Text Mining for Bioinformatics using Biomedical Literature

Andre Lamurias[a,b], Diana F. Sousa[a,*], Francisco M. Couto[a,]

[a]*LASIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisbon, Portugal*
[b]*NOVA LINCS, NOVA School of Science and Technology, Lisbon, Portugal*

Abstract

Biomedical literature is a large and rich source of information for various applications. Text mining tools aim at extracting information from the literature in an efficient manner since processing scientific texts is a complex task given the formal and highly specialized language. Text mining tools tackle these challenges using different approaches, such as rule-based methods and machine learning algorithms including deep learning. This document overviews the current biomedical text mining tools by describing their approaches, tasks (e.g. Named Entity Recognition, Relation Extraction, Event Extraction, Question Answering), available corpora, toolkits and applications, and community challenges.

Keywords: Biomedical Literature, Distant supervision, Event Extraction, Machine Learning, Named Entity Recognition, Normalization, Relation Extraction, Text Mining

1. Introduction

Biomedical literature is one of the primary sources of current biomedical knowledge. It is still the standard method researchers use to share their findings

*Accepted manuscript: https://doi.org/10.1016/B978-0-323-95502-7.00017-8
*Corresponding author
Email addresses: a.lamurias@fct.unl.pt (Andre Lamurias), dfsousa@lasige.di.fc.ul.pt (Diana F. Sousa), fjcouto@ciencias.ulisboa.pt (Francisco M. Couto)

in the form of articles, patents, and other types of written reports (Hearst, 1999). However, it is essential that a research group working on a given topic is aware of the work that has been done on the same topic by other research teams. This task requires manual effort and may take a long time to complete due to the large quantity of published literature. One of the largest sources of biomedical literature is the MEDLINE database, created in 1965 and accessible through PubMed. This database contains over 36 million references to journal articles in the life sciences, and more than 860,000 entries were added in 2022 [1]. Other document repositories are also relevant to biomedicine, such as the European Patent Office [2], ClinicalTrials.gov, and bioRxiv [3].

Automatic methods for Information Extraction (IE) aim at obtaining useful information from large datasets, where manual methods would be unfeasible. Text mining aims at using IE methods to process text documents. The main challenge of text mining is developing algorithms that can be applied to unstructured text to obtain valuable and structured information. Biomedical literature is particularly challenging for text mining algorithms. The writing style differs from other types of literature since it is more formal and specialized. Furthermore, different documents have different styles, depending on whether the document is a journal paper, patent, or clinical report (Friedman et al., 2002). Finally, different terms refer to genes, species, procedures, and techniques. Within each specific term, it is also common to have multiple spellings, abbreviations, and database identifiers. This complexity makes biomedical text mining a high-potential exploration field for developing IE tools (Cohen and Hunter, 2004).

The interactions found in the biomedical literature can be used to validate new research results or even to formulate new hypotheses to be tested experimentally. One of the first demonstrations of the hidden knowledge contained in

[1] `https://pubmed.ncbi.nlm.nih.gov/`

[2] `https://www.epo.org/searching-for-patents.html`

[3] https://www.biorxiv.org

a large literature was Swanson's ABC model (Swanson, 1990), which found that dietary fish oils might benefit patients with Raynaud's syndrome by connecting the information present in two different sets of articles that did not cite each other. Others have independently confirmed this inference in clinical trials (DiGiacomo et al., 1989). In the same study, Swanson provided two other examples of inferences that could not be drawn from a single article but only by combining the information from multiple articles. Since that study, the number of articles available has grown immensely. Intuitively, many new biomedical interactions might be extracted from this source of information.

Bioinformatics databases have adopted text mining tools to identify new entries more efficiently. MirTarBase (Huang et al., 2021) is a database of experimentally validated miRNA-target interactions published in journal papers. The curators of this database use a text mining system to identify new candidate entries for the database, which are then manually validated. This system was necessary due to the important role miRNAs have been found to play in human diseases over the last decade, leading to a high number of papers published about this subject. The introduction of the system as part of the workflow has led to a 42-fold increase in the number of interactions added to the database.

Text mining has generated much interest in the bioinformatics community in recent years. Several tools and applications have been developed based on adaptations of text mining techniques to diverse problems and domains. This paper provides a survey of biomedical text mining tools and applications that demonstrate the usefulness of text mining techniques. The rest of the paper consists of the following: Section 2 provides the basic concepts of text mining relevant to this article, Section 3 describes some toolkits that can be used to develop text mining tools, Section 4 describes the most used text mining tools, and Section 5 describes applications built using those tools that have been distributed to the general public. Section 6 provides a summary of the community challenges organized to evaluate biomedical text mining tools. Finally, Section 7 suggests future directions for biomedical text mining tools and applications, and Section 8 summarizes the article's main conclusions.

2. Background/Fundamentals

When developing and using text mining tools, it is necessary to define what type of information should be extracted. This decision will then influence the datasets to be considered, which text mining tasks to be explored, and which tools to be used. The objective of this section is to provide an overview of the options available to someone interested in developing a new text mining tool or using text mining for their work. The concepts introduced below are simple to understand and applicable to various problems.

2.1. NLP Concepts

Natural Language Processing (NLP) has been the focus of many researchers since the 1950's (Bates, 1995). The main difference between NLP and text mining is the objective of the tasks. While NLP techniques aim at making sense of the text, for example, determining its structure or sentiment, the objective of text mining tasks is to obtain concrete structured knowledge from text. However, there is an overlap between the two fields, and text mining tools usually use NLP concepts and tasks.

The following list defines NLP concepts relevant to text mining.

Token: a sequence of characters with some meaning, such as a word, number or symbol. The NLP task of identifying the tokens of a text is known as tokenization. It is of particular importance to text mining since most algorithms will not consider elements smaller than tokens [4].

Part-of-speech (POS): the lexical category of each token, for example, noun, adjective, or punctuation. The category imparts additional semantics to the tokens. Part-of-speech tagging is an NLP task that consists in classifying each token automatically.

[4] Recent text mining systems, such as BERT (Devlin et al., 2019), presented sub-tokens for vocabulary expansion

Lemma and stem: the base form of a word. The lemma represents the canonical form of the word, corresponding to a real word. The stem does not always correspond to a real word but only to the fragment of a word that never changes. For example, the lemma of the word "induces" is "induce" while the stem is "induc-".

Sentence splitting: the NLP task consists of identifying the sentence boundaries of a text. The methods used to accomplish this task should consider the difference between a period at the end of a sentence or at the end of an acronym or abbreviation. Breaking a document into sentences is desirable because they represent unique ideas. Although the context of the whole document is also important, extracting the knowledge of each sentence independently can provide useful results.

Entity: a segment of text with relevance to a specific domain. An entity may be composed of one or more tokens. Entity types relevant to biomedicine include genes, proteins, chemicals, cell lines, species, and biological processes.

2.2. Text Mining Tasks

Text mining tools focus on one or more text mining tasks. It is necessary to define these tasks properly so that it is possible to choose the type of tools that should be used for a given problem. Furthermore, these tasks and variants are used to evaluate the performance of a tool on community challenges. The text mining tasks presented here are common to all domains and sources of text, although the performance of the methods on different domains may differ, i.e., a method that has a good performance on patent documents may not perform as well on clinical reports due to the different characteristics of the text. The common final objective of these tasks, as to all text mining, is to extract useful knowledge from a high volume of written text, while the extracted knowledge can be useful for several applications, which will be described in section 5.

Topic modelling: the classification of documents according to their topics or themes. This task aims to organize a set of documents to identify which documents are more relevant to a given topic (Blei, 2012). Related tasks include document triage (Buchanan and Loizides, 2007) and document clustering.

