	1.0000	0.6667	0	43	0	43	0.0000	0.0000	0.0000	0.3333
MAK	83	0	3	0	0.9651	1.0000	0.9822	0	83	0	3	0.0000	0.0000	0.0000	0.4911
MI6	0	78	0	8	0.0000	0.0000	0.0000	78	0	8	0	0.9070	1.0000	0.9512	0.4756
MI	312	541	118	147	0.7256	0.6797	0.7019	541	312	147	118	0.7863	0.8209	0.8033	0.7526
MA					0.2791	0.3846	0.3183					0.4839	0.6154	0.5341	0.4262

* The names of the tags were abbreviated for conciseness of the table. Please refer to Table 4.4 for consulting the full name forms.

† The overall F1-score is the average between the F1-scores from the 'met' and 'not met' classes.

ular information from text with the aim of limiting document content to natural text. At the final stage of the pipeline, the pre-processing style, the classifier (heuristics or machine learning), and the training data (with or without additional MIMIC-III reports) are chosen so that the best combination is applied to the test set. Figure 4.1 shows the final overall system architecture.

Note that for the tag KETO-1YR, the machine learning models were not trained, due to the lack of training samples, being the output pre-defined to always be 'not met' in this case.

4.3.2 Results and discussion

In this section, we present several results obtained by applying different methods in the training and test sets. Performance in the training set was evaluated using 3-fold cross-validation in the case of supervised learning algorithms, whereas the rule-based classifiers were applied directly to the complete training set because they did not need to learn from labeled training samples.

We used two evaluation metrics proposed by the n2c2 organizers which take into ac-

Table 4.9: Overall averaged F1-scores in the 2018 n2c2 Track 1 training and test sets. The highest value in each row is highlighted in bold. Ada: AdaBoostClassifier. Bag: BaggingClassifier. DT: DecisionTreeClassifier. GB: GradientBoostingClassifier. XGB: XGBClassifier. FCNN: fully connected neural network. CNN: convolutional neural network. RB: rule-based classifier. MRB: modified rule-based classifier.

(a) Evaluation on the training set.

Tag	Classical machine learning					Deep learning		Rule-based	
	Ada	Bag	DT	GB	XGB	FCNN	CNN	RB	MRB
ABDOMINAL	**0.6071**	0.5024	0.5281	0.5868	0.5654	0.5717	0.4257		
ADVANCED-CAD	0.6780	0.6525	0.6034	0.7208	0.7611	0.4951	0.4977	**0.8251**	**0.8251**
ALCOHOL-ABUSE	0.4899	0.4912	0.4807	0.4847	0.4912	0.4912	0.4912	0.8598	**1.0000**
ASP-FOR-MI	0.5144	0.4843	0.4946	0.5025	0.4603	0.4466	0.4466	0.7916	**0.8625**
CREATININE	0.7760	0.7959	0.7258	0.7723	0.8042	0.4189	0.5846	0.8895	**0.9118**
DIETSUPP-2MOS	0.6526	0.6432	0.5937	0.6926	0.7126	0.6539	0.5308	**0.7975**	
DRUG-ABUSE	0.7123	0.5795	0.6802	0.7370	0.4873	0.4873	0.4873	0.7020	**1.0000**
ENGLISH	0.7780	0.7780	0.8837	0.8837	0.5795	0.4873	0.4873	0.9172	**1.0000**
HBA1C	0.5429	0.5139	0.5588	0.5279	0.5702	0.4568	0.4006	0.9374	**0.9601**
KETO-1YR	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	
MAJOR-DIABETES	0.7375	0.6929	0.6554	**0.7483**	0.7429	0.5656	0.5473		
MAKES-DECISIONS	0.4873	0.4899	0.6192	0.5706	0.4899	0.4899	0.4899	0.8256	**1.0000**
MI-6MOS	0.4753	0.5306	0.5936	0.5097	0.5730	0.4767	0.4767	0.8026	**0.8778**
Micro-averaged	0.8198	0.8108	0.7682	0.8222	**0.8355**	0.7813	0.7858		
Macro-averaged	0.6117	0.5888	0.6090	**0.6336**	0.5952	0.5031	0.4897		

(b) Evaluation on the test set.

Tag	Classical machine learning					Deep learning		Rule-based	
	Ada	Bag	DT	GB	XGB	FCNN	CNN	RB	MRB
ABDOMINAL	0.7574	0.5590	0.7079	**0.7807**	0.6334	0.4658	0.3944		
ADVANCED-CAD	0.7977	0.7889	0.6178	**0.8227**	0.8114	0.4113	0.6673	0.7832	0.8089
ALCOHOL-ABUSE	**0.5896**	0.4911	0.4850	**0.5896**	0.4911	0.4911	0.4911	0.4850	0.4850
ASP-FOR-MI	0.4847	0.4401	0.5271	0.5469	0.4908	0.4416	0.4416	0.7095	**0.7426**
CREATININE	0.7329	0.7219	0.5933	0.7219	0.7110	0.4948	0.5411	**0.8295**	0.7862
DIETSUPP-2MOS	0.6728	0.5597	0.5930	0.6510	0.6162	0.5390	0.5083	**0.7943**	
DRUG-ABUSE	0.4850	0.4911	0.6815	0.6601	0.4911	0.4911	0.4911	0.7312	**0.9255**
ENGLISH	0.7559	**0.7929**	0.7915	0.7929	0.5983	0.4591	0.4591	0.6554	0.6554
HBA1C	0.6098	0.6048	0.6210	0.5773	0.5773	0.3676	0.3723	**0.9382**	0.8439
KETO-1YR	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.4971	
MAJOR-DIABETES	0.7902	0.8023	0.6975	**0.8721**	0.8023	0.6044	0.5966		
MAKES-DECISIONS	0.4911	0.4881	0.6277	0.4850	**0.7440**	0.4911	0.4911	0.6067	0.4911
MI-6MOS	0.4724	0.4756	0.4625	0.4724	0.4756	0.4756	0.4756	0.7281	**0.8102**
Micro-averaged	0.8331	0.8134	0.7775	**0.8356**	0.8258	0.7566	0.7676		
Macro-averaged	0.6261	0.5935	0.6081	**0.6517**	0.6110	0.4794	0.4946		

Table 4.10: Overall averaged F1-scores in the 2018 n2c2 Track 1 dataset with classifiers trained with 100 additional 'met' training MIMIC-III reports. The highest value for each tag and subset is highlighted in bold. Ada: AdaBoostClassifier. Bag: BaggingClassifier. DT: DecisionTreeClassifier. GB: GradientBoostingClassifier. XGB: XGBClassifier.

Tag	Evaluation on the training set					Evaluation on the test set				
	Ada	Bag	DT	GB	XGB	Ada	Bag	DT	GB	XGB
ABDOMINAL	**0.6669**	0.5890	0.6075	0.6655	0.6581	0.7511	**0.7617**	0.6595	0.7352	0.7507
ALCOHOL-ABUSE	0.6729	**0.7949**	0.7040	0.7040	0.7159	**0.5753**	0.4911	0.4819	0.4911	0.4911
ASP-FOR-MI	**0.5848**	0.5365	0.4976	0.4951	0.4873	0.4673	0.5232	**0.5784**	0.4788	0.4379
DIETSUPP-2MOS	0.6977	0.6781	0.6219	**0.7335**	0.7106	0.6012	0.6007	0.5697	0.6510	**0.6977**
DRUG-ABUSE	**0.7228**	0.7093	0.6456	0.6962	0.7093	**0.7440**	0.6910	0.6601	0.6815	**0.7440**
MAJOR-DIABETES	0.7335	0.7214	0.6231	**0.7537**	0.7489	0.7790	0.6512	0.7089	**0.8372**	0.8140
MI-6MOS	0.6930	**0.7023**	0.7020	0.6842	0.6842	**0.4724**	**0.4724**	0.4625	**0.4724**	**0.4724**

count the dataset imbalance: overall micro- and macro-averaged F1-scores. This overall score is the average of the two F1-scores of the 'met' and 'not met' classes. The evaluation metrics were calculated for each tag, thus enabling the analysis of each criterion separately.

For a clear understanding and detailed exposition of all the calculated metrics, Table 4.8 presents the results from a baseline classifier which simply attributed the most frequent label in the training set to all test samples. This baseline classifier attained a micro-F1 of 0.7526 and a macro-F1 of 0.4262 on the test set. To simplify the presentation of the results, we only present the overall F1-scores in the next experiments. However, detailed results containing true positive, true negative, false positive, and false negative counts as presented in Table 4.8, were examined during the refinement of our approach.

Table 4.9 shows the results of machine learning and rule-based methods evaluated on the (a) training and (b) test sets. Looking only at the evaluation in the training set, one can see that for each tag where rules were implemented, the rule-based method was the best performing classifier. On the other hand, deep learning models produced the worst results.

The AdaBoostClassifier and the GradientBoostingClassifier achieved the two highest macro-average F1-scores both in training and test sets. For the tags where the modified rule-based classifier was implemented, this classifier achieved the best results in the training set, but the same was not verified for the test set which shows that certain rules were overfit to the training set. For the tags ABDOMINAL, ADVANCED-CAD, and MAJOR-DIABETES, the results obtained with classical machine learning significantly improved in the test set proving that training with more data helped to increase their generalization ability.

Table 4.11: Overall averaged F1-scores in the 2018 n2c2 Track 1 dataset with tabulated information removed from the raw texts. The highest value for each row and subset is highlighted in bold. Ada: AdaBoostClassifier. Bag: BaggingClassifier. DT: DecisionTreeClassifier. GB: GradientBoostingClassifier. XGB: XGBClassifier.

Tag	Evaluation on the training set					Evaluation on the test set				
	Ada	Bag	DT	GB	XGB	Ada	Bag	DT	GB	XGB
ABDOMINAL	**0.6325**	0.5466	0.5676	0.5957	0.5733	0.7399	0.5419	0.6765	**0.8294**	0.6052
ADVANCED-CAD	0.6668	0.6952	0.5761	**0.7405**	**0.7405**	0.8089	0.7531	0.6549	**0.8350**	**0.8350**
ALCOHOL-ABUSE	0.4899	**0.4912**	0.4794	0.4847	**0.4912**	**0.5753**	0.4911	**0.5753**	**0.5753**	0.4911
ASP-FOR-MI	0.5064	0.5076	**0.5190**	0.5084	0.4603	0.4908	0.4454	**0.5947**	0.5086	0.4342
CREATININE	0.7726	0.7726	0.6966	0.7864	**0.7978**	0.6718	0.6946	0.6142	0.7141	**0.7329**
DIETSUPP-2MOS	0.6261	0.6574	0.6081	0.6631	**0.6830**	**0.7089**	0.6073	0.5216	0.6158	0.6728
DRUG-ABUSE	**0.7123**	0.4873	0.5825	0.7093	0.4873	0.4819	0.4911	**0.6601**	**0.6601**	0.4911
ENGLISH	0.7780	**0.9079**	0.8422	0.8837	0.5795	0.7559	**0.7929**	0.7737	**0.7929**	0.5983
HBA1C	**0.6087**	0.5568	0.5462	0.5216	0.5816	**0.5951**	0.4574	0.5232	0.5681	0.5577
KETO-1YR	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000
MAJOR-DIABETES	0.7637	0.6859	0.6354	**0.7725**	0.7094	0.7906	0.7673	0.6510	**0.8604**	0.8136
MAKES-DECISIONS	0.4873	0.4899	**0.5629**	0.4886	0.4899	0.4881	0.4911	0.6546	0.4850	**0.7440**
MI-6MOS	0.4753	0.5306	0.5601	0.5097	**0.5730**	0.4724	**0.4756**	0.4658	0.4724	**0.4756**
Micro-averaged	0.8270	0.8185	0.7657	0.8263	**0.8306**	0.8298	0.8031	0.7675	**0.8336**	0.8304
Macro-averaged	0.6169	0.6022	0.5905	**0.6280**	0.5897	0.6215	0.5776	0.6051	**0.6475**	0.6117

Table 4.10 shows the results when 100 additional 'met' MIMIC-III reports are used for training. Because these additional reports were classified with pre-trained models on the full training dataset, these results are indirectly biased. This explains the fact that the results in the training set have more noticeable improvements, whereas the improvements in the test set are less significant, in comparison to the original results from Table 4.9.

Table 4.11 presents the results obtained when applying classical machine learning with tabulated information removed from the raw texts. Significant differences were not found when compared to the results presented in Table 4.9.

Finally, Table 4.12 shows the final results when the best combination was selected by inspecting the results in the training set. The best combination in the training set achieved a micro-F1 of 0.9143 and a macro-F1 of 0.8596 whereas in the test set it attained a micro-F1 of 0.8844 and a macro-F1 of 0.7271. These results show that there is a clear overfitting to the training set because the macro-F1 on the test set is around 13 percentage points lower. We highlight that, in this optimal configuration, rule-based methods were mostly selected.

Our system contains rule-based methods and machine learning algorithms that are accordingly selected to better classify each criterion. We developed handcrafted rules

Table 4.12: Overall averaged F1-scores in the 2018 n2c2 Track 1 dataset with the best combination of methods selected by inspecting the evaluation in the training set. The methods that provided the best results in the training set were applied. The highest value for each tag is highlighted in bold.

Selected method	Tag	Training	Test
AdaBoostClassifier with 100 additional training MIMIC reports	ABDOMINAL	0.6669	**0.7511**
Rule-based classifier	ADVANCED-CAD	**0.8251**	0.8089
Modified rule-based classifier	ALCOHOL-ABUSE	**1.0000**	0.4850
Modified rule-based classifier	ASP-FOR-MI	**0.8625**	0.7426
Modified rule-based classifier	CREATININE	**0.9118**	0.7862
Rule-based classifier	DIETSUPP-2MOS	**0.7975**	0.7943
Modified rule-based classifier	DRUG-ABUSE	**1.0000**	0.9255
Modified rule-based classifier	ENGLISH	**1.0000**	0.6554
Modified rule-based classifier	HBA1C	**0.9601**	0.8439
Rule-based classifier	KETO-1YR	**0.5000**	0.4971
GradientBoostingClassifier with tabulated information discarded	MAJOR-DIABETES	0.7725	**0.8604**
Modified rule-based classifier	MAKES-DECISIONS	**1.0000**	0.4911
Modified rule-based classifier	MI-6MOS	**0.8778**	0.8102
	Micro-averaged	**0.9143**	0.8844
	Macro-averaged	**0.8596**	0.7271

for almost all the criteria. However, the process of creating adequate rules is hard and cumbersome since it requires an analysis of the data, not excluding the medical expertise that is oftentimes required. Moreover, while rule-based methods achieved good results, these require the development of a distinct algorithm for each criterion while machine learning classifiers do not face this problem, being easier to re-use.

In this task, classical machine learning classifiers worked much better when compared to deep learning classifiers. In most cases, deep learning models predicted the same label for every sample, behaving similarly to a baseline classifier and demonstrating that the dataset had a reduced size. Our results also show that machine learning classifiers provided better results for criteria with balanced labels, evidencing that other criteria lack in training data.

4.4 Measuring clinical semantic textual similarity

For years, technology advancements have been applied in health care with the goal of preventing, diagnosing and treating diseases, as well as improving the quality of life of the general population. These technological breakthroughs have brought new tools and information sources to physicians, aiding them in clinical workflows like patient follow-

up and clinical decision making, whilst also playing an important role in the transition into the personalized medicine paradigm.

Electronic health records are an example of such widely adopted technologies in medicine, providing an electronic infrastructure that can aggregate a multitude of medical information and support the medical act. EHRs provide a longitudinal view of patient trajectory comprehending the past and present of a patient's health condition, and can also comprehend a future component of the patient trajectory if future treatments and appointments are considered. Despite these containing rich contextual information in structured data that could for instance be explored for prediction modelling purposes (Wu *et al.*, 2010; Ferrão *et al.*, 2013; Ferrão *et al.*, 2016), large amounts of valuable patient information are stored in unstructured notes, commonly referred to as free text, which are often underexplored due to difficulties in processing this type of text (Neustein *et al.*, 2014).

Due to the large extent of information contained in an EHR, physicians are provided with the key aspect of *context* when reasoning on their medical decisions, making EHR a key component of the patient-centered notion of health care. However, even though its wide adoption has enabled an improvement on healthcare quality, certain challenges also arose with it. An example of such issue is the facilitated process of replicating information in medical text reports through copy-paste actions or by using pseudo-templates. This has impacted on EHR data quality since it can lead to less concise documentation, and an increased chance of introducing erroneous information, that can consequently compromise the quality of the medical act (Cohen *et al.*, 2013; Singh *et al.*, 2013).

Due to the importance of reducing the dimension and redundancy in EHR data, solutions for annotating relevant data and summarizing clinical text (Pivovarov and Elhadad, 2015) have been the focus of much research, mostly targeting clinical natural language processing. A possible recently explored approach to reduce clinical text redundancy is by assessing the semantic textual similarity between different text excerpts from an EHR. Investigation on semantic textual similarity has attracted particular attention in past years. SemEval, an ongoing series of evaluations of computational semantic analysis systems, started a pilot STS challenge track in 2012 which attracted the attention of the research community (Agirre *et al.*, 2012) and, due to its success, progressively organized a series of STS challenge tracks from 2012 through 2017 (Agirre *et al.*, 2013, 2014, 2015, 2016; Cer *et al.*, 2017). However, these shared tasks had the major drawback of being centered in general-domain text, whereas clinical text is inherently different in its characteristics. Therefore, an additional effort for the clinical domain was required.

The increasing interest in exploring and pushing forward existing research with clinical data, to further improve healthcare quality, led to the creation of dedicated resources

and challenges. To that extent, in recent years, text corpora specifically focused on clinical text were created along with shared tasks on clinical STS that leverage from those corpora.

In this section we present an approach to measure the semantic textual similarity between clinical sentences. We explore neural networks and different types of text preprocessing pipelines, and evaluate the impact of using word embeddings or sentence embeddings. We present our results on the 2019 n2c2/OHNLP Track 1 shared task[5] dataset, perform an error analysis, and discuss obtained results along with possible future improvements.

Biomedical and clinical STS resources

Recent years have shown efforts to bring STS to the clinical domain. Motivated by the rapidly increasing availability of textual data in the biomedical domain, by the need to facilitate the retrieval and analysis of this data, and also by the lack of suitable datasets for the development of appropriate systems, Soğancıoğlu *et al.* (2017) created BIOSSES—a benchmark dataset containing 100 sentence pairs from biomedical literature where each sentence pair was scored in a scale from 0 to 4 regarding its semantic similarity. Despite being an interesting initial effort, as the type of text found in biomedical literature significantly differs from that of medical narratives, an additional effort was still required.

With the goal of exploring STS to reduce clinical text redundancy, Wang *et al.* (2018a) assembled the MedSTS dataset containing 174 629 pairs of clinical sentences extracted from a clinical corpus at Mayo Clinic. From this pool of clinical sentence pairs, 1068 pairs were annotated by two medical experts regarding their semantic similarity, who classified each pair with a value within 0 (dissimilar) and 5 (equivalent), resulting in the creation of the MedSTS_ann dataset. The authors used MedSTS_ann to compare the performance of existing STS approaches on general and clinical domain STS datasets, and as expected they observed that performances obtained on MedSTS_ann were in general lower, demonstrating the higher complexity of clinical text.

Clinical STS shared tasks

Driven by the interest of motivating the research community to solve real world clinical problems, Wang *et al.* (2018b) organized a shared task on clinical STS where they released MedSTS_ann, hence making it the first available resource for the study of clinical STS. The pioneering shared task was titled BioCreative/OHNLP Challenge

[5] https://portal.dbmi.hms.harvard.edu/projects/n2c2-2019-t1/

2018 Task 2: Clinical Semantic Textual Similarity[6] and attracted the participation of four teams who developed automatic systems to measure the semantic relation of sentence pairs from MedSTS_ann.

Submitted solutions explored various techniques ranging from conventional machine learning to deep learning models. The winning system (Chen *et al.*, 2018) used a regression model that evaluates the scores obtained by eight trained models based on traditional machine learning and neural networks using sentence encoders. This system attained the highest Pearson correlation of 0.8328. The second placing team (Xiong *et al.*, 2018) obtained its best correlation result (0.8143) with an ensemble that combined the predictions from an attention-based CNN and a bidirectional LSTM models. The third team (Liu *et al.*, 2018b) used sentence embeddings by performing a weighted average of word vectors and obtained a Pearson correlation of 0.7789. The final team attained a best Pearson correlation of 0.7090 and did not disclose a detailed approach.

On a more recent note, building upon the experience from the BioCreative/OHNLP shared task on clinical STS, a collaboration between n2c2 and OHNLP resulted in the 2019 n2c2/OHNLP Track 1 on clinical semantic textual similarity (Wang *et al.*, 2020) with the objective of expanding previous work by providing novel annotated data, thus allowing further development and evaluation of systems on previously unseen data.

Sentence representation: BioWordVec *versus* BioSentVec

Posterior to the BioCreative/OHNLP shared task, members from the winning team proceeded their work by exploring the field of sentence representation, focusing on word and sentence embeddings. Due to the limitations of traditional word embeddings, which are computed at the word-level and trained for general-domain text, Zhang *et al.* (2019b) decided to develop a set of biomedical word embeddings that combined subword information from biomedical text with the biomedical vocabulary MeSH. The resulting biomedical word embeddings, named BioWordVec, performed better than word2vec embeddings in medical word pair similarity benchmarks, and were made publicly available to the community. The authors further improved these embeddings by training them on a new corpus containing over 30 million documents from PubMed articles and clinical notes from the MIMIC-III clinical database.

Word embeddings capture representations at word-level, thus to represent a sentence it is necessary to decompose it in words and combine the representations from each word. However, it is also possible to represent sentence text in a more straightforward approach, using instead sentence embeddings. Similarly to the word embeddings scenario, despite the existence of pre-trained sentence encoders for the general-domain,

[6] https://sites.google.com/view/ohnlp2018/home

biomedical and clinical text remained an unexplored field, thus Chen *et al.* (2019b) trained a sentence embeddings model on the same dataset, combining PubMed and MIMIC-III text data, used for BioWordVec. The result was BioSentVec[7], a publicly available sentence embeddings model suitable for biomedical and clinical text applications.

To assess the adequacy and effectiveness of these sentence embeddings, Chen *et al.* (2019b) tested BioSentVec with BIOSSES and MedSTS_ann. Whilst using a much simpler 5-layer deep learning model that only received two sentence vectors generated by BioSentVec as input, the authors managed to obtain a Pearson correlation of 0.836 in MedSTS_ann, slightly improving the previous state-of-the-art. This demonstrated that BioSentVec embeddings can effectively capture sentence semantics in clinical text. The authors performed a further comparative evaluation in MedSTS_ann using various sentence similarity models, including the latest bidirectional transformers in the clinical domain, such as BioBERT (Lee *et al.*, 2020), and observed that (1) their simpler approach still obtained the best performance, and (2) embeddings trained on large corpora are the best solution to capture sentence semantics in small datasets (Chen *et al.*, 2019a).

4.4.1 Materials and methods

In this work we evaluate a machine learning system in a supervised setting for predicting the semantic relatedness between clinical sentences. For this task, a real value in the interval [0, 5] is attributed to each pair of sentences: if two sentences are completely unrelated they have a similarity score of 0 (minimum); otherwise if they share the same semantic meaning their similarity value is 5 (maximum). Alike previous challenges on STS (Wang *et al.*, 2018b; Cer *et al.*, 2017) the evaluation metric is the Pearson correlation coefficient, where two lists of similarity values—system predictions and ground-truth—are compared.

In the following sections we describe the dataset used in this work, the different approaches used to pre-processing clinical text, the process of feature representation, and finally the deep learning models employed in our simulations.

Dataset

The dataset contained a total of 2054 sentence pairs, and it was beforehand split by the n2c2 organizers into training and test subsets. The training data allowed for model development, whereas the test data was used solely for official evaluation in the challenge. Only afterwards, the ground-truth scores of the test set were made available to the participating teams. Table 4.13 presents some statistics about the dataset. Surprisingly,

[7] https://github.com/ncbi-nlp/BioSentVec

Table 4.13: Statistics of the 2019 n2c2/OHNLP Track 1 dataset detailing the number of pairs for different levels of similarity. N: number of sentence pairs.

Set	N	Similarity scores				
		[0, 1]	]1, 2]	]2, 3]	]3, 4]	]4, 5]
Training	1642	312	154	394	509	273
Test	412	238	46	32	62	34
Total	2054	550	200	426	571	307

the distribution of the similarity scores between the two subsets (training and test) are somewhat discrepant since the test data contains a higher number of pairs with scores in the interval [0, 1].

A list of example clinical sentence pairs is provided in Table 4.14, containing the sentence pairs, their respective similarity score, and an explanation of the criteria used to assign the score.

Clinical text pre-processing

Before converting sentences into embedding representations, it was necessary to apply a pre-processing step to the clinical sentence pairs. In this work, three different pipelines were tested.

Base pre-processing The baseline pre-processing pipeline was simple, consisting only of two steps: (1) lowercase conversion of the text, and (2) tokenization using the NLTK tokenizer[8].

Advanced pre-processing with full stop word removal This pipeline was inspired on the pre-processing used by Chen *et al.* (2018) in the BioCreative/OHNLP 2018 Task on Clinical Semantic Textual Similarity, and had the objective of retaining as much semantic information as possible in pre-processed sentences.

The pipeline started with the separation of number ranges (0.3-1.8 → 0.3 to 1.8) which are frequent in lab analysis data. The second step was to extend numbers into their textual counterpart (78 → seventy-eight; 0.9 → zero point nine). Next, words connected with slashes, dashes and dots where separated with spaces (yes/no → yes / no; point-of-care → point - of - care). Then starting white spaces are removed, and consecutive spaces are converted to single spaces.

[8] https://www.nltk.org/

Table 4.14: Examples of clinical sentence pairs and respective similarity scores with explanations from Wang *et al.* (2018a).

Score	Examples
5	*The two sentences are completely equivalent, as they mean the same thing.* S1: Albuterol [PROVENTIL/VENTOLIN] 90 mcg/Act HFA Aerosol 2 puffs by inhalation every 4 h as needed S2: Albuterol [PROVENTIL/VENTOLIN] 90 mcg/Act HFA Aerosol 1-2 puffs by inhalation every 4 h as needed #1 each
4	*The two sentences are mostly equivalent, but some unimportant details differ.* S1: Discussed goals, risks, alternatives, advanced directives, and the necessity of other members of the surgical team participating in the procedure with the patient S2: Discussed risks, goals, alternatives, advance directives, and the necessity of other members of the healthcare team participating in the procedure with the patient and his mother
3	*The two sentences are roughly equivalent, but some important information differs or is missing.* S1: Cardiovascular assessment findings include heart rate normal, Heart rhythm, atrial fibrillation with controlled ventricular response S2: Cardiovascular assessment findings include heart rate, bradycardic, Heart rhythm, first degree AV Block
2	*The two sentences are not equivalent, but share some details.* S1: Discussed risks, goals, alternatives, advance directives, and the necessity of other members of the healthcare team participating in the procedure with (patient) (legal representative and others present during the discussion) S2: We discussed the low likelihood that a blood transfusion would be required during the postoperative period and the necessity of other members of the surgical team participating in the procedure
1	*The two sentences are not equivalent, but are on the same topic.* S1: No: typical 'cold' symptoms; fever present (greater than or equal to 100.4 ºF or 38 ºC) or suspected fever; rash; white patches on lips, tongue or mouth (other than throat); blisters in the mouth; swollen or 'bull' neck; hoarseness or lost voice or ear pain S2: New wheezing or chest tightness, runny or blocked nose, or discharge down the back of the throat, hoarseness or lost voice
0	*The two sentences are completely dissimilar.* S1: The risks and benefits of the procedure were discussed, and the patient consented to this procedure S2: The content of this note has been reproduced, signed by an authorized physician in the space above, and mailed to the patient's parents, the patient's home care company

The resulting text is converted to lowercase and tokenized with the NLTK tokenizer. Finally, a stop word removal is performed using a complete stop word list for biomedical literature[9], punctuation is removed from the tokens, and tokens composed of a single character are discarded.

[9] https://www.ncbi.nlm.nih.gov/IRET/DATASET

Advanced pre-processing with partial stop word removal This approach is similar to the previous processing pipeline, differing only in the stop word removal part. Since the list of stop words for biomedical text contains terms that can be important for retaining the semantics in clinical text, a smaller stop word list from Luo *et al.* (2019) was used instead, being composed of the following terms: 's, a, an, any, her, his, patient, that, the, these, this, those, your.

Feature representation

A crucial step in machine learning is feature representation, which aims to transform any kind of data (text, image, and others) into a numeric representation. For this clinical STS task we evaluated, separately, the use of word embeddings and sentence embeddings. We employed the publicly available BioWordVec and BioSentVec models created by Chen *et al.* (2019b). To encode the sentences using word embeddings we normalized the sum of the embedding vectors of their constituent words.

Deep learning model

The Keras library (Chollet *et al.*, 2015) was used to implement and test different neural network models. Our proposed model was derived from other state-of-the-art works (Chen *et al.*, 2019b, 2018) that use word and sentence embeddings. We tested various types and configurations of deep learning models, yet simpler models yielded better results, similarly to what was observed by Chen *et al.* (2019b).

We present a neural network whose inputs are the embedding representations of the respective two sentences encoded by (1) word embeddings or (2) sentence embeddings. In both cases we also included its multiplication (element-wise) and dot product. However, in the latter case we additionally included the cosine similarity since the sentence embedding vectors were not normalized, and the cosine similarity provided valuable information (preliminary results were higher).

The neural network contained a first layer with 512 units and ReLU activation. Also, we used Xavier normal initialization, a bias constant of 0.01, and L2 regularization of 0.001. Afterwards, a dropout rate of 0.4 was set, and a final unit with sigmoid activation performed the predictions. The stochastic gradient descent optimizer was employed with a learning rate of 1.0, and the mean squared error as loss function. A simplified scheme of the neural network with sentence embeddings is presented in Figure 4.2.

For fine-tuning the hyperparameters and adapting the model architecture we used repeated K-fold cross-validation where for each repetition we applied cross-validation with the training data being split into three subsets: training, validation, and test. These allowed for consistent model development without biasing in regard to the test set. We

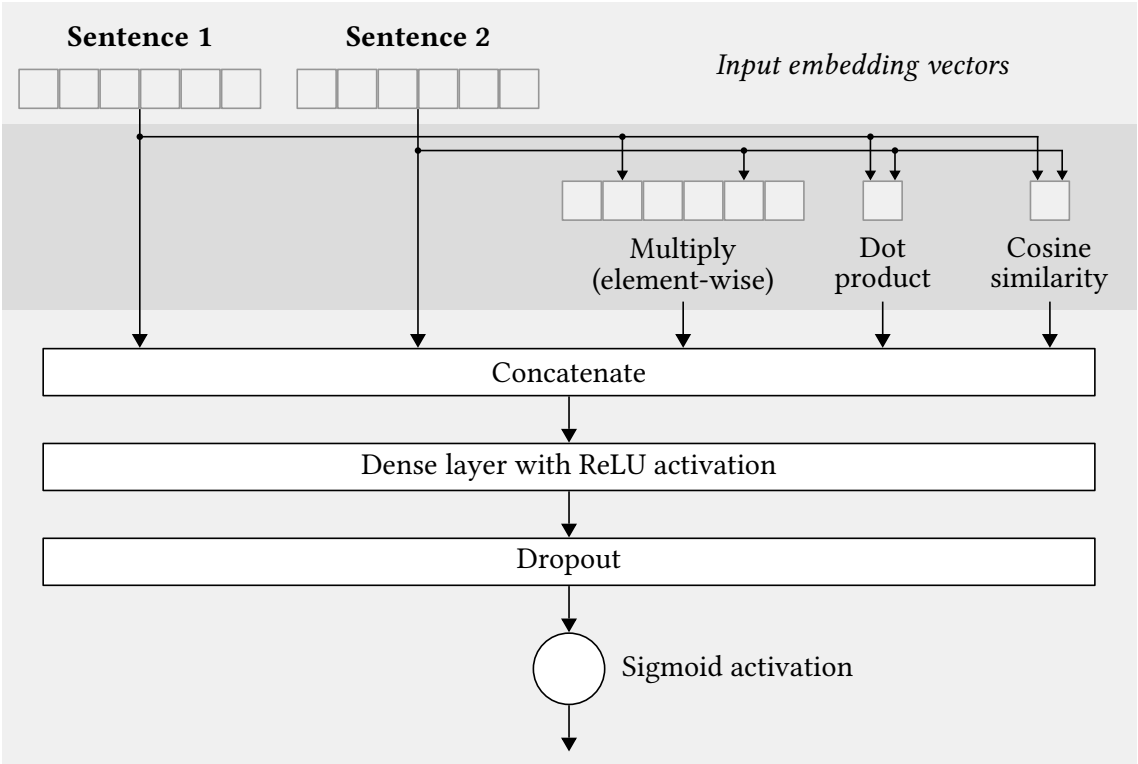

Figure 4.2: Neural network architecture using sentence embeddings for measuring semantic textual similarity. ReLU: rectified linear unit.

used the training subset to update the network weights, the validation subset to evaluate the model performance in each epoch, and finally the test subset was used for unbiased evaluation. The model that obtained the highest result in the validation subset was chosen to evaluate model performance in the test subset. Model training was halted when the performance in the validation subset did not improve for a period of 20 epochs.

After this intensive model refinement with thorough evaluation on training data, the configuration model was left unmodified to be applied on unseen test data. In the next section we evaluate the impact of using different pre-processing approaches, as well as using word embeddings or sentence embeddings as input vectors in the proposed neural network model.

4.4.2 Results and discussion

For gathering results we applied different text pre-processing pipelines and sentence representations: word embeddings *versus* sentence embeddings. We evaluated performance on (1) training data using repeated cross-validation and on (2) test data from predictions of a single evaluation. Results presented in the training set were obtained by averaging 30 separate scores, whereas evaluation on test data was slightly different: we

Table 4.15: Pearson correlation results obtained in the training and test sets of the 2019 n2c2/OHNLP Track 1 dataset by averaging, respectively, the results and the predictions of 30 individual models (number of folds × number of repetitions). For each distinct K-fold evaluation, the results of the configuration that obtained the highest performance in the training set, together with the corresponding results in the test set, are highlighted in bold. WE: word embeddings normalized sum. SE: sentence embeddings.

Set	Features	Text pre-processing	10-fold (3 repetitions)	5-fold (6 repetitions)	3-fold (10 repetitions)
Training	WE	Base	0.716	0.735	0.672
		Full	0.795	0.794	0.770
		Partial	0.771	0.773	0.728
	SE	Base	**0.811**	0.810	**0.792**
		Full	0.750	0.771	0.692
		Partial	0.809	**0.812**	0.791
Test	WE	Base	0.804	0.786	0.802
		Full	0.823	0.826	0.824
		Partial	0.819	0.818	0.819
	SE	Base	**0.837**	0.831	**0.836**
		Full	0.808	0.780	0.802
		Partial	0.818	**0.819**	0.826

averaged the predictions of 30 individual models. The idea behind this was to increase model robustness against the test data, because the final model is able to 'see', and learn from, the whole training data by using diverse folds for training and validation.

During model development on training data, a compromise between the sizes of training, validation, and test subsets sizes was necessary. Because of that, we explored the use of different K-fold split values to assess the impact of using less data for training and more for validation, and *vice versa*. We hypothesized that an optimal threshold considering enough training data and a solid evaluation on validation data could reflect an improvement on unseen data.

Table 4.15 presents detailed results from all these experiments. In general, the use of sentence embeddings provided superior results especially when using the base text pre-processing. Surprisingly, the full text pre-processing provided better results when using word embeddings, proving the benefit of stop words removal. To assess the effect of using different K-fold splits, we highlight the top scores according to the evaluation in the training set: for 10, 5, and 3-folds these were, respectively, 0.811, 0.812, and 0.792

in training data, and 0.837, 0.819, and 0.836 in test data. Therefore, we conclude that the use of different splits for training and validation did not affect significantly the results, and thus any of the number of folds we used would be acceptable.