Named Entity Recognition (NER): consists of identifying entities that are mentioned in the text. In most cases, the exact location of each entity in the text is required, given by the offset (position) of its first and last characters. In some cases, discontinuous entities may be considered, therefore requiring multiple offset pairs. The classification of entity properties such as its type (e.g., protein, cell line, chemical) can be included in this task (Nadeau and Sekine, 2007).

Normalization: consists of matching each entity to an identifier belonging to a knowledge base that unequivocally represents its concept. For example, a protein may be mentioned by its full name or an acronym; in this case, the normalization process should assign the same identifier to both occurrences. The identifiers can be provided by an external database or ontology (Tsuruoka et al., 2008). Related tasks include named entity disambiguation (Bunescu and Pasca, 2006), named entity linking (NEL), and harmonization.

Relation Extraction (RE): the identification of entities that participate in a relationship described in the text. Most tools consider relations between two entities in the same sentence, but some focus on n-ary relations (between more than two entities) across multiple sentences. Biomedical relations commonly extracted are gene-phenotype and drug-drug interactions, see (Segura-Bedmar et al., 2014), for example.

Event extraction: can be considered an extension of the relation extraction task, where the label of the relation and role of each participant is specified. The events extracted should represent the mechanisms described in

the text (Ananiadou et al., 2010). Related tasks include slot-filling and relation classification.

Question Answering (QA): is a task where we aim to automatically answer questions asked by humans in natural language using either an existing structured database or a collection of natural language documents (Calijorne Soares and Parreiras, 2020).

2.3. Text Mining Approaches

Text mining tools employ various approaches to accomplish the tasks described above. They may focus on one specific approach or combine several techniques according to their respective advantages, the latter being more common. Most approaches can also be adapted for performing more than one task.

Classic approaches: are approaches based on statistics that can be calculated on a large corpus of documents (Manning et al., 1999). Some of the most popular approaches are term frequency-inverse document frequency (tf-idf) for topic modelling and co-occurrence for relation extraction. These approaches preceded the popularization of machine learning algorithms, although most current approaches still have a statistical background.

Rule-based methods: consist of defining a set of rules to extract the desired information. These rules can be a list of terms, regular expressions or sentence structures. Due to the manual effort necessary to develop these rules, text mining tools based on this approach have limited applicability.

Machine learning (ML) algorithms: are used for automatically learning various tasks. In text mining, it is necessary to convert the text to a numeric representation, which is the expected input of these algorithms. Text mining tools using ML assemble models trained on a corpus that can subsequently be applied to other texts. In some cases, it may be possible to train additional models using other corpora. Several types of

ML approaches can be considered, for example, supervised learning, in which the labels of each instance of the training data are known and used to train the classifier, and unsupervised learning, in which the algorithm learns to classify the data without a labelled training set.

Distant supervision (DS): is a learning process which heuristically assigns labels to the data according to the information provided by a knowledge base. These annotations are prone to error, but using ML algorithms adapted to this method can provide effective classification models. Distant supervision is sometimes referred to as weak supervision.

Deep learning (DL): is an ML approach, based on artificial neural networks, that has become popular in the last years due to its performance in fields such as speech recognition, computer vision, and text mining (LeCun et al., 2015). In the case of text mining, deep learning is associated with word embeddings, which consist of vector representations of word frequencies that are used as inputs to the networks.

2.4. Biomedical Corpora

Biomedical corpora are necessary to develop and evaluate text mining tools. The simplest corpora consist of documents associated with a specific topic (e.g., disease, gene, or pathway). It is enough to know which documents are relevant for some tasks, such as simple topic modelling tasks. However, most ML algorithms require annotated text to train their models. The type of annotations necessary to evaluate a task should be similar to the type of annotations to be extracted by the tools. NER tasks require text annotated with relevant entities (e.g., BC5CDR (Li et al., 2016)), while relation extraction requires the relations between the entities described in the text to be annotated (e.g., PGR (Sousa et al., 2019, 2020)). Domain experts should manually curate the annotations according to established guidelines. Inter-annotator agreement measures, such as the kappa statistic Carletta (1996), can be used to assess the reliability of the

Table 1: Corpora relevant to biomedical text mining tasks

Name	Annotations	Document types	Reference
CRAFT	Biomedical entities	Full-text articles	Bada et al. (2012)
MedTag	Biomedical entities	PubMed abstracts	Smith et al. (2005)
Genia	Biomedical entities and events	PubMed abstracts	Kim et al. (2003)
CHEMDNER	Chemical compounds	PubMed abstracts	Krallinger et al. (2015a)
CHEMDNER-patents	Chemical compounds and proteins	Patent abstracts	Krallinger et al. (2015b)
BC5CDR	Chemicals, diseases, and chemical-disease interactions	PubMed abstracts	Li et al. (2016)
BioRED	Biomedical relations	PubMed abstracts	(Luo et al., 2022a)
PGR and PGR-crowd	Human phenotype-gene relations	PubMed abstracts	Sousa et al. (2019, 2020)
DDI	Drug-drug interactions	Drug descriptions and journal abstracts	Herrero-Zazo et al. (2013)
SeeDev	Seed development events	Full-text articles	Chaix et al. (2016)
Thyme	Events and time expressions	Clinical notes	Styler IV et al. (2014)
MLEE	Biological events	PubMed abstracts	Pyysalo et al. (2012)
BioASQ	Question-article pairs	PubMed articles	Tsatsaronis et al. (2015)
PubMedQA	Question-answer pairs	PubMed abstracts	Jin et al. (2019)
BiQA	Question-article pairs	Medical forums	Lamurias et al. (2020)

annotations. However, text mining tools may also help curators by providing automatic annotations as a baseline to be reviewed (Winnenburg et al., 2008).

The size of an annotated corpus is limited by the manual effort necessary to annotate the documents. More straightforward tasks, such as topic modelling, can be performed more quickly by human annotators, so developing an annotated corpus for this task is less expensive. Relation extraction requires that the annotators first identify the entities mentioned in the text and then the relationships described between the entities, which frequently requires multi-domain knowledge. For this reason, developing an annotated corpus for this task is more expensive. Biomedical text mining community challenges have contributed to releasing several annotated gold standards to evaluate different systems. Section 6 provides a summary of these challenges. Table 1 provides a list of annotated biomedical corpora relevant to various text mining tasks.

3. Text Mining Toolkits

Although biomedical text mining requires specialized approaches to deal with the characteristics of the biomedical literature, general text mining tools can be used as a starting point for more specialized approaches. These general tools can be adapted to specific domains by using models trained with biomedical datasets or by developing pre- and post-processing rules developed for this

type of text. Text mining toolkits are a type of software that can perform various NLP and text mining tasks. The objective of these toolkits is to provide general-purpose methods for performing various text mining tasks, which can be adapted to specific problems. Several toolkits can be used to pre-process the data, compare the performance of various tools and approaches, and select the best combination for a specific problem. This section provides a survey of well-known text mining toolkits used as frameworks for biomedical text mining tools. In addition to the toolkits presented here, tools can be developed from scratch using programming languages and libraries that implement specific algorithms.

One of the most widely used text mining toolkits is Stanford CoreNLP (Manning et al., 2014), which aggregates various tools developed by the Stanford NLP team for processing text data. Biomedical text mining tools may use Stanford CoreNLP to pre-process the data (e.g., for sentence splitting, tokenization, and co-reference resolution) and to generate features for machine learning classifiers (e.g., for POS tagging, lemmatization, and dependency parsing).

NLTK (Bird et al., 2009), another NLP toolkit, was implemented as a Python library. This toolkit provides interfaces to various NLP resources, such as WordNet, tokenizers, stopwords lists, and datasets from community challenges. It is often used by developers getting started in text mining due to its well-designed API and the availability of various online tutorials for this toolkit. SpaCy[5] is another Python-based toolkit, which has become a popular choice due to its focus computational performance and active development of new and improved features. This toolkit provides methods to easily prepare text data for state-of-the-art deep learning algorithms, and can also run on Graphical Processing Units (GPU).

ClearTK (Bethard et al., 2014) is a text mining toolkit based on machine learning and the Apache Unstructured Information Management Architecture (UIMA). This framework provides interfaces to several machine learning libraries and feature extractors.

[5] `https://spacy.io/`

GATE (Cunningham et al., 2013) is one of the few text mining toolkits with features specially designed for biomedical text mining. This toolkit provides plugins for bioinformatics resources such as Linked Life Data, other ontologies, and specialized biomedical NLP tools. Furthermore, it has a graphical user interface to visualize and edit the data and system architecture.

4. Biomedical Text Mining Tools

This section describes text mining tools commonly used in bioinformatics. These tools generally focus on one specific task, presenting novel approaches, and are evaluated on gold standards. We focus on tools described in the literature and freely available to the community. Even though the current trend is to make software available on code repositories such as GitHub, GitLab, and Bitbucket, this has not always been the case, and past works may not be accessible due to the privatization of the code. The tools described in this section have been used in community challenges. They may require considerable technical skill to apply to specific problems since the results provided by their developers often refer to gold standards rather than to real-world use cases. Usually, these tools are fine-tuned to work with English texts, but automatic translation techniques can be effective when using texts in other languages Campos et al. (2017). Table 2 provides a list of biomedical text mining tools available to the community.

4.1. NER and Normalization

Biomedical text mining tools can be organized in terms of the text mining tasks performed. The biomedical community challenges organized in the last two decades have motivated several teams to develop tools for bioinformatics and biomedical text mining. Initially, the main focus of these challenges has been on recognizing genes, proteins and chemical compounds mentioned in texts and linking those terms to databases. This focus leads to an imbalance in the quantity and variety of tools available for NER and normalization compared to other tasks.

Table 2: Text mining tools for bioinformatics and biomedical literature

Name	Tasks	Approaches	GUI	Reference
BANNER	NER	ML	N	Leaman et al. (2008)
ABNER	NER	ML	N	Settles (2005)
LingPipe	NER and Topic Modelling	ML and Rule-based	N	Carpenter (2007)
GNormPlus	NER and Normalization	ML	N	Wei et al. (2015)
DNorm	NER and Normalization	ML	N	Leaman et al. (2013)
tmChem	NER	ML	N	Leaman et al. (2015)
tmVar	NER	ML	N	Wei et al. (2013)
GENIA tagger	NER and POS tagging	ML	N	Tsuruoka and Tsujii (2005)
GENIA sentence splitter	Sentence splitting	ML	N	Sætre et al. (2007)
Acronime	Abbreviation resolution	Rule-based	Y	Okazaki and Ananiadou (2006)
@Note	NER, document retrieval	ML	Y	Lourenço et al. (2009)
MetaMap	NER and Normalization	Rule-based	Y	Aronson and Lang (2010)
LDPMap	Normalization	Rule-based	N	Ren et al. (2014)
SimSem	Normalization	ML and Rule-based	N	Stenetorp et al. (2011)
MER	NER	Rule-based	N	Couto et al. (2017)
IBEnt	NER and Normalization	ML and Rule-based	N	Lobo et al. (2017)
cTakes	NER, normalization, and RE	Rule-based	Y	Savova et al. (2010)
Neji	NER and Normalization	ML and Rule-based	Y	Campos et al. (2015)
jSRE	RE	ML	N	Giuliano et al. (2006)
DeepDive	RE	ML/DS	N	Zhang (2015)
IBRel	RE	ML/DS	N	Lamurias et al. (2017)
TEES	Event extraction	ML and Rule-based	N	Björne et al. (2011)
VERSE	Event extraction	ML	N	Lever and Jones (2016)
EventMine	Event extraction	ML	Y	Miwa et al. (2013)
Textpresso	NER and RE	Rule-based	Y	Müller et al. (2004)
BO-LSTM	RE	ML and Rule-based	N	Lamurias et al. (2019)
BiOnt and K-BiOnt	RE	ML and Rule-based	N	Sousa and Couto (2020, 2022)
K-RET	RE	ML	N	Sousa and Couto (2023)
MedQA	QA	Rule-based	Y	Lee et al. (2006)
askHERMES	QA	Rule-based	Y	Cao et al. (2011)
HealthQA	QA	ML	N	Zhu et al. (2019)
JPDRMM	QA	ML	N	Pappas et al. (2020)
SciBERT	Representation learning	ML	N	Beltagy et al. (2019)
BioBERT	Representation learning	ML	N	Lee et al. (2020)
PubMedBERT	Representation learning	ML	N	Gu et al. (2021)
BioGPT	Representation learning	ML	N	Luo et al. (2022b)

BANNER (Leaman et al., 2008) uses Conditional Random Fields (Sutton and McCallum, 2006) to perform NER of chemical compounds and genes. ABNER (Settles, 2005) and LingPipe (Carpenter, 2007) use similar approaches, each combining different techniques to improve the results on gold standards by optimizing the system architecture and feature selection. LingPipe also performs other NLP tasks, such as topic modelling and part-of-speech tagging, while all three provide ways to train models on new data. Other systems have combined machine learning algorithms and manual rules to achieve better results in the biomedical domain (Savova et al., 2010; Campos et al., 2015; Lobo et al., 2017).

GNormPlus (Wei et al., 2015) is a modular system for gene NER and normalization, performing mention simplification and abbreviation resolution to match each gene to an identifier with higher accuracy, even when more than one species is involved. It is part of a set of NER tools developed by NCBI for various entity types, which includes tmChem (Leaman et al., 2015), DNorm (Leaman et al., 2013) and tmVar (Wei et al., 2013). These tools are often evaluated in text mining community challenges.

The GENIA project is responsible for various contributions to biomedical text mining, including an annotated corpus (Kim et al., 2003) and various tools for text mining tasks. GENIA tagger (Tsuruoka and Tsujii, 2005) performs NER of several types of entities relevant to biomedicine (protein, DNA, RNA, cell line and cell types), as well as POS tagging. GENIA sentence splitter (Sætre et al., 2007) is an ML-based tool for identifying sentence boundaries in biomedical texts, trained on the GENIA corpus. Acromine (Okazaki and Ananiadou, 2006) is another tool developed by the same team to provide definitions for abbreviations found in MEDLINE abstracts.

Since the vocabulary used in clinical records differs from other biomedical texts, tools have been developed specifically for this type of document. These tools are based on the Unified Medical Language System (UMLS), a collection of vocabulary associated with the clinical domain. cTakes Savova et al. (2010) is a Java-based tool for processing clinical text, originally developed at the Mayo

Clinic, which performs several biomedical text mining tasks. It is possible to use this tool through a graphical user interface. Due to UMLS's large size and complex structure, tools have been specifically developed just to find UMLS concepts in documents. Such tools include MetaMap (Aronson and Lang, 2010) and LDPMap (Ren et al., 2014). SimSem (Stenetorp et al., 2011) is a tool for entity normalization, using string matching techniques and machine learning. This tool can match strings to various bioinformatics knowledge bases, such as ChEBI, Gene Ontology, Entrez Gene, and UMLS. Couto et al. (2017) introduced a system, MER (Minimal Entity Recognizer), which can be easily adapted to different entities. This system requires only a file with one entity per line and uses a simple matching algorithm to find those entities in text.

Ruas and Couto (2022) created a model to associate NIL (out-of-knowledge-base or unlinkable) entities, such as diseases, chemicals, anatomical parts, and biological processes, with the best available entry in biomedical knowledge bases such as MEDIC, CTD-Chemical, CTD-Anatomy, Gene Ontology - Biological Process, ChEBI ontology, and Human Phenotype Ontology.