The highest scoring model in training data (0.812), which consisted in using sentence embeddings with partial text pre-processing (5-fold setting), produced a correlation of 0.819 in test data. However, better results on test data were achieved when using sentence embeddings with the base text pre-processing (0.837, 0.831, and 0.836 for 10, 5, and 3-folds respectively). Furthermore, it is interesting to notice that word embeddings with full text pre-processing also attained good test results (0.823, 0.826, and 0.824 for 10, 5, and 3-folds respectively) similarly to those with sentence embeddings. Based on this, we believe that a careful combination of word and sentence embeddings may provide further improvement.

Overall, when analyzing training and test performances it is noteworthy that systems obtained higher scores on test data by a margin of approximately 0.02. We suspect that one of the reasons for this is the fact that evaluation on test data was performed by averaging several models (where each model was trained and validated on distinct data subsets) whilst the results reported on training data are the average of several simulations made by an individual model.

Finally, we compare our highest test result (0.837) with those achieved during the 2019 n2c2/OHNLP clinical STS task (Wang *et al.*, 2020) where a total of 87 valid system predictions were submitted: final aggregated results (Pearson correlation) presented a mean correlation of 0.712, a median of 0.829, and a maximum of 0.901. We consider that our model attained a positive performance given its simplicity, being slightly above the median score, but also that there still exists large margin for progress given the maximum result achieved.

Error analysis

A detailed error analysis was performed to better understand model behaviour—that is, which similarity levels the model predicted more correctly—and to perceive if this could be another reason that made results on the test set higher than those on the training set (due to the dataset imbalance). As such, we used the same similarity intervals as expressed in the dataset statistics (Table 4.13): [0, 1], ...,]4, 5].

To perform this evaluation, we started by computing the number of true positives, false positives, and false negatives to calculate precision, recall, and F1-score. A prediction was considered a true positive if the ground-truth was in the same similarity interval. Otherwise, the prediction was assigned as false positive and false negative in the corresponding intervals. To demonstrate this procedure, assume the model predic-

Table 4.16: Error analysis of predictions on the 2019 n2c2/OHNLP Track 1 dataset. Performance evaluation for different similarity levels using the following model from Table 4.15: sentence embeddings, base text pre-processing, and 10-fold evaluation. The three highest F1-scores in each set are highlighted in bold.

Metric	Training					Test				
	[0, 1]	]1, 2]	]2, 3]	]3, 4]	]4, 5]	[0, 1]	]1, 2]	]2, 3]	]3, 4]	]4, 5]
Precision	0.826	0.206	0.361	0.488	0.491	0.853	0.115	0.097	0.365	0.318
Recall	0.416	0.362	0.244	0.625	0.532	0.269	0.391	0.188	0.307	0.618
F1-score	**0.553**	0.263	0.291	**0.548**	**0.511**	**0.409**	0.177	0.128	**0.333**	**0.420**

tion is 3.7 whereas the ground-truth is 4.1. In such case, a false positive and a false negative would be added in the similarity intervals]3, 4] and]4, 5], respectively.

Table 4.16 presents this detailed error analysis in training and test data for a model with the following configuration: sentence embeddings, base text pre-processing pipeline, and 10-fold evaluation. Overall, it is noticeable that the system had more difficulty in correctly predicting sentence pairs with scores in the interval]1, 3].

It is interesting to note that the highest precision (around 0.8) was observed in the interval [0, 1] showing the model's ability to correctly detect completely dissimilar sentences. Since the majority of samples (around 58%) in test data were in this interval, this also supports our assumption that test results were higher because of their scores imbalance. Additionally, one can observe that F1-scores were higher in extreme similarity intervals, corroborating the assumption from Wang *et al.* (2018b) that machines can succeed at distinguishing completely similar or dissimilar sentence pairs but, alike humans, they struggle in distinguishing less clear relations of semantic similarity.

Finally, despite the fact that F1-scores on test data are somewhat smaller to those on training, we emphasize this type of error analysis does not elucidate directly the model performance, since (1) the test set is highly imbalanced and (2) near-correct scores can be misinterpreted as complete failures as they fall in a different interval by a slight margin (for example the model predicts a score of 0.9 whereas the ground-truth is 1.1).

4.5 Summary

In this chapter we presented methods for text classification and semantic textual similarity measurement. Regarding text classification, we described two different systems: one for identifying biomedical scientific abstracts that describe genetic mutations affecting protein–protein interactions; and a second one to classify certain criteria according

to patients' clinical records. For text similarity measurement we proposed a system that calculates the semantic agreement between clinical sentences.

We obtained competitive results with deep learning models in text classification when there was enough training data. However, in the case of classifying clinical narratives the dataset was relatively small and our results with deep learning were significantly lower which emphasizes the need of having sufficient labeled data. In that case, we had to create handcrafted rules or use traditional machine learning. For text similarity measurement, deep learning performed relatively well with the use of word embeddings or sentence embeddings.

Text classification and measurement of semantic textual similarity were studied in this same chapter since it is our understanding that these tasks partly overlap. For instance, documents with disparate semantic similarity may correspond to different topics in a text classification task. However, we also recognize that text similarity measurement has use beyond text classification and can be applied in other tasks such as sense disambiguation in which contexts of ambiguous terms are compared.

We stress that pre-selecting appropriate documents for mining a particular type of biomedical knowledge is a crucial step oftentimes performed manually by domain experts and that text classification algorithms may assist, or even surpass, human annotation requiring reduced manual effort. Thereby, techniques for improving biomedical text classification must continue to be investigated.

Chapter 5

Biomedical relation extraction

A vast amount of information is recorded electronically in natural language text containing knowledge about many concepts and their relationships. However, from a computational point of view, this text is considered *unstructured* because its information is not organized in strucured forms such as databases (McCallum, 2005). The human manual action of reading and interpreting the text to populate and enrich databases, with information in a computer readable form, is time consuming and burdensome. Besides, the abundant number of text documents[1] makes the manual extraction process, for knowledge base population (KBP), an impractical task—KBP consists in augmenting a knowledge base with discovered new facts about entities from a large text corpus (Balog, 2018). Therefore, it is mandatory to develop automatic methods to efficiently and accurately identify relationships in *unstructured text*, that can improve the quality and coverage of the databases, and match the continued growth of text knowledge sources (Ohta *et al.*, 1997; Temkin and Gilder, 2003). The computational task to address this problem is known as *relation extraction* (RE) and it is a fundamental piece in text data mining pipelines, that consists in identifying relationships between entities recorded in *unstructured text* (Bach and Badaskar, 2007; Pawar *et al.*, 2017; Liu, 2020). Relation extraction is relevant to KBP since discovered relationships can be used to complete missing information in a knowledge base. A reference example is the Stanford KBP system[2] which uses a relation extraction system as its workhorse (Surdeanu *et al.*, 2012; Angeli *et al.*, 2014).

Traditionally, the information to be extracted is usually limited to a particular domain with specific and pre-defined types of entities and relations—as Balog (2018) mentions, this paradigm may be referred to as *closed* information extraction. On the other hand, *open* information extraction operates in a domain-independent manner and does

[1] In the recent years, almost one million scientific publications have been indexed per year in MEDLINE, as illustrated in Figure 1.1.

[2] `https://nlp.stanford.edu/projects/kbp/`

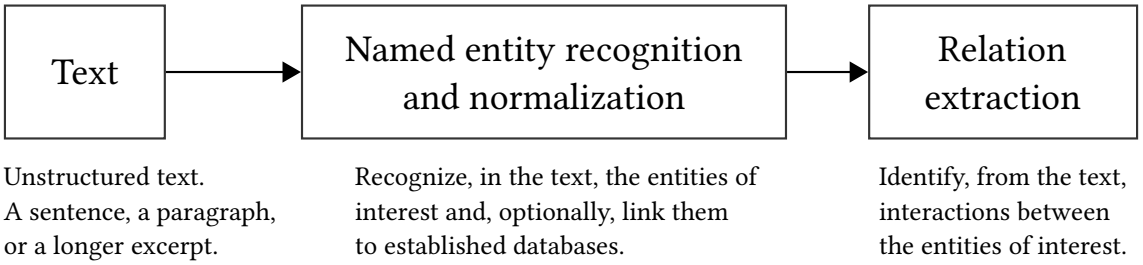

Figure 5.1: Entity and relation extraction pipeline.

not assume beforehand a particular set of entity and relation types (Banko *et al.*, 2007; Soderland *et al.*, 2010; Balog, 2018). In this chapter, we are interested in the former scenario, where a set of entity and relation classes, from a specialized domain, is known in advance. Specifically, our work is focused on studying RE methodologies to identify interactions between biomedical terms, such as chemicals and genes, in the life-sciences scientific literature. The automatic extraction of this information assists expert curators, in maitaining biological databases updated, and alleviates the demand for their work (Yeh *et al.*, 2003; Cotton *et al.*, 2007; Singhal *et al.*, 2016b). These databases foster knowledge discovery, helping researchers to test different hypotheses and deduce new facts. As an example, the extraction of relationships within the biomedical domain helps to gather evidence about diseases and adverse drug reactions (Gonzalez-Hernandez *et al.*, 2022). For instance, studies have shown that patients with diabetes have an increased risk of developing Alzheimer disease, but the underlying biological mechanisms responsible for that correlation are not well understood (Sims-Robinson *et al.*, 2010), and text mining–based tools can deliver key insights (Saik and Klimontov, 2021).

Relation extraction can also boost applications such as question answering (Wang *et al.*, 2012) and information retrieval (McDonald *et al.*, 2005). Generally, it follows a *named entity recognition* step where entity mentions are identified in the text and grounded to standardized ontologies or vocabularies if a normalization method is employed. Although the *normalization* task is often disregarded, considering a joint resolution for all the three tasks—entity recognition, entity normalization, and relation extraction—is more useful in a real-world scenario (Yaseen *et al.*, 2019) since researchers benefit from these systems for their text mining tasks (Kim *et al.*, 2019). Figure 5.1 illustrates the classical RE pipeline.

This chapter is focused on biomedical relation extraction from the scientific literature. We first give a brief overview of the development of the RE task enumerating biomedical NLP worldwide challenges and shared-tasks, and describe commonly used datasets. Then, we introduce a deep neural network model for identifying chemical–protein interactions in PubMed scientific abstracts. Extensive experiments were per-

formed to train the neural network under different configurations including model architecture, hyper-parameters, and training data. Finally, we discuss the limitations of our method, perform a detailed error analysis, and offer viable future research directions.

5.1 Background

A major research effort that tackled the relation extraction task was the ACE (Automatic Content Extration)[3] program conducted by the National Institute of Standards and Technology (NIST) of the United States, and the Linguistic Data Consortium (LDC) at the University of Pennsylvania (Doddington *et al.*, 2004). This challenge included the recognition of entities (such as persons, organizations, or geographical locations) and relations (such as employment or affiliation between a person and an organization, or a location relationship) in English, Chinese, and Arabic texts. The LDC's ACE annotators tagged broadcast transcripts, newswire, and newspaper data producing training and test sets for common research task evaluations. The ACE program succeeded the Message Understanding Conferences (MUCs), that similarly addressed the extraction of information related to person and organization entities (Sundheim, 1996; Grishman and Sundheim, 1996). Initiated in 2008, the Text Analysis Conference (TAC)[4] followed the ACE research program consisting in a series of workshops for large-scale evaluation of NLP systems. The first TAC cycle addressed tasks such as question answering and summarization, but knowledge base population was the main task in later editions (Ji *et al.*, 2010, 2011; Ji and Grishman, 2011; Ellis *et al.*, 2012; Surdeanu, 2013). The aforementioned challenges focused on the development of NLP systems for the general-domain, although in recent years TAC has also been targeting biomedical information extraction tasks. In 2017, TAC organized a task for identifying adverse drug reactions (ADRs) found in drug labels (Roberts *et al.*, 2017). Following up, in the 2018 and 2019 editions, challenges for extracting drug–drug interactions (DDIs) from drug labels were conducted (Demner-Fushman *et al.*, 2018; Goodwin *et al.*, 2019).

Likewise, other biomedical text mining efforts, particularly for relation extraction, took place during the past years. The LLL05 (Learning Language in Logic 2005 workshop) challenge focused on extracting gene–protein interactions in biology abstracts from the MEDLINE bibliographic database (Nédellec, 2005).

BioCreative II introduced the protein–protein interaction (PPI) extraction task where protein interaction pairs had to be predicted from PubMed records (Krallinger *et al.*, 2007, 2008). BioCreative V proposed a challenge task for automatic extraction of chemical–

[3] https://www.ldc.upenn.edu/collaborations/past-projects/ace
[4] https://tac.nist.gov

disease relations (CDRs), from PubMed abstracts, which aimed to support biocuration, new drug discovery, and drug safety surveillance (Wei *et al.*, 2015, 2016). BioCreative VI addressed two relation extraction tasks, relevant for precision medicine, both using PubMed abstracts: (1) the aim of track 4 was to identify experimentally verified PPIs affected by genetic mutations (Doğan *et al.*, 2017, 2019); and (2) track 5 promoted the development of systems for detecting relations between chemicals and GPROs (gene and protein related objects) (Krallinger *et al.*, 2017a). Similarly, built with experience from the past edition, BioCreative VII prepared a challenge task for text mining chemical–protein interactions (CPIs) but used a higher number of PubMed records for training and evaluation, and considered the identification of more interaction types (Miranda *et al.*, 2021).

In 2010, i2b2[5] partnered with VA (Veterans Affairs) Salt Lake City Health Care System to promote a task for detecting relations between medical problems, tests, and treatments in patient clinical reports (Uzuner *et al.*, 2011). Similar work had also focused on the extraction of disease–treatment semantic relations (*cure*, *prevent*, or *side effect*) but using biomedical abstracts from the MEDLINE bibliographic database (Rosario and Hearst, 2004; Frunza and Inkpen, 2010). The 2012 i2b2 challenge focused on temporal relations between clinical events (problems, tests, or treatments) and temporal expressions (dates, times, or durations) documented in clinical discharge summaries (Sun *et al.*, 2013). The 2018 n2c2 shared task focused on the extraction of adverse drug events (ADEs) from clinical records (Henry *et al.*, 2021).

The BioNLP Shared Task (BioNLP-ST) series started in 2009 toward fine-grained information extraction in the biomedical domain. The first edition, based on the GENIA corpus (Ohta *et al.*, 2002; Kim *et al.*, 2003; Kim *et al.*, 2008), was focused on the extraction of molecular events involving proteins and genes (Kim *et al.*, 2009, 2011a). This task, then renamed *Genia event extraction*, was hosted again in the second, third, and fourth editions of the BioNLP-ST. The 2011 edition aimed to evaluate generalization of the systems to full-text papers (Kim *et al.*, 2011c), whereas in the 2013 and 2016 editions the task was extended toward knowledge base construction (Kim *et al.*, 2013a, 2015, 2016). The *Bacteria Biotopes* task introduced in the BioNLP-ST 2011 consisted in extracting bacteria localization events, identifying the habitats of bacteria, from textbook documents (Bossy *et al.*, 2011, 2012). Later editions refined this task by (1) considering a more comprehensive and fine-grained categorization (normalization) of the entity mentions and respective events to domain knowledge sources such as the NCBI (National Center for Biotechnology Information) taxonomy (Federhen, 2012) and the OntoBiotope ontology (Nédellec

[5] The i2b2 (Informatics for Integrating Biology and the Bedside) NLP challenges for clinical data are now housed in the Department of Biomedical Informatics at Harvard Medical School as n2c2 (National NLP Clinical Challenges).

et al., 2018), and (2) using scientific literature such as abstracts and full-text articles from the PubMed database (Bossy *et al.*, 2013b; Deléger *et al.*, 2016; Bossy *et al.*, 2019). Also, the BioNLP-ST 2011 (Kim *et al.*, 2011b; Pyysalo *et al.*, 2012) organized other information extraction challenges including (1) the *Epigenetics and Post-translational Modifications* task which focused on events of epigenetics interest (Ohta *et al.*, 2011), (2) the *Infectious Diseases* task targeting biomolecular mechanisms of infectious diseases (Pyysalo *et al.*, 2011a), and (3) the *Entity Relations* task concerning part-of relations between a gene or protein and an associated entity (Pyysalo *et al.*, 2011b). Likewise, the BioNLP-ST 2013 (Nédellec *et al.*, 2013; Pyysalo *et al.*, 2015; Bossy *et al.*, 2015) addressed other challenges including (1) the *Cancer Genetics* task that focused on the extraction of events relevant to cancer (Pyysalo *et al.*, 2013), (2) the *Pathway Curation* task concerning the extraction of biomolecular reactions for supporting the development of biomolecular pathway models (Ohta *et al.*, 2013), and (3) the *Gene Regulation Ontology* and *Genic Regulation Network* tasks which addressed the identification of events and relations relevant to gene regulation (Kim *et al.*, 2013b; Bossy *et al.*, 2013a). The BioNLP-OST (Open Shared Task)[6] 2019 also organized an extraction task addressing mutation–disease relationships to support knowledge discovery for drug repurposing (Wang *et al.*, 2019).

The SemEval-2013 Task 9, DDIExtraction 2013, evaluated the extraction of drug–drug interactions from biomedical texts (Segura-Bedmar *et al.*, 2013) which followed the DDIExtraction-2011 challenge task (Segura-Bedmar *et al.*, 2011).

Much of the past work on biomedical information extraction relied on hand-labeled data to enable automatic training of machine learning models in a supervised fashion, but the manual annotation of ground-truth information, such as entity mentions and their interactions, in text documents is a labored and expensive task requiring expert professionals (Baumgartner *et al.*, 2007; Howe *et al.*, 2008; Winnenburg *et al.*, 2008; Karp, 2016). Knowledge-based, unsupervised, and semi-supervised approaches aim to counteract the need of gold-standard data in developing competitive relation extraction systems. These type of techniques have also been addressed in past work. For instance, Craven and Kumlien (1999) present a distant supervision approach to extract information from text using knowledge bases, where they train classifiers using *weakly labeled* training data. The authors observed that for many information extraction tasks there are external knowledge sources that could be coupled with documents to create what they called *weakly labeled* training examples. They coined the term "weak" because each training instance does not consist of a precise annotated document, but rather it consists of a known fact—from a knowlege base—that may be asserted in a particular

[6] In 2019, the BioNLP-OST was organized as a reformulation of the previous efforts around the BioNLP-ST.

document.

Another example is the work of Mintz *et al.* (2009) that continued to investigate the *distant supervision* mechanism to create *weakly labeled* data of any size. The authors relied on Freebase, a large semantic database, containing thousands of relationships (Bollacker *et al.*, 2008). For every pair of entities present in the database, the authors found all sentences that contained those entities in an unlabeled corpus, and used these to train a relation classifier. Distant supervision is now a mature technique and has been widely used in relation extraction systems that can benefit from knowledge bases as a source of training data (Smirnova and Cudré-Mauroux, 2019).

Early work on biomedical relation extraction focused on protein–protein interactions (Blaschke *et al.*, 1999) and relationships betweens genes and drugs relevant to cancer (Rindflesch *et al.*, 1999). Temkin and Gilder (2003) proposed a system for extracting protein interaction information from unstructured text. Their approach is based on an external database containing ditictionaries that are then used by a lexical analyzer followed by a parser, constructed around a context-free grammar (Chomsky, 1956; Aho *et al.*, 2007), that identifies interactions based on the rules of the grammar. The rules of the context-free grammar were manually derived from 500 PubMed abstracts, whereas their final system was evaluated using a test corpus of 100 PubMed abstracts.

Airola *et al.* (2008) propose a graph kernel–based approach for PPI extraction. They assessed their method on five PPI annotated datasets, which as the authors state, provided the most comprehensive evaluation for a machine learning–based PPI-extraction system, and achieved a 0.564 F-score. The authors performed cross-corpus evaluation for understanding how a trained model on a specific dataset with its own characteristics generalizes to a different dataset. They also emphasize that the comparison of RE systems must be done carefully because the use of different evaluation strategies and resources make results incomparable.

Previous research on biomedical relation extraction have been focusing on protein–protein interactions (Krallinger *et al.*, 2011) and relations between drugs, genes and diseases (Frijters *et al.*, 2010; Krallinger *et al.*, 2017b). Machine learning methods combined with kernel functions to calculate similarities between instances given some representation, were shown to achieve good results in textual relation extraction.

As opposed to the traditional machine learning methods employed in initial works, deep learning techniques eliminate the need for feature engineering, instead using multiple data transformation layers that apply simple non-linear functions to obtain different levels of representation of the input data, intrinsically learning complex classification functions (LeCun *et al.*, 2015). These strengths have brought much attention with significant successes in NLP tasks, including word sense disambiguation (Jimeno-Yepes,

2017), text classification (Kim, 2014; Kowsari *et al.*, 2017), and named entity recognition (Habibi *et al.*, 2017; Lyu *et al.*, 2017).

Several works have demonstrated the use of deep neural networks for biomedical relation extraction and classification. For example, Nguyen and Grishman (2015) used a CNN (convolutional neural network) with pre-trained word embeddings, outperforming previous state-of-the-art systems for relation classification. Nonetheless, the sequential nature of natural texts can be better modeled by recurrent networks, which contain a feedback loop that allows the network to use information regarding the previous state. LSTM (long short-term memory) networks are a special type of recurrent neural networks (RNNs) in which a set of information gates is introduced in the processing unit that allow these networks to memorize long-term dependencies while avoiding the vanishing gradient problem. Wang *et al.* (2017b) used BiLSTM (bidirectional LSTM) networks and features from the dependency structure of the sentences obtaining an F1-score of 0.720 in the DDIExtraction 2013 corpus. Zhang *et al.* (2017) also used BiLSTM models for extracting drug–drug interactions (DDIs) achieving a state-of-the-art F1-score of 0.729 in the same dataset. They integrated the shortest dependency path (SDP) and sentence sequences, and concatenated word, part-of-speech and position embeddings into a unique embedding, and an attention mechanism was employed to give more weight to more relevant words.

Methods for extracting chemical-disease relations were evaluated in the BioCreative V CDR task, in which participants were required to identify disease and chemical entities and relations between them (Wei *et al.*, 2016). Using the provided gold-standard entities, Zhou *et al.* (2016) achieved an F1-score of 0.560 with a hybrid system consisting of a feature-based SVM (support vector machine) model, a tree kernel-based model using dependency features and a LSTM network to generate semantic representations. This result was improved to 0.613 by inclusion of post-processing rules. The same result was achieved by Gu *et al.* (2017), also with an hybrid system combining a maximum entropy model with linguistic features, a CNN using dependency parsing information, and heuristic rules.

Regarding chemical–protein relation extraction, the state-of-the-art results were achieved by teams participating in the BioCreative VI ChemProt challenge (Krallinger *et al.*, 2017a), with some improvements described in follow-up works. The best participating team achieved an F1-score of 0.641 using a stacking ensemble combining a SVM, a CNN, and a BiLSTM (Peng *et al.*, 2017, 2018). Lemmatization, PoS (part-of-speech), and chunk labels from the surrounding entity mentions and from the shortest dependency path were used as features for the SVM classifier. For the CNN and BiLSTM, the sentence and shortest path sequences were used, where each word was represented

by a concatenation of several embeddings (PoS tags, dependencies, named entities, and others). Corbett and Boyle (2017b) achieved an F1-score of 0.614 using pre-trained word embeddings and a network model with multiple LSTM layers, with the ChemListem NER system used for tokenization (Corbett and Boyle, 2017a). This result was improved to an F1-score of 0.626 in post-challenge experiments (Corbett and Boyle, 2018). Mehryary *et al.* (2017) proposed two different systems: a SVM classifier and an ensemble of neural networks that use LSTM layers. Both systems took features from the dependency parsing graph, although the SVM required more feature engineering. They combined the predictions of the two systems, yet the SVM alone produced the best F1-score (0.610). After the challenge they achieved an F1-score of 0.631 by using their improved artificial neural network (ANN) (Mehryary *et al.*, 2018). Lim and Kang (2017) used ensembles of tree-LSTM networks, achieving an F-score of 0.585 during the challenge. They later improved this result to 0.637 with a revised pre-processing and by using more members in the ensemble, and equaled the best challenge F1-score (0.641) using a shift-reduce parser based network architecture (Lim and Kang, 2018). Lung *et al.* (2017, 2019) achieved an F1-score of 0.567 using traditional machine learning. Neural networks with attention mechanisms were also followed by Liu *et al.* (2017, 2018a) and Verga and McCallum (2017), but achieved lower results. However, the use of attention layers (Bahdanau *et al.*, 2014; Vaswani *et al.*, 2017) has been shown to be effective in different information extraction tasks such as document classification (Yang *et al.*, 2016) and relation extraction (Shen and Huang, 2016), being an interesting direction to explore.

Zhang and Lu (2019) present a semi-supervised approach based on a variational autoencoder for biomedical relation extraction. They evaluated their method in the ChemProt dataset experimenting with different number of labeled samples, showing that adding unlabeled data improves the relation extraction mainly when there are only a few hundred training samples. Using 4000 (from a total of 25 071) labeled training instances together with unlabeled data taken from the remaining training instances (with true labels removed), their semi-supervised method achieved an F-score of 0.509.

Lastly, a recent work by Zhang *et al.* (2019c) achieved the state-of-the-art F-score of 0.659 using BiLSTM models with deep context representation (providing superior sentence representation compared to traditional word embeddings) and multi-head attention.

Huang *et al.* (2017) used a support vector machine and LSTM networks to identify drug–drug interactions, but most recent works often solely use neural networks. For example, Asada *et al.* (2021) used SciBERT—a tranformer model trained on biomedical text (Beltagy *et al.*, 2019)—and external database information to boost DDI extraction performance.

For more information on related works and challenges about extracting relations from the biomedical literature we point the reader to major review works (Hirschman *et al.*, 2002; Krallinger *et al.*, 2005; Ananiadou *et al.*, 2006; Huang and Lu, 2016; Zhang *et al.*, 2019d; Zhao *et al.*, 2021). Zhang *et al.* (2019d) present an extensive review of neural network–based approaches for biomedical RE classification. They discuss works identifying PPIs and DDIs in the biomedical literature, and methods including CNNs and RNNs. They present several corpora, and discuss different word embeddings models. Finally, they detail current challenges and present potential directions to further improve the performance of the biomedical RE task.

Regarding the availability of high-quality datasets for assessing the performance of biomedical relation extraction systems many corpora have been manually annotated over the past years (Table 5.1). Some of these were developed for worldwide challenges and shared-tasks whereas others were published by particular research groups which allowed other researchers to assess and compare their methods with previous results.

Table 5.1: Datasets for biomedical relation extraction, presented in chronological order. ADE: adverse drug effect. CDR: chemical–disease relation. DDI: drug–drug interaction. IPS: interaction pair subtask. PPI: protein–protein interaction.

Resource	Description
BioText (Rosario and Hearst, 2004)	This dataset concentrates on disease–treatment semantic relations (*cure*, *prevent*, or *side effect*) using biomedical text found in the titles and abstracts from the MEDLINE 2001 database. An annotator with a biological background performed the labeling. `https://biotext.berkeley.edu/data/dis_treat_data.html`
AIMed (Bunescu *et al.*, 2005) (Bunescu and Mooney, 2005b) (Bunescu, 2007)	The corpus is annotated with human protein names (genes and proteins are interchangeable) and their interactions. It consists of 225 MEDLINE abstracts, containing 4084 protein mentions and around 1000 interactions. `https://www.cs.utexas.edu/ftp/mooney/bio-data/`
BioInfer (Pyysalo *et al.*, 2007)	The resource contains 1100 sentences from abstracts of biomedical research articles. These are annotated with named entities of the protein, gene, and RNA types, their relationships, and syntactic dependencies. The corpus contains a total of 6349 entities and 2662 relationships.

Resource	Description
BioCreative II PPI IPS (Krallinger *et al.*, 2007) (Krallinger *et al.*, 2008)	A corpus of full-text articles annotated with binary protein–protein interaction pairs. It is split into two sets: a *training* collection of 740 articles and a smaller *test* set of 358 articles. http://biocreative.sourceforge.net/bc2_ppi_ips.html
2010 i2b2/VA Challenge (Uzuner *et al.*, 2011)	A corpus of patient reports focused on medical concept extraction and relation classification between medical problems, tests, and treatments. It contains a total of 394 *training* reports and 477 *test* reports manually annotated. https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/
ADE corpus (Gurulingappa *et al.*, 2012b) (Gurulingappa *et al.*, 2012a)	A corpus containing 2972 MEDLINE case reports annotated with two types of relations: drug-related adverse events (signs, symptoms, diseases, disorders, and others) and drug–dosage information (such as quantitative measurements or frequency mentions). https://sites.google.com/site/adecorpus/
DDI corpus (Herrero-Zazo *et al.*, 2013) (Segura-Bedmar *et al.*, 2013) (Segura-Bedmar *et al.*, 2014)	A corpus containing 792 texts from the DrugBank database and other 233 PubMed abstracts. These are annotated with 18 502 pharmacological substances and 5028 drug–drug interactions. https://github.com/isegura/DDICorpus
CDR corpus (Li *et al.*, 2016)	A collection of 1500 PubMed abstracts manually annotated with 3116 chemical–disease relations and the respective named entities (4409 chemicals and 5818 diseases). The entity annotations also contain normalized MeSH concept identifiers. https://biocreative.bioinformatics.udel.edu/tasks/biocreative-v/track-3-cdr/
ChemProt (Krallinger *et al.*, 2017a)	It contains a total of 2432 PubMed abstracts split into *training*, *development*, and *test* subsets. With 31 831 chemical and 30 316 protein annotations, there are a total of 15 739 chemical–protein interactions. Five different relation types are annotated: *activation*, *inhibition*, *agonist*, *antagonist*, and *substrate*. https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vi/track-5/

Resource	Description
2018 n2c2 Track 2 (Henry *et al.*, 2021)	A collection of 505 narrative discharge summaries from the MIMIC-III clinical care database. These are annotated with concepts related to medications (strengths and dosages, duration and frequency of administration, route of administration, reason for administration, and adverse drug effects), and interactions between them. There are a total of 83 869 named entities and 59 810 relations. https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/
DrugProt (Miranda *et al.*, 2021)	Built from the existing ChemProt corpus, it includes more PubMed abstracts and is also split into three subsets (*training*, *development*, and *test*). A total of 3500 documents with over 100 thousand annotated entities and interactions. Moreover, it contains more relation types—a total of 13 distinct chemical–protein interactions. https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-1/
BioRED (Luo *et al.*, 2022)	This corpus contains a set of 600 PubMed abstracts annotated with multiple entity types (genes, diseases, chemicals) and relation pairs (such as gene–disease and chemical–chemical). Each relation is further labeled as describing either a novel finding or background knowledge. It contains a total of 20 419 entity mentions and 6503 relations. https://ftp.ncbi.nlm.nih.gov/pub/lu/BioRED/
ChemDisGene (Zhang *et al.*, 2022a)	A distant supervision corpus for extracting multi-class multi-label relations between chemicals, diseases, and genes. The dataset contains around 80 thousand PubMed abstracts and is split into two portions: (1) one curated by human experts intended for *evaluation* containing 523 documents, and (2) another intended for *training* which was distantly labeleled via the Comparative Toxicogenomics Database (CTD). The dataset is annotated with 18 relation types. https://github.com/chanzuckerberg/ChemDisGene

5.2 Text mining chemical–protein interactions

The scientific literature contains large amounts of information on genes, proteins, chemicals, and their interactions. Extraction and integration of this information in curated knowledge bases helps researchers support their experimental results, leading to new hypotheses and discoveries. This is especially relevant for precision medicine, which aims to understand the individual variability across patient groups in order to select the most appropriate treatments. Methods for improved retrieval and automatic relation extraction from biomedical literature are therefore required for collecting structured information from the growing number of published works.

As the knowledge of how biological systems work at different structural levels grows, more possibilities arise for applying it in diagnosing and treating common and complex diseases. Furthermore, exploiting the large amounts of biomolecular data from -omics studies and patient-level information recorded in electronic health records (EHRs) offers prospects for precision and personalized medicine (Wu *et al.*, 2017). Nonetheless, relevant fine-grained information is constantly being communicated in the form of natural language through scientific publications. To exploit this source of updated knowledge, several methods have been proposed for retrieving relevant articles for database curation (Wang *et al.*, 2016a), and for extracting from the unstructured texts information such as entity mentions (Campos *et al.*, 2013; Nunes *et al.*, 2013), biomolecular interactions and events (Ananiadou *et al.*, 2015; Krallinger *et al.*, 2011), or the clinical and pharmacological impact of genetic mutations (Singhal *et al.*, 2016b). These methods have proven essential for collecting the most recent research results and for expediting database curation (Krallinger *et al.*, 2017b).

The BioCreative VI ChemProt challenge stimulated the development of systems for extracting interactions between chemical compounds (drugs) and GPROs (gene and protein related objects) from running text, given their importance for precision medicine, drug discovery and basic biomedical research (Krallinger *et al.*, 2017a). An example illustrating various relations that can be extracted from a single sentence in a publication is shown in Figure 5.2. The development of systems able to automatically extract such relations may expedite curation work and contribute to the amount of information available in structured annotation databases, in a form that is easily searched and retrieved by researchers.

Data for the ChemProt task was composed of PubMed abstracts in which gold-standard entities were provided, and the aim was to detect chemical–protein pairs that expressed a certain interaction. Therefore, the biomedical named entity recognition step was out of the scope of this task. The organizers defined ten groups of chemical–protein relations (CPRs) that shared some underlying biological properties, in which only five

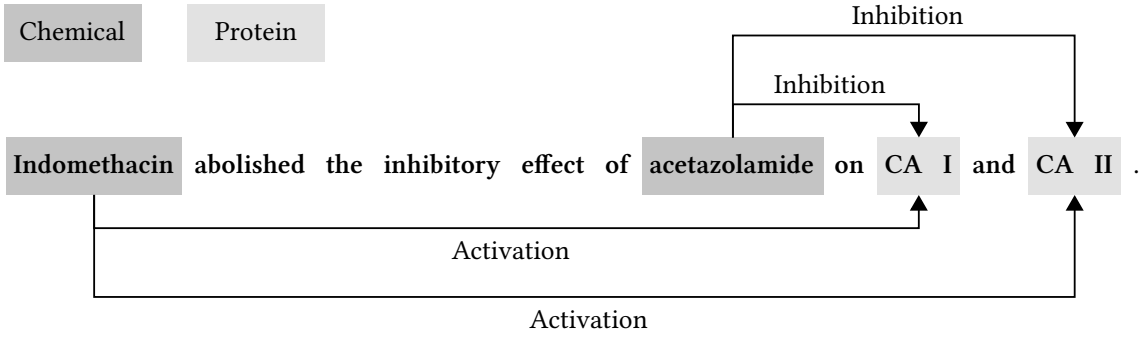

Figure 5.2: Example sentence illustrating biochemical entities and their relations from the ChemProt training dataset (PMID 8667211).

of them (up-regulation, down-regulation, agonist, antagonist, and substrate) were used for evaluation purposes. More detail about the data is presented in Section 5.2.1.

In this section, we present a deep learning approach for extracting mentions of chemical–protein interactions from biomedical articles, based on various enhancements following our participation in the BioCreative VI ChemProt task. A significant aspect of our best method is the use of a simple deep learning model together with a very narrow representation of the relation instances, using only up to ten words from the shortest dependency path and the respective dependency edges. BiLSTM recurrent networks or CNNs are used to build the deep learning models.

We report the results of several experiments and show that our best model is competitive with more complex sentence representations or network structures, achieving an F1-score of 0.6306 on the test set. The source code of our work, along with detailed statistics, is publicly available at:

`https://github.com/ruiantunes/biocreative-vi-track-5-chemprot`.

5.2.1 Materials and methods

This section describes the resources used, the evaluation metric employed, and the methods implemented.