4.2. Relation and Event Extraction

Most tools use ML algorithms for RE to classify which pairs of entities mentioned in the text constitute a relation. In this task, kernel methods and Support Vector Machines were initially popular. jSRE (Giuliano et al., 2006) uses a shallow linguistic kernel that considers the tokens, POS, and lemmas around each entity of the pair. It has been used for various problems, including drug-drug interaction extraction (Segura-Bedmar et al., 2011).

Distant supervision has become particularly relevant to RE tasks because of the cost of developing a corpus annotated with relations. Mallory et al. (2016) developed an approach to gene RE using DeepDive, a general-purpose system for training distantly supervised RE models. They applied this approach to a corpus of full-text documents from three journals, using the BioGRID and Negatome databases as a reference. Another DS-based tool, IBRel (Lamurias et al., 2017), uses TransmiR, a database of miRNA-gene associations, to extract

the same type of relations from text.

BO-LSTM (Lamurias et al., 2019), BiOnt (Sousa and Couto, 2020), and K-BiOnt (Sousa and Couto, 2022) are systems based on bidirectional long short-term memory networks (LSTM) allied with the addition of knowledge external to the training data itself, such as domain-ontologies. The knowledge is linked to the entities in the candidate relation, and entities can be of various types, such as human phenotypes and genes. Moreover, K-RET (Sousa and Couto, 2023) performs entity knowledge injection directly into the text data, taking advantage of the latest advancements in pre-trained language representation models.

Biomedical event extraction is complex, but some tools have been developed. TEES (Turku Event Extraction System) (Björne et al., 2011) identifies complex events based on trigger words and graph methods. This system has been evaluated on multiple community challenges for event extraction and RE tasks, such as the BioNLP-ST 2011 event extraction task. In the 2016 edition of BioNLP-ST, Lever and Jones (2016) presented VERSE, a system for extracting relations and events from text, and evaluated it on three subtasks. This system is based on ML algorithms and has the advantage of being able to extract relations between entities in different sentences.

Textpresso (Müller et al., 2004) is a system for biomedical information extraction based on regular expressions and ontologies. This system has been applied to various domains. It is available through a web application to search the results obtained on each domain.

4.3. Question Answering

QA approaches are varied in that they can answer a question in multiple forms that require different tool architectures. An answer can be a simple YES/NO/MAYBE, a sentence retrieved from a document or a sentence resulting from multiple-text processing, or a relevant document or set of documents.

Biomedical QA was initially tackled by using rule-based models and other complex modular pipelines on small-scale datasets (Jin et al., 2022). MedQA

(Lee et al., 2006) and askHERMES (Cao et al., 2011) are examples of traditional QA approaches. MedQA integrates information retrieval, extraction, and summarizing techniques that answer user questions by generating paragraph-size answers. askHERMES implements innovative approaches in question analysis, summarization, and answer presentation in a sentence format by naturally entering a question.

Recent approaches are mostly ML-based, such as HealthQA (Zhu et al., 2019) and JPDRMM (Pappas et al., 2020). HealthQA focuses on recommending relevant documents for the question proposed using a deep attention mechanism at word, sentence, and document levels for the retrieval of both factoid and non-factoid queries. JPDRMM is a neural re-ranking model that receives the top N documents retrieved by a conventional information retrieval engine and is trained to jointly re-rank the top N documents and their snippets to produce an answer.

4.4. Representation Learning

One of the latest pushes towards improving biomedical NER, normalization and several other NLP tasks, such as RE and QA, were contextualized word embeddings generated by pre-trained language models, also known as Large Language Models (LLMs). SciBERT (Beltagy et al., 2019), BioBERT (Lee et al., 2020), PubMedBERT (Gu et al., 2021), and BioGPT (Luo et al., 2022b) are examples that specifically target the biomedical domain, but several other models/LLMs trained or fine-tuned on biomedical corpora exist (Kim et al., 2019).

There are also general NLP tools that do not target the biomedical domain specifically but have gained traction given the possibility of using them without the need for domain adaption, which are mostly LLMs such as PaLM 2 (Anil et al., 2023) and GPT (Brown et al., 2020). We call this usage zero- or few-shot learning, where we make predictions on data different from the data used for training. While some works have demonstrated their ability to perform simple biomedical tasks, most still struggle to grasp the full complexity of the domain

Table 3: Bioinformatics applications that either use text mining tools or their results, accessible from the web

Name	API	Reference
Whatizit	Y	Rebholz-Schuhmann et al. (2008)
BeCAS	Y	Nunes et al. (2013)
miRetrieve	N	Friedrich et al. (2021)
MER	Y	Couto et al. (2017)
PubTator	Y	Wei et al. (2019)
SciLite	Y	Venkatesan et al. (2016)
BEST	N	Lee et al. (2016)
STRING	Y	Szklarczyk et al. (2019)
STITCH	Y	Szklarczyk et al. (2016)
FACTA+	N	Tsuruoka et al. (2011)
PolySearch2	Y	Liu et al. (2015)
EVEX	Y	Hakala et al. (2013)
MEDIE	N	Miyao et al. (2006)

and perform better when fine-tuned (Jahan et al., 2023). It is also necessary to guarantee a rigorous validation and verification protocol of these tools, given the inherent risk of misinformation, lack of transparency, and biased interpretations prevalent in generative artificial intelligence.

5. Applications

Even though it is important to develop methods for specific tasks, those methods will only benefit the community if they can be easily used to help address biomedical problems. Since recent text mining tools have obtained good performance on evaluation corpora, efforts have been made to deliver these tools to the general public. In this section, we present a survey of text mining applications available in the form of web pages and APIs focusing on the user experience. Table 5 provides a summary of these applications.

Some biomedical text mining applications provide access to a text mining tool via a web application. The user uploads one or more documents to be processed by the tool on a server. Then, the tool delivers or displays the results to the user. Even though this is a significant effort, a web application assumes user pre-selection of documents to be processed and depends on downstream applications to use the results. Whatizit (Rebholz-Schuhmann et al., 2008) is a text mining application that can be used to identify biomedical entities in text using a web browser or API. This application is a rule-based text mining system which annotates the documents submitted by users. The entities correspond to entries in biomedical knowledge bases, such as ChEBI and UniProt. The results are presented as a web page, where each entity type is marked differently. A similar application is BeCAS (Nunes et al., 2013), based on the Neji tool. This application can also access the results through a web browser or the API, which can then be exported to various file formats. miRetrieve (Friedrich et al., 2021) is an R package and web application for miRNA text mining. Designed for PubMed abstracts, this tool is able to extract miRNA names and their associated terms.

Other text mining applications provide pre-processed results, reducing the time necessary to obtain results. For example, PubTator (Wei et al., 2019) contains every PubMed abstract, annotated with the NCBI NER tools, and it is updated as new abstracts are added to PubMed. Users can search for a list of abstracts or by keyword. Also, it is possible to create a collection of abstracts, manually fix annotation errors, and download the results. PubTator provides access to the results through an API for integration with other applications. For example, the Mark2Cure crowdsourcing project uses this API to provide users with a baseline of automatic annotations. At the same time, the HuGE navigator knowledge base (Yu et al., 2008) relies on PubTator to improve its weekly update process. Another application based on pre-processed results is SciLite, a platform for displaying text mining annotations integrated with the Europe PMC database (Venkatesan et al., 2016). This application shows a list of biomedical terms associated with each document, allowing users to endorse

and report incorrect annotations to improve the text mining method. Biomedical Entity Search Tool (BEST) (Lee et al., 2016) uses text mining techniques to retrieve entities relevant to user queries. BEST is updated daily with the abstracts added to PubMed and can identify up to ten types of entities in each document.

The STRING database stores information about protein-protein interaction networks (Szklarczyk et al., 2017). It contains information obtained through various methods, including text mining. The interactions extracted using text mining methods are obtained from PubMed and a collection of full-text documents. The RE method used is based on the co-occurrence of proteins in the same document and the presence of trigger words such as "binding" and "phosphorylation by". A related database, STITCH (Szklarczyk et al., 2016), uses a similar method to identify chemical-protein interactions based on the biomedical literature.

FACTA+ (Tsuruoka et al., 2011) is a text mining application for identifying biomedical events described in PubMed abstracts. It uses both co-occurrence and machine learning approaches to extract relations from text. The user can perform a keyword search to obtain associated documents and biomedical entities, such as genes, diseases, and drugs. Furthermore, FACTA+ can identify indirect relations between a concept and a type of biomedical entity. For example, it is possible to search for a disease name and obtain genes indirectly associated with that disease through an intermediary disease, ranked by a novelty and reliability score.

PolySearch2 (Liu et al., 2015) can also identify relations between biomedical concepts based on co-occurrence at the sentence level. With this application, it is possible to obtain all the entities of a specific type associated with the input query. The corpora and databases used by this application are stored locally and updated daily to ensure that the complete information is available to the users.

EVEX (Hakala et al., 2013) is a database of biomolecular events extracted from abstracts and full-text articles using text mining tools such as BANNER

and TEES. This database contains more than 40 million associations between genes and proteins, and its data can be downloaded and accessed through an API, although it is not updated regularly. MEDIE (Miyao et al., 2006) contains biomolecular events extracted from MEDLINE. Each event comprises a subject, a verb, and an object. Using MEDIE, it is possible to search by subject, verb or object (or a combination of the three) and obtain all matching events extracted from the abstracts.

Recently, there has not been much focus on transitioning from building a tool to making it available via a web page or API. This lack of focus on direct usage is rooted primarily in the increasing facilitation of direct application of tools via code repositories and Docker images. However, with the rise and popularity of generative Artificial Intelligence (AI) based web applications, such as BARD[6] and ChatGPT[7], the integration of biomedical tools into applications has appeal to more companies (Song, 2023), with Viz.ai[8] and PathAI[9] being just some of the examples. These applications come with the caveat of being black boxes; in a sense that most lack the support of a research paper, the source code and model weights are not shared, and they can only be used through an API, whose results may differ with time for the same input and cannot be run locally.

6. Community Challenges

The scientific community organises text mining challenges regularly to evaluate the performance of text mining tools. These text mining challenges are open to the community, meaning that any academic or industry team can participate. Usually, each challenge comprises several tasks (sometimes called tracks), each with a specific motivation, objective and gold standard. Each team may submit results to one or more tasks. Furthermore, the teams may develop their own tools or adapt existing tools to the proposed task.

[6] https://bard.google.com/
[7] https://chat.openai.com/
[8] https://www.viz.ai/
[9] https://www.pathai.com/

The task organizers announce their objectives on the challenge's official websites and through mailing lists. Since there are various data file formats used in text mining, a sample of the data may be provided to the participants simultaneously with the announcement. This is also the case of datasets that require data use agreements. Afterwards, the training set is provided to the participants, consisting of documents and annotations. This training set is used to develop or adapt tools and systems to the task. A development set may also be provided, similar in size to the training set, to improve the systems further. During the final phase of the challenge, a testing set is sent to the teams without the gold standard annotations. The teams have a deadline to submit the annotations obtained with their tools, which are then compared to the gold standard by the organizers. Each task has a defined set of measures to perform this evaluation and rank the teams. The results are then published on the challenge website and in a task overview paper.

One of the earliest NLP challenges, TREC, mainly focuses on the news domain, but it has included a bioinformatics task in some of its editions (TREC Genomics and TREC Chemistry). In 2003, this challenge had a task for retrieving documents related to gene functions (Hersh and Bhupatiraju, 2003), while in later years, more complex tasks have also been proposed (Hersh and Voorhees, 2009). Other NLP challenges, such as KDD Cup (Yeh et al., 2002) and CoNLL (Farkas et al., 2010), also include bioinformatics tasks. SemEval is a series of semantic analysis evaluations organized yearly, and in some editions, there has been one task relevant to bioinformatics (Segura Bedmar et al., 2013; Elhadad et al., 2015; Bethard et al., 2016).

Due to increasing interest in biomedical NLP and text mining, community challenges for this domain have been organized. BioCreative was first organized in 2004, and it consisted of identifying gene mentions and Gene Ontology terms in articles and gene name normalization (Hirschman et al., 2005). Since then, six more editions of this challenge have been organized, with various tasks. BioNLP has organized various biomedical IE tasks, usually focused on a specific biological system such as seed development (Chaix et al., 2016),

epigenetics and post-translational modifications (Ohta et al., 2011), cancer genetics (Pyysalo et al., 2015), and general clinical NLP (Demner-Fushman et al., 2022). Other community challenges relevant to biomedical text mining include JNLPBA (Kim et al., 2004), BioASQ (Nentidis et al., 2022), i2b2 (Sun et al., 2013), and ShARe/CLEF eHealth (Kelly et al., 2014; Nakov et al., 2022). Huang and Lu (2016) provides an overview of the community challenges organized over 12 years.

7. Future Directions

While recent advances in biomedical text mining are exciting and open new opportunities for data exploration and comprehension, the emergence of deep learning methods catapulted the necessity for explainability (Frisoni et al., 2021). The lack of explainability of DL models makes it hard to trust predictions, specifically when these target the highly complex biomedical domain. The injection of knowledge into DL systems can contribute to more explainable predictions (Moradi and Samwald, 2021; Aisopos and Paliouras, 2023). However, we can still not fully follow the path from input to prediction when dealing with DL. The path forward passes by not only explaining predictions but also defining what degree of explainability is adequate and necessary for each end-user.

Further, ethical AI is a field that aims that AI systems go towards greater ecological integrity and social justice (van Wynsberghe, 2021; Vinuesa et al., 2020). A study by Strubell et al. (2019) illustrated that training a single DL, NLP model can lead to approximately 600,000 lb of carbon dioxide emissions. Reproducibility is also an ongoing issue in the field (Digan et al., 2021), given that there are systems whose code is not shared. Frequently, authors justify this choice through privacy claims. However, there should always be a way to at least partially replicate the system. All community members' continued sharing and discussion of open-sourced systems contribute to significant advances in the field. These and other concerns, such as biased datasets (Bender and Friedman, 2018; Ray, 2023), should be addressed and prioritised by the groups working on

biomedical text mining.

8. Closing Remarks

There has been considerable effort by the text mining community to develop and release efficient tools and applications for helping biomedical researchers find what they need from the vast amount of knowledge being published. This article presents various text mining tools and applications using different approaches and addressing different tasks, and successfully real-world use cases. The evolution of biomedical text mining methods has led to more efficient processing of biomedical literature. These advances should affect how databases are created and maintained and how search engines index documents.

Besides its complexity, biomedical knowledge is constantly evolving and is highly dependent on the context, i.e. it is hard to find a generic answer to a biomedical question that perfectly fits the need of all users. Thus, like in precision medicine, future bioinformatics search engines may take into account the context of the user and provide him with multiple scientifically and ethically medical perspectives that help him decide the best solution for his problem.

9. Acknowledgement

This work has been supported by FCT through Deep Semantic Tagger (DeST) Project under Grant PTDC/CCIBIO/28685/2017 (http://dest.rd.ciencias.ulisboa.pt/), in part by LASIGE Research Unit under Grants UID/CEC/00408/2013, UIDB/00408/2020, and UIDP/00408/2020, and in part by FCT and FSE through Ph.D. Scholarships under Grants PD/BD/106083/2015 and SFRH/BD/145221/2019.

References

Aisopos, F., Paliouras, G., 2023. Comparing methods for drug–gene interaction prediction on the biomedical literature knowledge graph: performance versus explainability. BMC bioinformatics 24, 272.