Dataset

The ChemProt corpus was created by the BioCreative VI organizers (Krallinger *et al.*, 2017a), being composed of three distinct sets: training, development, and test (Table 5.2). During the challenge, to hinder manual corrections and to ensure that systems could annotate larger datasets, the organizers included 2599 extra documents in the test set, which were not used for evaluation.

Table 5.2: ChemProt dataset statistics.

			Training	Development	Test
Abstracts	Total		1020	612	800
	With any relation		767	443	620
	With evaluated relations		635	376	514
Entities	Chemical		13 017	8004	10 810
	Protein		12 735	7563	10 018
	Total		25 752	15 567	20 828
Relations	Activation	(CPR:3)	768	550	665
	Inhibition	(CPR:4)	2254	1094	1661
	Agonist	(CPR:5)	173	116	195
	Antagonist	(CPR:6)	235	199	293
	Substrate	(CPR:9)	727	457	644
	Total		6437	3558	5744

Each document, containing the title and the abstract of a PubMed article, was annotated by expert curators with chemical and protein entity mentions, and their relations. The annotation guidelines considered ten groups of biological interactions, which were designated as chemical–protein relation groups. However, for this task, only five classes were considered for evaluation purposes: activation (CPR:3), inhibition (CPR:4), agonist (CPR:5), antagonist (CPR:6), and substrate (CPR:9). Table 5.2 presents detailed dataset statistics.

One can see from Table 5.2 that not all abstracts contain annotated relations, although all abstracts were annotated with entity mentions. Nevertheless, abstracts without evaluated relations are useful as they can be used to create negative instances for training the system. Only 1525 documents of 2432 (63%) are annotated with evaluated relations. This suggests that it could be a reasonable idea to first apply a document triage step to ignore documents that probably are not relevant for extracting chemical–protein interactions, reducing the number of false positive relations, while still considering them for generating negative instances to feed the deep learning model. Though, we did not follow this possibility leaving it as possible future work. Similar binary approaches were followed by Lung *et al.* (2017, 2019) and Warikoo *et al.* (2018) who start by predicting if a CPR pair is positive.

A more scrupulous analysis of the corpus shows that there are some relations between overlapped entities (for example, a protein entity containing a chemical entity), as well as some cross-sentence relations. However, cross-sentence relations appear in a

very small number and were deliberately discarded. Also, despite some ChemProt relations were classified with more than one CPR group we considered only one label, since these are rare, simplifying the task as a multi-class problem.

Performance evaluation

The BioCreative VI ChemProt organizers considered the micro-averaged precision, recall, and balanced micro F1-score for evaluation purposes (Krallinger *et al.*, 2017a). Micro F1-score was the official metric used to evaluate and compare the teams' submissions. This metric was integrated in our pipeline, for measuring the neural network performance at each training epoch, allowing to develop and select the best model dynamically for this specific task.

Pre-processing

We pre-processed the entire ChemProt dataset using the Turku Event Extraction System (TEES) (Björne and Salakoski, 2015) applying a pipeline composed with the GENIA sentence splitter (Sætre *et al.*, 2007), the BLLIP parser (Charniak and Johnson, 2005) using the McClosky and Charniak (McCC) biomedical parsing model (McClosky and Charniak, 2008), and the Stanford dependency parser (Chen and Manning, 2014) (version 3.8.0, released on 2017-06-09). This pre-processing performs sentence splitting, tokenization, part-of-speech tagging, and dependency parsing. Sentence splitting is required to obtain all the chemical–protein pair candidates in the same sentence, since these are the only ones we considered. The yielded tokens, PoS tags, and dependency labels are encoded using embedding vectors (more detail in the next sections). The dependency parsing structure is also used to find the shortest dependency path between the two entities, since previous work had already proven its value for relation extraction (Bunescu and Mooney, 2005a).

For every chemical–protein pair in each sentence, we obtain five sequences using the TEES output: the SDP and the sequences containing the left text and the right text of the chemical and protein entities (Figure 5.3). Like the work of Mehryary *et al.* (2017, 2018), our system traverses the shortest dependency path always from the chemical entity to the protein entity. For entities spanning more than one word, we obtain the shortest path starting from the head word, as indicated by the TEES result. For each chemical–protein pair candidate instance, the chemical and protein entities (in cause) are replaced respectively by the placeholders '#chemical' and '#gene', except when the chemical–protein pair comes only from a single token (overlapped entities) which in this case is replaced by '#chemical#gene'. While in the SDP the dependency features were obtained traversing the path, in the four left and right sequences the incoming edge of each token

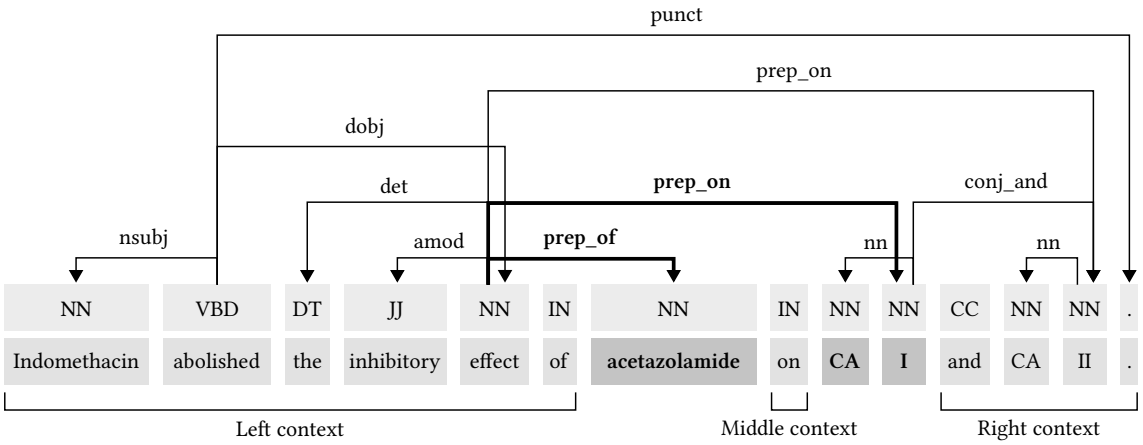

Figure 5.3: Example illustrating the dependency parsing structure of a sentence from the ChemProt training dataset (PMID 8667211). In this example, we considered the relation between the 'acetazolamide' chemical mention and the 'CA I' protein mention. The shortest dependency path is highlighted in bold. Penn Treebank part-of-speech tags (Marcus *et al.*, 1993) used in this example: coordinating conjunction (CC); determiner (DT); preposition or subordinating conjunction (IN); adjective (JJ); noun, singular or mass (NN); verb, past tense (VBD); sentence final punctuation (.). Stanford dependencies (de Marneffe and Manning, 2016) used in this example: adjectival modifier (amod); conjunction and (conj_and); determiner (det); direct object (dobj); noun compound modifier (nn); nominal subject (nsubj); prepositional modifier of (prep_of); prepositional modifier on (prep_on); punctuation (punct).

was used as dependency features. If a token did not have an incoming edge or it was the last token in the SDP then the dependency feature was set to '#none'. Each one of the five sequences is therefore represented by a sequence of tokens, PoS tags, and dependency edge labels.

Taking the sentence in Figure 5.3 as example, and considering the chemical–protein pair ['acetazolamide', 'CA I'], the five extracted sequences (containing the tokens, PoS tags, and dependency edges) are:

1. **Shortest dependency path:** #chemical / NN / prep_of — effect / NN / prep_on — #gene / NN / #none;

2. **Chemical left text:** Indomethacin / NN / nsubj — abolished / VBD / #none — the / DT / det — inhibitory / JJ / amod — effect / NN / dobj — of / IN / #none;

3. **Chemical right text:** on / IN / #none;

4. **Protein left text:** in this case, it is the same as the chemical right text;

5. **Protein right text:** and / CC / #none — CA / NN / nn — II / NN / prep_on — . / . / punct.

The SDP together with the left and right sequences are fed to the neural network through embedding layers, as explained in the following sections.

Word embeddings

For text based tasks, it is necessary to encode the input data in a way that it can be used by the deep network classifier. This can be achieved by representing words as embedding vectors of a relatively small dimension, rather than using the large feature space resulting from the traditional one-hot encoding. Word embeddings is a technique that consists in deriving vector representations of words, such that words with similar semantics are represented by vectors that are close to one another in the vector space (Bengio *et al.*, 2003). This way, each document is represented by a sequence of word vectors which are fed directly to the network. Efficient calculation of word embeddings, such as provided by word2vec (Mikolov *et al.*, 2013a), allow inferring word representations from large unannotated corpora.

We applied the word2vec implementation from the Gensim framework (Řehůřek and Sojka, 2010) to obtain word embeddings from 15 million PubMed abstracts in English language from the years 1900 to 2015. In previous research we created six models, with vector sizes of 100 and 300 features and windows of 5, 20, and 50. The models contain around 775 thousand distinct words (stop words were removed). These pre-trained word embeddings models showed their value achieving favorable results both in biomedical document triage (Matos and Antunes, 2017b) and biomedical word sense disambiguation (Antunes and Matos, 2017c). In this work we use the word embeddings models with a window size of 50, which are available in our online repository.

Another successor technique for creating word embeddings, from large unlabeled corpora, with subword information was proposed by Bojanowski *et al.* (2017). Their library, fastText, was used by Chen *et al.* (2019b) to create biomedical word embeddings (vector size of 200, and window of 20) from PubMed articles and MIMIC-III clinical notes (Johnson *et al.*, 2016). We included these publicly available word embeddings in our simulations to compare to our own models.

Furthermore, we created PoS and dependency embeddings from the ChemProt dataset applying different vector sizes (20, 50, 100) and windows (3, 5, 10). The training, development, and test sets are used, with 1020, 612, and 800 documents respectively (Table 5.2). However, we acknowledge the inclusion of the test set adds a slight bias. (A lapse that we do not find it worth for repeating all our simulations.) This could be over-

come, possibly improving the overall results, by including: (1) PubMed abstracts outside the ChemProt dataset or (2) the remaining 2599 abstracts that initially existed, in the test set, to avoid manual annotations. Based on preliminary experiments on the training and development sets, we decide to use the pre-trained embedding vectors, with a window size of 3, which are kept fixed during training. We tested using randomly initialized PoS and dependency embeddings being adapted during training, but the results were similar and the runtime was higher.

Different tools—Gensim (Řehůřek and Sojka, 2010), fastText (Bojanowski *et al.*, 2017) and TEES (Björne and Salakoski, 2015)—were used for tokenization in the word embeddings creation and in the ChemProt dataset. Therefore, we created a mapping between the dataset vocabulary and its embedding vectors: each word of the ChemProt vocabulary was tokenized according to the word embeddings vocabulary, and its word vector was calculated using the L2-normalized sum of the constituent words. With this approach, the dataset vocabulary was strongly reduced (the respective PoS tags and dependency edges were also removed) because some uninformative tokens are not present in the word embeddings model. Preliminarily, this showed to be profitable since stop words or out-of-vocabulary words were discarded from start.

We chose a fixed maximum length of 10 tokens (or 9 hops) for the shortest dependency path, and a maximum length of 20 tokens for each of the left and right sequences. These values were manually chosen according to the distribution of maximum lengths in the training set. We had tested using the length of the longer sequence, but this did not show to be advantageous since results were not better and implied a much higher training time. In the few cases in which the distance between the two entities is too long causing the extracted sequences to have more tokens than the pre-defined maximum, the sequences are truncated (the remaining tokens are discarded). In the opposite case, when there are less tokens than the maximum length allowed, the input vectors are padded with zeros to keep the same input vector size.

Deep neural network

Figure 5.4 shows the general structure of the neural network used in this work. Similarly to other works in relation extraction (Zhang *et al.*, 2017; Peng *et al.*, 2018; Mehryary *et al.*, 2018), the different representations of a relation instance, namely the linear and SDP representations, are handled by two separate sub-networks, the results of which are concatenated at later stages.

Initially, each token in each one of the five extracted sequences (SDP, left and right texts) is represented by the concatenation of the embedding vectors from the word, PoS, and dependency embedding matrices. Furthermore, the four left and right sequences

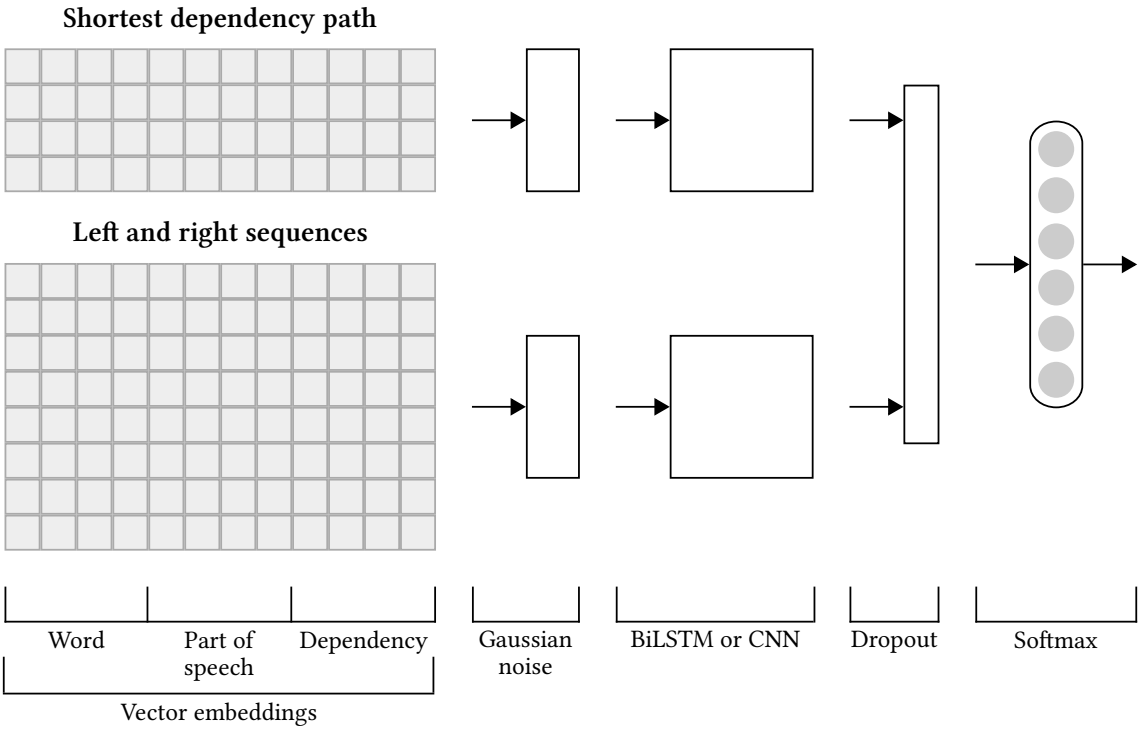

Figure 5.4: Neural network structure for the ChemProt relation extraction task. The inner model can be a BiLSTM or CNN.

corresponding to the linear representation are concatenated into a single input. For each of these two inputs (SDP and linear), Gaussian noise is added up, followed by a BiLSTM model or a CNN model (several convolution layers with multiple window sizes followed by global max pooling). Then, the two obtained outputs are concatenated and dropout is applied. At the final stage, a fully connected layer with softmax activation outputs the prediction probabilities. As can be seen in Figure 5.4, the neural network model only differs in an intermediate step (BiLSTM or CNN). We implemented these deep learning models in the Keras framework (Chollet *et al.*, 2015) and the TensorFlow backend (Abadi *et al.*, 2016) using the Python programming language (Chollet, 2017).

An important consideration when defining and training deep network models is related to overfitting, which means that the network learns the "best" data representation but is not able to generalize to new data. Various strategies have been proposed and are commonly employed to address this problem. In our experiments, we applied common strategies to avoid overfitting, namely random data augmentation (Gaussian noise addition), dropout, and early stopping. Early stopping looks at the value of a specific evaluation metric in a validation subset and stops the training process when this value stops improving for a pre-specified number of training epochs (patience value). Also, early stopping brings a gain in total training time since the "best" model is usually selected after a few epochs instead of training for a fixed, usually larger, number of epochs.

Table 5.3: System parameters for the ChemProt relation extraction task.

Gaussian noise standard deviation	0.01
LSTM units	128
LSTM recurrent dropout	0.4
LSTM dropout	0.4
Convolution filters	64
Convolution window sizes	[3, 4, 5]
Dropout rate	0.4
Optimizer	RMSprop (Tieleman and Hinton, 2012)
Loss	Categorical cross entropy
Batch size	128
Maximum number of epochs	500
Early stopping patience	30
Early stopping monitor	Validation micro F1-score
Validation split	0.3

This is an important aspect especially when running several simulations to test different network structures and parameters. According to preliminary results, we decided to fix 30% of the training data as validation subset, and calculated the F1-score at each epoch for monitoring the quality of the model. Similarly, when creating the final model to apply to the test data, we merged the training and development sets and used respectively 70% for training and 30% for validation and early stopping.

Table 5.3 shows the network hyper-parameters and other variables used in our system (default values were used in unmentioned parameters). Despite that we did not perform an exhaustive grid-search for the best parameters, these were iteratively adjusted according to several experiments using the training and development sets. Class weights inversely proportional to their frequency in the training set were used to weight the input instances.

Additional methods

To improve the generalization ability of our system and to reduce the fluctuation of the results due to the random initialization, all the results were obtained by averaging the prediction probabilities of three simulations using different random states. The use of a different random state means that a different random initialization was made in the neural network weights, and that distinct subsets of the training data were effectively used for training and validation.

Another crucial method in our system is the balancing between precision and recall to maximize the F1-score, achieved by adjusting the classification threshold at each training epoch. The training data is used in this process to avoid biasing the test results. A similar experiment was performed by Corbett and Boyle (2018) where they also used a threshold value to maximize the F1-score on the development set.

Additionally, we pre-processed an external dataset from the BioGRID database containing chemical–protein interactions (Chatr-aryamontri *et al.*, 2017). This dataset supplied further 1102 PubMed abstracts for training, annotated with 2155 chemicals, 2190 proteins, and 2277 relations between them.

In the next section we present and discuss the obtained results using the methods mentioned in this section.

5.2.2 Results and discussion

As noted in the previous section, the use of different random states generates different training and validation subsets which in turn results in different trained models (network weights and optimal classification threshold). This approach allows using a large amount of data for early stopping, which in our preliminary experiments proved important for improving generalization, while still using most of the available data for training. Thereby, the results presented in this section are obtained by averaging the probabilities from three simulations.

Table 5.4 presents a detailed gathering of results obtained on the development set by the BiLSTM and CNN models combining different inputs: sequences (SDP, left and right sequences), features (words, PoS, dependencies), and embedding models. The three best results on the development set (F1-scores: 0.6496, 0.6473, and 0.6385) were obtained by the BiLSTM model using only the shortest dependency path with word and dependency features where different embedding models are used, being the highest result achieved with the biomedical word embeddings created by Chen *et al.* (2019b).

The results show that, in general, the left and right sequences generated much lower results, and when combining them with the SDP, the results were worst than using only the SDP. We believe this may be due to the way the left and right sequences are combined and encoded into the neural network, and also because the larger number of tokens (80 *versus* 10 in the SDP) may contribute with more noise by means of uninformative tokens. It is possible that different approaches for incorporating the linear sequence information could improve the final results.

As expected, words were the more informative type of feature, while the PoS tags were the less informative being worthless in some configurations. For example, in the majority of the cases, combining the PoS tags with words and dependencies worsened

Table 5.4: F1-score results on the ChemProt development set using the BiLSTM and CNN models. WS: word embeddings size. PS: part-of-speech embeddings size. DS: dependency embeddings size. SDP: shortest dependency path sequence. LR: left and right sequences. NN: neural network. BiLSTM: bidirectional long short-term memory. CNN: convolutional neural network. W: words. P: part-of-speech tags. D: dependency edges. The highest value in each row is highlighted in bold; the best overall value is underlined.

(WS, PS, DS)	Features	NN	W	P	D	W+P	W+D	P+D	W+P+D
(100, 20, 20)*	SDP	BiLSTM	0.6007	0.1695	0.2609	0.5971	**0.6385**	0.2991	0.6351
		CNN	0.5594	0.1628	0.2832	0.5622	0.5978	0.3102	**0.6010**
	LR	BiLSTM	0.4967	0.2003	0.2059	0.4906	**0.5149**	0.2106	0.5043
		CNN	**0.4371**	0.1902	0.1635	0.4131	0.4193	0.1683	0.3984
	SDP+LR	BiLSTM	0.5857	0.2271	0.3044	0.5776	**0.6000**	0.2807	0.5979
		CNN	0.5243	0.2332	0.2594	0.5268	0.5381	0.2361	**0.5403**
(300, 100, 100)*	SDP	BiLSTM	0.6161	0.1601	0.2920	0.6002	**0.6473**	0.3228	0.6310
		CNN	0.5642	0.1595	0.3019	0.5782	**0.6141**	0.2991	0.6092
	LR	BiLSTM	0.5135	0.2093	0.1910	0.5133	0.5209	0.1847	**0.5227**
		CNN	0.4293	0.1962	0.1550	**0.4576**	0.4321	0.1448	0.4216
	SDP+LR	BiLSTM	0.5914	0.2176	0.2873	0.5812	**0.6036**	0.2692	0.6015
		CNN	0.5572	0.2152	0.2519	0.5618	0.5672	0.2340	**0.5819**
(200, 50, 50)†	SDP	BiLSTM	0.6229	0.1530	0.2806	0.6192	**0.6496**	0.3087	0.6453
		CNN	0.5804	0.1555	0.2867	0.5841	**0.6259**	0.3182	0.6205
	LR	BiLSTM	0.5030	0.2353	0.2096	**0.5158**	0.5060	0.2166	0.4849
		CNN	**0.4813**	0.1827	0.1681	0.4504	0.4201	0.2130	0.4291
	SDP+LR	BiLSTM	0.5943	0.2428	0.2918	0.5993	**0.6126**	0.2715	0.5824
		CNN	0.5690	0.1966	0.2413	0.5440	**0.5760**	0.2645	0.5605

* Our PubMed-based word embeddings.

† Pre-trained word embeddings by Chen *et al.* (2019b).

results. Interestingly, the dependency edge labels showed to be much more informative than the PoS tags, effectively improving performance in several configurations. Essentially, the highest results were achieved by combining words and dependency features.

Different embedding models were also explored (Table 5.4). We used larger embedding sizes for words, giving greater importance to word semantics, and smaller embedding sizes for PoS tags and dependency labels. The results show, in the case of our PubMed-based word2vec embeddings, that using larger encoding vectors—(300, 100, 100) *versus* (100, 20, 20)—leads to slightly improved results. Nonetheless, the best overall results were obtained with the fastText embeddings by Chen *et al.* (2019b), although

Table 5.5: Detailed results on the ChemProt development and test sets using distinct approaches. The best configuration from the results in the development set (Table 5.4) was employed. WS: word embeddings size. PS: part-of-speech embeddings size. DS: dependency embeddings size. P: precision. R: recall. F: F1-score. The highest value in each column is highlighted in bold.

			Development			Test		
(WS, PS, DS)			P	R	F	P	R	F
(300, 200, 300)*,†	Best official run		0.4999	0.6074	0.5470	0.5738	0.4722	0.5181
(300, 100, 100)†	Baseline§	BiLSTM	0.6737	0.6229	0.6473	0.7089	0.5480	0.6182
		CNN	0.7059	0.5435	0.6141	**0.7423**	0.4939	0.5932
(200, 50, 50)‡	Baseline§	BiLSTM	0.6908	0.6130	**0.6496**	0.6812	0.5870	0.6306
		CNN	**0.7252**	0.5505	0.6259	0.7182	0.5093	0.5959
	BioGRID¶	BiLSTM	0.5337	**0.6523**	0.5871	0.5881	**0.6050**	0.5964
		CNN	0.5913	0.5642	0.5774	0.6323	0.5191	0.5701
	No validation|	BiLSTM	0.6867	0.6068	0.6443	0.6791	0.5980	**0.6360**
		CNN	0.6247	0.4988	0.5547	0.6091	0.5160	0.5586

* Our official evaluated run (Krallinger *et al.*, 2017a; Matos, 2017).

† Our PubMed-based word embeddings.

‡ Word embeddings by Chen *et al.* (2019b).

§ Results on the development set are the same as reported in Table 5.4.

¶ 30% of the training data (BioGRID excluded) used for validation.

| Model trained during 500 epochs (without monitoring).

these use a smaller vector size. This result highlights that the incorporation of subword information in the embedding vectors is beneficial for biomedical information extraction.

For collecting the final results (on the test set) we applied our described approach, but with two additional arrangements: (1) adding BioGRID external training data, and (2) using no validation data (the validation split was set to 0.0). Table 5.5 presents these results using the best configuration based on the results obtained on the development set (Table 5.4), which consisted in using the shortest dependency path with word embeddings of size 200 (fastText model by Chen *et al.* (2019b)) and dependency features encoded by embedding vectors of size 50. For better comparison we also include in Table 5.5 the results of our best official run (during the challenge) and the baseline results using our PubMed-based word embeddings.

Inclusion of the dataset from BioGRID as additional training data deteriorated F1-score results when compared to not using it, in both BiLSTM (development: 0.5871 *versus* 0.6496, test: 0.5964 *versus* 0.6306) and CNN models (development: 0.5774 *versus* 0.6259,

test: 0.5701 *versus* 0.5959). This suggests that these data diverge from the ChemProt guidelines and that some kind of heuristics would be required to decide which instances to include. Other approaches such as multi-instance (Surdeanu *et al.*, 2012; Lamurias *et al.*, 2017; Eberts and Ulges, 2021) or adversarial learning (Qin *et al.*, 2018) could also be applied.

Inspection of the training and validation F1-score for each epoch indicated that the BiLSTM model suffered less from overfitting than the CNN model. Therefore we performed an experiment where models were trained for 500 epochs without early stopping, since this has the advantage of training each model (in the three simulations) using all of the available training data. Overall, the highest F1-score on the test set was achieved following this approach (0.6360 *versus* 0.6306 in the baseline) showing that the BiLSTM model was in fact very resistant to overfitting. Conversely, the CNN performed much worst when early stopping, and therefore validation data, was not used (0.5586 *versus* 0.5959). Even when trained with the external dataset from BioGRID, where validation data was used, the CNN model obtained better results compared to those obtained without validation monitoring (0.5701 *versus* 0.5586). Despite 0.6360 being the highest F1-score in the test set, we consider our best F-score is 0.6306 since it was selected according to the best method in the development set (Table 5.5), which represents an improvement of 11 percentage points compared to our best official F1-score (0.5181).

From the results in Tables 5.4 and 5.5, we conclude that a solid benefit of our approach is that the best method uses at most 10 tokens from the SDP to classify the chemical–protein relation, using a small representation vector and therefore reducing training time. For instance, on a Intel i3-4160T (dual-core, 3.10 GHz) CPU, training the BiLSTM and CNN models for one epoch with 70% of the training set (word and dependency embeddings with sizes 100 and 20), takes respectively around 5 and 2 seconds (the additional cost of balancing precision and recall is excluded). Also, another positive remark is that our BiLSTM model is resistant to overfitting, since the results obtained in the baseline approach are similar to those reported without using validation data, and the results in the development and test sets are similar. On the other hand, overfitting is evident when using the CNN model, since training it for 500 epochs grossly declined the results (development: 0.6259 *versus* 0.5547, test: 0.5959 *versus* 0.5586). This overfitting also helps to explain the higher precision seen for the CNN model as compared to the BiLSTM model, since the network is better capable of identifying with high confidence those test instances that are very similar to instances seen during training.

Table 5.6: Comparison between participating teams in the ChemProt challenge (F1-score results on the test set). CNN: convolutional neural network. ML: machine learning. RNN: recurrent neural network. SVM: support vector machine.

R*	Work	Classifiers	Challenge	PC†
1	Peng *et al.* (2017, 2018)	SVM, CNN and RNN	**0.6410**	
2	Corbett and Boyle (2017a, 2018)	RNN and CNN	0.6141	0.6258
3	Mehryary *et al.* (2017, 2018)	SVM and RNN	0.6099	0.6310
4	Lim and Kang (2017, 2018)	Tree-structured RNN	0.5853	0.6410
5	Lung *et al.* (2017, 2019)	Traditional ML	0.5671	
6	Ours (Matos, 2017; Antunes and Matos, 2019)	RNN and CNN	0.5181	0.6306
7	Liu *et al.* (2017, 2018a)	CNN and attention-based RNN	0.4948	0.5270
8	Verga and McCallum (2017)	Bi-affine attention network	0.4582	
9	Wang *et al.* (2017a)	RNN	0.3839	
10	Tripodi *et al.* (2017)	Traditional ML and neural networks	0.3700	
11	Warikoo *et al.* (2017, 2018)	Tree kernel	0.3092	0.3654
12	Sun (Krallinger *et al.*, 2017a)	-	0.2195	
13	Yüksel *et al.* (2017)	CNN	0.1864	

* R: rank. Teams ranked according to the official evaluation.

† PC: post-challenge. Improved results due to post-challenge enhancements.

Comparison with other participating teams

Table 5.6 compares our results with other works presented during the ChemProt challenge as well as post-challenge improvements. All the top performing teams used recurrent neural networks showing their strength in this chemical–protein relation extraction task. Also, SVMs and CNNs are amongst some of the classifiers used by other works.

Similarly to our work, Corbett and Boyle (2017b, 2018) used LSTM and CNN layers. They achieved a best F1-score of 0.6258 on the test set, which is in line with our result (0.6306). However, their network structure is larger being composed of more layers. Mehryary *et al.* (2017) applied a similar pre-processing pipeline as described in this work, using the TEES tool to perform tokenization, part-of-speech tagging, and dependency parsing. They achieved a top F1-score of 0.6099 with a combination of SVMs and LSTM networks. This result was improved to 0.6310 following the challenge (Mehryary *et al.*, 2018). Using the ANN alone, with whole sentence tokens and features from the SDP, they achieved an F1-score of 0.6001 in the test set, while our BiLSTM model achieves an F1-score of 0.6306 by only using features from the SDP. Lim and Kang (2017, 2018) used a tree-structured RNN exploiting syntactic parse information and obtained an F1-score of 0.6410, equalling the best official result.

Differently from the works cited above, Lung *et al.* (2019) used traditional machine learning algorithms with handcrafted features, achieving an F1-score of 0.5671. As part

Table 5.7: Confusion matrix in the ChemProt test set (F1-score 0.6306) obtained by the BiLSTM model that achieved the highest F1-score in the development set, as reported in Table 5.5. The light-gray cells show correct classifications (true positives); mid-gray cells show false negatives (first row) and misclassifications between classes; and dark-gray cells show false positives.

Predicted	Gold-standard						
	Negative	CPR:3 Activation	CPR:4 Inhibition	CPR:5 Agonist	CPR:6 Antagonist	CPR:9 Substrate	Sum
Negative		238	524	97	124	341	1324
Activation	263	382	19	5	0	0	669
Inhibition	401	45	1107	14	2	2	1571
Agonist	45	0	2	79	6	0	132
Antagonist	56	0	1	0	161	0	218
Substrate	185	0	8	0	0	301	494
Sum	950	665	1661	195	293	644	

True positives	2030
False negatives	1428
False positives	950

of their approach, the authors manually built a dictionary with 1155 interaction words, which where mapped to the corresponding CPR type, to create chemical–protein interaction (CPI) triplets.

Error analysis

In this section we evaluate, making a detailed error analysis, the predictions obtained in the test set using the baseline approach with the fastText word embeddings and the BiLSTM model (Tables 5.7 to 5.10). The confusion matrix, presented in Table 5.7, follows the official evaluation script and reflects the same results reported in Table 5.5. We observe that the "activation" and "inhibition" relation classes were the ones most difficult to discriminate, with 19 "inhibition" relations predicted as "activation" and 45 "activation" relations predicted as "inhibition".

Tables 5.8 and 5.9 show, respectively, heatmaps of the precision and recall values in function of the numbers of gold-standard entities per sentence and gold-standard relations per sentence. Numbers in the cells show the amount of correct classifications (true positives) and incorrect (false positives) or missed classifications (false negatives). This representation makes it easier to understand which type of sentences are more difficult

Table 5.8: Heatmap representing the precision values obtained by the BiLSTM model (the best in the development set) applied to the ChemProt test set. True positives (TP) and false positives (FP) are displayed as TP/FP. X-axis: number of gold-standard entities per sentence. Y-axis: number of gold-standard evaluated relations per sentence. Axes are truncated for conciseness.

Y	X								
	2	3	4	5	6	7	8	9	10
1	148/10	88/19	62/35	20/15	13/21	4/ 1	1/ 0	4/10	0/ 0
2		182/14	121/18	86/31	33/12	29/18	6/ 5	8/21	1/ 3
3			89/ 9	60/ 9	36/15	15/14	14/15	11/ 3	10/13
4			58/ 4	96/ 4	62/11	40/14	17/ 6	21/18	8/ 8
5				5/ 0	41/ 7	52/ 9	6/ 7	9/ 3	8/ 1
6				50/ 0	26/ 3	50/ 9	39/17	9/ 1	2/ 5
7					2/ 2	6/ 0	14/ 0	21/ 1	0/ 0
8					32/ 9	14/ 3	5/ 0	26/ 4	14/ 0
9					17/ 0	14/ 0	0/ 0	18/10	9/ 6
10						6/ 0	10/ 5	0/ 0	10/ 0

for our model to 'interpret'. In Table 5.8 we see a clear and somewhat expected trend with lower precision when the number of entities in a sentence is high but the number of existing relations in that sentence is low. This is intuitive since many chemical–protein pair candidates are generated, potentially leading to several false positive relations. From Table 5.9 we verify that the majority of the sentences in the corpus have only a few number of entities and relations. Sentences with many entities are rare, and these may have few or many relations. Interestingly, the results in Table 5.9 indicate that, although the worst results in terms of recall are obtained for sentences containing many entities, there is a considerable number of unidentified relations from sentences containing up to four entities.

We present a detailed error analysis showing concrete cases where the model failed

Table 5.9: Heatmap representing the recall values obtained by the BiLSTM model (the best in the development set) applied to the ChemProt test set. True positives (TP) and false negatives (FN) are displayed as TP/FN. X-axis: number of gold-standard entities per sentence. Y-axis: number of gold-standard evaluated relations per sentence. Axes are truncated for conciseness.

Y \ X	2	3	4	5	6	7	8	9	10
1	148/98	88/ 70	62/42	20/19	13/10	4/ 7	1/ 1	4/ 4	0/ 1
2		182/132	121/83	86/50	33/25	29/ 7	6/ 8	8/10	1/ 3
3			89/76	60/58	36/12	15/ 6	14/28	11/ 4	10/ 5
4			58/50	96/48	62/42	40/20	17/19	21/11	8/ 0
5				5/ 5	41/19	52/13	6/ 4	9/ 6	8/ 2
6				50/16	26/ 4	50/16	39/27	9/ 3	2/10
7					2/ 5	6/ 1	14/14	21/ 0	0/ 7
8					32/24	14/18	5/11	26/22	14/ 2
9					17/ 1	14/13	0/ 0	18/ 9	9/ 9
10						6/ 4	10/ 0	0/ 0	10/ 0

to predict (Table 5.10). A comprehensive list with all the predictions can be found in the online repository. We enumerate different causes for the analyzed frequent errors:

- Limited or incorrect instance representation. Information obtained exclusively from the SDP is, often, insufficient or faulty since essential words may be missing or misleading words may be present. Examples 1, 2, and 3 in Table 5.10 show cases where crucial terms such as "agonistic" and "antagonist" are not included in the SDP. On the other hand, examples 4, 5, 6 include words, such as "downregulation", "activation", and "inhibition", that are frequently related with other relation classes and caused incorrect classification in these cases.

- Misinterpretation of negation. In some cases, there is a term giving the opposite

Table 5.10: Error analysis: examples of incorrect predictions in the ChemProt test set obtained by the BiLSTM model (the best in the development set). The chemical–protein pairs are presented with information from the sentence and the shortest dependency path (SDP). The chemical and protein named entities are annotated in dark-gray and light-gray colored boxes, respectively. For simplicity, the chemical and gene placeholders were omitted in the list of words from the SDP. Sentences are from PMIDs 23265901, 17082235, 10611634, 10909982, 10701840, and 12897749.

Example	Correct	Predicted	Full sentence	Words in the SDP
1	Agonist	Activation	The introduction of the amino group resulted in not only improved water solubility but also enhanced TLR7 agonistic activity.	group introduction resulted activity
2	Agonist	Inhibition	Our work shows that sulfonylureas and glinides additionally bind to PPARgamma and exhibit PPARgamma agonistic activity.	exhibit activity
3	Antagonist	Agonist	In guinea pigs, antagonist actions of yohimbine at 5-HT(1B) receptors are revealed by blockade of hypothermia evoked by the 5-HT(1B) agonist, GR46,611.	receptors
4	Activation	Inhibition	Impaired expression of the uncoupling protein-3 gene in skeletal muscle during lactation: fibrates and troglitazone reverse lactation-induced downregulation of the uncoupling protein-3 gene.	reverse downregulation gene
5	Inhibition	Activation	Geldanamycin also disrupts the T-cell receptor-mediated activation of nuclear factor of activated T-cells (NF-AT).	disrupts activation
6	Substrate	Inhibition	Blockade of LTC4 synthesis caused by additive inhibition of gIV-PLA2 phosphorylation: Effect of salmeterol and PDE4 inhibition in human eosinophils.	synthesis caused inhibition phosphorylation

meaning to the textual sequence. However, these terms are not correctly handled by our model. For example, cases 4 and 5 have, in the SDP, the expressions "reverse downregulation" and "disrupts activation" which should direct to the true relation classes, namely activation and inhibition.

- Complex sentences, requiring expert interpretation. Some cases, as in example 6, are not easily interpreted without domain knowledge or more context.

To counteract these errors, we hypothesize that improved feature representations and more training data may alleviate these issues. Also, we suspect that building a system for multi-label classification would improve recall, and could improve the final results, since there are failed predicted relations that count simultaneously as a false

positive and a false negative.

Another limitation of our model is that for each chemical–protein pair only information from the respective sentence is being used. We suspect more context would prove helpful, and could facilitate the extraction of cross-sentence relations.

5.3 Summary

In this chapter we introduced background work on biomedical relation extraction, enumerated several common evaluation datasets, and presented a deep learning model for identifying chemical–protein interactions, in PubMed abstracts, based on recurrent or convolutional neural networks. We mapped chemical–protein interactions from the BioGRID database to add as additional training data, but inclusion of these data did not improve results and we believe that a more accurate handling of these data could prove effective.

We recognize that the *relation extraction* task is far from being solved, and there is plenty room for improvement. The recent transformer models such as BERT (Devlin *et al.*, 2019) have been dictating the state-of-the-art performance not only in relation extraction but in various biomedical NLP tasks (Peng *et al.*, 2019; Gu *et al.*, 2021). Competitions and shared-task venues often launch to thrive the development of information extraction solutions, and these have been targeted with heavy neural network and transformer–based models which are only possible to train due to the improvement and advance of computer hardware resources.

Chapter 6

Conclusions

In this thesis we presented a plethora of biomedical information extraction systems composed of machine learning models, knowledge-based methods, or rule-based approaches built with handcrafted heuristics. These were applied and evaluated in textual data from biomedical literature and clinical reports. Particularly, we investigated solutions for the NLP tasks of word sense disambiguation, entity linking, document classification, text similarity measurement, and relation extraction. The application of these techniques contributes to improved information extraction from the unstructured text found in biomedical literature which is relevant for molecular biology, biomedicine, and chemistry (Krallinger *et al.*, 2005, 2017b). For example, finding biological entities such as gene and protein names, and their relationships, in the millions of articles that exist in the literature helps to unveil hidden information and provide hints for new discoveries.

In this last chapter, we highlight the key contributions of our thesis work, show limitations of our proposed methods, and point the reader to future research directions.

6.1 Key contributions

The main contributions of this thesis are supported by extensive and detailed experiments that we performed in the different levels, or tasks, of a complete pipelined information extraction system. These contributions, in many biomedical text mining tasks, are summarized below:

- We proposed a new method based on external knowledge captured from standard medical terminologies to improve biomedical word sense disambiguation in scientific articles. Moreover, we compared this approach with supervised learning classifiers and verified that the use of ground-truth training instances allows to achieve higher accuracies, but knowledge-based systems have the ability to adapt

(generalize) to different ambiguous concepts without the need of training samples.

- We developed a knowledge-based method for medical concept normalization where entity mentions in clinical text such as medications, disorders, and medical treatments are linked to unique codes from standard medical terminologies. The system, based on pre-trained biomedical word embeddings, consists in a straightforward yet effective computation: the cosine similarity between (1) the vector embedding of the target entity mention and (2) every pre-calculated concept embedding from the training subset (of the corpus being used) and the UMLS database (considering only the RxNorm and SNOMED CT vocabularies in the specific work that was conducted).

- We investigated the use of classical machine learning and convolutional recurrent neural networks for document triage. Our deep learning models showed competitive performance in selecting PubMed abstracts that contain protein–protein interactions affected by genetic mutations.

- Regarding clinical text classification, we created a hybrid system for automatic patient cohort selection where clinical records were used to find if the patients met certain selection criteria for clinical trials. We concluded that, due to the small size of the dataset and its high imbalance in labels, handcrafted rules performed overall better, while for some, more balanced criteria, machine learning models proved effective.

- We studied the use of word and sentence embeddings with neural network models to quantify the semantic textual similarity between clinical sentences. Results with sentence embeddings achieved better performance in comparison to word embeddings, which made us conclude that sentence vectors generated by BioSentVec (Chen *et al.*, 2019b) provided a superior semantic representation. Also, we observed that pre-processing the sentences with different levels of granularity, such as stop words removal or converting the numbers to their textual form, had a considerable impact with word embeddings but deteriorated performance when using sentence embeddings. Therefore we established that sentence embeddings not only provide a preferable representation but also they require less effort in 'fine-tuning' the input textual data.

- Regarding the task of relation extraction we investigated and evaluated the use of recurrent and convolutional neural networks for identifying interactions between chemicals and proteins in PubMed abstracts. Our best method uses a small feature vector (narrow instance representation), from up to only 10 words per each candi-

date pair. Our BiLSTM network showed improved performance and comparable results to other works. Finally, we performed an extensive error analysis.

In short, in this thesis we proposed different methods for various information extraction tasks applied within the biomedical domain. The majority of them were machine learning–based (some being deep neural networks), and the remaining consisted of knowledge-based systems or approaches with heuristics. We conclude that machine learning methods are in general 'easier to teach'—require less feature engineering, model handcrafting and fine-tuning—and can obtain better results but at the cost of sufficient and high-quality labeled data frequently annotated in advance by domain experts.

The existence of annotated datasets and lexical resources is in fact a primary requirement for information extraction. Without access to the free text found in biomedical literature or eletronic health records this work would be impractical. A related requisite is the existence of ground-truth or high-quality data, relevant for biomedical text mining, such as (1) annotated datasets, and (2) curated vocabularies, terminologies, and databases. These resources are fundamental because they endow the researchers with the data needed for creating their information extraction systems.

Annotated datasets by domain experts, such as documents with biomedical entities and their interactions identified, are important for assessing and comparing the developed methods, and help to find out what approaches work best. Since these corpora are associated with gold-standard labels (annotations) they are favored for supervised machine learning models that learn from training data. However, a dataset may be considered small and not provide enough training samples for a machine learning model to achieve an acceptable performance; in these cases other approaches based on heuristics or external knowledge should be explored.

Fortunately, corpora for biomedical text mining is becoming less scarce due to the increasing interest for automatic solutions that can extract information from the text (Rosário-Ferreira *et al.*, 2021). Creation of these datasets is often performed manually by domain expert curators, such as biologists, chemists, or pharmacologists, with the help of software tools for text annotation (Neves and Leser, 2014; Neves and Ševa, 2021)—this is especially the case when entity mentions and relations need to be marked in a document. On the other hand, some researchers already proposed (semi-)automatic procedures to alleviate the manual efforts of curation. For instance, Jimeno-Yepes *et al.* (2011) present an automatic method to create a dataset for biomedical WSD, though they took benefit from the manual MeSH indexing that is routinely made by NLM experts. In another work, Pérez-Pérez *et al.* (2022) proposed a semi-automatic workflow for supporting a biomedical relation extraction curation task: they implement a deep learning model that learns from the decisions of the curators in iterative annotation rounds and show

potential relevant relations in next rounds.

In this thesis, despite of our broad use of gold-standard datasets for biomedical information extraction we did not investigate the creation of these since it was not one of the goals of this work—furthermore it would require interdisciplinary cooperation and qualified manpower for expert curation. Nevertheless we affirm that a straightforward approach for improving the performance of our methods would be the production and availability of more high-quality annotated data; since, in machine learning, a higher number of training samples usually leads to better performance and better generalization to unseen data. However, this forms an unrealistic or difficult scenario because data itself is limited and further manual annotation is a long, arduous, and costly task. Also, the use of larger datasets for training machine learning models (such as neural networks) would imply an increase in the training time and likely require more potent computer resources.

6.2 Limitations

As any developed solution for biomedical text mining our proposed methods have shortcomings or requirements that may restrict their suitability or reduce their performance in real-world applications. In this section we present specific limitations or drawbacks of our methods:

- Regarding the biomedical WSD task, the MSH WSD dataset contains PubMed abstracts in which the ambiguous terms appear. However, we presume that in some cases access to the full-text article could prove relevant for finding the correct sense of an ambiguous term. We did not explore this, yet both our approaches—supervised learning and knowledge-based—could be adapted to be applied to the full-text.

  Similarly, our word embedding models were generated using only the text from PubMed abstracts; and we did not investigate if adding full-text articles from PubMed Central (PMC) would be beneficial.

  Also, our knowledge-based method relies on textual definitions, extracted from UMLS knowledge sources, for every CUI (Concept Unique Identifier) to create *concepts embeddings*. During our preliminary experiments we inspected some of these descriptions and found that some were short (containing only a few words), and therefore we hypothesize that the embedding representations, and the overall method, would improve with more complete and correct definitions.

- In the medical concept normalization task our proposed system only uses the men-

tion text of the target entity to be normalized, which in some cases is not self-explanatory. We hypothesize that the context around the entities would provide more information and improve the normalization accuracy. Luo *et al.* (2019), the authors of the MCN corpus, also clarified that contextual information affects the results of normalization because annotators may interpret the context differently. They further explained that they only required the annotators to use contextual information when the mention itself did not provide enough information.

In clinical text, our model directly uses the vector embeddings of abbreviated terms, considering their (lowercased) surface form. This may not provide sufficient information, and a hybrid approach combining (1) word embeddings and (2) external dictionaries of abbreviations with their respective long forms could be helpful for disambiguation and normalization.

- We proposed the use of supervised learning methods for biomedical document triage. The aim was to detect if PubMed abstracts were relevant, or not, for extracting protein–protein interactions affected by genetic mutations. We experimented with adding more training data employing the BioCreative III PPI corpus (Krallinger *et al.*, 2011) in a self-training approach, but the results only improved by a tiny margin (0.12% F1-score in the official test set). We believe that including external corpora for training can be more beneficial but further investigation is required.

- In the task of patient cohort selection for clinical trials, we found that rule-based methods were more adequate given the relatively small size of the dataset. However, the results show that our heuristics were severely overfit to the training set and could be improved with unbiased and specialized knowledge from physicians or clinical experts.

  Additionally, we tested removing tabular information from the clinical documents to restrict their content to free text, but we did not find significant differences in the results. Extracting the tabular information and customizing its analysis, instead of discarding it, could be a more viable approach for some criteria because tables may contain clinical tests' measurements, or dosages of medications, relevant for inferring the patient health status.

- When measuring the semantic textual similarity between clinical sentences, we observed that our model could more easily identify pairs of highly similar or dissimilar sentences, but struggled with sentences that were not equivalent but shared identical portions or were about the same subject.

- For our chemical–protein relation extraction system, we pre-processed the BioGRID database (Chatr-aryamontri *et al.*, 2017) for extracting additional chemical–protein interactions for training the model but the results deteriorated and we believe that a more rigorous treatment of the data would be necessary.

  Also, we note that our model is limited to extracting relations within sentences and requires that chemical and protein entities are previously identified, since performing named entity recognition was out of scope in this task.

6.3 Future research

As highlighted in the previous section, our proposed methods have some limitations and unexplored details that can be further studied or addressed in upcoming research. We now point out other aspects that can be investigated for improving biomedical text mining according to state-of-the-art research.

The use of word embeddings was investigated in every biomedical NLP task presented in this thesis. In our initial experiments, we pre-trained our word embedding models employing the word2vec algorithm from the Gensim library (Řehůřek and Sojka, 2010), using the continuous bag-of-words and skip-gram architectures proposed by Mikolov *et al.* (2013a). Then, we tested the BioWordVec model (Chen *et al.*, 2019b) that consists of word vectors trained on biomedical and clinical text and is based on the fastText algorithm that takes into account subword information (Bojanowski *et al.*, 2017). From our experiments we concluded that BioWordVec provided superior representations, also having the advantage of computing out-of-vocabulary words since fastText exploits subword information.

Word embeddings pre-trained using the word2vec and fastText approaches return fixed vectors—a word has the same vector regardless of the context in which it is inserted. This is a known limitation of these models because they attribute the same vector to a word that may have different meanings depending on which context appears. Lately, to counteract this issue, more advanced word representations have been proposed. These are known as *contextualized word representations* where the calculation of word vectors is made on-the-fly to take into account the context in which the words are inserted.

ELMo (Embeddings from Language Models) proposed by Peters *et al.* (2018) and BERT (Bidirectional Encoder Representations from Transformers) proposed by Devlin *et al.* (2019) are arguably the two most used contextualized representations and have improved results considerably for a variety of NLP tasks. Since the publication of BERT, many variants have been pre-trained on biomedical and clinical textual data including BioBERT (Lee *et al.*, 2020), PubMedBERT (Gu *et al.*, 2021), and ClinicalBERT (Alsentzer

et al., 2019; Huang *et al.*, 2019). The exploration of these models is in our opinion a viable research direction. For instance, Peng *et al.* (2019) made an extensive evaluation of BERT and ELMo on ten datasets for biomedical NLP presenting several improvements over the state-of-the-art.

Another interesting idea is to explore token-free models that do not require the tokenization step and operate directly at the byte- or character-level. A recent example of such models is the ByT5 architecture (Xue *et al.*, 2022) where the authors use a standard transformer architecture to process byte sequences. These models have the advantage of easily process text in any language and remove errors from the text pre-processing pipeline.

Our studies in biomedical information extraction were restricted to the English language because research is commonly most updated for the English idiom and there is a lag in developing resources such as annotated corpora, curated databases, and word embeddings models in other languages. We consider that the resolution of biomedical NLP tasks in other languages is an under-researched area. To the best of our knowledge, Ferreira (2011) presents the first information extraction system to process clinical records written in European Portuguese. Schneider *et al.* (2020) transfer-learned information from a multilingual BERT to a corpora of clinical and biomedical text in Brazilian Portuguese releasing the BioBERTpt model. Silva e Oliveira *et al.* (2022) present the first available Brazilian Portuguese corpus for clinical NLP tasks (SemClinBr). And more recently, Miranda-Escalada *et al.* (2022) organized a shared task to promote the development of automatic methods for the recognition and normalization of disease mentions in Spanish clinical narratives. Therefore we consider that targeting biomedical NLP tasks in other languages is a relevant future research direction.

Another emerging research area is the study of information extraction from text found on social media platforms such as Twitter and Reddit. Although social media content might be about any topic, some of the users' posts may contain relevant clinical information such as adverse drug effects. Users may share and discuss their current health status after taking a medicine or have undergone a medical procedure. However, the processing of social media text poses particular challenges because text may be clumsy and contain misspellings, abbreviations, slang terms, and emojis. One example of recent research on mining social media text is the detection of medication mentions from tweets (Weissenbacher *et al.*, 2019, 2021; Zhang *et al.*, 2022b). Hence we argue that analysis of social media textual data is also relevant for biomedical discoveries and has potential for future research.

Finally, we believe that a robust idea that could improve our information extraction systems would be the use of neural network–based joint learning approaches, where

multiple NLP tasks are trained and learned simultaneously, which minimizes error propagation from initial steps. For example, the tasks of named entity recognition and relation extraction could be addressed together through joint learning as shown in previous research (Bekoulis *et al.*, 2018b; Luo *et al.*, 2020a).

This thesis presented several ideas and methods for a wide range of NLP problems involving information extraction in the biomedical domain. We believe that there is still much room for improvement and that biomedical text mining will continue to benefit greatly from deep learning breakthroughs and more curated resources.

References

Abadi, Martín, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng (Nov. 2016). "TensorFlow: a system for large-scale machine learning." In: *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* (Savannah, Georgia, USA). USENIX Association, pp. 265–283.
URL: https://www.usenix.org/conference/osdi16/program (cit. on pp. 67, 113).

Abney, Steven P. (1991). "Parsing by chunks." In: *Principle-Based Parsing: Computation and Psycholinguistics*. Ed. by Robert C. Berwick, Steven P. Abney, and Carol Tenny. Studies in Linguistics and Philosophy. Springer Nature, pp. 257–278.
URL: https://doi.org/10.1007/978-94-011-3474-3_10 (cit. on p. 14).

Adel, Heike and Hinrich Schütze (Sept. 2017). "Global normalization of convolutional neural networks for joint entity and relation classification." In: *2017 Conference on Empirical Methods in Natural Language Processing* (Copenhagen, Denmark). Association for Computational Linguistics, pp. 1723–1729.
URL: https://doi.org/10.18653/v1/D17-1181 (cit. on p. 24).

Aggarwal, Charu C. and ChengXiang Zhai (2012). "A survey of text classification algorithms." In: *Mining Text Data*. Ed. by Charu C. Aggarwal and ChengXiang Zhai. Springer Nature, pp. 163–222.
URL: https://doi.org/10.1007/978-1-4614-3223-4_6 (cit. on p. 63).

Agirre, Eneko, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe (June 2015). "SemEval-2015 task 2: semantic textual similarity, English, Spanish and pilot on interpretability." In: *9th International Workshop on Semantic Evaluation (SemEval 2015)* (Denver, Colorado, USA). Association for Computational Linguistics, pp. 252–263.
URL: https://doi.org/10.18653/v1/S15-2045 (cit. on pp. 63, 82).

Agirre, Eneko, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe (Aug. 2014). "SemEval-2014 task 10: multilingual semantic textual similarity." In: *8th International Workshop on Semantic Evaluation (SemEval 2014)* (Dublin, Ireland). Association for Computational Linguistics, pp. 81–91.
URL: https://doi.org/10.3115/v1/S14-2010 (cit. on pp. 63, 82).

Agirre, Eneko, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe (June 2016). "SemEval-2016 task 1: semantic textual similarity, monolingual and cross-lingual evaluation." In: *10th International Workshop on Semantic Evaluation (SemEval-2016)* (San Diego, California, USA). Association for Computational Linguistics, pp. 497–511.
URL: https://doi.org/10.18653/v1/S16-1081 (cit. on pp. 63, 82).

Agirre, Eneko, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo (June 2013). "*SEM 2013 shared task: semantic textual similarity." In: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Main Conference and the Shared Task: Semantic Textual Similarity* (Atlanta, Georgia, USA). Association for Computational Linguistics, pp. 32–43.
URL: https://aclanthology.org/S13-1004 (cit. on pp. 63, 82).

Agirre, Eneko, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre (June 2012). "SemEval-2012 task 6: a pilot on semantic textual similarity." In: *First Joint Conference on Lexical and Computational Semantics - Volume 1: Main Conference and the Shared Task, and Volume 2: Sixth International Workshop on Semantic Evaluation* (Montreal, Quebec, Canada). Association for Computational Linguistics, pp. 385–393.
URL: http://dl.acm.org/citation.cfm?id=2387636.2387697 (cit. on pp. 63, 82).

Aho, Alfred V., Monica S. Lam, Ravi Sethi, and Jeffrey D. Ullman (2007). *Compilers: principles, techniques, and tools*. 2nd ed. Pearson.
URL: https://suif.stanford.edu/dragonbook/ (cit. on p. 100).

Airola, Antti, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski (Nov. 2008). "All-paths graph kernel for protein–protein interaction extraction with evaluation of cross-corpus learning." In: *BMC Bioinformatics* 9.11. BioMed Central Ltd, S2.
URL: https://doi.org/10.1186/1471-2105-9-S11-S2 (cit. on p. 100).

Akbik, Alan, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf (June 2019). "FLAIR: an easy-to-use framework for state-of-the-art NLP." In: *2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)* (Minneapolis, Minnesota). Association for Com-

putational Linguistics, pp. 54–59.
URL: https://doi.org/10.18653/v1/n19-4010 (cit. on p. 33).

Aliguliyev, Ramiz M. (May 2009). "A new sentence similarity measure and sentence based extractive technique for automatic text summarization." In: *Expert Systems with Applications* 36.4. Elsevier, pp. 7764–7772.
URL: https://doi.org/10.1016/j.eswa.2008.11.022 (cit. on p. 60).

Allahyari, Mehdi, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut (July 2017). *A brief survey of text mining: classification, clustering and extraction techniques.* arXiv:1707.02919.
URL: https://arxiv.org/abs/1707.02919 (cit. on p. 1).

Allot, Alexis, Kyubum Lee, Qingyu Chen, Ling Luo, and Zhiyong Lu (July 2021). "LitSuggest: a web-based system for literature recommendation and curation using machine learning." In: *Nucleic Acids Research* 49.W1. Oxford University Press, W352–W358.
URL: https://doi.org/10.1093/nar/gkab326 (cit. on p. 33).

Almeida, Tiago, Rui Antunes, João F. Silva, João R. Almeida, and Sérgio Matos (July 2022). "Chemical identification and indexing in PubMed full-text articles using deep learning and heuristics." In: *Database* 2022 (baac047). Oxford University Press.
URL: https://doi.org/10.1093/database/baac047 (cit. on p. 8).

Almeida, Tiago, Rui Antunes, João Figueira Silva, João Rafael Almeida, and Sérgio Matos (Nov. 2021). "Chemical detection and indexing in PubMed full text articles using deep learning and rule-based methods." In: *BioCreative VII Challenge Evaluation Workshop* (Online), pp. 119–123.
URL: https://biocreative.bioinformatics.udel.edu/resources/publications/bc-vii-workshop-proceedings/ (cit. on p. 8).

Alsentzer, Emily, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott (June 2019). "Publicly available clinical BERT embeddings." In: *2nd Clinical Natural Language Processing Workshop* (Minneapolis, Minnesota, USA). Association for Computational Linguistics, pp. 72–78.
URL: https://doi.org/10.18653/v1/W19-1909 (cit. on pp. 63, 130).

Altınel, Berna and Murat Can Ganiz (Nov. 2018). "Semantic text classification: a survey of past and recent advances." In: *Information Processing & Management* 54.6. Elsevier, pp. 1129–1153.
URL: https://doi.org/10.1016/j.ipm.2018.08.001 (cit. on p. 63).

Altınel, Berna, Zehra Melce Hüsünbeyi, and Arzucan Özgür (Oct. 2017). "Text classification using ontology and semantic values of terms for mining protein interactions and mutations." In: *BioCreative VI Workshop* (Bethesda, Maryland, USA), pp. 110–114.

URL: https://biocreative.bioinformatics.udel.edu/events/biocreative-vi/workshop/ (cit. on p. 69).

Amos, Brandon (May 2019). "Differentiable optimization-based modeling for machine learning." PhD thesis. Carnegie Mellon University.
URL: http://reports-archive.adm.cs.cmu.edu/anon/2019/abstracts/19-109.html (cit. on p. 9).

Ananiadou, Sophia, Douglas B. Kell, and Jun-ichi Tsujii (Dec. 2006). "Text mining and its potential applications in systems biology." In: *Trends in Biotechnology* 24.12. Elsevier, pp. 571–579.
URL: https://doi.org/10.1016/j.tibtech.2006.10.002 (cit. on pp. 32, 103).

Ananiadou, Sophia, Paul Thompson, Raheel Nawaz, John McNaught, and Douglas B. Kell (May 2015). "Event-based text mining for biology and functional genomics." In: *Briefings in Functional Genomics* 14.3. Oxford University Press, pp. 213–230.
URL: https://doi.org/10.1093/bfgp/elu015 (cit. on p. 106).

Angeli, Gabor, Julie Tibshirani, Jean Wu, and Christopher D. Manning (Oct. 2014). "Combining distant and partial supervision for relation extraction." In: *2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha, Qatar). Association for Computational Linguistics, pp. 1556–1567.
URL: https://doi.org/10.3115/v1/d14-1164 (cit. on p. 95).

Antunes, Rui, Tiago Almeida, João Figueira Silva, and Sérgio Matos (Nov. 2021). "Chemical–protein relation extraction in PubMed abstracts using BERT and neural networks." In: *BioCreative VII Challenge Evaluation Workshop* (Online), pp. 76–79.
URL: https://biocreative.bioinformatics.udel.edu/resources/publications/bc-vii-workshop-proceedings/ (cit. on p. 8).

Antunes, Rui and Sérgio Matos (Oct. 2016). "Machine learning with word embeddings applied to biomedical concept disambiguation." In: *22nd Portuguese Conference on Pattern Recognition* (Aveiro, Portugal). University of Aveiro, pp. 77–78.
URL: http://hdl.handle.net/10773/25118 (cit. on pp. 5, 43, 44).

Antunes, Rui and Sérgio Matos (June 2017a). "Biomedical word sense disambiguation with word embeddings." In: *11th International Conference on Practical Applications of Computational Biology & Bioinformatics* (Porto, Portugal). Springer Nature, pp. 273–279.
URL: http://hdl.handle.net/10773/25112 (cit. on pp. 5, 6, 43, 45).

Antunes, Rui and Sérgio Matos (Oct. 2017b). "Evaluation of word embedding vector averaging functions for biomedical word sense disambiguation." In: *9th INForum — Informatics Symposium* (Aveiro, Portugal). University of Aveiro, pp. 25–30.
URL: http://hdl.handle.net/10773/25119 (cit. on pp. 5, 6, 43, 45).

Antunes, Rui and Sérgio Matos (Dec. 2017c). "Supervised learning and knowledge-based approaches applied to biomedical word sense disambiguation." In: *Journal of Integrative Bioinformatics* 14.4. De Gruyter.
URL: https://doi.org/10.1515/jib-2017-0051 (cit. on pp. 6, 43, 45, 52, 111).

Antunes, Rui and Sérgio Matos (Oct. 2019). "Extraction of chemical–protein interactions from the literature using neural networks and narrow instance representation." In: *Database* 2019 (baz095). Oxford University Press.
URL: https://doi.org/10.1093/database/baz095 (cit. on pp. 7, 8, 119).

Antunes, Rui, João Figueira Silva, and Sérgio Matos (Mar. 2020). "Evaluating semantic textual similarity in clinical sentences using deep learning and sentence embeddings." In: *35th Annual ACM Symposium on Applied Computing* (Online). ACM, pp. 662–669.
URL: http://hdl.handle.net/10773/31473 (cit. on p. 7).

Antunes, Rui, João Figueira Silva, Arnaldo Pereira, and Sérgio Matos (Feb. 2019). "Rule-based and machine learning hybrid system for patient cohort selection." In: *12th International Joint Conference on Biomedical Engineering Systems and Technologies* (Prague, Czech Republic). SciTePress, pp. 59–67.
URL: https://doi.org/10.5220/0007349300590067 (cit. on pp. 7, 8).

Asada, Masaki, Makoto Miwa, and Yutaka Sasaki (June 2021). "Using drug descriptions and molecular structures for drug–drug interaction extraction from literature." In: *Bioinformatics* 37.12. Oxford University Press, pp. 1739–1746.
URL: https://doi.org/10.1093/bioinformatics/btaa907 (cit. on p. 102).

Bach, Nguyen and Sameer Badaskar (2007). *A review of relation extraction*. Literature review for Language and Statistics II. Canergie Mellon University, School of Computer Science, Language Technologies Institute.
URL: https://www.cs.cmu.edu/~nbach/papers/A-survey-on-Relation-Extraction.pdf (cit. on p. 95).

Baeza-Yates, Ricardo and Berthier Ribeiro-Neto (1999). *Modern information retrieval*. Addison-Wesley Longman Publishing Co., Inc.
URL: https://dl.acm.org/citation.cfm?id=553876 (cit. on p. 2).

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (Sept. 2014). *Neural machine translation by jointly learning to align and translate*. arXiv:1409.0473.
URL: https://arxiv.org/abs/1409.0473 (cit. on p. 102).

Baker, Simon (Sept. 2017). "Semantic text classification for cancer text mining." PhD thesis. University of Cambridge.
URL: https://doi.org/10.17863/CAM.23105 (cit. on p. 9).

Balog, Krisztian (Oct. 2018). "Populating knowledge bases." In: *Entity-Oriented Search*. Ed. by Krisztian Balog. Vol. 39. The Information Retrieval Series. Springer Nature, pp. 189–222.
URL: https://doi.org/10.1007/978-3-319-93935-3_6 (cit. on pp. 59, 95, 96).

Banerjee, Satanjeev and Ted Pedersen (2002). "An adapted Lesk algorithm for word sense disambiguation using WordNet." In: *Computational Linguistics and Intelligent Text Processing* (Mexico City, Mexico). Ed. by Alexander Gelbukh. Springer Nature, pp. 136–145.
URL: https://doi.org/10.1007/3-540-45715-1_11 (cit. on p. 39).

Banko, Michele, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni (Jan. 2007). "Open information extraction from the web." In: *Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07)* (Hyderabad, India), pp. 2670–2676.
URL: https://www.ijcai.org/proceedings/2007 (cit. on pp. 28, 96).

Baumgartner, William A., K. Bretonnel Cohen, Lynne M. Fox, George Acquaah-Mensah, and Lawrence Hunter (July 2007). "Manual curation is not sufficient for annotation of genomic databases." In: *Bioinformatics* 23.13. Oxford University Press, pp. i41–i48.
URL: https://doi.org/10.1093/bioinformatics/btm229 (cit. on pp. 41, 99).

Bekoulis, Giannis, Johannes Deleu, Thomas Demeester, and Chris Develder (Aug. 2018a). *Adversarial training for multi-context joint entity and relation extraction*. arXiv:1808.06876.
URL: https://arxiv.org/abs/1808.06876 (cit. on p. 24).

Bekoulis, Giannis, Johannes Deleu, Thomas Demeester, and Chris Develder (Dec. 2018b). "Joint entity recognition and relation extraction as a multi-head selection problem." In: *Expert Systems with Applications* 114. Elsevier, pp. 34–45.
URL: https://doi.org/10.1016/j.eswa.2018.07.032 (cit. on pp. 24, 132).

Beltagy, Iz, Kyle Lo, and Arman Cohan (Nov. 2019). "SciBERT: a pretrained language model for scientific text." In: *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong, China). Association for Computational Linguistics, pp. 3615–3620.
URL: https://doi.org/10.18653/v1/d19-1371 (cit. on p. 102).

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin (Feb. 2003). "A neural probabilistic language model." In: *Journal of Machine Learning Research* 3, pp. 1137–1155.
URL: https://jmlr.org/papers/v3/bengio03a.html (cit. on p. 111).

Bhowmick, Alexy and Shyamanta M. Hazarika (2018). "E-mail spam filtering: a review of techniques and trends." In: *Advances in Electronics, Communication and Computing*. Ed. by Akhtar Kalam, Swagatam Das, and Kalpana Sharma. Springer Nature, pp. 583–590.
URL: https://doi.org/10.1007/978-981-10-4765-7_61 (cit. on p. 59).

Bird, Steven, Ewan Klein, and Edward Loper (2009). *Natural language processing with Python*. 1st ed. O'Reilly Media, Inc.
URL: https://dl.acm.org/citation.cfm?id=1717171 (cit. on p. 11).

Björne, Jari and Tapio Salakoski (Oct. 2015). "TEES 2.2: biomedical event extraction for diverse corpora." In: *BMC Bioinformatics* 16.16. BioMed Central Ltd, S4.
URL: https://doi.org/10.1186/1471-2105-16-S16-S4 (cit. on pp. 109, 112).

Blaschke, Christian, Miguel A. Andrade, Christos Ouzounis, and Alfonso Valencia (Aug. 1999). "Automatic extraction of biological information from scientific text: protein-protein interactions." In: *Seventh International Conference on Intelligent Systems for Molecular Biology* (Heidelberg, Germany). Association for the Advancement of Artificial Intelligence, pp. 60–67.
URL: https://www.aaai.org/Library/ISMB/1999/ismb99-008.php (cit. on p. 100).

Bodenreider, Oliver, Ronald Cornet, and Daniel J. Vreeman (Aug. 2018). "Recent developments in clinical terminologies — SNOMED CT, LOINC, and RxNorm." In: *Yearbook of Medical Informatics* 27.01, pp. 129–139.
URL: https://doi.org/10.1055/s-0038-1667077 (cit. on pp. 39, 42).

Bodenreider, Olivier (Jan. 2004). "The unified medical language system (UMLS): integrating biomedical terminology." In: *Nucleic Acids Research* 32 (Suppl 1). Oxford University Press, p. D267.
URL: https://doi.org/10.1093/nar/gkh061 (cit. on p. 39).

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017). "Enriching word vectors with subword information." In: *Transactions of the Association for Computational Linguistics* 5.1. Association for Computational Linguistics, pp. 135–146.
URL: https://doi.org/10.1162/tacl_a_00051 (cit. on pp. 56, 111, 112, 130).

Bollacker, Kurt, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor (June 2008). "Freebase: a collaboratively created graph database for structuring human knowledge." In: *2008 ACM SIGMOD international conference on Management of data* (Vancouver, British Columbia, Canada). ACM, pp. 1247–1250.
URL: https://doi.org/10.1145/1376616.1376746 (cit. on p. 100).

Bossy, Robert, Philippe Bessières, and Claire Nédellec (Aug. 2013a). "BioNLP Shared Task 2013 – An overview of the Genic Regulation Network Task." In: *BioNLP Shared Task*

2013 Workshop (Sofia, Bulgaria). Association for Computational Linguistics, pp. 153–160.
URL: https://aclanthology.org/W13-2023 (cit. on p. 99).

Bossy, Robert, Louise Deléger, Estelle Chaix, Mouhamadou Ba, and Claire Nédellec (Nov. 2019). "Bacteria Biotope at BioNLP Open Shared Tasks 2019." In: *5th Workshop on BioNLP Open Shared Tasks* (Hong Kong, China). Association for Computational Linguistics, pp. 121–131.
URL: https://doi.org/10.18653/v1/d19-5719 (cit. on p. 99).

Bossy, Robert, Wiktoria Golik, Zorana Ratkovic, Philippe Bessières, and Claire Nédellec (Aug. 2013b). "BioNLP Shared Task 2013 – an overview of the Bacteria Biotope Task." In: *BioNLP Shared Task 2013 Workshop* (Sofia, Bulgaria). Association for Computational Linguistics, pp. 161–169.
URL: https://aclanthology.org/W13-2024 (cit. on p. 99).

Bossy, Robert, Wiktoria Golik, Zorana Ratkovic, Dialekti Valsamou, Philippe Bessières, and Claire Nédellec (June 2015). "Overview of the gene regulation network and the bacteria biotope tasks in BioNLP'13 shared task." In: *BMC Bioinformatics* 16.10. BioMed Central Ltd, S1.
URL: https://doi.org/10.1186/1471-2105-16-S10-S1 (cit. on p. 99).

Bossy, Robert, Julien Jourde, Philippe Bessières, Maarten van de Guchte, and Claire Nédellec (June 2011). "BioNLP Shared Task 2011 - Bacteria Biotope." In: *BioNLP Shared Task 2011 Workshop* (Portland, Oregon, USA). Association for Computational Linguistics, pp. 56–64.
URL: https://aclanthology.org/W11-1809 (cit. on p. 98).