Nomenclature

AI	Artificial Intelligence
DL	Deep Learning
DS	Distant Supervision
GPU	Graphical Processing Unit
IE	Information Extraction
ML	Machine Learning
NER	Named Entity Recognition
NLP	Natural Language Processing
POS	Part-of-speech
QA	Question Answering
RE	Relation Extraction
$UMLS$	Unified Medical Language System

Ananiadou, S., Pyysalo, S., Tsujii, J., Kell, D.B., 2010. Event extraction for systems biology by text mining the literature. Trends in Biotechnology 28, 381–390. doi:10.1016/j.tibtech.2010.04.005.

Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al., 2023. Palm 2 technical report. arXiv preprint arXiv:2305.10403 .

Aronson, A.R., Lang, F.M., 2010. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association 17, 229–236.

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W.A., Cohen, K.B., Verspoor, K., Blake, J.A., et al., 2012. Concept annotation in the CRAFT corpus. BMC bioinformatics 13, 161.

Bates, M., 1995. Models of natural language understanding. Proceedings of the National Academy of Sciences 92, 9977–9982.

Beltagy, I., Lo, K., Cohan, A., 2019. SciBERT: A pretrained language model for scientific text, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3615–3620.

Bender, E.M., Friedman, B., 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics 6, 587–604.

Bethard, S., Ogren, P., Becker, L., 2014. ClearTK 2.0: Design patterns for machine learning in UIMA, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland. pp. 3289–3293. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/218_Paper.pdf.

Bethard, S., Savova, G., Chen, W.T., Derczynski, L., Pustejovsky, J., Verhagen, M., 2016. Semeval-2016 task 12: Clinical tempeval. Proceedings of SemEval , 1052–1062.

Bird, S., Klein, E., Loper, E., 2009. Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.".

Björne, J., Heimonen, J., Ginter, F., Airola, A., Pahikkala, T., Salakoski, T., 2011. Extracting contextualized complex biological events with rich graph-based feature sets. Computational Intelligence 27, 541–557.

Blei, D.M., 2012. Probabilistic topic models. Communications of the ACM 55, 77–84.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901.

Buchanan, G., Loizides, F., 2007. Investigating document triage on paper and electronic media. Research and Advanced Technology for Digital Libraries , 416–427.

Bunescu, R.C., Pasca, M., 2006. Using encyclopedic knowledge for named entity disambiguation., in: Eacl, pp. 9–16.

Calijorne Soares, M.A., Parreiras, F.S., 2020. A literature review on question answering techniques, paradigms and systems. Journal of King Saud University - Computer and Information Sciences 32, 635–646. doi:https://doi.org/10.1016/j.jksuci.2018.08.005.

Campos, D., Matos, S., Oliveira, J.L., 2015. A document processing pipeline for annotating chemical entities in scientific documents. J. Cheminformatics 7, S7.

Campos, L., Pedro, V., Couto, F., 2017. Impact of translation on named-entity recognition in radiology texts. Database 2017.

Cao, Y., Liu, F., Simpson, P., Antieau, L., Bennett, A., Cimino, J.J., Ely, J., Yu, H., 2011. AskHERMES: An online question answering system for complex clinical questions. Journal of biomedical informatics 44, 277–288.

Carletta, J., 1996. Assessing agreement on classification tasks: the kappa statistic. Computational linguistics 22, 249–254.

Carpenter, B., 2007. LingPipe for 99.99% recall of gene mentions, in: Proceedings of the Second BioCreative Challenge Evaluation Workshop, pp. 307–309.

Chaix, E., Dubreucq, B., Fatihi, A., Valsamou, D., Bossy, R., Ba, M., Deléger, L., Zweigenbaum, P., Bessieres, P., Lepiniec, L., et al., 2016. Overview of

the regulatory network of plant seed development (seedev) task at the bionlp shared task 2016, in: Proceedings of the 4th BioNLP shared task workshop. Berlin: Association for Computational Linguistic, pp. 1–11.

Cohen, K.B., Hunter, L., 2004. Natural language processing and systems biology, in: Artificial intelligence methods and tools for systems biology. Springer, pp. 147–173.

Couto, F., Campos, L., Lamurias, A., 2017. MER: a minimal named-entity recognition tagger and annotation server, in: BioCreative V.5 Challenge Evaluation.

Cunningham, H., Tablan, V., Roberts, A., Bontcheva, K., 2013. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. PLoS Comput Biol 9, e1002854.

Demner-Fushman, D., Cohen, K.B., Ananiadou, S., Tsujii, J., 2022. Proceedings of the 21st workshop on biomedical language processing, in: Proceedings of the 21st Workshop on Biomedical Language Processing.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186.

Digan, W., Névéol, A., Neuraz, A., Wack, M., Baudoin, D., Burgun, A., Rance, B., 2021. Can reproducibility be improved in clinical natural language processing? a study of 7 clinical NLP suites. Journal of the American Medical Informatics Association 28, 504–515.

DiGiacomo, R.A., Kremer, J.M., Shah, D.M., 1989. Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study. The American journal of medicine 86, 158–164.

Elhadad, N., Pradhan, S., Chapman, W., Manandhar, S., Savova, G., 2015. Semeval-2015 task 14: Analysis of clinical text, in: Proc of Workshop on Semantic Evaluation. Association for Computational Linguistics, pp. 303–10.

Farkas, R., Vincze, V., Móra, G., Csirik, J., Szarvas, G., 2010. The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text, in: Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task, Association for Computational Linguistics. pp. 1–12.

Friedman, C., Kra, P., Rzhetsky, A., 2002. Two biomedical sublanguages: a description based on the theories of Zellig Harris. Journal of biomedical informatics 35, 222–235.

Friedrich, J., Hammes, H.P., Krenning, G., 2021. miRetrieve—an r package and web application for mirna text mining. NAR genomics and bioinformatics 3, lqab117.

Frisoni, G., Moro, G., Carbonaro, A., 2021. A survey on event extraction for natural language understanding: Riding the biomedical literature wave. IEEE Access 9, 160721–160757.

Giuliano, C., Lavelli, A., Romano, L., 2006. Exploiting shallow linguistic information for relation extraction from biomedical literature., in: EACL, Citeseer. pp. 401–408.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H., 2021. Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) 3, 1–23.

Hakala, K., Van Landeghem, S., Salakoski, T., Van de Peer, Y., Ginter, F., 2013. EVEX in ST'13: Application of a large-scale text mining resource to event extraction and network construction, in: Proceedings of the BioNLP

Shared Task 2013 Workshop, Association for Computational Linguistics. pp. 26–34.

Hearst, M.A., 1999. Untangling text data mining, in: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, Association for Computational Linguistics. pp. 3–10.

Herrero-Zazo, M., Segura-Bedmar, I., Martínez, P., Declerck, T., 2013. The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. Journal of biomedical informatics 46, 914–920.

Hersh, W., Voorhees, E., 2009. TREC genomics special issue overview. Information Retrieval 12, 1–15.

Hersh, W.R., Bhupatiraju, R.T., 2003. TREC genomics track overview., in: TREC, pp. 14–23.

Hirschman, L., Yeh, A., Blaschke, C., Valencia, A., 2005. Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC bioinformatics 6, S1.

Huang, C.C., Lu, Z., 2016. Community challenges in biomedical text mining over 10 years: Success, failure and the future. Briefings in Bioinformatics 17, 132–144. doi:10.1093/bib/bbv024.

Huang, H.Y., Lin, Y.C.D., Cui, S., Huang, Y., Tang, Y., Xu, J., Bao, J., Li, Y., Wen, J., Zuo, H., et al., 2021. miRTarBase update 2022: an informative resource for experimentally validated miRNA–target interactions. Nucleic Acids Research 50, D222–D230. doi:10.1093/nar/gkab1079.