Bossy, Robert, Julien Jourde, Alain-Pierre Manine, Philippe Veber, Erick Alphonse, Maarten van de Guchte, Philippe Bessières, and Claire Nédellec (June 2012). "BioNLP Shared Task - The Bacteria Track." In: *BMC Bioinformatics* 13.11. BioMed Central Ltd, S3.
URL: https://doi.org/10.1186/1471-2105-13-S11-S3 (cit. on p. 98).

Bouazizi, Mondher and Tomoaki Ohtsuki (May 2016). "Sentiment analysis: from binary to multi-class classification: a pattern-based approach for multi-class sentiment analysis in Twitter." In: *2016 IEEE International Conference on Communications (ICC)* (Kuala Lumpur, Malaysia). IEEE, pp. 1–6.
URL: https://doi.org/10.1109/icc.2016.7511392 (cit. on p. 59).

Bunescu, Razvan, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun K. Ramani, and Yuk Wah Wong (Feb. 2005). "Comparative experiments on learning information extractors for proteins and their interactions." In: *Artificial In-*

telligence in Medicine 33.2. Elsevier, pp. 139–155.
URL: https://doi.org/10.1016/j.artmed.2004.07.016 (cit. on p. 103).

Bunescu, Razvan C. and Raymond J. Mooney (Oct. 2005a). "A shortest path dependency kernel for relation extraction." In: *Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (Vancouver, British Columbia, Canada). Association for Computational Linguistics, pp. 724–731.
URL: https://doi.org/10.3115/1220575.1220666 (cit. on p. 109).

Bunescu, Razvan C. and Raymond J. Mooney (Dec. 2005b). "Subsequence kernels for relation extraction." In: *18th Conference on Neural Information Processing Systems (NIPS 2005)* (Vancouver, British Columbia, Canada). The MIT Press, pp. 171–178.
URL: http://papers.nips.cc/paper/2787-subsequence-kernels-for-relation-extraction (cit. on p. 103).

Bunescu, Razvan Constantin (Aug. 2007). "Learning for information extraction: from named entity recognition and disambiguation to relation extraction." PhD thesis. The University of Texas at Austin.
URL: http://www.cs.utexas.edu/users/ai-lab/?bunescu:phd07 (cit. on p. 103).

Cambria, Erik, Björn Schuller, Yunqing Xia, and Catherine Havasi (Mar. 2013). "New avenues in opinion mining and sentiment analysis." In: *IEEE Intelligent Systems* 28.2. IEEE, pp. 15–21.
URL: https://doi.org/10.1109/MIS.2013.30 (cit. on p. 19).

Campos, David (2013). "Mining biomedical information from scientific literature." PhD thesis. University of Aveiro.
URL: http://hdl.handle.net/10773/12853 (cit. on p. 9).

Campos, David, Quoc-Chinh Bui, Sérgio Matos, and José Luís Oliveira (Jan. 2014). "TrigNER: automatically optimized biomedical event trigger recognition on scientific documents." In: *Source Code for Biology and Medicine* 9.1. Springer Nature, p. 1.
URL: https://doi.org/10.1186/1751-0473-9-1 (cit. on p. 22).

Campos, David, Sérgio Matos, and José Luís Oliveira (Sept. 2013). "A modular framework for biomedical concept recognition." In: *BMC Bioinformatics* 14.1. BioMed Central Ltd, p. 281.
URL: https://doi.org/10.1186/1471-2105-14-281 (cit. on p. 106).

Cases, M., L. I. Furlong, J. Albanell, R. B. Altman, R. Bellazzi, S. Boyer, A. Brand, A. J. Brookes, S. Brunak, T. W. Clark, J. Gea, P. Ghazal, N. Graf, R. Guigó, T. E. Klein, N. López-Bigas, V. Maojo, B. Mons, M. Musen, J. L. Oliveira, A. Rowe, P. Ruch, A. Shabo, E. H. Shortliffe, A. Valencia, J. van der Lei, M. A. Mayer, and F. Sanz (Oct. 2013). "Improving data and knowledge management to better integrate health care

and research." In: *Journal of Internal Medicine* 274.4. Wiley, pp. 321–328.
URL: https://doi.org/10.1111/joim.12105 (cit. on p. 32).

Cer, Daniel, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia (July 2017). *SemEval-2017 Task 1: semantic textual similarity - multilingual and cross-lingual focused evaluation.* arXiv:1708.00055.
URL: https://arxiv.org/abs/1708.00055 (cit. on pp. 82, 85).

Chandrasekaran, Dhivya and Vijay Mago (Mar. 2022). "Evolution of Semantic Similarity—A Survey." In: *ACM Computing Surveys* 54.2. ACM, 41:1–41:37.
URL: https://doi.org/10.1145/3440755 (cit. on p. 63).

Charniak, Eugene and Mark Johnson (June 2005). "Coarse-to-fine n-best parsing and MaxEnt discriminative reranking." In: *43rd Annual Meeting on Association for Computational Linguistics* (Ann Arbor, Michigan). Association for Computational Linguistics, pp. 173–180.
URL: https://doi.org/10.3115/1219840.1219862 (cit. on p. 109).

Chatr-aryamontri, Andrew, Rose Oughtred, Lorrie Boucher, Jennifer Rust, Christie Chang, Nadine K. Kolas, Lara O'Donnell, Sara Oster, Chandra Theesfeld, Adnane Sellam, Chris Stark, Bobby-Joe Breitkreutz, Kara Dolinski, and Mike Tyers (Jan. 2017). "The BioGRID interaction database: 2017 update." In: *Nucleic Acids Research* 45.D1. Oxford University Press, pp. D369–D379.
URL: https://doi.org/10.1093/nar/gkw1102 (cit. on pp. 34, 115, 130).

Chen, Danqi and Christopher D. Manning (Oct. 2014). "A fast and accurate dependency parser using neural networks." In: *2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha, Qatar). Association for Computational Linguistics, pp. 740–750.
URL: https://aclanthology.org/D14-1082 (cit. on p. 109).

Chen, Eric Zhe-You, Onkar Singh, Toni Rose Jue, Chen-Kai Wang, Jitendra Jonnagaddala, and Hong-Jie Dai (Oct. 2017a). "Identifying mutation-induced protein-protein interactions in scientific literature." In: *BioCreative VI Workshop* (Bethesda, Maryland, USA), pp. 123–126.
URL: https://biocreative.bioinformatics.udel.edu/events/biocreative-vi/workshop/ (cit. on p. 69).

Chen, Qingyu, Alexis Allot, Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu (Nov. 2021a). "Overview of the BioCreative VII LitCovid track: multi-label topic classification for COVID-19 literature annotation." In: *BioCreative VII Challenge Evaluation Workshop* (Online), pp. 266–271.
URL: https://biocreative.bioinformatics.udel.edu/resources/publications/bc-vii-workshop-proceedings/ (cit. on p. 63).

Chen, Qingyu, Alexis Allot, Robert Leaman, Rezarta Islamaj, Jingcheng Du, Li Fang, Kai Wang, Shuo Xu, Yuefu Zhang, Parsa Bagherzadeh, Sabine Bergler, Aakash Bhatnagar, Nidhir Bhavsar, Yung-Chun Chang, Sheng-Jie Lin, Wentai Tang, Hongtong Zhang, Ilija Tavchioski, Senja Pollak, Shubo Tian, Jinfeng Zhang, Yulia Otmakhova, Antonio Jimeno Yepes, Hang Dong, Honghan Wu, Richard Dufour, Yanis Labrak, Niladri Chatterjee, Kushagri Tandon, Fréjus A. A. Laleye, Loïc Rakotoson, Emmanuele Chersoni, Jinghang Gu, Annemarie Friedrich, Subhash Chandra Pujari, Mariia Chizhikova, Naveen Sivadasan, Saipradeep VG, and Zhiyong Lu (Aug. 2022). "Multi-label classification for biomedical literature: an overview of the BioCreative VII LitCovid Track for COVID-19 literature topic annotations." In: *Database* 2022. Oxford University Press, baac069.
URL: https://doi.org/10.1093/database/baac069 (cit. on p. 63).

Chen, Qingyu, Alexis Allot, and Zhiyong Lu (Jan. 2021b). "LitCovid: an open database of COVID-19 literature." In: *Nucleic Acids Research* 49.D1. Oxford University Press, pp. D1534–D1540.
URL: https://doi.org/10.1093/nar/gkaa952 (cit. on p. 63).

Chen, Qingyu, Jingcheng Du, Sun Kim, W. John Wilbur, and Zhiyong Lu (Aug. 2018). "Combining rich features and deep learning for finding similar sentences in electronic medical records." In: *BioCreative/OHNLP Challenge 2018* (Washington, DC, USA).
URL: https://sites.google.com/view/ohnlp2018 (cit. on pp. 84, 86, 88).

Chen, Qingyu, Jingcheng Du, Sun Kim, W. John Wilbur, and Zhiyong Lu (Sept. 2019a). "Evaluation of five sentence similarity models on electronic medical records." In: *10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* (Niagara Falls, New York, USA). ACM, p. 533.
URL: https://doi.org/10.1145/3307339.3343239 (cit. on p. 85).

Chen, Qingyu, Nagesh C. Panyam, Aparna Elangovan, Melissa Davis, and Karin Verspoor (Oct. 2017b). "Document triage and relation extraction for protein-protein interactions affected by mutations." In: *BioCreative VI Workshop* (Bethesda, Maryland, USA), pp. 102–105.
URL: https://biocreative.bioinformatics.udel.edu/events/biocreative-vi/workshop/ (cit. on p. 69).

Chen, Qingyu, Yifan Peng, and Zhiyong Lu (June 2019b). "BioSentVec: creating sentence embeddings for biomedical texts." In: *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, pp. 1–5.
URL: https://doi.org/10.1109/ichi.2019.8904728 (cit. on pp. 33, 56, 63, 85, 88, 111, 115–117, 126, 130).

Chen, Qingyu, Alex Rankine, Yifan Peng, Elaheh Aghaarabi, and Zhiyong Lu (Dec. 2021c). "Benchmarking effectiveness and efficiency of deep learning models for semantic textual similarity in the clinical domain: validation study." In: *JMIR Medical Informatics* 9.12. JMIR Publications, e27386.
URL: https://doi.org/10.2196/27386 (cit. on p. 63).

Chen, Tianqi and Carlos Guestrin (Aug. 2016). "XGBoost: a scalable tree boosting system." In: *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA). ACM, pp. 785–794.
URL: https://doi.acm.org/10.1145/2939672.2939785 (cit. on p. 71).

Chinchor, Nancy and Beth Sundheim (Aug. 1993). "MUC-5 evaluation metrics." In: *Fifth Message Understanding Conference (MUC-5)* (Baltimore, Maryland, USA). Association for Computational Linguistics, pp. 69–78.
URL: https://aclanthology.org/M93-1007 (cit. on p. 29).

Cho, Han-Cheol, Naoaki Okazaki, Makoto Miwa, and Jun'ichi Tsujii (July 2013). "Named entity recognition with multiple segment representations." In: *Information Processing & Management* 49.4. Elsevier, pp. 954–965.
URL: https://doi.org/10.1016/j.ipm.2013.03.002 (cit. on pp. 20, 21).

Chollet, François *et al.* (Mar. 2015). "Keras." In:
URL: https://keras.io/ (cit. on pp. 67, 71, 88, 113).

Chollet, François (2017). *Deep learning with Python*. Manning Publications Co.
URL: https://www.manning.com/books/deep-learning-with-python (cit. on p. 113).

Chomsky, N. (Sept. 1956). "Three models for the description of language." In: *IRE Transactions on Information Theory* 2.3. IEEE, pp. 113–124.
URL: https://doi.org/10.1109/tit.1956.1056813 (cit. on p. 100).

Cohen, Aaron M. and William R. Hersh (Mar. 2005). "A survey of current work in biomedical text mining." In: *Briefings in Bioinformatics* 6.1. Oxford University Press, p. 57.
URL: https://doi.org/10.1093/bib/6.1.57 (cit. on p. 61).

Cohen, Raphael, Michael Elhadad, and Noémie Elhadad (Jan. 2013). "Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies." In: *BMC Bioinformatics* 14.10. BioMed Central Ltd.
URL: https://doi.org/10.1186/1471-2105-14-10 (cit. on p. 82).

Comeau, Donald C., Rezarta Islamaj Doğan, Paolo Ciccarese, Kevin Bretonnel Cohen, Martin Krallinger, Florian Leitner, Zhiyong Lu, Yifan Peng, Fabio Rinaldi, Manabu Torii, Alfonso Valencia, Karin Verspoor, Thomas C. Wiegers, Cathy H. Wu, and W. John Wilbur (Jan. 2013). "BioC: a minimalist approach to interoperability for biomedical text processing." In: *Database* 2013. Oxford University Press, bat064.
URL: https://doi.org/10.1093/database/bat064 (cit. on p. 33).

Corbett, P. and J. Boyle (Apr. 2017a). "Chemlistem - chemical named entity recognition using recurrent neural networks." In: *BioCreative V.5 Challenge Evaluation Workshop* (Barcelona, Spain), pp. 61–68.
URL: https://biocreative.bioinformatics.udel.edu/resources/publications/bcv5_proceedings/ (cit. on pp. 102, 119).

Corbett, P. and J. Boyle (Oct. 2017b). "Improving the learning of chemical–protein interactions from literature using transfer learning and word embeddings." In: *BioCreative VI Workshop* (Bethesda, Maryland, USA), pp. 180–183.
URL: https://biocreative.bioinformatics.udel.edu/events/biocreative-vi/workshop/ (cit. on pp. 102, 119).

Corbett, P. and J. Boyle (Jan. 2018). "Improving the learning of chemical–protein interactions from literature using transfer learning and specialized word embeddings." In: *Database* 2018. Oxford University Press, bay066.
URL: https://doi.org/10.1093/database/bay066 (cit. on pp. 102, 115, 119).

Cornet, Ronald and Nicolette de Keizer (Oct. 2008). "Forty years of SNOMED: a literature review." In: *BMC Medical Informatics and Decision Making* 8.1. BioMed Central Ltd, S2.
URL: https://doi.org/10.1186/1472-6947-8-S1-S2 (cit. on pp. 34, 42).

Cote, Roger A. (Oct. 1986). "Architecture of SNOMED: its contribution to medical language processing." In: *Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association, pp. 74–80.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245000/ (cit. on p. 42).

Cotton, Richard G. H., Kate Phillips, and Ourania Horaitis (Mar. 2007). "A survey of locus-specific database curation." In: *Journal of Medical Genetics* 44.4. BMJ Publishing Group Ltd, e72.
URL: https://doi.org/10.1136/jmg.2006.044081 (cit. on p. 96).

Crammer, Koby, Mark Dredze, Kuzman Ganchev, Partha Pratim Talukdar, and Steven Carroll (June 2007). "Automatic code assignment to medical text." In: *Biological, translational, and clinical language processing* (Prague, Czech Republic). Association for Computational Linguistics, pp. 129–136.
URL: https://aclanthology.org/W07-1017 (cit. on p. 62).

Craven, Mark and Johan Kumlien (Aug. 1999). "Constructing biological knowledge bases by extracting information from text sources." In: *Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB-99)* (Heidelberg, Germany). Association for the Advancement of Artificial Intelligence, pp. 77–86.
URL: http://aaai.org/Library/ISMB/ismb99contents.php (cit. on pp. 29, 60, 99).

Dai, Hong-Jie, Po-Ting Lai, Yung-Chun Chang, and Richard Tzong-Han Tsai (Jan. 2015). "Enhancing of chemical compound and drug name recognition using representative tag scheme and fine-grained tokenization." In: *Journal of Cheminformatics* 7.1. BioMed Central Ltd, S14.
URL: https://doi.org/10.1186/1758-2946-7-S1-S14 (cit. on p. 20).

Dalianis, Hercules (2018). "Evaluation metrics and evaluation." In: *Clinical Text Mining: Secondary Use of Electronic Patient Records*. Ed. by Hercules Dalianis. Springer Nature, pp. 45–53.
URL: https://doi.org/10.1007/978-3-319-78503-5_6 (cit. on p. 29).

Davis, Allan Peter, Cynthia J. Grondin, Robin J. Johnson, Daniela Sciaky, Benjamin L. King, Roy McMorran, Jolene Wiegers, Thomas C. Wiegers, and Carolyn J. Mattingly (Jan. 2017). "The comparative toxicogenomics database: update 2017." In: *Nucleic Acids Research* 45.D1. Oxford University Press, p. D972.
URL: https://doi.org/10.1093/nar/gkw838 (cit. on p. 34).

Deléger, Louise, Robert Bossy, Estelle Chaix, Mouhamadou Ba, Arnaud Ferré, Philippe Bassières, and Claire Nédellec (Aug. 2016). "Overview of the Bacteria Biotope task at BioNLP shared task 2016." In: *4th BioNLP Shared Task Workshop* (Berlin, Germany). Association for Computational Linguistics, pp. 12–22.
URL: https://aclanthology.org/W16-3002 (cit. on p. 99).

De Marneffe, Marie-Catherine and Christopher D. Manning (Aug. 2008). "The Stanford typed dependencies representation." In: *Coling 2008: Workshop on Cross-Framework and Cross-Domain Parser Evaluation* (Stroudsburg, Pennsylvania, USA). Association for Computational Linguistics, pp. 1–8.
URL: http://dl.acm.org/citation.cfm?id=1608858.1608859 (cit. on p. 14).

De Marneffe, Marie-Catherine and Christopher D. Manning (Sept. 2016). *Stanford typed dependencies manual*. Stanford University.
URL: https://nlp.stanford.edu/software/stanford-dependencies.shtml (cit. on pp. 14, 110).

Demner-Fushman, Dina, Kin Wah Fung, Phong Do, Richard D. Boyce, and Goodwin Travis (Nov. 2018). "Overview of the TAC 2018 drug-drug interaction extraction from drug labels track." In: *Text Analysis Conference (TAC 2018)* (Gaithersburg, Maryland, USA).
URL: https://tac.nist.gov/publications/2018/papers.html (cit. on p. 97).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "BERT: pre-training of deep bidirectional transformers for language understanding." In: *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapo-

lis, Minnesota, USA). Association for Computational Linguistics, pp. 4171–4186. URL: https://doi.org/10.18653/v1/N19-1423 (cit. on pp. 24, 124, 130).

Diao, Yanlei, Hongjun Lu, and Dekai Wu (Apr. 2000). "A comparative study of classification based personal e-mail filtering." In: *Knowledge Discovery and Data Mining. Current Issues and New Applications: 4th Pacific-Asia Conference* (Kyoto, Japan). Ed. by Takao Terano, Huan Liu, and Arbee L. P. Chen. Springer Nature, pp. 408–419. URL: https://doi.org/10.1007/3-540-45571-x_48 (cit. on p. 59).

Doddington, George, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel (May 2004). "The automatic content extraction (ACE) program — tasks, data, and evaluation." In: *Fourth International Conference on Language Resources and Evaluation (LREC'04)* (Lisbon, Portugal). European Language Resources Association (ELRA), pp. 837–840. URL: https://aclanthology.org/L04-1011 (cit. on pp. 23, 97).

Doğan, Rezarta Islamaj, Sun Kim, Andrew Chatr-aryamontri, Chih-Hsuan Wei, Donald C. Comeau, Rui Antunes, Sérgio Matos, Qingyu Chen, Aparna Elangovan, Nagesh C. Panyam, Karin Verspoor, Hongfang Liu, Yanshan Wang, Zhuang Liu, Berna Altınel, Zehra Melce Hüsünbeyi, Arzucan Özgür, Aris Fergadis, Chen-Kai Wang, Hong-Jie Dai, Tung Tran, Ramakanth Kavuluru, Ling Luo, Albert Steppi, Jinfeng Zhang, Jinchan Qu, and Zhiyong Lu (Jan. 2019). "Overview of the BioCreative VI Precision Medicine Track: mining protein interactions and mutations for precision medicine." In: *Database* 2019 (bay147). Oxford University Press. URL: https://doi.org/10.1093/database/bay147 (cit. on pp. 7, 64, 69, 98).

Doğan, Rezarta Islamaj, Sun Kim, Andrew Chatr-Aryamontri, Chih-Hsuan Wei, Donald C. Comeau, and Zhiyong Lu (Oct. 2017). "Overview of the BioCreative VI Precision Medicine Track." In: *BioCreative VI Workshop* (Bethesda, Maryland, USA), pp. 83–87. URL: https://biocreative.bioinformatics.udel.edu/events/biocreative-vi/workshop/ (cit. on pp. 64, 98).

Doğan, Rezarta Islamaj, Robert Leaman, and Zhiyong Lu (Feb. 2014). "NCBI disease corpus: a resource for disease name recognition and concept normalization." In: *Journal of Biomedical Informatics* 47. Elsevier, pp. 1–10. URL: https://doi.org/10.1016/j.jbi.2013.12.006 (cit. on p. 42).

Doğan, Rezarta Islamaj and Zhiyong Lu (June 2012). "An improved corpus of disease mentions in PubMed citations." In: *BioNLP 2012: Workshop on Biomedical Natural Language Processing* (Montréal, Canada). Association for Computational Linguistics, pp. 91–99. URL: https://aclanthology.org/W12-2411 (cit. on p. 42).

Donaldson, Ian, Joel Martin, Berry de Bruijn, Cheryl Wolting, Vicki Lay, Brigitte Tuekam, Shudong Zhang, Berivan Baskin, Gary D. Bader, Katerina Michalickova, Tony Pawson, and Christopher WV Hogue (Mar. 2003). "PreBIND and Textomy – mining the biomedical literature for protein-protein interactions using a support vector machine." In: *BMC Bioinformatics* 4.1. BioMed Central Ltd, p. 11.
URL: https://doi.org/10.1186/1471-2105-4-11 (cit. on p. 61).

Duque, Andres, Mark Stevenson, Juan Martinez-Romo, and Lourdes Araujo (May 2018). "Co-occurrence graphs for word sense disambiguation in the biomedical domain." In: *Artificial Intelligence in Medicine* 87. Elsevier, pp. 9–19.
URL: https://doi.org/10.1016/j.artmed.2018.03.002 (cit. on pp. 40, 52).

Eberts, Markus and Adrian Ulges (Sept. 2019). *Span-based joint entity and relation extraction with transformer pre-training*. arXiv:1909.07755.
URL: https://arxiv.org/abs/1909.07755 (cit. on p. 24).

Eberts, Markus and Adrian Ulges (Apr. 2021). "An end-to-end model for entity-level relation extraction using multi-instance learning." In: *16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (Online). Association for Computational Linguistics, pp. 3650–3660.
URL: https://doi.org/10.18653/v1/2021.eacl-main.319 (cit. on p. 118).

Elhadad, Noémie, Sameer Pradhan, Sharon Gorman, Suresh Manandhar, Wendy Chapman, and Guergana Savova (June 2015). "SemEval-2015 Task 14: analysis of clinical text." In: *9th International Workshop on Semantic Evaluation (SemEval 2015)* (Denver, Colorado, USA). Association for Computational Linguistics, pp. 303–310.
URL: https://doi.org/10.18653/v1/S15-2051 (cit. on pp. 42, 54).

Ellis, Joe, Xuansong Li, Kira Griffitt, Stephanie M. Strassel, and Jonathan Wright (Nov. 2012). "Linguistic resources for 2012 knowledge base population evaluation." In: *Fifth Text Analysis Conference (TAC 2012)* (Gaithersburg, Maryland, USA).
URL: https://tac.nist.gov/publications/2012/papers.html (cit. on p. 97).

Etzioni, Oren (Aug. 2011). "Search needs a shake-up." In: *Nature* 476. Springer Nature, pp. 25–26.
URL: https://doi.org/10.1038/476025a (cit. on p. 28).

Fader, Anthony, Stephen Soderland, and Oren Etzioni (July 2011). "Identifying relations for open information extraction." In: *2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Edinburgh, Scotland, United Kingdom). Association for Computational Linguistics, pp. 1535–1545.
URL: https://dl.acm.org/citation.cfm?id=2145596 (cit. on p. 28).

Fader, Anthony, Luke Zettlemoyer, and Oren Etzioni (Aug. 2014). "Open question answering over curated and extracted knowledge bases." In: *20th ACM SIGKDD Inter-*

national Conference on Knowledge Discovery and Data Mining (New York, USA). ACM, pp. 1156–1165.
URL: https://doi.org/10.1145/2623330.2623677 (cit. on p. 28).

Fang, Ruihua, Gary Schindelman, Kimberly Van Auken, Jolene Fernandes, Wen Chen, Xiaodong Wang, Paul Davis, Mary Ann Tuli, Steven J. Marygold, Gillian Millburn, Beverley Matthews, Haiyan Zhang, Nick Brown, William M. Gelbart, and Paul W. Sternberg (Jan. 2012). "Automatic categorization of diverse experimental information in the bioscience literature." In: *BMC Bioinformatics* 13.1. BioMed Central Ltd, p. 16.
URL: https://doi.org/10.1186/1471-2105-13-16 (cit. on p. 63).

Farkas, Richárd and György Szarvas (Apr. 2008). "Automatic construction of rule-based ICD-9-CM coding systems." In: *BMC Bioinformatics* 9.3. Springer Nature, S10.
URL: https://doi.org/10.1186/1471-2105-9-S3-S10 (cit. on p. 62).

Fawcett, Tom (June 2006). "An introduction to ROC analysis." In: *Pattern Recognition Letters* 27.8. Elsevier, pp. 861–874.
URL: https://doi.org/10.1016/j.patrec.2005.10.010 (cit. on p. 29).

Federhen, Scott (Jan. 2012). "The NCBI Taxonomy database." In: *Nucleic Acids Research* 40.D1. Oxford University Press, pp. D136–D143.
URL: https://doi.org/10.1093/nar/gkr1178 (cit. on p. 98).

Feldman, Ronen (Apr. 2013). "Techniques and applications for sentiment analysis." In: *Communications of the ACM* 56.4. ACM, pp. 82–89.
URL: https://doi.org/10.1145/2436256.2436274 (cit. on p. 59).

Feldman, Ronen, Yizhar Regev, Eyal Hurvitz, and Michal Finkelstein-Landau (May 2003). "Mining the biomedical literature using semantic analysis and natural language processing techniques." In: *BIOSILICO* 1.2. Elsevier, pp. 69–80.
URL: https://doi.org/10.1016/s1478-5382(03)02330-8 (cit. on p. 60).

Fellbaum, Christiane (May 1998). *WordNet: an electronic lexical database*. The MIT Press.
URL: http://mitpress.mit.edu/books/wordnet (cit. on p. 39).

Fergadis, Aris, Christos Baziotis, Dimitris Pappas, Haris Papageorgiou, and Alexandros Potamianos (Oct. 2017). "Hierarchical bidirectional attention-based RNN in BioCreative VI Precision Medicine Track, document triage task." In: *BioCreative VI Workshop* (Bethesda, Maryland, USA), pp. 94–98.
URL: https://biocreative.bioinformatics.udel.edu/events/biocreative-vi/workshop/ (cit. on p. 69).

Fergadis, Aris, Christos Baziotis, Dimitris Pappas, Haris Papageorgiou, and Alexandros Potamianos (Aug. 2018). "Hierarchical bi-directional attention-based RNNs for supporting document classification on protein–protein interactions affected by genetic

mutations." In: *Database* 2018. Oxford University Press, bay076.
URL: https://doi.org/10.1093/database/bay076 (cit. on p. 19).

Ferrão, J. C., F. Janela, M. D. Oliveira, and H. M. G. Martins (Sept. 2013). "Using structured EHR data and SVM to support ICD-9-CM coding." In: *2013 IEEE International Conference on Healthcare Informatics* (Philadelphia, Pennsylvania, USA). IEEE, pp. 511–516.
URL: https://doi.org/10.1109/ICHI.2013.79 (cit. on p. 82).

Ferrão, José Carlos, Mónica Duarte Oliveira, Filipe Janela, and Henrique M. G. Martins (Dec. 2016). "Preprocessing structured clinical data for predictive modeling and decision support." In: *Applied Clinical Informatics* 07.04. Schattauer, pp. 1135–1153.
URL: https://doi.org/10.4338/ACI-2016-03-SOA-0035 (cit. on p. 82).

Ferreira, Liliana da Silva (July 2011). "Medical information extraction in European Portuguese." PhD thesis. University of Aveiro.
URL: http://hdl.handle.net/10773/7678 (cit. on p. 131).

Frijters, Raoul, Marianne van Vugt, Ruben Smeets, René van Schaik, Jacob de Vlieg, and Wynand Alkema (Sept. 2010). "Literature mining for the discovery of hidden connections between drugs, genes and diseases." In: *PLOS Computational Biology* 6.9. Public Library of Science, pp. 1–11.
URL: https://doi.org/10.1371%2Fjournal.pcbi.1000943 (cit. on p. 100).

Frunza, Oana and Diana Inkpen (July 2010). "Extraction of disease-treatment semantic relations from biomedical sentences." In: *2010 Workshop on Biomedical Natural Language Processing* (Uppsala, Sweden). Association for Computational Linguistics, pp. 91–98.
URL: https://aclanthology.org/W10-1912 (cit. on p. 98).

Gal, Yarin (Sept. 2016). "Uncertainty in deep learning." PhD thesis. University of Cambridge.
URL: https://idiscover.lib.cam.ac.uk/permalink/f/t9gok8/44CAM_ALMA21582084170003606 (cit. on p. 9).

Garla, Vijay N. and Cynthia Brandt (Sept. 2013). "Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification." In: *Journal of the American Medical Informatics Association* 20.5. Oxford University Press, p. 882.
URL: https://doi.org/10.1136/amiajnl-2012-001350 (cit. on pp. 40, 52, 53).

Gildea, Daniel and Daniel Jurafsky (Oct. 2000). "Automatic labeling of semantic roles." In: *38th Annual Meeting of the Association for Computational Linguistics* (Hong Kong). Association for Computational Linguistics, pp. 512–520.
URL: https://doi.org/10.3115/1075218.1075283 (cit. on p. 14).

Gildea, Daniel and Daniel Jurafsky (Sept. 2002). "Automatic labeling of semantic roles." In: *Computational Linguistics* 28.3. Association for Computational Linguistics, pp. 245–288.
URL: https://doi.org/10.1162/089120102760275983 (cit. on p. 14).

Gonzalez-Hernandez, Graciela, Martin Krallinger, Monica Muñoz, Raul Rodriguez-Esteban, Özlem Uzuner, and Lynette Hirschman (Sept. 2022). "Challenges and opportunities for mining adverse drug reactions: perspectives from pharma, regulatory agencies, healthcare providers and consumers." In: *Database* 2022. Oxford University Press, baac071.
URL: https://doi.org/10.1093/database/baac071 (cit. on p. 96).

Goodwin, Travis R., Dina Demner-Fushman, Kin Wah Fung, and Phong Do (Nov. 2019). "Overview of the TAC 2019 track on drug-drug interaction extraction from drug labels." In: *Text Analysis Conference (TAC 2019)* (Gaithersburg, Maryland, USA).
URL: https://tac.nist.gov/publications/2019/papers.html (cit. on p. 97).

Graves, Alex (Feb. 2012). "Long short-term memory." In: *Supervised sequence labelling with recurrent neural networks*. Springer Nature, pp. 37–45.
URL: https://doi.org/10.1007/978-3-642-24797-2_4 (cit. on p. 66).

Grishman, Ralph (July 1997). "Information extraction: techniques and challenges." In: *International Summer School on Information Extraction* (Frascati, Italy). Springer Nature, pp. 10–27.
URL: https://doi.org/10.1007/3-540-63438-X_2 (cit. on p. 1).

Grishman, Ralph (Sept. 2015). "Information extraction." In: *IEEE Intelligent Systems* 30.5. IEEE, pp. 8–15.
URL: https://doi.org/10.1109/mis.2015.68 (cit. on p. 1).

Grishman, Ralph (Nov. 2019). "Twenty-five years of information extraction." In: *Natural Language Engineering* 25.6. Cambridge University Press, pp. 677–692.
URL: https://doi.org/10.1017/S1351324919000512 (cit. on p. 1).

Grishman, Ralph and Beth Sundheim (Aug. 1996). "Message Understanding Conference - 6: a brief history." In: *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics* (Copenhagen, Denmark).
URL: https://aclanthology.org/C96-1079 (cit. on p. 97).

Gu, Jinghang, Fuqing Sun, Longhua Qian, and Guodong Zhou (Jan. 2017). "Chemical-induced disease relation extraction via convolutional neural network." In: *Database* 2017. Oxford University Press, bax024.
URL: https://doi.org/10.1093/database/bax024 (cit. on p. 101).

Gu, Yu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon (Oct. 2021). "Domain-specific language

model pretraining for biomedical natural language processing." In: *ACM Transactions on Computing for Healthcare* 3.1. ACM, 2:1–2:23.
URL: https://doi.org/10.1145/3458754 (cit. on pp. 33, 124, 130).

Gupta, Pankaj, Hinrich Schütze, and Bernt Andrassy (Dec. 2016). "Table filling multi-task recurrent neural network for joint entity and relation extraction." In: *COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (Osaka, Japan). The COLING 2016 Organizing Committee, pp. 2537–2547.
URL: https://aclanthology.org/C16-1239 (cit. on p. 24).

Gurulingappa, Harsha, Abdul Mateen-Rajpu, and Luca Toldo (Dec. 2012a). "Extraction of potential adverse drug events from medical case reports." In: *Journal of Biomedical Semantics* 3.15. BioMed Central Ltd.
URL: https://doi.org/10.1186/2041-1480-3-15 (cit. on p. 104).

Gurulingappa, Harsha, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo (Oct. 2012b). "Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports." In: *Journal of Biomedical Informatics* 45.5. Elsevier, pp. 885–892.
URL: https://doi.org/10.1016/j.jbi.2012.04.008 (cit. on pp. 25, 104).

Habibi, Maryam, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser (July 2017). "Deep learning with word embeddings improves biomedical named entity recognition." In: *Bioinformatics* 33.14. Oxford University Press, pp. i37–i48.
URL: https://doi.org/10.1093/bioinformatics/btx228 (cit. on p. 101).

Hearst, Marti A. (June 1999). "Untangling text data mining." In: *37th Annual Meeting of the Association for Computational Linguistics* (College Park, Maryland, USA). Association for Computational Linguistics, pp. 3–10.
URL: https://doi.org/10.3115/1034678.1034679 (cit. on pp. 1, 2, 59).

Henry, Sam, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner (Jan. 2021). "2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records." In: *Journal of the American Medical Informatics Association* 27.1. Oxford University Press, pp. 3–12.
URL: https://doi.org/10.1093/jamia/ocz166 (cit. on pp. 98, 105).

Herrero-Zazo, María, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck (Oct. 2013). "The DDI corpus: an annotated corpus with pharmacological substances and drug–drug interactions." In: *Journal of Biomedical Informatics* 46.5. Elsevier, pp. 914–920.
URL: https://doi.org/10.1016/j.jbi.2013.07.011 (cit. on p. 104).

Hersh, William R. (June 2005). "Report on the TREC 2004 Genomics Track." In: *ACM SIGIR Forum* 39.1. ACM, pp. 21–24.
URL: https://doi.org/10.1145/1067268.1067273 (cit. on p. 61).

Hersh, William R., Ravi Teja Bhuptiraju, Laura Ross, Phoebe Johnson, Aaron M. Cohen, and Dale F. Kraemer (Nov. 2004). "TREC 2004 Genomics Track overview." In: *The Thirteenth Text Retrieval Conference (TREC 2004)* (Gaithersburg, Maryland, USA).
URL: https://trec.nist.gov/pubs/trec13/t13_proceedings.html (cit. on p. 61).

Hinton, Geoffrey E. (Sept. 1992). "How neural networks learn from experience." In: *Scientific American* 267.3. Scientific American, a division of Nature America, Inc., pp. 144–151.
URL: https://www.jstor.org/stable/24939221 (cit. on p. 27).

Hirschman, Lynette, Jong C. Park, Junichi Tsujii, Limsoon Wong, and Cathy H. Wu (Dec. 2002). "Accomplishments and challenges in literature data mining for biology." In: *Bioinformatics* 18.12. Oxford University Press, pp. 1553–1561.
URL: https://doi.org/10.1093/bioinformatics/18.12.1553 (cit. on p. 103).

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). "Long short-term memory." In: *Neural Computation* 9.8. The MIT Press, pp. 1735–1780.
URL: https://doi.org/10.1162/neco.1997.9.8.1735 (cit. on p. 66).

Hotho, Andreas, A. Nürnberger, and Gerhard Paaß (2005). "A brief survey of text mining." In: *LDV Forum* 20. German Society for Computational Linguistics and Language Technology, pp. 19–62.
URL: https://jlcl.org/content/2-allissues/24-Heft1-2005/19-62_HothoNuernbergerPaass.pdf (cit. on p. 1).

Howe, Doug, Maria Costanzo, Petra Fey, Takashi Gojobori, Linda Hannick, Winston Hide, David P. Hill, Renate Kania, Mary Schaeffer, Susan St Pierre, Simon Twigger, Owen White, and Seung Yon Rhee (Sept. 2008). "The future of biocuration." In: *Nature* 455. Springer Nature, pp. 47–50.
URL: https://doi.org/10.1038/455047a (cit. on pp. 41, 99).

Huang, Chung-Chi and Zhiyong Lu (Jan. 2016). "Community challenges in biomedical text mining over 10 years: success, failure and the future." In: *Briefings in Bioinformatics* 17.1. Oxford University Press, pp. 132–144.
URL: https://doi.org/10.1093/bib/bbv024 (cit. on pp. 35, 62, 103).

Huang, Degen, Zhenchao Jiang, Li Zou, and Lishuang Li (Nov. 2017). "Drug–drug interaction extraction from biomedical literature using support vector machine and long short term memory networks." In: *Information Sciences* 415–416. Elsevier, pp. 100–109.
URL: https://doi.org/10.1016/j.ins.2017.06.021 (cit. on p. 102).

Huang, Kexin, Jaan Altosaar, and Rajesh Ranganath (Oct. 2019). *ClinicalBERT: modeling clinical notes and predicting hospital readmission*. arXiv:1904.05342.
URL: http://arxiv.org/abs/1904.05342 (cit. on p. 131).

Ide, Nancy and Jean Véronis (Mar. 1998). "Introduction to the special issue on word sense disambiguation: the state of the art." In: *Computational Linguistics* 24.1. Association for Computational Linguistics, pp. 1–40.
URL: https://aclanthology.org/J98-1001 (cit. on pp. 37, 39).

Indurkhya, Nitin and Fred J. Damerau (2010). *Handbook of natural language processing*. 2nd ed. Chapman & Hall/CRC.
URL: https://dl.acm.org/citation.cfm?id=1738958 (cit. on p. 1).

Ingersoll, Grant S., Thomas S. Morton, and Andrew L. Farris (2013). *Taming text: how to find, organize, and manipulate it*. Manning Publications Co.
URL: https://www.manning.com/books/taming-text (cit. on p. 11).

Islamaj, Rezarta, Robert Leaman, David Cissel, Meng Cheng, Cathleen Coss, Joseph Denicola, Carol Fisher, Rob Guzman, Preeti Gokal Kochar, Nicholas Miliaras, Zoe Punske, Keiko Sekiya, Dorothy Trinh, Deborah Whitman, Susan Schmidt, and Zhiyong Lu (Nov. 2021a). "The chemical corpus of the NLM-Chem BioCreative VII track: full-text chemical identification and indexing in PubMed articles." In: *BioCreative VII Challenge Evaluation Workshop* (Online), pp. 114–118.
URL: https://biocreative.bioinformatics.udel.edu/resources/publications/bc-vii-workshop-proceedings/ (cit. on p. 42).

Islamaj, Rezarta, Robert Leaman, Sun Kim, Dongseop Kwon, Chih-Hsuan Wei, Donald C. Comeau, Yifan Peng, David Cissel, Cathleen Coss, Carol Fisher, Rob Guzman, Preeti Gokal Kochar, Stella Koppel, Dorothy Trinh, Keiko Sekiya, Janice Ward, Deborah Whitman, Susan Schmidt, and Zhiyong Lu (Mar. 2021b). "NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature." In: *Scientific Data* 8.91. Springer Nature.
URL: https://doi.org/10.1038/s41597-021-00875-1 (cit. on pp. 38, 42).

Ji, Heng and Ralph Grishman (June 2011). "Knowledge base population: successful approaches and challenges." In: *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Portland, Oregon, USA). Association for Computational Linguistics, pp. 1148–1158.
URL: https://aclanthology.org/P11-1115 (cit. on p. 97).

Ji, Heng, Ralph Grishman, and Hoa Trang Dang (Nov. 2011). "Overview of the TAC 2011 knowledge base population track." In: *Fourth Text Analysis Conference (TAC 2011)* (Gaithersburg, Maryland, USA).
URL: https://tac.nist.gov/publications/2011/papers.html (cit. on p. 97).

Ji, Heng, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis (Nov. 2010). "Overview of the TAC 2010 knowledge base population track." In: *Third Text Analysis Conference (TAC 2010)* (Gaithersburg, Maryland, USA).
URL: https://tac.nist.gov//publications/2010/papers.html (cit. on p. 97).

Jimeno-Yepes, Antonio (Sept. 2017). "Word embeddings and recurrent neural networks based on long-short term memory nodes in supervised biomedical word sense disambiguation." In: *Journal of Biomedical Informatics* 73 (Suppl C). Elsevier, pp. 137–147.
URL: https://doi.org/10.1016/j.jbi.2017.08.001 (cit. on pp. 40, 52, 53, 100).

Jimeno-Yepes, Antonio and Rafael Berlanga (Feb. 2015). "Knowledge based word-concept model estimation and refinement for biomedical text mining." In: *Journal of Biomedical Informatics* 53. Elsevier, pp. 300–307.
URL: https://doi.org/10.1016/j.jbi.2014.11.015 (cit. on pp. 52, 53).

Jimeno-Yepes, Antonio, Bridget T. McInnes, and Alan R. Aronson (June 2011). "Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation." In: *BMC Bioinformatics* 12.1. BioMed Central Ltd, p. 223.
URL: https://doi.org/10.1186/1471-2105-12-223 (cit. on pp. 41, 43, 52, 127).

Jimeno-Yepes, Antonio J. and Alan R. Aronson (Nov. 2010). "Knowledge-based biomedical word sense disambiguation: comparison of approaches." In: *BMC Bioinformatics* 11.1. BioMed Central Ltd, p. 569.
URL: https://doi.org/10.1186/1471-2105-11-569 (cit. on p. 40).

Joachims, Thorsten (Apr. 1998). "Text categorization with Support Vector Machines: learning with many relevant features." In: *Machine Learning: ECML-98, 10th European Conference on Machine Learning* (Chemnitz, Germany). Springer Nature, pp. 137–142.
URL: https://doi.org/10.1007/bfb0026683 (cit. on p. 60).

Johnson, Alistair E. W., Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark (May 2016). "MIMIC-III, a freely accessible critical care database." In: *Scientific Data* 3.160035. Springer Nature.
URL: https://doi.org/10.1038/sdata.2016.35 (cit. on pp. 34, 56, 72, 111).

Jurafsky, Daniel and James H. Martin (2008). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition.* 2nd ed. Prentice Hall.
URL: https://home.cs.colorado.edu/~martin/slp.html (cit. on p. 1).

Jurafsky, Daniel and James H. Martin (2018). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition.*

3rd draft.
URL: http://web.stanford.edu/~jurafsky/slp3/ (cit. on pp. 11, 14).

Kadhim, Ammar Ismael (June 2019). "Survey on supervised machine learning techniques for automatic text classification." In: *Artificial Intelligence Review* 52.1. Springer Nature, pp. 273–292.
URL: https://doi.org/10.1007/s10462-018-09677-1 (cit. on p. 66).

Kamath, Cannannore Nidhi, Syed Saqib Bukhari, and Andreas Dengel (Aug. 2018). "Comparative study between traditional machine learning and deep learning approaches for text classification." In: *ACM Symposium on Document Engineering 2018* (Halifax, Nova Scotia, Canada). ACM, pp. 1–11.
URL: https://doi.org/10.1145/3209280.3209526 (cit. on p. 66).

Karp, Peter D. (Dec. 2016). "Can we replace curation with information extraction software?" In: *Database* 2016 (baw150). Oxford University Press.
URL: https://doi.org/10.1093/database/baw150 (cit. on pp. 41, 63, 99).

Katiyar, Arzoo and Claire Cardie (July 2017). "Going out on a limb: joint extraction of entity mentions and relations without dependency trees." In: *55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vancouver, British Columbia, Canada). Association for Computational Linguistics, pp. 917–928.
URL: https://doi.org/10.18653/v1/P17-1085 (cit. on p. 23).

Kilgarriff, Adam (Mar. 1997). ""I don't believe in word senses"." In: *Computers and the Humanities* 31.2. Springer Nature, pp. 91–113.
URL: https://doi.org/10.1023/A:1000583911091 (cit. on p. 39).

Kilgarriff, Adam (2007). "Word senses." In: *Word sense disambiguation: algorithms and applications*. Vol. 33. Springer Nature, pp. 29–46.
URL: https://doi.org/10.1007/978-1-4020-4809-8_2 (cit. on p. 39).

Kim, Donghyeon, Jinhyuk Lee, Chan Ho So, Hwisang Jeon, Minbyul Jeong, Yonghwa Choi, Wonjin Yoon, Mujeen Sung, and Jaewoo Kang (June 2019). "A neural named entity recognition and multi-type normalization tool for biomedical text mining." In: *IEEE Access* 7. IEEE, pp. 73729–73740.
URL: https://doi.org/10.1109/access.2019.2920708 (cit. on p. 96).

Kim, J.-D., T. Ohta, Y. Tateisi, and J. Tsujii (July 2003). "GENIA corpus—a semantically annotated corpus for bio-textmining." In: *Bioinformatics* 19 (Suppl 1). Oxford University Press, p. i180.
URL: https://doi.org/10.1093/bioinformatics/btg1023 (cit. on pp. 25, 98).

Kim, Jin-Dong, Jung-jae Kim, Xu Han, and Dietrich Rebholz-Schuhmann (June 2015). "Extending the evaluation of Genia Event task toward knowledge base construction and comparison to Gene Regulation Ontology task." In: *BMC Bioinformatics* 16.10.

BioMed Central Ltd, S3.
URL: https://doi.org/10.1186/1471-2105-16-S10-S3 (cit. on p. 98).

Kim, Jin-Dong, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii (June 2009). "Overview of BioNLP'09 shared task on event extraction." In: *Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task* (Boulder, Colorado, USA). Association for Computational Linguistics, pp. 1–9.
URL: https://dl.acm.org/citation.cfm?id=1572340.1572342 (cit. on p. 98).

Kim, Jin-Dong, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii (Nov. 2011a). "Extracting bio-molecular events from literature—the BioNLP'09 shared task." In: *Computational Intelligence* 27.4. Wiley, pp. 513–540.
URL: https://doi.org/10.1111/j.1467-8640.2011.00398.x (cit. on p. 98).

Kim, Jin-Dong, Tomoko Ohta, and Jun'ichi Tsujii (Jan. 2008). "Corpus annotation for mining biomedical events from literature." In: *BMC Bioinformatics* 9.1. BioMed Central Ltd, p. 10.
URL: https://doi.org/10.1186/1471-2105-9-10 (cit. on p. 98).

Kim, Jin-Dong, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii (June 2011b). "Overview of BioNLP Shared Task 2011." In: *BioNLP Shared Task 2011 Workshop* (Portland, Oregon, USA). Association for Computational Linguistics, pp. 1–6.
URL: https://aclanthology.org/W11-1801 (cit. on p. 99).

Kim, Jin-Dong, Yue Wang, Nicola Colic, Seung Han Beak, Yong Hwan Kim, and Min Song (Aug. 2016). "Refactoring the Genia Event Extraction Shared Task toward a general framework for IE-driven KB development." In: *4th BioNLP Shared Task Workshop* (Berlin, Germany). Association for Computational Linguistics, pp. 23–31.
URL: https://doi.org/10.18653/v1/W16-3003 (cit. on p. 98).

Kim, Jin-Dong, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa (June 2011c). "Overview of Genia event task in BioNLP Shared Task 2011." In: *BioNLP Shared Task 2011 Workshop* (Portland, Oregon, USA). Association for Computational Linguistics, pp. 7–15.
URL: https://aclanthology.org/W11-1802 (cit. on p. 98).

Kim, Jin-Dong, Yue Wang, and Yamamoto Yasunori (Aug. 2013a). "The Genia Event Extraction Shared Task, 2013 edition - overview." In: *BioNLP Shared Task 2013 Workshop* (Sofia, Bulgaria). Association for Computational Linguistics, pp. 8–15.
URL: https://aclanthology.org/W13-2002 (cit. on p. 98).

Kim, Jung-jae, Xu Han, Vivian Lee, and Dietrich Rebholz-Schuhmann (Aug. 2013b). "GRO task: populating the Gene Regulation Ontology with events and relations." In: *BioNLP Shared Task 2013 Workshop* (Sofia, Bulgaria). Association for Computational Linguis-

tics, pp. 50–57.
URL: https://aclanthology.org/W13-2007 (cit. on p. 99).

Kim, Sun and W. John Wilbur (Sept. 2010). "Improving protein-protein interaction article classification performance by utilizing grammatical relations." In: *BioCreative III Workshop* (Bethesda, Maryland, USA), pp. 77–82.
URL: https://biocreative.bioinformatics.udel.edu/resources/publications/bc-iii-workshop-proceedings/ (cit. on p. 64).

Kim, Sun and W. John Wilbur (Oct. 2011). "Classifying protein-protein interaction articles using word and syntactic features." In: *BMC Bioinformatics* 12.8. BioMed Central Ltd, S9.
URL: https://doi.org/10.1186/1471-2105-12-S8-S9 (cit. on p. 64).

Kim, Yoon (Oct. 2014). "Convolutional neural networks for sentence classification." In: *2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha, Qatar). Association for Computational Linguistics, pp. 1746–1751.
URL: https://doi.org/10.3115/v1/D14-1181 (cit. on p. 101).

Kowsari, K., D. E. Brown, M. Heidarysafa, K. Jafari Meimandi, M. S. Gerber, and L. E. Barnes (Dec. 2017). "HDLTex: hierarchical deep learning for text classification." In: *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)* (Cancun, Mexico). IEEE, pp. 364–371.
URL: https://doi.org/10.1109/icmla.2017.0-134 (cit. on p. 101).

Kowsari, Kamran, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown (Apr. 2019). "Text classification algorithms: a survey." In: *Information* 10.4. MDPI, p. 150.
URL: https://doi.org/10.3390/info10040150 (cit. on p. 63).

Krallinger, Martin, Ramon Alonso-Allende Erhardt, and Alfonso Valencia (Mar. 2005). "Text-mining approaches in molecular biology and biomedicine." In: *Drug Discovery Today* 10.6. Elsevier, pp. 439–445.
URL: https://doi.org/10.1016/S1359-6446(05)03376-3 (cit. on pp. 32, 103, 125).

Krallinger, Martin, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, and Alfonso Valencia (Jan. 2015). "CHEMDNER: the drugs and chemical names extraction challenge." In: *Journal of Cheminformatics* 7.1. BioMed Central Ltd, S1.
URL: https://doi.org/10.1186/1758-2946-7-S1-S1 (cit. on p. 35).

Krallinger, Martin, Florian Leitner, Carlos Rodriguez-Penagos, and Alfonso Valencia (Sept. 2008). "Overview of the protein-protein interaction annotation extraction task of BioCreative II." In: *Genome Biology* 9.2. Springer Nature, S4.
URL: https://doi.org/10.1186/gb-2008-9-s2-s4 (cit. on pp. 97, 104).

Krallinger, Martin, Florian Leitner, and Alfonso Valencia (Apr. 2007). "Assessment of the Second BioCreative PPI task: automatic extraction of protein-protein interactions." In: *Second BioCreative Challenge Evaluation Workshop* (Madrid, Spain), pp. 41–54.
URL: https://biocreative.bioinformatics.udel.edu/resources/publications/ (cit. on pp. 97, 104).

Krallinger, Martin, Obdulia Rabal, Saber A. Akhondi, Martín Pérez Pérez, Jesús Santamaría, Gael Pérez Rodríguez, Georgios Tsatsaronis, Ander Intxaurrondo, José Antonio López, Umesh Nandal, Erin Van Buel, Akileshwari Chandrasekhar, Marleen Rodenburg, Astrid Laegreid, Marius Doornenbal, Julen Oyarzabal, Anália Lourenço, and Alfonso Valencia (Oct. 2017a). "Overview of the BioCreative VI chemical–protein interaction track." In: *BioCreative VI Workshop* (Bethesda, Maryland, USA), pp. 141–146.
URL: https://biocreative.bioinformatics.udel.edu/events/biocreative-vi/workshop/ (cit. on pp. 98, 101, 104, 106, 107, 109, 117, 119).

Krallinger, Martin, Obdulia Rabal, Anália Lourenço, Julen Oyarzabal, and Alfonso Valencia (May 2017b). "Information retrieval and text mining technologies for chemistry." In: *Chemical Reviews* 117.12. American Chemical Society, pp. 7673–7761.
URL: https://doi.org/10.1021/acs.chemrev.6b00851 (cit. on pp. 100, 106, 125).

Krallinger, Martin, Miguel Vazquez, Florian Leitner, David Salgado, Andrew Chatr-aryamontri, Andrew Winter, Livia Perfetto, Leonardo Briganti, Luana Licata, Marta Iannuccelli, Luisa Castagnoli, Gianni Cesareni, Mike Tyers, Gerold Schneider, Fabio Rinaldi, Robert Leaman, Graciela Gonzalez, Sérgio Matos, Sun Kim, W. John Wilbur, Luis Rocha, Hagit Shatkay, Ashish V. Tendulkar, Shashank Agarwal, Feifan Liu, Xinglong Wang, Rafal Rak, Keith Noto, Charles Elkan, Zhiyong Lu, Rezarta Islamaj Dogan, Jean-Fred Fontaine, Miguel A. Andrade-Navarro, and Alfonso Valencia (Oct. 2011). "The protein–protein interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text." In: *BMC Bioinformatics* 12.8. BioMed Central Ltd, S3.
URL: https://doi.org/10.1186/1471-2105-12-S8-S3 (cit. on pp. 35, 62, 64, 65, 100, 106, 129).

Krallinger, Martin, Miguel Vazquez, Florian Leitner, David Salgado, and Alfonso Valencia (Sept. 2010). "Results of the BioCreative III (Interaction) Article Classification Task." In: *BioCreative III Workshop* (Bethesda, Maryland, USA), pp. 13–19.
URL: https://biocreative.bioinformatics.udel.edu/resources/publications/bc-iii-workshop-proceedings/ (cit. on pp. 62, 64, 65).

Krauthammer, Michael and Goran Nenadic (Dec. 2004). "Term identification in the biomedical literature." In: *Journal of Biomedical Informatics* 37.6. Elsevier, pp. 512–

526.
URL: https://doi.org/10.1016/j.jbi.2004.08.004 (cit. on p. 40).

Krovetz, R. and W. B. Croft (May 1989). "Word sense disambiguation using machine-readable dictionaries." In: *12th Annual International Conference on Research Development in Information Retrieval* (Cambridge, Massachusetts, USA). ACM, pp. 127–136.
URL: https://doi.org/10.1145/75334.75349 (cit. on p. 39).

Krovetz, Robert (July 1997). "Homonymy and polysemy in information retrieval." In: *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics* (Madrid, Spain). Association for Computational Linguistics, pp. 72–79.
URL: https://doi.org/10.3115/976909.979627 (cit. on p. 37).

Kudo, Taku and Yuji Matsumoto (June 2001). "Chunking with support vector machines." In: *Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies* (Stroudsburg, Pennsylvania, USA). Association for Computational Linguistics, pp. 1–8.
URL: https://doi.org/10.3115/1073336.1073361 (cit. on p. 20).

Kulkarni, Chaitanya, Wei Xu, Alan Ritter, and Raghu Machiraju (June 2018). "An annotated corpus for machine reading of instructions in wet lab protocols." In: *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (New Orleans, Louisiana, USA). Association for Computational Linguistics, pp. 97–106.
URL: https://doi.org/10.18653/v1/N18-2016 (cit. on p. 25).

Lafferty, John D., Andrew McCallum, and Fernando C. N. Pereira (June 2001). "Conditional random fields: probabilistic models for segmenting and labeling sequence data." In: *Eighteenth International Conference on Machine Learning* (Williamstown, Massachusetts, USA). Morgan Kaufmann Publishers Inc., pp. 282–289.
URL: http://dl.acm.org/citation.cfm?id=645530.655813 (cit. on p. 26).

Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer (Mar. 2016). *Neural architectures for named entity recognition.* arXiv:1603.01360.
URL: https://arxiv.org/abs/1603.01360 (cit. on pp. 20, 26).

Lamurias, Andre, Luka A. Clarke, and Francisco M. Couto (Mar. 2017). "Extracting microRNA–gene relations from biomedical literature using distant supervision." In: *PLOS ONE* 12.3. Public Library of Science, pp. 1–20.
URL: https://doi.org/10.1371/journal.pone.0171929 (cit. on p. 118).

Leaman, Robert, Rezarta Islamaj Doğan, and Zhiyong Lu (Nov. 2013). "DNorm: disease name normalization with pairwise learning to rank." In: *Bioinformatics* 29.22. Oxford

University Press, pp. 2909–2917.
URL: https://doi.org/10.1093/bioinformatics/btt474 (cit. on p. 42).

Leaman, Robert, Rezarta Islamaj, Virginia Adams, Mohammed A Alliheedi, João Rafael Almeida, Rui Antunes, Robert Bevan, Yung-Chun Chang, Arslan Erdengasileng, Matthew Hodgskiss, Ryuki Ida, Hyunjae Kim, Keqiao Li, Robert E Mercer, Lukrécia Mertová, Ghadeer Mobasher, Hoo-Chang Shin, Mujeen Sung, Tomoki Tsujimura, Wen-Chao Yeh, and Zhiyong Lu (Mar. 2023). "Chemical identification and indexing in full-text articles: an overview of the NLM-Chem track at BioCreative VII." In: *Database* 2023. Oxford University Press, baad005.
URL: https://doi.org/10.1093/database/baad005 (cit. on p. 8).

Leaman, Robert and Zhiyong Lu (Sept. 2016). "TaggerOne: joint named entity recognition and normalization with semi-Markov Models." In: *Bioinformatics* 32.18. Oxford University Press, pp. 2839–2846.
URL: https://doi.org/10.1093/bioinformatics/btw343 (cit. on p. 38).

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (May 2015). "Deep learning." In: *Nature* 521.7553. Springer Nature, pp. 436–444.
URL: https://doi.org/10.1038/nature14539 (cit. on p. 100).

Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang (Feb. 2020). "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." In: *Bioinformatics* 36.4. Oxford University Press, pp. 1234–1240.
URL: https://doi.org/10.1093/bioinformatics/btz682 (cit. on pp. 33, 63, 85, 130).

Lesk, Michael (1986). "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone." In: *5th Annual International Conference on Systems Documentation* (Toronto, Ontario, Canada). ACM, pp. 24–26.
URL: https://doi.org/10.1145/318723.318728 (cit. on p. 39).

Lewis, David D. (Feb. 1992). "Feature selection and feature extraction for text categorization." In: *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992* (Harriman, New York). Morgan Kaufmann Publishers.
URL: https://aclanthology.org/H92-1041 (cit. on p. 60).

Lewis, David D. (July 1995). "Evaluating and optimizing autonomous text classification systems." In: *18th annual international ACM SIGIR conference on Research and development in information retrieval* (Seattle, Washington, USA). ACM, pp. 246–254.
URL: https://doi.org/10.1145/215206.215366 (cit. on p. 60).

Li, Fei, Meishan Zhang, Guohong Fu, and Donghong Ji (Mar. 2017). "A neural joint model for entity and relation extraction from biomedical text." In: *BMC Bioinformatics* 18.1.

BioMed Central Ltd, p. 198.
URL: https://doi.org/10.1186/s12859-017-1609-9 (cit. on p. 24).

Li, Jiao, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu (Sept. 2015). "Annotating chemicals, diseases and their interactions in biomedical literature." In: *BioCreative V Workshop* (Sevilla, Spain), pp. 173–182.
URL: https://biocreative.bioinformatics.udel.edu/resources/publications/bcv_proceedings/ (cit. on p. 42).

Li, Jiao, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu (May 2016). "BioCreative V CDR task corpus: a resource for chemical disease relation extraction." In: *Database* 2016. Oxford University Press, baw068.
URL: https://doi.org/10.1093/database/baw068 (cit. on pp. 42, 104).

Li, Qi and Heng Ji (June 2014). "Incremental joint extraction of entity mentions and relations." In: *52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Baltimore, Maryland, USA). Association for Computational Linguistics, pp. 402–412.
URL: https://doi.org/10.3115/v1/P14-1038 (cit. on pp. 22, 23).

Li, Zhi, Fan Yang, and Yaoru Luo (May 2019). "Context embedding based on bi-LSTM in semi-supervised biomedical word sense disambiguation." In: *IEEE Access* 7. IEEE, pp. 72928–72935.
URL: https://doi.org/10.1109/access.2019.2912584 (cit. on pp. 40, 53).

Lim, Sangrak and Jaewoo Kang (Oct. 2017). "Chemical-gene relation extraction using recursive neural network." In: *BioCreative VI Workshop* (Bethesda, Maryland, USA), pp. 190–193.
URL: https://biocreative.bioinformatics.udel.edu/events/biocreative-vi/workshop/ (cit. on pp. 102, 119).

Lim, Sangrak and Jaewoo Kang (Jan. 2018). "Chemical–gene relation extraction using recursive neural network." In: *Database* 2018. Oxford University Press, bay060.
URL: https://doi.org/10.1093/database/bay060 (cit. on pp. 102, 119).

Lipscomb, Carolyn E. (July 2000). "Medical subject headings (MeSH)." In: *Bulletin of the Medical Library Association* 88.3. Medical Library Association, pp. 265–266.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC35238/ (cit. on pp. 34, 40).

Liu, Kang (Oct. 2020). "A survey on neural relation extraction." In: *Science China Technological Sciences* 63.10. Springer Nature, pp. 1971–1989.
URL: https://doi.org/10.1007/s11431-020-1673-6 (cit. on p. 95).

Liu, S., Wei Ma, R. Moore, V. Ganesan, and S. Nelson (Oct. 2005). "RxNorm: prescription for electronic drug information exchange." In: *IT Professional* 7.5. IEEE, pp. 17–23.
URL: https://doi.org/10.1109/mitp.2005.122 (cit. on p. 42).

Liu, Shuhua Monica and Jiun-Hung Chen (Feb. 2015). "A multi-label classification based approach for sentiment classification." In: *Expert Systems with Applications* 42.3. Elsevier, pp. 1083–1093.
URL: https://doi.org/10.1016/j.eswa.2014.08.036 (cit. on p. 59).

Liu, Sijia, Feichen Shen, Ravikumar Komandur Elayavilli, Yanshan Wang, Majid Rastegar-Mojarad, Vipin Chaudhary, and Hongfang Liu (Oct. 2018a). "Extracting chemical–protein relations using attention-based neural networks." In: *Database* 2018. Oxford University Press, bay102.
URL: https://doi.org/10.1093/database/bay102 (cit. on pp. 102, 119).

Liu, Sijia, Feichen Shen, Yanshan Wang, Majid Rastegar-Mojarad, Ravikumar Komandur Elayavilli, Vipin Chaundary, and Hongfang Liu (Oct. 2017). "Attention-based neural networks for chemical protein relation extraction." In: *BioCreative VI Workshop* (Bethesda, Maryland, USA), pp. 155–158.
URL: https://biocreative.bioinformatics.udel.edu/events/biocreative-vi/workshop/ (cit. on pp. 102, 119).

Liu, Tianlin, João Sedoc, and Lyle Ungar (Aug. 2018b). "Correcting the common discourse bias in linear representation of sentences using conceptors." In: *BioCreative/OHNLP Challenge 2018* (Washington, DC, USA).
URL: https://sites.google.com/view/ohnlp2018 (cit. on p. 84).

Loper, Edward and Steven Bird (May 2002). *NLTK: the natural language toolkit.* arXiv:cs/0205028.
URL: https://arxiv.org/abs/cs/0205028 (cit. on p. 33).

Loureiro, Daniel and Alípio Jorge (July 2019). "Language modelling makes sense: propagating representations through WordNet for full-coverage word sense disambiguation." In: *57th Annual Meeting of the Association for Computational Linguistics* (Florence, Italy). Association for Computational Linguistics, pp. 5682–5691.
URL: https://doi.org/10.18653/v1/P19-1569 (cit. on p. 39).

Luan, Yi, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi (Aug. 2018). *Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction.* arXiv:1808.09602.
URL: https://arxiv.org/abs/1808.09602 (cit. on p. 25).

Luan, Yi, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi (June 2019). "A general framework for information extraction using dynamic span graphs." In: *2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (Minneapolis, Minnesota, USA). Association for Computational Linguistics, pp. 3036–3046.
URL: https://doi.org/10.18653/v1/N19-1308 (cit. on p. 25).

Ludvigsson, Jonas F., Jyotishman Pathak, Sean Murphy, Matthew Durski, Phillip S. Kirsch, Christophe G. Chute, Euijung Ryu, and Joseph A. Murray (Dec. 2013). "Use of computerized algorithm to identify individuals in need of testing for celiac disease." In: *Journal of the American Medical Informatics Association* 20.e2. Oxford University Press, e306–e310.
URL: https://doi.org/10.1136/amiajnl-2013-001924 (cit. on p. 70).

Lukashenko, Romans, Vita Graudina, and Janis Grundspenkis (June 2007). "Computer-based plagiarism detection methods and tools: an overview." In: *2007 International Conference on Computer Systems and Technologies* (Bulgaria). ACM, pp. 1–6.
URL: https://doi.org/10.1145/1330598.1330642 (cit. on p. 60).

Lung, Pei-Yau, Zhe He, Tingting Zhao, Disa Yu, and Jinfeng Zhang (Jan. 2019). "Extracting chemical–protein interactions from literature using sentence structure analysis and feature engineering." In: *Database* 2019. Oxford University Press, bay138.
URL: https://doi.org/10.1093/database/bay138 (cit. on pp. 102, 108, 119).

Lung, Pei-Yau, Tingting Zhao, Zhe He, and Jinfeng Zhang (Oct. 2017). "Extracting chemical–protein interactions from literature." In: *BioCreative VI Workshop* (Bethesda, Maryland, USA), pp. 159–162.
URL: https://biocreative.bioinformatics.udel.edu/events/biocreative-vi/workshop/ (cit. on pp. 102, 108, 119).

Luo, Ling, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N. Arighi, and Zhiyong Lu (July 2022). "BioRED: a rich biomedical relation extraction dataset." In: *Briefings in Bioinformatics*. Oxford University Press, bbac282.
URL: https://doi.org/10.1093/bib/bbac282 (cit. on p. 105).

Luo, Ling, Zhihao Yang, Mingyu Cao, Lei Wang, Yin Zhang, and Hongfei Lin (Mar. 2020a). "A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature." In: *Journal of Biomedical Informatics* 103.103384. Elsevier.
URL: https://doi.org/10.1016/j.jbi.2020.103384 (cit. on pp. 25, 132).

Luo, Ling, Zhihao Yang, Hongfei Lin, and Jian Wang (Oct. 2017). "DUTIR at the BioCreative VI Precision Medicine Track: document triage for identifying PPIs affected by genetic mutations." In: *BioCreative VI Workshop* (Bethesda, Maryland, USA), pp. 119–122.

URL: https://biocreative.bioinformatics.udel.edu/events/biocreative-vi/workshop/ (cit. on p. 69).

Luo, Yen-Fu, Sam Henry, Yanshan Wang, Feichen Shen, Ozlem Uzuner, and Anna Rumshisky (Oct. 2020b). "The 2019 n2c2/UMass Lowell shared task on clinical concept normalization." In: *Journal of the American Medical Informatics Association* 27.10. Oxford University Press, 1529–e1.
URL: https://doi.org/10.1093/jamia/ocaa106 (cit. on pp. 42, 54, 56, 57).

Luo, Yen-Fu, Weiyi Sun, and Anna Rumshisky (Apr. 2019). "MCN: a comprehensive corpus for medical concept normalization." In: *Journal of Biomedical Informatics* 92. Elsevier, p. 103132.
URL: https://doi.org/10.1016/j.jbi.2019.103132 (cit. on pp. 42, 43, 54, 56, 88, 129).

Lyu, Chen, Bo Chen, Yafeng Ren, and Donghong Ji (Oct. 2017). "Long short-term memory RNN for biomedical named entity recognition." In: *BMC Bioinformatics* 18.1. BioMed Central Ltd, p. 462.
URL: https://doi.org/10.1186/s12859-017-1868-5 (cit. on p. 101).

Manning, Christopher D. (Dec. 2015). "Computational linguistics and deep learning." In: *Computational Linguistics* 41.4. The MIT Press, pp. 701–707.
URL: https://doi.org/10.1162/COLI_a_00239 (cit. on p. 12).

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to information retrieval*. Cambridge University Press.
URL: https://nlp.stanford.edu/IR-book/ (cit. on pp. 11, 60).

Marcus, Mitchell, Beatrice Santorini, and Mary Ann Marcinkiewicz (Oct. 1993). *Building a large annotated corpus of English: the Penn Treebank*. MS-CIS-93-87. University of Pennsylvania, Department of Computer and Information Science.
URL: https://repository.upenn.edu/cis_reports/237/ (cit. on pp. 14, 110).

Marquez, Lluís and Jordi Girona Salgado (July 2000). *Machine learning and natural language processing*. Polytechnic University of Catalonia.
URL: https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.3498 (cit. on p. 12).

Martinez-Cruz, Carmen, Ignacio J. Blanco, and M. Amparo Vila (Dec. 2012). "Ontologies versus relational databases: are they so different? A comparison." In: *Artificial Intelligence Review* 38.4. Springer Nature, pp. 271–290.
URL: https://doi.org/10.1007/s10462-011-9251-9 (cit. on p. 33).

Matos, Sérgio (Oct. 2017). "Extracting chemical–protein interactions using long short-term memory networks." In: *BioCreative VI Workshop* (Bethesda, Maryland, USA), pp. 151–154.

URL: https://biocreative.bioinformatics.udel.edu/events/biocreative-vi/workshop/ (cit. on pp. 117, 119).

Matos, Sérgio and Rui Antunes (Oct. 2017a). "Identifying relevant literature for precision medicine using deep neural networks." In: *BioCreative VI Workshop* (Bethesda, Maryland, USA), pp. 99–101.
URL: https://biocreative.bioinformatics.udel.edu/events/biocreative-vi/workshop/ (cit. on pp. 6, 69).

Matos, Sérgio and Rui Antunes (Dec. 2017b). "Protein–protein interaction article classification using a convolutional recurrent neural network with pre-trained word embeddings." In: *Journal of Integrative Bioinformatics* 14.4. De Gruyter.
URL: https://doi.org/10.1515/jib-2017-0055 (cit. on pp. 7, 111).

McCallum, Andrew (Nov. 2005). "Information extraction: distilling structured data from unstructured text." In: *Queue* 3.9. ACM, pp. 48–57.
URL: https://doi.org/10.1145/1105664.1105679 (cit. on p. 95).

McClosky, David and Eugene Charniak (June 2008). "Self-training for biomedical parsing." In: *46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers* (Columbus, Ohio). Association for Computational Linguistics, pp. 101–104.
URL: http://dl.acm.org/citation.cfm?id=1557690.1557717 (cit. on p. 109).