Jahan, I., Laskar, M.T.R., Peng, C., Huang, J., 2023. Evaluation of ChatGPT on biomedical tasks: A zero-shot comparison with fine-tuned generative transformers, in: Demner-fushman, D., Ananiadou, S., Cohen, K. (Eds.), The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared

Tasks, Association for Computational Linguistics, Toronto, Canada. pp. 326–336. URL: https://aclanthology.org/2023.bionlp-1.30, doi:10.18653/v1/2023.bionlp-1.30.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W., Lu, X., 2019. PubMedQA: A dataset for biomedical research question answering, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2567–2577.

Jin, Q., Yuan, Z., Xiong, G., Yu, Q., Ying, H., Tan, C., Chen, M., Huang, S., Liu, X., Yu, S., 2022. Biomedical question answering: A survey of approaches and challenges. ACM Computing Surveys (CSUR) 55, 1–36.

Kelly, L., Goeuriot, L., Suominen, H., Schreck, T., Leroy, G., Mowery, D.L., Velupillai, S., Chapman, W.W., Martinez, D., Zuccon, G., et al., 2014. Overview of the share/clef ehealth evaluation lab 2014, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer. pp. 172–191.

Kim, D., Lee, J., So, C.H., Jeon, H., Jeong, M., Choi, Y., Yoon, W., Sung, M., Kang, J., 2019. A neural named entity recognition and multi-type normalization tool for biomedical text mining. IEEE Access 7, 73729–73740.

Kim, J.D., Ohta, T., Tateisi, Y., Tsujii, J., 2003. GENIA corpus—a semantically annotated corpus for bio-textmining. Bioinformatics 19, i180–i182.

Kim, J.D., Ohta, T., Tsuruoka, Y., Tateisi, Y., Collier, N., 2004. Introduction to the bio-entity recognition task at JNLPBA, in: Proceedings of the international joint workshop on natural language processing in biomedicine and its applications, Association for Computational Linguistics. pp. 70–75.

Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., Leaman, R., Lu, Y., Ji, D., Lowe, D.M., et al., 2015a. The CHEMDNER corpus

of chemicals and drugs and its annotation principles. Journal of cheminformatics 7, S2.

Krallinger, M., Rabal, O., Lourenço, A., Perez, M.P., Rodriguez, G.P., Vazquez, M., Leitner, F., Oyarzabal, J., Valencia, A., 2015b. Overview of the CHEMDNER patents task, in: Proceedings of the fifth BioCreative challenge evaluation workshop, pp. 63–75.

Lamurias, A., Clarke, L., Couto, F., 2017. Extracting microRNA-gene relations from biomedical literature using distant supervision. PLoS ONE 12. doi:http://dx.doi.org/10.1371/journal.pone.0171929, pmid:http://www.ncbi.nlm.nih.gov/pubmed/28263989.

Lamurias, A., Sousa, D., Clarke, L.A., Couto, F.M., 2019. BO-LSTM: classifying relations via long short-term memory networks along biomedical ontologies. BMC bioinformatics 20, 1–12.

Lamurias, A., Sousa, D., Couto, F.M., 2020. Generating biomedical question answering corpora from Q&A forums. IEEE Access 8, 161042–161051. doi:10.1109/ACCESS.2020.3020868.

Leaman, R., Gonzalez, G., et al., 2008. BANNER: an executable survey of advances in biomedical named entity recognition., in: Pacific symposium on biocomputing, pp. 652–663.

Leaman, R., Islamaj Doğan, R., Lu, Z., 2013. DNorm: disease name normalization with pairwise learning to rank. Bioinformatics 29, 2909–2917.

Leaman, R., Wei, C.H., Lu, Z., 2015. tmChem: a high performance approach for chemical named entity recognition and normalization. Journal of cheminformatics 7, S3.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J., 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36, 1234–1240.

Lee, M., Cimino, J., Zhu, H.R., Sable, C., Shanker, V., Ely, J., Yu, H., 2006. Beyond information retrieval—medical question answering, in: AMIA annual symposium proceedings, American Medical Informatics Association. p. 469.

Lee, S., Kim, D., Lee, K., Choi, J., Kim, S., Jeon, M., Lim, S., Choi, D., Kim, S., Tan, A.C., et al., 2016. BEST: next-generation biomedical entity search tool for knowledge discovery from biomedical literature. PloS one 11, e0164680.

Lever, J., Jones, S.J., 2016. VERSE: Event and relation extraction in the BioNLP 2016 shared task. ACL 2016 , 42.

Li, J., Sun, Y., Johnson, R.J., Sciaky, D., Wei, C.H., Leaman, R., Davis, A.P., Mattingly, C.J., Wiegers, T.C., Lu, Z., 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database 2016.

Liu, Y., Liang, Y., Wishart, D., 2015. PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. Nucleic acids research 43, W535–W542.

Lobo, M., Lamurias, A., Couto, F.M., 2017. Identifying human phenotype terms by combining machine learning and validation rules. BioMed Research International 2017.

Lourenço, A., Carreira, R., Carneiro, S., Maia, P., Glez-Peña, D., Fdez-Riverola, F., Ferreira, E.C., Rocha, I., Rocha, M., 2009. @ note: a workbench for biomedical text mining. Journal of biomedical informatics 42, 710–720.

Luo, L., Lai, P.T., Wei, C.H., Arighi, C.N., Lu, Z., 2022a. BioRED: a rich biomedical relation extraction dataset. Briefings in Bioinformatics 23, bbac282.

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., Liu, T.Y., 2022b. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics 23, bbac409.

Mallory, E.K., Zhang, C., Ré, C., Altman, R.B., 2016. Large-scale extraction of gene interactions from full-text literature using DeepDive. Bioinformatics 32, 106–113.

Manning, C.D., Schütze, H., et al., 1999. Foundations of statistical natural language processing. volume 999. MIT Press.

Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D., 2014. The Stanford CoreNLP natural language processing toolkit, in: Association for Computational Linguistics (ACL) System Demonstrations, pp. 55–60. URL: http://www.aclweb.org/anthology/P/P14/P14-5010.

Miwa, M., Pyysalo, S., Ohta, T., Ananiadou, S., 2013. Wide coverage biomedical event extraction using multiple partially overlapping corpora. BMC bioinformatics 14, 175.

Miyao, Y., Ohta, T., Masuda, K., Tsuruoka, Y., Yoshida, K., Ninomiya, T., Tsujii, J., 2006. Semantic retrieval for the accurate identification of relational concepts in massive textbases, in: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, Association for Computational Linguistics. pp. 1017–1024.

Moradi, M., Samwald, M., 2021. Explaining black-box models for biomedical text classification. IEEE journal of biomedical and health informatics 25, 3112–3120.

Müller, H.M.M., Kenny, E.E., Sternberg, P.W., 2004. Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biology 2, e309. doi:10.1371/journal.pbio.0020309.

Nadeau, D., Sekine, S., 2007. A survey of named entity recognition and classification. Lingvisticae Investigationes 30, 3–26.

Nakov, P., Barrón-Cedeño, A., da San Martino, G., Alam, F., Struß, J.M., Mandl, T., Míguez, R., Caselli, T., Kutlu, M., Zaghouani, W., et al., 2022.

Overview of the clef–2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer. pp. 495–520.

Nentidis, A., Katsimpras, G., Vandorou, E., Krithara, A., Miranda-Escalada, A., Gasco, L., Krallinger, M., Paliouras, G., 2022. Overview of bioasq 2022: The tenth bioasq challenge on large-scale biomedical semantic indexing and question answering, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer. pp. 337–361.

Nunes, T., Campos, D., Matos, S., Oliveira, J.L., 2013. BeCAS: biomedical concept recognition services and visualization. Bioinformatics 29, 1915–1916.

Ohta, T., Pyysalo, S., Tsujii, J., 2011. Overview of the epigenetics and post-translational modifications (EPI) task of BioNLP shared task 2011, in: Proceedings of the BioNLP Shared Task 2011 Workshop, Association for Computational Linguistics. pp. 16–25.

Okazaki, N., Ananiadou, S., 2006. Building an abbreviation dictionary using a term recognition approach. Bioinformatics 22, 3089–3095.