McCray, Alexa T. (Nov. 1989). "The UMLS semantic network." In: *Symposium on Computer Applications in Medical Care* (Washington, DC, USA). American Medical Informatics Association, pp. 503–507.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245676/ (cit. on p. 34).

McDonald, Ryan, Fernando Pereira, Seth Kulick, Scott Winters, Yang Jin, and Pete White (June 2005). "Simple algorithms for complex relation extraction with applications to biomedical IE." In: *43rd Annual Meeting on Association for Computational Linguistics* (Ann Arbor, Michigan, USA). Association for Computational Linguistics, pp. 491–498.
URL: https://doi.org/10.3115/1219840.1219901 (cit. on p. 96).

McInnes, Bridget T. and Ted Pedersen (Dec. 2013). "Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text." In: *Journal of Biomedical Informatics* 46.6. Elsevier, pp. 1116–1124.
URL: https://doi.org/10.1016/j.jbi.2013.08.008 (cit. on pp. 40, 52, 53).

McInnes, Bridget T., Ted Pedersen, Ying Liu, Genevieve B. Melton, and Serguei V. Pakhomov (Oct. 2011). "Knowledge-based method for determining the meaning of ambiguous biomedical terms using information content measures of similarity." In: *AMIA Annual Symposium* (Washington, DC, USA). American Medical Informatics Associ-

ation, pp. 895–904.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243213/ (cit. on p. 40).

McInnes, Bridget T. and Mark Stevenson (Feb. 2014). "Determining the difficulty of word sense disambiguation." In: *Journal of Biomedical Informatics* 47. Elsevier, pp. 83–90.
URL: https://doi.org/10.1016/j.jbi.2013.09.009 (cit. on pp. 40, 52, 53).

Mehryary, Farrokh, Jari Björne, Tapio Salakoski, and Filip Ginter (Oct. 2017). "Combining support vector machines and LSTM networks for chemical–protein relation extraction." In: *BioCreative VI Workshop* (Bethesda, Maryland, USA), pp. 175–179.
URL: https://biocreative.bioinformatics.udel.edu/events/biocreative-vi/workshop/ (cit. on pp. 102, 109, 119).

Mehryary, Farrokh, Jari Björne, Tapio Salakoski, and Filip Ginter (Jan. 2018). "Potent pairing: ensemble of long short-term memory networks and support vector machine for chemical–protein relation extraction." In: *Database* 2018. Oxford University Press, bay120.
URL: https://doi.org/10.1093/database/bay120 (cit. on pp. 102, 109, 112, 119).

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (Jan. 2013a). *Efficient estimation of word representations in vector space*. arXiv:1301.3781.
URL: https://arxiv.org/abs/1301.3781 (cit. on pp. 17, 46, 76, 111, 130).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean (Dec. 2013b). "Distributed representations of words and phrases and their compositionality." In: *27th Conference on Neural Information Processing Systems (NIPS 2013)* (Lake Tahoe, Nevada, USA). Curran Associates, Inc., pp. 3111–3119.
URL: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality (cit. on p. 17).

Miller, George A. (Nov. 1995). "WordNet: a lexical database for English." In: *Communications of the ACM* 38.11. ACM, pp. 39–41.
URL: https://doi.org/10.1145/219717.219748 (cit. on p. 39).

Minaee, Shervin, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao (Apr. 2022). "Deep Learning–based text classification: a comprehensive review." In: *ACM Computing Surveys* 54.3. ACM, 62:1–62:40.
URL: https://doi.org/10.1145/3439726 (cit. on p. 63).

Mintz, Mike, Steven Bills, Rion Snow, and Dan Jurafsky (Aug. 2009). "Distant supervision for relation extraction without labeled data." In: *Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2* (Suntec, Singapore). Association for Computational Linguistics, pp. 1003–1011.
URL: https://dl.acm.org/citation.cfm?id=1690219.1690287 (cit. on p. 100).

Miranda, Antonio, Farrokh Mehryary, Jouni Luoma, Sampo Pyysalo, Alfonso Valencia, and Martin Krallinger (Nov. 2021). "Overview of DrugProt BioCreative VII track: quality evaluation and large scale text mining of drug-gene/protein relations." In: *BioCreative VII Challenge Evaluation Workshop* (Online), pp. 11–21.
URL: https://biocreative.bioinformatics.udel.edu/resources/publications/bc-vii-workshop-proceedings/ (cit. on pp. 98, 105).

Miranda-Escalada, Antonio, Luis Gascó, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger (Sept. 2022). "Overview of DisTEMIST at BioASQ: automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources." In: *CLEF 2022 Working Notes* (Bologna, Italy). CEUR Workshop Proceedings, pp. 179–203.
URL: http://ceur-ws.org/Vol-3180/ (cit. on p. 131).

Mirończuk, Marcin Michał and Jarosław Protasiewicz (Sept. 2018). "A recent overview of the state-of-the-art elements of text classification." In: *Expert Systems with Applications* 106. Elsevier, pp. 36–54.
URL: https://doi.org/10.1016/j.eswa.2018.03.058 (cit. on p. 63).

Miwa, Makoto and Mohit Bansal (Aug. 2016). "End-to-end relation extraction using LSTMs on sequences and tree structures." In: *54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Berlin, Germany). Association for Computational Linguistics, pp. 1105–1116.
URL: https://doi.org/10.18653/v1/P16-1105 (cit. on p. 23).

Miwa, Makoto and Yutaka Sasaki (Oct. 2014). "Modeling joint entity and relation extraction with table representation." In: *2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha, Qatar). Association for Computational Linguistics, pp. 1858–1869.
URL: https://doi.org/10.3115/v1/D14-1200 (cit. on pp. 23, 24).

Moon, Sungrim, Serguei Pakhomov, Nathan Liu, James O. Ryan, and Genevieve B. Melton (Mar. 2014). "A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources." In: *Journal of the American Medical Informatics Association* 21. Oxford University Press, pp. 299–307.
URL: https://doi.org/10.1136/amiajnl-2012-001506 (cit. on p. 41).

Nadkarni, Prakash M., Lucila Ohno-Machado, and Wendy W. Chapman (Sept. 2011). "Natural language processing: an introduction." In: *Journal of the American Medical Informatics Association* 18.5. Oxford University Press, pp. 544–551.
URL: https://doi.org/10.1136/amiajnl-2011-000464 (cit. on pp. 11, 12).

Navigli, Roberto (Feb. 2009). "Word sense disambiguation: a survey." In: *ACM Computing Surveys* 41.2. ACM, 10:1–10:69.
URL: https://doi.org/10.1145/1459352.1459355 (cit. on p. 37).

Nédellec, Claire (Aug. 2005). "Learning language in logic - genic interaction extraction challenge." In: *Learning Language in Logic workshop (LLL05)* (Born, Germany), pp. 31–37.
URL: https://hal.inrae.fr/hal-02762818 (cit. on p. 97).

Nédellec, Claire, Robert Bossy, Estelle Chaix, and Louise Deléger (May 2018). *Text-mining and ontologies: new approaches to knowledge discovery of microbial diversity*. arXiv:1805.04107.
URL: http://arxiv.org/abs/1805.04107 (cit. on p. 98).

Nédellec, Claire, Robert Bossy, Jin-Dong Kim, Jung-jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum (Aug. 2013). "Overview of BioNLP Shared Task 2013." In: *BioNLP Shared Task 2013 Workshop* (Sofia, Bulgaria). Association for Computational Linguistics, pp. 1–7.
URL: https://aclanthology.org/W13-2001 (cit. on p. 99).

Nelson, Stuart J., Kelly Zeng, John Kilbourne, Tammy Powell, and Robin Moore (July 2011). "Normalized names for clinical drugs: RxNorm at 6 years." In: *Journal of the American Medical Informatics Association* 18.4. Oxford University Press, p. 441.
URL: https://doi.org/10.1136/amiajnl-2011-000116 (cit. on pp. 42, 55).

Neustein, Amy, S. Sagar Imambi, Mário Rodrigues, António Teixeira, and Liliana Ferreira (Oct. 2014). "Application of text mining to biomedical knowledge extraction: analyzing clinical narratives and medical literature." In: *Text mining of web-based medical content*. De Gruyter, pp. 3–32.
URL: https://doi.org/10.1515/9781614513902.3 (cit. on p. 82).

Neves, Mariana and Ulf Leser (Mar. 2014). "A survey on annotation tools for the biomedical literature." In: *Briefings in Bioinformatics* 15.2. Oxford University Press, pp. 327–340.
URL: https://doi.org/10.1093/bib/bbs084 (cit. on p. 127).

Neves, Mariana and Jurica Ševa (Jan. 2021). "An extensive review of tools for manual annotation of documents." In: *Briefings in Bioinformatics* 22.1. Oxford University Press, pp. 146–163.
URL: https://doi.org/10.1093/bib/bbz130 (cit. on pp. 33, 127).

Newman-Griffis, Denis, Albert M. Lai, and Eric Fosler-Lussier (July 2018). "Jointly embedding entities and text with distant supervision." In: *The Third Workshop on Representation Learning for NLP* (Melbourne, Australia). Association for Computational

Linguistics, pp. 195–206.
URL: https://doi.org/10.18653/v1/w18-3026 (cit. on p. 40).

Nguyen, Thien Huu and Ralph Grishman (June 2015). "Relation extraction: perspective from convolutional neural networks." In: *1st Workshop on Vector Space Modeling for Natural Language Processing* (Denver, Colorado, USA). Association for Computational Linguistics, pp. 39–48.
URL: https://aclanthology.org/W15-1506 (cit. on p. 101).

Nunes, Tiago, David Campos, Sérgio Matos, and José Luís Oliveira (Aug. 2013). "BeCAS: biomedical concept recognition services and visualization." In: *Bioinformatics* 29.15. Oxford University Press, p. 1915.
URL: https://doi.org/10.1093/bioinformatics/btt317 (cit. on p. 106).

Ohta, Tomoko, Sampo Pyysalo, Rafal Rak, Andrew Rowley, Hong-Woo Chun, Sung-Jae Jung, Sung-Pil Choi, Sophia Ananiadou, and Jun'ichi Tsujii (Aug. 2013). "Overview of the Pathway Curation (PC) task of BioNLP Shared Task 2013." In: *BioNLP Shared Task 2013 Workshop* (Sofia, Bulgaria). Association for Computational Linguistics, pp. 67–75.
URL: https://aclanthology.org/W13-2009 (cit. on p. 99).

Ohta, Tomoko, Sampo Pyysalo, and Jun'ichi Tsujii (June 2011). "Overview of the Epigenetics and Post-translational Modifications (EPI) task of BioNLP Shared Task 2011." In: *BioNLP Shared Task 2011 Workshop* (Portland, Oregon, USA). Association for Computational Linguistics, pp. 16–25.
URL: https://aclanthology.org/W11-1803 (cit. on p. 99).

Ohta, Tomoko, Yuka Tateisi, and Jin-Dong Kim (Mar. 2002). "The GENIA corpus: an annotated research abstract corpus in molecular biology domain." In: *Second International Conference on Human Language Technology Research* (San Francisco, California, USA). Morgan Kaufmann Publishers Inc., pp. 82–86.
URL: https://dl.acm.org/doi/abs/10.5555/1289189.1289260 (cit. on p. 98).

Ohta, Yoshihiro, Yasunori Yamamoto, Tomoko Okazaki, Ikuo Uchiyama, and Toshihisa Takagi (June 1997). "Automatic construction of knowledge base from biological papers." In: *Fifth International Conference on Intelligent Systems for Molecular Biology* (Halkidiki, Greece). Association for the Advancement of Artificial Intelligence, pp. 218–225.
URL: https://www.aaai.org/Library/ISMB/1997/ismb97-033.php (cit. on p. 95).

Oleynik, Michel (June 2020). "Leveraging word embeddings for biomedical natural language processing." PhD thesis. Medical University of Graz.
URL: https://online.medunigraz.at/mug_online/ee/ui/ca2/app/desktop/#/pl/u

i/$ctx/wbAbs.showThesis?$ctx=design=ca2;header=max;lang=en&pThesisNr=58784 (cit. on p. 9).

Oliveira e Silva, Tomás (Apr. 1994). "Sobre os filtros de Kautz e sua utilização na aproximação de sistemas lineares invariantes no tempo." PhD thesis. University of Aveiro. URL: http://hdl.handle.net/10773/4641 (cit. on p. 9).

Opitz, Juri and Sebastian Burst (Nov. 2019). *Macro F1 and macro F1*. arXiv:1911.03347. URL: https://arxiv.org/abs/1911.03347 (cit. on p. 31).

Panchenko, Alexander and Olga Morozova (Apr. 2012). "A study of hybrid similarity measures for semantic relation extraction." In: *Workshop on Innovative Hybrid Approaches to the Processing of Textual Data* (Avignon, France). Association for Computational Linguistics, pp. 10–18. URL: https://aclanthology.org/W12-0502 (cit. on p. 60).

Pathak, Jyotishman, Abel N. Kho, and Joshua C. Denny (Dec. 2013). "Electronic health records–driven phenotyping: challenges, recent advances, and perspectives." In: *Journal of the American Medical Informatics Association* 20.e2. Oxford University Press, e206–e211. URL: https://doi.org/10.1136/amiajnl-2013-002428 (cit. on p. 70).

Pawar, Sachin, Girish K. Palshikar, and Pushpak Bhattacharyya (Dec. 2017). *Relation extraction : a survey*. arXiv:1712.05191. URL: https://arxiv.org/abs/1712.05191 (cit. on p. 95).

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (Oct. 2011). "Scikit-learn: machine learning in Python." In: *Journal of Machine Learning Research* 12, pp. 2825–2830. URL: https://jmlr.org/papers/v12/pedregosa11a.html (cit. on pp. 44, 47, 65, 71).

Peng, Yifan, Anthony Rios, Ramakanth Kavuluru, and Zhiyong Lu (July 2018). "Extracting chemical–protein relations with ensembles of SVM and deep learning models." In: *Database* 2018. Oxford University Press, bay073. URL: https://doi.org/10.1093/database/bay073 (cit. on pp. 101, 112, 119).

Peng, Yifan, Anthony Rios, Kavuluru Ramakanth, and Zhiyong Lu (Oct. 2017). "Chemical–protein relation extraction with ensembles of SVM, CNN, and RNN models." In: *BioCreative VI Workshop* (Bethesda, Maryland, USA), pp. 147–150. URL: https://biocreative.bioinformatics.udel.edu/events/biocreative-vi/workshop/ (cit. on pp. 101, 119).

Peng, Yifan, Shankai Yan, and Zhiyong Lu (Aug. 2019). "Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets." In: *18th BioNLP Workshop and Shared Task* (Florence, Italy). Association for

Computational Linguistics, pp. 58–65.
URL: https://doi.org/10.18653/v1/W19-5006 (cit. on pp. 63, 124, 131).

Pennington, Jeffrey, Richard Socher, and Christopher Manning (Oct. 2014). "Glove: global vectors for word representation." In: *2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha, Qatar). Association for Computational Linguistics, pp. 1532–1543.
URL: https://aclanthology.org/D14-1162 (cit. on p. 17).

Pereira, Vitor, Sérgio Matos, and José Luís Oliveira (Oct. 2018). "Automated ICD-9-CM medical coding of diabetic patient's clinical reports." In: *First International Conference on Data Science, E-learning and Information Systems* (Madrid, Spain). ACM, pp. 1–6.
URL: https://doi.org/10.1145/3279996.3280019 (cit. on p. 62).

Pérez-Pérez, Martín, Tânia Ferreira, Gilberto Igrejas, and Florentino Fdez-Riverola (June 2022). "A deep learning relation extraction approach to support a biomedical semi-automatic curation task: the case of the gluten bibliome." In: *Expert Systems with Applications* 195. Elsevier, p. 116616.
URL: https://doi.org/10.1016/j.eswa.2022.116616 (cit. on p. 127).

Pesaranghader, Ahmad, Stan Matwin, Marina Sokolova, and Ali Pesaranghader (May 2019). "deepBioWSD: effective deep neural word sense disambiguation of biomedical text data." In: *Journal of the American Medical Informatics Association* 26.5. Oxford University Press, pp. 438–446.
URL: https://doi.org/10.1093/jamia/ocy189 (cit. on pp. 52, 53).

Pestian, John P., Chris Brew, Pawel Matykiewicz, DJ Hovermale, Neil Johnson, K. Bretonnel Cohen, and Wlodzislaw Duch (June 2007). "A shared task involving multi-label classification of clinical free text." In: *Biological, translational, and clinical language processing* (Prague, Czech Republic). Association for Computational Linguistics, pp. 97–104.
URL: https://aclanthology.org/W07-1013 (cit. on p. 62).

Peters, Matthew, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (June 2018). "Deep contextualized word representations." In: *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (New Orleans, Louisiana, USA). Association for Computational Linguistics, pp. 2227–2237.
URL: https://doi.org/10.18653/v1/N18-1202 (cit. on p. 130).

Pivovarov, Rimma and Noémie Elhadad (Sept. 2015). "Automated methods for the summarization of electronic health records." In: *Journal of the American Medical Informatics Association* 22.5. Oxford University Press, pp. 938–947.
URL: https://doi.org/10.1093/jamia/ocv032 (cit. on p. 82).

Porter, M. F. (Mar. 1980). "An algorithm for suffix stripping." In: *Program: electronic library and information systems* 14.3. Emerald Group Publishing Limited, pp. 130–137. URL: https://doi.org/10.1108/eb046814 (cit. on p. 13).

Pradhan, Sameer, Noemie Elhadad, Brett R. South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, Wendy W. Chapman, and Guergana Savova (Sept. 2013). "Task 1: ShARe/CLEF eHealth evaluation lab 2013." In: *CLEF 2013 Working Notes* (Valencia, Spain). CEUR Workshop Proceedings. URL: http://ceur-ws.org/Vol-1179/ (cit. on p. 54).

Pradhan, Sameer, Noémie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova (Aug. 2014). "SemEval-2014 Task 7: analysis of clinical text." In: *8th International Workshop on Semantic Evaluation (SemEval 2014)* (Dublin, Ireland). Association for Computational Linguistics, pp. 54–62. URL: https://doi.org/10.3115/v1/S14-2007 (cit. on p. 54).

Pradhan, Sameer, Noémie Elhadad, Brett R. South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, Wendy W. Chapman, and Guergana Savova (Jan. 2015). "Evaluating the state of the art in disorder recognition and normalization of the clinical narrative." In: *Journal of the American Medical Informatics Association* 22.1. Oxford University Press, pp. 143–154. URL: https://doi.org/10.1136/amiajnl-2013-002544 (cit. on p. 42).

Przybyła, Piotr, Matthew Shardlow, Sophie Aubin, Robert Bossy, Richard Eckart de Castilho, Stelios Piperidis, John McNaught, and Sophia Ananiadou (Sept. 2016). "Text mining resources for the life sciences." In: *Database* 2016. Oxford University Press, baw145. URL: https://doi.org/10.1093/database/baw145 (cit. on p. 32).

Pyysalo, Sampo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski (Feb. 2007). "BioInfer: a corpus for information extraction in the biomedical domain." In: *BMC Bioinformatics* 8.1. BioMed Central Ltd, p. 50. URL: https://doi.org/10.1186/1471-2105-8-50 (cit. on p. 103).

Pyysalo, Sampo, Tomoko Ohta, and Sophia Ananiadou (Aug. 2013). "Overview of the Cancer Genetics (CG) task of BioNLP Shared Task 2013." In: *BioNLP Shared Task 2013 Workshop* (Sofia, Bulgaria). Association for Computational Linguistics, pp. 58–66. URL: https://aclanthology.org/W13-2008 (cit. on p. 99).

Pyysalo, Sampo, Tomoko Ohta, Rafal Rak, Andrew Rowley, Hong-Woo Chun, Sung-Jae Jung, Sung-Pil Choi, Jun'ichi Tsujii, and Sophia Ananiadou (June 2015). "Overview of the Cancer Genetics and Pathway Curation tasks of BioNLP Shared Task 2013." In: *BMC Bioinformatics* 16.10. BioMed Central Ltd, S2. URL: https://doi.org/10.1186/1471-2105-16-S10-S2 (cit. on p. 99).

Pyysalo, Sampo, Tomoko Ohta, Rafal Rak, Dan Sullivan, Chunhong Mao, Chunxia Wang, Bruno Sobral, Jun'ichi Tsujii, and Sophia Ananiadou (June 2011a). "Overview of the Infectious Diseases (ID) task of BioNLP Shared Task 2011." In: *BioNLP Shared Task 2011 Workshop* (Portland, Oregon, USA). Association for Computational Linguistics, pp. 26–35.
URL: https://aclanthology.org/W11-1804 (cit. on p. 99).

Pyysalo, Sampo, Tomoko Ohta, Rafal Rak, Dan Sullivan, Chunhong Mao, Chunxia Wang, Bruno Sobral, Jun'ichi Tsujii, and Sophia Ananiadou (June 2012). "Overview of the ID, EPI and REL tasks of BioNLP Shared Task 2011." In: *BMC Bioinformatics* 13.11. Springer Nature, S2.
URL: https://doi.org/10.1186/1471-2105-13-S11-S2 (cit. on p. 99).

Pyysalo, Sampo, Tomoko Ohta, and Jun'ichi Tsujii (June 2011b). "Overview of the Entity Relations (REL) supporting task of BioNLP Shared Task 2011." In: *BioNLP Shared Task 2011 Workshop* (Portland, Oregon, USA). Association for Computational Linguistics, pp. 83–88.
URL: https://aclanthology.org/W11-1812 (cit. on p. 99).

Qin, Pengda, Weiran Xu, and William Yang Wang (July 2018). "DSGAN: generative adversarial training for distant supervision relation extraction." In: *56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Melbourne, Australia). Association for Computational Linguistics, pp. 496–505.
URL: https://doi.org/10.18653/v1/P18-1046 (cit. on p. 118).

Qu, Jinchan, Albert Steppi, Jie Hao, Jian Wang, Pei-Yau Lung, Tingting Zhao, Zhe He, and Jinfeng Zhang (Oct. 2017). "Mining protein interactions affected by mutations using a NLP based machine learning approach." In: *BioCreative VI Workshop* (Bethesda, Maryland, USA), pp. 131–134.
URL: https://biocreative.bioinformatics.udel.edu/events/biocreative-vi/workshop/ (cit. on p. 69).

El-Rab, Wessam Gad, Osmar R. Zaïane, and Mohammad El-Hajj (Aug. 2013). "Biomedical text disambiguation using UMLS." In: *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (Niagara, Ontario, Canada). ACM, pp. 943–947.
URL: https://dl.acm.org/citation.cfm?doid=2492517.2500251 (cit. on p. 40).

Ramshaw, Lance and Mitch Marcus (June 1995). "Text chunking using transformation-based learning." In: *Third Workshop on Very Large Corpora* (Cambridge, Massachusetts, USA). Association for Computational Linguistics, pp. 82–94.
URL: https://aclanthology.org/W95-0107 (cit. on p. 20).

Ratinov, Lev and Dan Roth (June 2009). "Design challenges and misconceptions in named entity recognition." In: *Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)* (Boulder, Colorado, USA). Association for Computational Linguistics, pp. 147–155.
URL: https://aclanthology.org/W09-1119 (cit. on p. 20).

Ratnaparkhi, Adwait (1998). "Maximum entropy models for natural language ambiguity resolution." PhD thesis. University of Pennsylvania.
URL: https://dl.acm.org/citation.cfm?id=927230 (cit. on p. 20).

Rawat, Waseem and Zenghui Wang (Sept. 2017). "Deep convolutional neural networks for image classification: a comprehensive review." In: *Neural Computation* 29.9. IEEE, pp. 2352–2449.
URL: https://doi.org/10.1162/neco_a_00990 (cit. on p. 66).

Rebholz-Schuhmann, Dietrich, Anika Oellrich, and Robert Hoehndorf (Nov. 2012). "Text-mining solutions for biomedical research: enabling integrative biology." In: *Nature Reviews Genetics* 13.12. Springer Nature, pp. 829–839.
URL: https://doi.org/10.1038/nrg3337 (cit. on p. 32).

Regev, Yizhar, Michal Finkelstein-Landau, Ronen Feldman, Maya Gorodetsky, Xin Zheng, Samuel Levy, Rosane Charlab, Charles Lawrence, Ross A. Lippert, Qing Zhang, and Hagit Shatkay (Dec. 2002). "Rule-based extraction of experimental evidence in the biomedical domain: the KDD Cup 2002 (task 1)." In: *ACM SIGKDD Explorations Newsletter* 4.2. ACM, pp. 90–92.
URL: https://doi.org/10.1145/772862.772874 (cit. on p. 61).

Řehůřek, Radim and Petr Sojka (May 2010). "Software framework for topic modelling with large corpora." In: *LREC 2010 Workshop on New Challenges for NLP Frameworks* (Valletta, Malta). University of Malta, pp. 46–50.
URL: http://is.muni.cz/publication/884893/en (cit. on pp. 33, 44, 46, 65, 76, 111, 112, 130).

Ren, Xiang, Zeqiu Wu, Wenqi He, Meng Qu, Clare R. Voss, Heng Ji, Tarek F. Abdelzaher, and Jiawei Han (Apr. 2017). "CoType: joint extraction of typed entities and relations with knowledge bases." In: *26th International Conference on World Wide Web* (Perth, Australia). International World Wide Web Conferences Steering Committee, pp. 1015–1024.
URL: https://doi.org/10.1145/3038912.3052708 (cit. on pp. 23, 24).

Rindflesch, Thomas C. (Mar. 1996). "Natural language processing." In: *Annual Review of Applied Linguistics* 16. Cambridge University Press, pp. 70–85.
URL: https://doi.org/10.1017/s0267190500001446 (cit. on p. 12).

Rindflesch, Thomas C., Lorraine Tanabe, John N. Weinstein, and Lawrence Hunter (1999). "EDGAR: extraction of drugs, genes and relations from the biomedical literature." In: *Biocomputing 2000*. World Scientific, pp. 517–528.
URL: https://doi.org/10.1142/9789814447331_0049 (cit. on p. 100).

Rios, Anthony and Ramakanth Kavuluru (Sept. 2015). "Convolutional neural networks for biomedical text classification: application in indexing biomedical articles." In: *6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics* (Atlanta, Georgia, USA). ACM, pp. 258–267.
URL: https://doi.org/10.1145/2808719.2808746 (cit. on p. 66).

Roberts, Kirk, Dina Demner-Fushman, and Joseph M. Tonning (Nov. 2017). "Overview of the TAC 2017 adverse reaction extraction from drug labels track." In: *Tenth Text Analysis Conference (TAC 2017)* (Gaithersburg, Maryland, USA).
URL: https://tac.nist.gov/publications/2017/index.html (cit. on p. 97).

Rosario, Barbara and Marti A. Hearst (July 2004). "Classifying semantic relations in bioscience texts." In: *42nd Annual Meeting on Association for Computational Linguistics* (Barcelona, Spain). Association for Computational Linguistics, pp. 430–437.
URL: https://doi.org/10.3115/1218955.1219010 (cit. on pp. 98, 103).

Rosário-Ferreira, Nícia, Catarina Marques-Pereira, Manuel Pires, Daniel Ramalhão, Nádia Pereira, Victor Guimarães, Vítor Santos Costa, and Irina Sousa Moreira (Sept. 2021). "The treasury chest of text mining: piling available resources for powerful biomedical text mining." In: *BioChem* 1.2. MDPI, pp. 60–80.
URL: https://doi.org/10.3390/biochem1020007 (cit. on pp. 32, 127).

Roth, Dan and Wen-tau Yih (May 2004). "A linear programming formulation for global inference in natural language tasks." In: *Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004* (Boston, Massachusetts, USA). Association for Computational Linguistics, pp. 1–8.
URL: https://aclanthology.org/W04-2401 (cit. on p. 23).

Russell, Stuart and Peter Norvig (Dec. 2009). *Artificial intelligence: a modern approach*. 3rd ed. Pearson.
URL: http://aima.cs.berkeley.edu/ (cit. on p. 1).

Sabbir, A. K. M., Antonio Jimeno-Yepes, and Ramakanth Kavuluru (Oct. 2016). *Knowledge-based biomedical word sense disambiguation with neural concept embeddings*. arXiv:1610.08557.
URL: https://arxiv.org/abs/1610.08557 (cit. on pp. 40, 52, 53).

Sætre, Rune, Kazuhiro Yoshida, Akane Yakushiji, Yusuke Miyao, Yuichiro Matsubayashi, and Tomoko Ohta (Apr. 2007). "AKANE system: protein–protein interaction pairs in the BioCreAtIvE2 challenge, PPI-IPS subtask." In: *Second BioCreative Challenge*

Evaluation Workshop (Madrid, Spain), pp. 209–211.
URL: https://biocreative.bioinformatics.udel.edu/resources/publications/ (cit. on p. 109).

Saif, Hassan, Miriam Fernandez, Yulan He, and Harith Alani (May 2014). "On stopwords, filtering and data sparsity for sentiment analysis of Twitter." In: *Ninth International Conference on Language Resources and Evaluation (LREC'14)* (Reykjavik, Iceland). European Language Resources Association (ELRA), pp. 810–817.
URL: https://aclanthology.org/L14-1265 (cit. on p. 13).

Saik, Olga V. and Vadim V. Klimontov (Nov. 2021). "Hypoglycemia, vascular disease and cognitive dysfunction in diabetes: insights from text mining-based reconstruction and bioinformatics analysis of the gene networks." In: *International Journal of Molecular Sciences* 22.22. MDPI, p. 12419.
URL: https://doi.org/10.3390/ijms222212419 (cit. on p. 96).

Salgado, David, Martin Krallinger, Marc Depaule, Elodie Drula, Ashish V. Tendulkar, Florian Leitner, Alfonso Valencia, and Christophe Marcelle (Sept. 2012). "MyMiner: a web application for computer-assisted biocuration and text annotation." In: *Bioinformatics* 28.17, pp. 2285–2287.
URL: https://doi.org/10.1093/bioinformatics/bts435 (cit. on p. 33).

Sanderson, Mark (July 1994). "Word sense disambiguation and information retrieval." In: *17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland). Springer Nature, pp. 142–151.
URL: https://doi.org/10.1007/978-1-4471-2099-5_15 (cit. on p. 39).

Sanderson, Mark (Sept. 1996). "Word sense disambiguation and information retrieval." PhD thesis. University of Glasgow.
URL: https://theses.gla.ac.uk/4463/ (cit. on p. 39).

Sang, Erik F. Tjong Kim and Jorn Veenstra (June 1999). "Representing text chunks." In: *Ninth Conference on European Chapter of the Association for Computational Linguistics (EACL '99)* (Bergen, Norway). Association for Computational Linguistics, pp. 173–179.
URL: https://doi.org/10.3115/977035.977059 (cit. on pp. 14, 20).

Santorini, Beatrice (July 1990). *Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision)*. MS-CIS-90-47. University of Pennsylvania, Department of Computer and Information Science.
URL: https://repository.upenn.edu/cis_reports/570/ (cit. on p. 14).

Sarawagi, Sunita (Nov. 2008). "Information extraction." In: *Foundations and Trends® in Databases* 1.3. Now Publishers, Inc., pp. 261–377.
URL: https://doi.org/10.1561/1900000003 (cit. on pp. 1, 59).

Sasaki, Yutaka, Brian Rea, and Sophia Ananiadou (Nov. 2007). "Multi-topic aspects in clinical text classification." In: *2007 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2007)*. IEEE, pp. 62–70.
URL: https://doi.org/10.1109/bibm.2007.23 (cit. on p. 62).

Savova, Guergana K., James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute (Sept. 2010). "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications." In: *Journal of the American Medical Informatics Association* 17.5. Oxford University Press, pp. 507–513.
URL: https://doi.org/10.1136/jamia.2009.001560 (cit. on p. 33).

Sayers, Eric W., Jeffrey Beck, Evan E. Bolton, Devon Bourexis, James R. Brister, Kathi Canese, Donald C. Comeau, Kathryn Funk, Sunghwan Kim, William Klimke, Aron Marchler-Bauer, Melissa Landrum, Stacy Lathrop, Zhiyong Lu, Thomas L. Madden, Nuala O'Leary, Lon Phan, Sanjida H. Rangwala, Valerie A. Schneider, Yuri Skripchenko, Jiyao Wang, Jian Ye, Barton W. Trawick, Kim D. Pruitt, and Stephen T. Sherry (Jan. 2021). "Database resources of the National Center for Biotechnology Information." In: *Nucleic Acids Research* 49.D1. Oxford University Press, pp. D10–D17.
URL: https://doi.org/10.1093/nar/gkaa892 (cit. on p. 34).

Schneider, Elisa Terumi Rubel, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra (Nov. 2020). "BioBERTpt - a Portuguese neural language model for clinical named entity recognition." In: *3rd Clinical Natural Language Processing Workshop* (Online). Association for Computational Linguistics, pp. 65–72.
URL: https://doi.org/10.18653/v1/2020.clinicalnlp-1.7 (cit. on p. 131).

Schriver, Karen A. (Nov. 1989). "Document design from 1980 to 1989: challenges that remain." In: *Technical Communication* 36.4. Society for Technical Communication, pp. 316–331.
URL: http://www.jstor.org/stable/43095665 (cit. on p. 9).

Schuemie, Martijn J., Jan A. Kors, and Barend Mons (June 2005). "Word sense disambiguation in the biomedical domain: an overview." In: *Journal of Computational Biology* 12.5. Mary Ann Liebert, pp. 554–565.
URL: https://doi.org/10.1089/cmb.2005.12.554 (cit. on p. 40).

Schuster, Sebastian and Christopher D. Manning (May 2016). "Enhanced English Universal Dependencies: an improved representation for natural language understanding tasks." In: *Tenth International Conference on Language Resources and Evaluation (LREC'16)* (Portorož, Slovenia). European Language Resources Association (ELRA),

pp. 2371–2378.
URL: https://aclanthology.org/L16-1376 (cit. on p. 14).

Segura-Bedmar, Isabel, Paloma Martínez, and María Herrero-Zazo (June 2013). "SemEval-2013 task 9 : extraction of drug–drug interactions from biomedical texts (DDIExtraction 2013)." In: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (Atlanta, Georgia, USA). Association for Computational Linguistics, pp. 341–350.
URL: https://aclanthology.org/S13-2056 (cit. on pp. 99, 104).

Segura-Bedmar, Isabel, Paloma Martínez, and María Herrero-Zazo (Oct. 2014). "Lessons learnt from the DDIExtraction-2013 shared task." In: *Journal of Biomedical Informatics* 51. Elsevier, pp. 152–164.
URL: https://doi.org/10.1016/j.jbi.2014.05.007 (cit. on p. 104).

Segura-Bedmar, Isabel, Paloma Martínez, and Daniel Sánchez-Cisneros (Sept. 2011). "The 1st DDIExtraction-2011 challenge task: extraction of drug-drug interactions from biomedical texts." In: *First Challenge Task on Drug-Drug Interaction Extraction 2011* (Huelva, Spain). Vol. 761. CEUR Workshop Proceedings.
URL: http://ceur-ws.org/Vol-761/ (cit. on p. 99).

Shah, Kanish, Henil Patel, Devanshi Sanghvi, and Manan Shah (Mar. 2020). "A comparative analysis of logistic regression, random forest and KNN models for the text classification." In: *Augmented Human Research* 5.1. Springer Nature, p. 12.
URL: https://doi.org/10.1007/s41133-020-00032-0 (cit. on p. 66).

Shen, Yatian and Xuanjing Huang (Dec. 2016). "Attention-based convolutional neural network for semantic relation extraction." In: *COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (Osaka, Japan). The COLING 2016 Organizing Committee, pp. 2526–2536.
URL: https://aclanthology.org/C16-1238 (cit. on p. 102).

Shivade, Chaitanya, Preethi Raghavan, Eric Fosler-Lussier, Peter J. Embi, Noemie Elhadad, Stephen B. Johnson, and Albert M. Lai (Mar. 2014). "A review of approaches to identifying patient phenotype cohorts using electronic health records." In: *Journal of the American Medical Informatics Association* 21.2. Oxford University Press, pp. 221–230.
URL: https://doi.org/10.1136/amiajnl-2013-001935 (cit. on p. 70).

Silva, João Figueira, Tiago Almeida, Rui Antunes, João Rafael Almeida, and Sérgio Matos (Nov. 2021a). "Drug mention recognition in Twitter posts using a deep learning approach." In: *BioCreative VII Challenge Evaluation Workshop* (Online), pp. 210–213.

URL: https://biocreative.bioinformatics.udel.edu/resources/publications/bc-vii-workshop-proceedings/ (cit. on p. 8).

Silva, João Figueira, Rui Antunes, João Rafael Almeida, and Sérgio Matos (Apr. 2020). "Clinical concept normalization on medical records using word embeddings and heuristics." In: *30th Medical Informatics Europe Conference* (Canceled). IOS Press, pp. 93–97.
URL: http://hdl.handle.net/10773/29190 (cit. on pp. 6, 54).

Silva, Jorge Miguel, Diogo Pratas, Rui Antunes, Sérgio Matos, and Armando J. Pinho (June 2021b). "Automatic analysis of artistic paintings using information-based measures." In: *Pattern Recognition* 114. Elsevier, p. 107864.
URL: https://doi.org/10.1016/j.patcog.2021.107864 (cit. on p. 8).

Silva e Oliveira, Lucas Emanuel, Ana Carolina Peters, Adalniza Moura Pucca da Silva, Caroline Pilatti Gebeluca, Yohan Bonescki Gumiel, Lilian Mie Mukai Cintho, Deborah Ribeiro Carvalho, Sadid Al Hasan, and Claudia Maria Cabral Moro (May 2022). "SemClinBr - a multi-institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks." In: *Journal of Biomedical Semantics* 13.1, p. 13.
URL: https://doi.org/10.1186/s13326-022-00269-1 (cit. on p. 131).

Simpson, Matthew S. and Dina Demner-Fushman (Jan. 2012). "Biomedical text mining: a survey of recent progress." In: *Mining Text Data*. Ed. by Charu C. Aggarwal and ChengXiang Zhai. Springer Nature, pp. 465–517.
URL: https://doi.org/10.1007/978-1-4614-3223-4_14 (cit. on p. 32).

Sims-Robinson, Catrina, Bhumsoo Kim, Andrew Rosko, and Eva L. Feldman (Oct. 2010). "How does diabetes accelerate Alzheimer disease pathology?" In: *Nature Reviews Neurology* 6.10. Springer Nature, pp. 551–559.
URL: https://doi.org/10.1038/nrneurol.2010.130 (cit. on p. 96).

Singh, Hardeep, Traber Davis Giardina, Ashley N. D. Meyer, Samuel N. Forjuoh, Michael D. Reis, and Eric J. Thomas (Mar. 2013). "Types and origins of diagnostic errors in primary care settings." In: *JAMA Internal Medicine* 173.6. American Medical Association, pp. 418–425.
URL: https://doi.org/10.1001/jamainternmed.2013.2777 (cit. on p. 82).

Singhal, Ayush, Robert Leaman, Natalie Catlett, Thomas Lemberger, Johanna McEntyre, Shawn Polson, Ioannis Xenarios, Cecilia Arighi, and Zhiyong Lu (Jan. 2016a). "Pressing needs of biomedical text mining in biocuration and beyond: opportunities and challenges." In: *Database* 2016. Oxford University Press, baw161.
URL: https://doi.org/10.1093/database/baw161 (cit. on p. 12).

Singhal, Ayush, Michael Simmons, and Zhiyong Lu (Nov. 2016b). "Text mining genotype–phenotype relationships from biomedical literature for database curation and

precision medicine." In: *PLOS Computational Biology* 12.11. Public Library of Science, pp. 1–19.
URL: https://doi.org/10.1371/journal.pcbi.1005017 (cit. on pp. 96, 106).

Siu, Amy, Patrick Ernst, and Gerhard Weikum (Aug. 2016). "Disambiguation of entities in MEDLINE abstracts by combining MeSH terms with knowledge." In: *15th Workshop on Biomedical Natural Language Processing* (Berlin, Germany). Association for Computational Linguistics, pp. 72–76.
URL: https://doi.org/10.18653/v1/W16-2909 (cit. on p. 40).

Smirnova, Alisa and Philippe Cudré-Mauroux (Jan. 2019). "Relation extraction using distant supervision: a survey." In: *ACM Computing Surveys* 51.5. ACM, 106:1–106:35.
URL: http://doi.org/10.1145/3241741 (cit. on p. 100).

Smith, Barry, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J. Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J. Mungall, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H. Scheuermann, Nigam Shah, Patricia L. Whetzel, and Suzanna Lewis (Nov. 2007). "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration." In: *Nature Biotechnology* 25.11. Springer Nature, pp. 1251–1255.
URL: https://doi.org/10.1038/nbt1346 (cit. on p. 34).

Smith, Larry, Lorraine K. Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M. Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig A. Struble, Richard J. Povinelli, Andreas Vlachos, William A. Baumgartner, Lawrence Hunter, Bob Carpenter, Richard Tzong-Han Tsai, Hong-Jie Dai, Feng Liu, Yifei Chen, Chengjie Sun, Sophia Katrenko, Pieter Adriaans, Christian Blaschke, Rafael Torres, Mariana Neves, Preslav Nakov, Anna Divoli, Manuel Maña-López, Jacinto Mata, and W. John Wilbur (Sept. 2008). "Overview of BioCreative II gene mention recognition." In: *Genome Biology* 9.2. BioMed Central Ltd, S2.
URL: https://doi.org/10.1186/gb-2008-9-s2-s2 (cit. on p. 35).

Soderland, Stephen, Brendan Roof, Bo Qin, Shi Xu, Mausam, and Oren Etzioni (July 2010). "Adapting open information extraction to domain-specific relations." In: *AI Magazine* 31.3. Association for the Advancement of Artificial Intelligence, pp. 93–102.
URL: https://doi.org/10.1609/aimag.v31i3.2305 (cit. on pp. 28, 96).

Soğancıoğlu, Gizem, Hakime Öztürk, and Arzucan Özgür (July 2017). "BIOSSES: a semantic sentence similarity estimation system for the biomedical domain." In: *Bioinformatics* 33.14. Oxford University Press, pp. i49–i58.
URL: https://doi.org/10.1093/bioinformatics/btx238 (cit. on pp. 63, 83).

Stearns, Michael Q., Colin Price, Kent A. Spackman, and Amy Y. Wang (Nov. 2001). "SNOMED clinical terms: overview of the development process and project status." In: *AMIA Symposium* (Washington, DC, USA). American Medical Informatics Association, pp. 662–666.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243297/ (cit. on pp. 42, 55).

Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii (Apr. 2012). "BRAT: a web-based tool for NLP-assisted text annotation." In: *Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics* (Avignon, France). Association for Computational Linguistics, pp. 102–107.
URL: https://aclanthology.org/E12-2021 (cit. on p. 33).

Stevenson, Mark and Yorick Wilks (Sept. 2001). "The interaction of knowledge sources in word sense disambiguation." In: *Computational Linguistics* 27.3. The MIT Press, pp. 321–349.
URL: https://doi.org/10.1162/089120101317066104 (cit. on p. 39).

Stubbs, Amber, Michele Filannino, Ergin Soysal, Samuel Henry, and Özlem Uzuner (Nov. 2019). "Cohort selection for clinical trials: n2c2 2018 shared task track 1." In: *Journal of the American Medical Informatics Association* 26.11. Oxford University Press, pp. 1163–1171.
URL: https://doi.org/10.1093/jamia/ocz163 (cit. on pp. 70, 72).

Sun, Weiyi, Anna Rumshisky, and Ozlem Uzuner (Sept. 2013). "Evaluating temporal relations in clinical text: 2012 i2b2 Challenge." In: *Journal of the American Medical Informatics Association* 20.5. Oxford University Press, pp. 806–813.
URL: https://doi.org/10.1136/amiajnl-2013-001628 (cit. on p. 98).

Sundheim, Beth M. (May 1996). "The Message Understanding Conferences." In: *TIPSTER TEXT PROGRAM PHASE II: Proceedings of a Workshop held at Vienna, Virginia, May 6-8, 1996* (Vienna, Virginia, USA). Association for Computational Linguistics, pp. 35–37.
URL: https://doi.org/10.3115/1119018.1119025 (cit. on p. 97).

Surdeanu, Mihai (Nov. 2013). "Overview of the TAC2013 knowledge base population evaluation: english slot filling and temporal slot filling." In: *Text Analysis Conference (TAC 2013)* (Gaithersburg, Maryland, USA).
URL: https://tac.nist.gov/publications/2013/papers.html (cit. on p. 97).

Surdeanu, Mihai, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning (July 2012). "Multi-instance multi-label learning for relation extraction." In: *2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Nat-*

ural Language Learning (Jeju Island, Korea). Association for Computational Linguistics, pp. 455–465.
URL: https://aclanthology.org/D12-1042/ (cit. on pp. 95, 118).

Tao, Jie and Xing Fang (Jan. 2020). "Toward multi-label sentiment analysis: a transfer learning based approach." In: *Journal of Big Data* 7.1. Springer Nature.
URL: https://doi.org/10.1186/s40537-019-0278-0 (cit. on p. 59).

Telg, Ricky and Ashley McLeod-Morin (Apr. 2021). "Document design: WC127, rev. 3/2021." In: *EDIS* 2021.2. University of Florida.
URL: https://doi.org/10.32473/edis-wc127-2021 (cit. on p. 9).

Temkin, Joshua M. and Mark R. Gilder (Nov. 2003). "Extraction of protein interaction information from unstructured text using a context-free grammar." In: *Bioinformatics* 19.16, pp. 2046–2053.
URL: https://doi.org/10.1093/bioinformatics/btg279 (cit. on pp. 95, 100).

The FlyBase Consortium (Jan. 2002). "The FlyBase database of the Drosophila genome projects and community literature." In: *Nucleic Acids Research* 30.1. Oxford University Press, pp. 106–108.
URL: https://doi.org/10.1093/nar/30.1.106 (cit. on p. 61).

The UniProt Consortium (Jan. 2017). "UniProt: the universal protein knowledgebase." In: *Nucleic Acids Research* 45.D1. Oxford University Press, p. D158.
URL: https://doi.org/10.1093/nar/gkw1099 (cit. on p. 34).

Tieleman, Tijmen and Geoffrey Hinton (2012). "Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude." In: *COURSERA: neural networks for machine learning*. COURSERA: neural networks for machine learning.
URL: https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf (cit. on pp. 67, 114).

Tran, Tung and Ramakanth Kavuluru (Oct. 2017). "Exploring a deep learning pipeline for the BioCreative VI Precision Medicine task." In: *BioCreative VI Workshop* (Bethesda, Maryland, USA), pp. 106–109.
URL: https://biocreative.bioinformatics.udel.edu/events/biocreative-vi/workshop/ (cit. on p. 69).

Trifan, Alina, Rui Antunes, Sérgio Matos, and Jose Luís Oliveira (Apr. 2020a). "Understanding depression from psycholinguistic patterns in social media texts." In: *42nd European Conference on Information Retrieval* (Online). Springer Nature, pp. 402–409.
URL: https://doi.org/10.1007/978-3-030-45442-5_50 (cit. on p. 8).

Trifan, Alina, Rui Antunes, and José Luís Oliveira (Oct. 2020b). "Machine learning for depression screening in online communities." In: *14th International Conference on Practical Applications of Computational Biology & Bioinformatics* (Online). Springer

Nature, pp. 102–111.
URL: https://doi.org/10.1007/978-3-030-54568-0_11 (cit. on p. 8).

Tripodi, Ignacio, Mayla Boguslav, Negacy Hailu, and Lawrence E. Hunter (Oct. 2017). "Knowledge-base-enriched relation extraction." In: *BioCreative VI Workshop* (Bethesda, Maryland, USA), pp. 163–166.
URL: https://biocreative.bioinformatics.udel.edu/events/biocreative-vi/workshop/ (cit. on p. 119).

Tsai, Chen-Tse and Dan Roth (Apr. 2016). "Concept grounding to multiple knowledge bases via indirect supervision." In: *Transactions of the Association for Computational Linguistics* 4. Association for Computational Linguistics, pp. 141–154.
URL: https://doi.org/10.1162/tacl_a_00089 (cit. on p. 40).

Tsatsaronis, George, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artiéres, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras (Apr. 2015). "An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition." In: *BMC Bioinformatics* 16.1. BioMed Central Ltd, p. 138.
URL: https://doi.org/10.1186/s12859-015-0564-6 (cit. on p. 36).

Tsoumakas, Grigorios, Ioannis Katakis, and Ioannis Vlahavas (2009). "Mining multi-label data." In: *Data Mining and Knowledge Discovery Handbook*. Ed. by Oded Maimon and Lior Rokach. Springer Nature, pp. 667–685.
URL: https://doi.org/10.1007/978-0-387-09823-4_34 (cit. on p. 31).

Tulkens, Stéphan, Simon Šuster, and Walter Daelemans (Aug. 2016). "Using distributed representations to disambiguate biomedical and clinical concepts." In: *15th Workshop on Biomedical Natural Language Processing* (Berlin, Germany). Association for Computational Linguistics, pp. 77–82.
URL: https://doi.org/10.18653/v1/W16-2910 (cit. on pp. 52, 53).

Turing, Alan Mathison (Oct. 1950). "Computing machinery and intelligence." In: *Mind* LIX.236. Oxford University Press, pp. 433–460.
URL: https://doi.org/10.1093/mind/LIX.236.433 (cit. on p. 11).

Uzuner, Özlem (July 2009). "Recognizing obesity and comorbidities in sparse data." In: *Journal of the American Medical Informatics Association* 16.4. Oxford University Press, p. 561.
URL: https://doi.org/10.1197/jamia.M3115 (cit. on p. 35).

Uzuner, Özlem, Yuan Luo, and Peter Szolovits (Sept. 2007). "Evaluating the state-of-the-art in automatic de-identification." In: *Journal of the American Medical Informatics Association* 14.5. Oxford University Press, pp. 550–563.
URL: https://doi.org/10.1197/jamia.M2444 (cit. on p. 35).

Uzuner, Özlem, Imre Solti, and Eithon Cadag (Sept. 2010). "Extracting medication information from clinical text." In: *Journal of the American Medical Informatics Association* 17.5. Oxford University Press, p. 514.
URL: https://doi.org/10.1136/jamia.2010.003947 (cit. on p. 35).

Uzuner, Özlem, Brett R. South, Shuying Shen, and Scott L. DuVall (Sept. 2011). "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text." In: *Journal of the American Medical Informatics Association* 18.5. Oxford University Press, pp. 552–556.
URL: https://doi.org/10.1136/amiajnl-2011-000203 (cit. on pp. 98, 104).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin (Dec. 2017). "Attention is all you need." In: *31st Conference on Neural Information Processing Systems (NIPS 2017)* (Long Beach, California, USA). Curran Associates, Inc., pp. 5998–6008.
URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need (cit. on p. 102).

Velavan, Thirumalaisamy P. and Christian G. Meyer (Mar. 2020). "The COVID-19 epidemic." In: *Tropical Medicine & International Health* 25.3. Wiley, pp. 278–280.
URL: https://doi.org/10.1111/tmi.13383 (cit. on p. 3).

Verga, Patrick and Andrew McCallum (Oct. 2017). "Predicting chemical protein relations with biaffine relation attention networks." In: *BioCreative VI Workshop* (Bethesda, Maryland, USA), pp. 187–189.
URL: https://biocreative.bioinformatics.udel.edu/events/biocreative-vi/workshop/ (cit. on pp. 102, 119).

Verga, Patrick, Emma Strubell, and Andrew McCallum (June 2018). "Simultaneously self-attending to all mentions for full-abstract biological relation extraction." In: *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (New Orleans, Louisiana, USA). Association for Computational Linguistics, pp. 872–884.
URL: https://doi.org/10.18653/v1/N18-1080 (cit. on p. 24).

Veronis, Jean and Nancy M. Ide (1990). "Word sense disambiguation with very large neural networks extracted from machine readable dictionaries." In: *COLING 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics.*
URL: https://aclanthology.org/C90-2067 (cit. on p. 39).

Vicente, Agustín and Ingrid L. Falkum (July 2021). *Polysemy*. In: *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.
URL: https://doi.org/10.1093/acrefore/9780199384655.013.325 (cit. on p. 37).

Voorhees, Ellen M. (Nov. 2004). "Overview of TREC 2004." In: *The Thirteenth Text Retrieval Conference (TREC 2004)* (Gaithersburg, Maryland, USA).
URL: https://trec.nist.gov/pubs/trec13/t13_proceedings.html (cit. on p. 61).

Wadden, David, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi (Sept. 2019). *Entity, relation, and event extraction with contextualized span representations*. arXiv:1909.03546.
URL: https://arxiv.org/abs/1909.03546 (cit. on p. 25).

Wang, C., A. Kalyanpur, J. Fan, B. K. Boguraev, and D. C. Gondek (May 2012). "Relation extraction and scoring in DeepQA." In: *IBM Journal of Research and Development* 56.3.4. IEEE, 9:1–9:12.
URL: https://doi.org/10.1147/jrd.2012.2187239 (cit. on p. 96).

Wang, Jiapeng and Yihong Dong (Aug. 2020). "Measurement of text similarity: a survey." In: *Information* 11.9. MDPI, p. 421.
URL: https://www.mdpi.com/2078-2489/11/9/421 (cit. on p. 63).

Wang, Qinghua, Shabbir S. Abdul, Lara Almeida, Sophia Ananiadou, Yalbi I. Balderas-Martínez, Riza Batista-Navarro, David Campos, Lucy Chilton, Hui-Jou Chou, Gabriela Contreras, Laurel Cooper, Hong-Jie Dai, Barbra Ferrell, Juliane Fluck, Socorro Gama-Castro, Nancy George, Georgios Gkoutos, Afroza K. Irin, Lars J. Jensen, Silvia Jimenez, Toni R. Jue, Ingrid Keseler, Sumit Madan, Sérgio Matos, Peter McQuilton, Marija Milacic, Matthew Mort, Jeyakumar Natarajan, Evangelos Pafilis, Emiliano Pereira, Shruti Rao, Fabio Rinaldi, Karen Rothfels, David Salgado, Raquel M. Silva, Onkar Singh, Raymund Stefancsik, Chu-Hsien Su, Suresh Subramani, Hamsa D. Tadepally, Loukia Tsaprouni, Nicole Vasilevsky, Xiaodong Wang, Andrew Chatr-Aryamontri, Stanley J. F. Laulederkind, Sherri Matis-Mitchell, Johanna McEntyre, Sandra Orchard, Sangya Pundir, Raul Rodriguez-Esteban, Kimberly Van Auken, Zhiyong Lu, Mary Schaeffer, Cathy H. Wu, Lynette Hirschman, and Cecilia N. Arighi (Jan. 2016a). "Overview of the interactive task in BioCreative V." In: *Database* 2016. Oxford University Press, baw119.
URL: https://doi.org/10.1093/database/baw119 (cit. on p. 106).

Wang, Wei, Xi Yang, Yuting Xing, Chengkun Wu, and Zhuo Song (Oct. 2017a). "Extracting chemical–protein interactions via bidirectional long short-term memory network." In: *BioCreative VI Workshop* (Bethesda, Maryland, USA), pp. 171–174.
URL: https://biocreative.bioinformatics.udel.edu/events/biocreative-vi/workshop/ (cit. on p. 119).

Wang, Wei, Xi Yang, Canqun Yang, Xiaowei Guo, Xiang Zhang, and Chengkun Wu (Dec. 2017b). "Dependency-based long short term memory network for drug–drug interaction extraction." In: *BMC Bioinformatics* 18.16. BioMed Central Ltd, p. 578.
URL: https://doi.org/10.1186/s12859-017-1962-8 (cit. on p. 101).

Wang, Yanshan, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu (Oct. 2018a). "MedSTS: a resource for clinical semantic textual similarity." In: *Language Resources and Evaluation*. Springer Nature.
URL: https://doi.org/10.1007/s10579-018-9431-1 (cit. on pp. 63, 83, 87).

Wang, Yanshan, Naveed Afzal, Sijia Liu, Majid Rastegar-Mojarad, Liwei Wang Feichen Shen, Sunyang Fu, and Hongfang Liu (Aug. 2018b). "Overview of BioCreative/OHNLP Challenge 2018 Task 2: clinical semantic textual similarity." In: *BioCreative/OHNLP Challenge 2018* (Washington, DC, USA).
URL: https://sites.google.com/view/ohnlp2018 (cit. on pp. 63, 83, 85, 92).

Wang, Yanshan, Sunyang Fu, Feichen Shen, Sam Henry, Ozlem Uzuner, and Hongfang Liu (Nov. 2020). "The 2019 n2c2/OHNLP track on clinical semantic textual similarity: overview." In: *JMIR Medical Informatics* 8.11. JMIR Publications, e23375.
URL: https://doi.org/10.2196/23375 (cit. on pp. 63, 84, 91).

Wang, Yanshan, Feichen Shen, Ravikumar Komandur Elayavilli, Sijia Liu, Majid Rastegar-Mojarad, and Hongfang Liu (Oct. 2017c). "MayoNLP at the BioCreative VI PM Track: entity-enhanced hierarchical attention neural networks for mining protein interactions from biomedical text." In: *BioCreative VI Workshop* (Bethesda, Maryland, USA), pp. 127–130.
URL: https://biocreative.bioinformatics.udel.edu/events/biocreative-vi/workshop/ (cit. on p. 69).

Wang, Yue, Kai Zheng, Hua Xu, and Qiaozhu Mei (Nov. 2016b). "Clinical word sense disambiguation with interactive search and classification." In: *AMIA Annual Symposium* (Chicago, Illinois, USA). American Medical Informatics Association, pp. 2062–2071.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5333264/ (cit. on p. 41).

Wang, Yue, Kai Zheng, Hua Xu, and Qiaozhu Mei (July 2018c). "Interactive medical word sense disambiguation through informed learning." In: *Journal of the American Medical Informatics Association* 25.7. Oxford University Press, pp. 800–808.
URL: https://doi.org/10.1093/jamia/ocy013 (cit. on p. 41).

Wang, Yuxing, Kaiyin Zhou, Mina Gachloo, and Jingbo Xia (Nov. 2019). "An overview of the active gene annotation corpus and the BioNLP OST 2019 AGAC track tasks." In: *5th Workshop on BioNLP Open Shared Tasks* (Hong Kong, China). Association for Computational Linguistics, pp. 62–71.
URL: https://doi.org/10.18653/v1/d19-5710 (cit. on p. 99).

Warikoo, Neha, Yung-Chun Chang, and Wen-Lian Hsu (Jan. 2018). "LPTK: a linguistic pattern-aware dependency tree kernel approach for the BioCreative VI CHEMPROT task." In: *Database* 2018. Oxford University Press, bay108.
URL: https://doi.org/10.1093/database/bay108 (cit. on pp. 108, 119).

Warikoo, Neha, Yung-Chun Chang, Po-Ting Lai, and Wen-Lian Hsu (Oct. 2017). "CTCPI — convolution tree kernel-based chemical–protein interaction detection." In: *BioCreative VI Workshop* (Bethesda, Maryland, USA), pp. 167–170.
URL: https://biocreative.bioinformatics.udel.edu/events/biocreative-vi/workshop/ (cit. on p. 119).

Weeber, M., J. G. Mork, and A. R. Aronson (Nov. 2001). "Developing a test collection for biomedical word sense disambiguation." In: *AMIA Symposium* (Washington, DC, USA). American Medical Informatics Association, pp. 746–750.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243574/ (cit. on p. 41).

Wei, Chih-Hsuan, Hung-Yu Kao, and Zhiyong Lu (July 2013). "PubTator: a web-based text mining tool for assisting biocuration." In: *Nucleic Acids Research* 41.W1. Oxford University Press, W518–W522.
URL: https://doi.org/10.1093/nar/gkt441 (cit. on p. 33).

Wei, Chih-Hsuan, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Jiao Li, Thomas C. Wiegers, and Zhiyong Lu (Sept. 2015). "Overview of the BioCreative V chemical disease relation (CDR) task." In: *BioCreative V Workshop* (Sevilla, Spain), pp. 154–166.
URL: https://biocreative.bioinformatics.udel.edu/resources/publications/bcv_proceedings/ (cit. on p. 98).

Wei, Chih-Hsuan, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Jiao Li, Thomas C. Wiegers, and Zhiyong Lu (Jan. 2016). "Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical–disease relation (CDR) task." In: *Database* 2016. Oxford University Press, baw032.
URL: https://doi.org/10.1093/database/baw032 (cit. on pp. 98, 101).

Weissenbacher, Davy, Karen O'Connor, Siddharth Rawal, and Graciela Gonzalez-Hernandez (Nov. 2021). "BioCreative VII – Task 3: automatic extraction of medication names in tweets." In: *BioCreative VII Challenge Evaluation Workshop* (Online), pp. 163–167.
URL: https://biocreative.bioinformatics.udel.edu/resources/publications/bc-vii-workshop-proceedings/ (cit. on p. 131).

Weissenbacher, Davy, Abeed Sarker, Ari Klein, Karen O'Connor, Arjun Magge, and Graciela Gonzalez-Hernandez (Dec. 2019). "Deep neural networks ensemble for detecting medication mentions in tweets." In: *Journal of the American Medical Informatics*

Association 26.12. Oxford University Press, pp. 1618–1626.
URL: https://doi.org/10.1093/jamia/ocz156 (cit. on p. 131).

Winnenburg, Rainer, Thomas Wächter, Conrad Plake, Andreas Doms, and Michael Schroeder (Nov. 2008). "Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies?" In: *Briefings in Bioinformatics* 9.6. Oxford University Press, pp. 466–478.
URL: https://doi.org/10.1093/bib/bbn043 (cit. on p. 99).

Wishart, David S., Yannick D. Feunang, An C. Guo, Elvis J. Lo, Ana Marcu, Jason R. Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson (Jan. 2018). "DrugBank 5.0: a major update to the DrugBank database for 2018." In: *Nucleic Acids Research* 46.D1. Oxford University Press, pp. D1074–D1082.
URL: https://doi.org/10.1093/nar/gkx1037 (cit. on pp. 34, 75).

Wu, Jionglin, Jason Roy, and Walter F. Stewart (June 2010). "Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches." In: *Medical Care* 48.6. Lippincott Williams & Wilkins, S106–S113.
URL: https://doi.org/10.1097/MLR.0b013e3181de9e17 (cit. on p. 82).

Wu, Po-Yen, Chih-Wen Cheng, Chanchala D. Kaddi, Janani Venugopalan, Ryan Hoffman, and May D. Wang (Oct. 2017). "–Omic and electronic health record big data analytics for precision medicine." In: *IEEE Transactions on Biomedical Engineering* 64.2. IEEE, pp. 263–273.
URL: https://doi.org/10.1109/tbme.2016.2573285 (cit. on p. 106).

Wu, Yonghui, Joshua Denny, S. Trent Rosenbloom, Randolph A. Miller, Dario A. Giuse, Min Song, and Hua Xu (Nov. 2013). "A prototype application for real-time recognition and disambiguation of clinical abbreviations." In: *7th International Workshop on Data and Text Mining in Biomedical Informatics* (San Francisco, California, USA). ACM, pp. 7–8.
URL: https://doi.acm.org/10.1145/2512089.2512096 (cit. on p. 41).

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean (Sept. 2016). *Google's neural machine translation system: bridging the gap between human and machine translation.* arXiv:1609.08144.
URL: https://arxiv.org/abs/1609.08144 (cit. on p. 13).

Wu, Yonghui, Jun Xu, Yaoyun Zhang, and Hua Xu (July 2015). "Clinical abbreviation disambiguation using neural word embeddings." In: *2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015)* (Beijing, China). Association for Computational Linguistics, pp. 171–176.
URL: https://doi.org/10.18653/v1/W15-3822 (cit. on p. 41).

Xiong, Ying, Shuai Chen, Yedan Shen, Xiaolong Wang, Qingcai Chen, Jun Yan, and Buzhou Tang (Aug. 2018). "A hybrid system for clinical semantic textual similarity." In: *BioCreative/OHNLP Challenge 2018* (Washington, DC, USA).
URL: https://sites.google.com/view/ohnlp2018 (cit. on p. 84).

Xue, Linting, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel (Mar. 2022). "ByT5: towards a token-free future with pre-trained byte-to-byte models." In: *Transactions of the Association for Computational Linguistics* 10. The MIT Press, pp. 291–306.
URL: https://doi.org/10.1162/tacl_a_00461 (cit. on p. 131).

Yang, Yiming (Apr. 1999). "An evaluation of statistical approaches to text categorization." In: *Information Retrieval* 1.1. Springer Nature, pp. 69–90.
URL: https://doi.org/10.1023/A:1009982220290 (cit. on p. 31).

Yang, Zichao, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy (June 2016). "Hierarchical attention networks for document classification." In: *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (San Diego, California, USA). Association for Computational Linguistics, pp. 1480–1489.
URL: http://doi.org/10.18653/v1/N16-1174 (cit. on pp. 19, 102).

Yarowsky, David (June 1995). "Unsupervised word sense disambiguation rivaling supervised methods." In: *33rd Annual Meeting of the Association for Computational Linguistics* (Cambridge, Massachusetts, USA). Association for Computational Linguistics, pp. 189–196.
URL: https://doi.org/10.3115/981658.981684 (cit. on p. 39).

Yaseen, Usama, Pankaj Gupta, and Hinrich Schütze (Nov. 2019). "Linguistically informed relation extraction and neural architectures for nested named entity recognition in BioNLP-OST 2019." In: *5th Workshop on BioNLP Open Shared Tasks* (Hong Kong, China). Association for Computational Linguistics, pp. 132–142.
URL: https://doi.org/10.18653/v1/d19-5720 (cit. on p. 96).

Yates, Andrew, Arman Cohan, and Nazli Goharian (Sept. 2017). *Depression and self-harm risk assessment in online forums*. arXiv:1709.01848.
URL: https://arxiv.org/abs/1709.01848 (cit. on p. 19).

Yeh, Alexander S., Lynette Hirschman, and Alexander A. Morgan (July 2003). "Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup." In: *Bioinformatics* 19 (suppl_1). Oxford University Press, pp. i331–i339.
URL: https://doi.org/10.1093/bioinformatics/btg1046 (cit. on pp. 61, 96).

Yüksel, Atakan, Hakime Öztürk, Elif Ozkirimli, and Arzucan Özgür (Oct. 2017). "CNN-based chemical–protein interactions classification." In: *BioCreative VI Workshop* (Bethesda, Maryland, USA), pp. 184–186.
URL: https://biocreative.bioinformatics.udel.edu/events/biocreative-vi/workshop/ (cit. on p. 119).

Zeng, Min, Min Li, Zhihui Fei, Ying Yu, Yi Pan, and Jianxin Wang (Jan. 2019). "Automatic ICD-9 coding via deep transfer learning." In: *Neurocomputing* 324. Elsevier, pp. 43–50.
URL: https://doi.org/10.1016/j.neucom.2018.04.081 (cit. on p. 62).

Zhang, Canlin, Daniel Biś, Xiuwen Liu, and Zhe He (Dec. 2019a). "Biomedical word sense disambiguation with bidirectional long short-term memory and attention-based neural networks." In: *BMC Bioinformatics* 20.16. Springer Nature, p. 502.
URL: https://doi.org/10.1186/s12859-019-3079-8 (cit. on pp. 52, 53).

Zhang, Dongxu, Sunil Mohan, Michaela Torkar, and Andrew McCallum (Apr. 2022a). *A distant supervision corpus for extracting biomedical relationships between chemicals, diseases and genes*. arXiv:2204.06584.
URL: https://arxiv.org/abs/2204.06584 (cit. on p. 105).

Zhang, Le, Jingbo Zhu, and Tianshun Yao (Dec. 2004). "An evaluation of statistical spam filtering techniques." In: *ACM Transactions on Asian Language Information Processing* 3.4. ACM, pp. 243–269.
URL: https://doi.org/10.1145/1039621.1039625 (cit. on p. 59).

Zhang, Yijia, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu (May 2019b). "BioWordVec, improving biomedical word embeddings with subword information and MeSH." In: *Scientific Data* 6.1. Springer Nature, p. 52.
URL: https://doi.org/10.1038/s41597-019-0055-0 (cit. on pp. 33, 84).

Zhang, Yijia, Hongfei Lin, Zhihao Yang, Jian Wang, and Yuanyuan Sun (May 2019c). "Chemical–protein interaction extraction via contextualized word representations and multihead attention." In: *Database* 2019. Oxford University Press.
URL: https://doi.org/10.1093/database/baz054 (cit. on p. 102).

Zhang, Yijia, Hongfei Lin, Zhihao Yang, Jian Wang, Yuanyuan Sun, Bo Xu, and Zhehuan Zhao (Nov. 2019d). "Neural network-based approaches for biomedical relation classification: a review." In: *Journal of Biomedical Informatics* 99. Elsevier, p. 103294.
URL: https://doi.org/10.1016/j.jbi.2019.103294 (cit. on p. 103).

Zhang, Yijia and Zhiyong Lu (Feb. 2019). "Exploring semi-supervised variational autoencoders for biomedical relation extraction." In: *Methods*. Elsevier.
URL: https://doi.org/10.1016/j.ymeth.2019.02.021 (cit. on p. 102).

Zhang, Yijia, Wei Zheng, Hongfei Lin, Jian Wang, Zhihao Yang, and Michel Dumontier (Mar. 2017). "Drug–drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths." In: *Bioinformatics* 34.5. Oxford University Press, btx659.
URL: https://doi.org/10.1093/bioinformatics/btx659 (cit. on pp. 101, 112).

Zhang, Yu, Jong Kang Lee, Jen-Chieh Han, and Richard Tzong-Han Tsai (Aug. 2022b). "Task reformulation and data-centric approach for Twitter medication name extraction." In: *Database* 2022. Oxford University Press, baac067.
URL: https://doi.org/10.1093/database/baac067 (cit. on p. 131).

Zhao, Sendong, Chang Su, Zhiyong Lu, and Fei Wang (May 2021). "Recent advances in biomedical literature mining." In: *Briefings in Bioinformatics* 22.3. Oxford University Press, bbaa057.
URL: https://doi.org/10.1093/bib/bbaa057 (cit. on p. 103).

Zheng, Suncong, Yuexing Hao, Dongyuan Lu, Hongyun Bao, Jiaming Xu, Hongwei Hao, and Bo Xu (Sept. 2017a). "Joint entity and relation extraction based on a hybrid neural network." In: *Neurocomputing* 257. Elsevier, pp. 59–66.
URL: https://doi.org/10.1016/j.neucom.2016.12.075 (cit. on p. 24).

Zheng, Suncong, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu (July 2017b). "Joint extraction of entities and relations based on a novel tagging scheme." In: *55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vancouver, British Columbia, Canada). Association for Computational Linguistics, pp. 1227–1236.
URL: https://doi.org/10.18653/v1/P17-1113 (cit. on p. 23).

Zhou, Huiwei, Huijie Deng, Long Chen, Yunlong Yang, Chen Jia, and Degen Huang (Jan. 2016). "Exploiting syntactic and semantics information for chemical–disease relation extraction." In: *Database* 2016. Oxford University Press, baw048.
URL: https://doi.org/10.1093/database/baw048 (cit. on p. 101).

Zhu, Fei, Preecha Patumcharoenpol, Cheng Zhang, Yang Yang, Jonathan Chan, Asawin Meechai, Wanwipa Vongsangnak, and Bairong Shen (Apr. 2013). "Biomedical text mining and its applications in cancer research." In: *Journal of Biomedical Informatics* 46.2. Elsevier, pp. 200–211.
URL: https://doi.org/10.1016/j.jbi.2012.10.007 (cit. on p. 32).

Zobel, Justin (2014). *Writing for computer science*. 3rd ed. Springer Nature.
URL: https://doi.org/10.1007/978-1-4471-6639-9 (cit. on p. 9).