Pappas, D., Stavropoulos, P., Androutsopoulos, I., 2020. AUEB-NLP at BioASQ 8: Biomedical document and snippet retrieval., in: CLEF (Working Notes).

Pyysalo, S., Ohta, T., Miwa, M., Cho, H.C., Tsujii, J., Ananiadou, S., 2012. Event extraction across multiple levels of biological organization. Bioinformatics 28, i575–i581.

Pyysalo, S., Ohta, T., Rak, R., Rowley, A., Chun, H.W., Jung, S.J., Choi, S.P., Tsujii, J., Ananiadou, S., 2015. Overview of the cancer genetics and pathway curation tasks of bionlp shared task 2013. BMC bioinformatics 16, S2.

Ray, P.P., 2023. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems .

Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H., Jimeno, A., 2008. Text processing through web services: calling Whatizit. Bioinformatics 24, 296–298.

Ren, K., Lai, A.M., Mukhopadhyay, A., Machiraju, R., Huang, K., Xiang, Y., 2014. Effectively processing medical term queries on the UMLS metathesaurus by layered dynamic programming. BMC medical genomics 7, S11.

Ruas, P., Couto, F.M., 2022. NILINKER: Attention-based approach to NIL entity linking. Journal of Biomedical Informatics 132, 104137.

Sætre, R., Yoshida, K., Yakushiji, A., Miyao, Y., Matsubayashi, Y., Ohta, T., 2007. AKANE system: protein-protein interaction pairs in BioCreAtIvE2 challenge, PPI-IPS subtask, in: Proceedings of the Second BioCreative Challenge Workshop, Madrid. pp. 209–212.

Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G., 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association : JAMIA 17, 507–513. doi:10.1136/jamia.2009.001560.

Segura Bedmar, I., Martínez, P., Herrero Zazo, M., 2013. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013), in: Proceedings of the Seventh International Workshop on Semantic Evaluation, Association for Computational Linguistics.

Segura-Bedmar, I., Martínez, P., Herrero-Zazo, M., 2014. Lessons learnt from the DDIExtraction-2013 shared task. Journal of biomedical informatics 51, 152–164.

Segura-Bedmar, I., Martinez, P., de Pablo-Sánchez, C., 2011. Using a shallow linguistic kernel for drug–drug interaction extraction. Journal of biomedical informatics 44, 789–804.

Settles, B., 2005. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics 21, 3191–3192.

Smith, L.H., Tanabe, L., Rindflesch, T., Wilbur, W.J., 2005. MedTag: a collection of biomedical annotations, in: Proceedings of the ACL-ISMB workshop on linking biological literature, ontologies and databases: mining biological semantics, Association for Computational Linguistics. pp. 32–37.

Song, Y., 2023. Artificial intelligence algorithms in biomedical application, in: 2023 International Conference on Intelligent Supercomputing and BioPharma (ISBP), pp. 42–47. doi:10.1109/ISBP57705.2023.10061317.

Sousa, D., Couto, F.M., 2020. BiOnt: deep learning using multiple biomedical ontologies for relation extraction, in: European Conference on Information Retrieval, Springer. pp. 367–374.

Sousa, D., Couto, F.M., 2022. Biomedical relation extraction with knowledge graph-based recommendations. IEEE Journal of Biomedical and Health Informatics .

Sousa, D., Lamúrias, A., Couto, F.M., 2019. A silver standard corpus of human phenotype-gene relations, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 1487–1492.

Sousa, D., Lamurias, A., Couto, F.M., 2020. A hybrid approach toward biomedical relation extraction training corpora: combining distant supervision with crowdsourcing. Database 2020.

Sousa, D.F., Couto, F.M., 2023. K-RET: knowledgeable biomedical relation extraction system. Bioinformatics 39, btad174.

Stenetorp, P., Pyysalo, S., Tsujii, J., 2011. SimSem: Fast approximate string matching in relation to semantic category disambiguation, in: Proceedings of BioNLP 2011 Workshop, Association for Computational Linguis-

tics, Portland, Oregon, USA. pp. 136–145. URL: http://www.aclweb.org/anthology/W11-0218.

Strubell, E., Ganesh, A., McCallum, A., 2019. Energy and policy considerations for deep learning in NLP, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3645–3650.

Styler IV, W.F., Bethard, S., Finan, S., Palmer, M., Pradhan, S., de Groen, P.C., Erickson, B., Miller, T., Lin, C., Savova, G., et al., 2014. Temporal annotation in the clinical domain. Transactions of the Association for Computational Linguistics 2, 143–154.

Sun, W., Rumshisky, A., Uzuner, O., 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. Journal of the American Medical Informatics Association 20, 806–813.

Sutton, C., McCallum, A., 2006. An introduction to conditional random fields for relational learning. Introduction to statistical relational learning , 93–128.

Swanson, D.R., 1990. Medical literature as a potential source of new knowledge. Bulletin of the Medical Library Association 78, 29.

Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al., 2019. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic acids research 47, D607–D613.

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., et al., 2017. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic acids research 45, D362–D368.

Szklarczyk, D., Santos, A., von Mering, C., Jensen, L.J., Bork, P., Kuhn, M., 2016. STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. Nucleic acids research 44, D380–D384.

Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M.R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., et al., 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC bioinformatics 16, 1–28.

Tsuruoka, Y., McNaught, J., Ananiadou, S., 2008. Normalizing biomedical terms by minimizing ambiguity and variability. BMC bioinformatics 9, S2.

Tsuruoka, Y., Miwa, M., Hamamoto, K., Tsujii, J., Ananiadou, S., 2011. Discovering and visualizing indirect associations between biomedical concepts. Bioinformatics 27, 111–119. doi:10.1093/bioinformatics/btr214.

Tsuruoka, Y., Tsujii, J., 2005. Bidirectional inference with the easiest-first strategy for tagging sequence data, in: Proceedings of the conference on human language technology and empirical methods in natural language processing, Association for Computational Linguistics. pp. 467–474.

Venkatesan, A., Kim, J.H., Talo, F., Ide-Smith, M., Gobeill, J., Carter, J., Batista-Navarro, R., Ananiadou, S., Ruch, P., McEntyre, J., Venkatesan, A., Kim, J.H., Talo, F., Ide-Smith, M., Gobeill, J., Carter, J., Batista-Navarro, R., Ananiadou, S., Ruch, P., McEntyre, J., 2016. SciLite: a platform for displaying text-mined annotations as a means to link research articles with biological data. Wellcome Open Research 1, 25. doi:10.12688/wellcomeopenres.10210.1.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S.D., Tegmark, M., Fuso Nerini, F., 2020. The role of artificial intelligence in achieving the sustainable development goals. Nature communications 11, 1–10.

Wei, C.H., Allot, A., Leaman, R., Lu, Z., 2019. PubTator central: automated concept annotation for biomedical full text articles. Nucleic acids research 47, W587–W593.

Wei, C.H., Harris, B.R., Kao, H.Y., Lu, Z., 2013. tmVar: a text mining approach for extracting sequence variants in biomedical literature. Bioinformatics , btt156.

Wei, C.H., Kao, H.Y., Lu, Z., 2015. GNormPlus: an integrative approach for tagging genes, gene families, and protein domains. BioMed research international 2015.

Winnenburg, R., Wächter, T., Plake, C., Doms, A., Schroeder, M., 2008. Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies? Briefings in Bioinformatics 9, 466–478. URL: http://www.ncbi.nlm.nih.gov/pubmed/19060303, doi:10.1093/bib/bbn043.

van Wynsberghe, A., 2021. Sustainable AI: AI for sustainability and the sustainability of AI. AI and Ethics 1, 213–218.

Yeh, A., Hirschman, L., Morgan, A., 2002. Background and overview for KDD cup 2002 task 1: Information extraction from biomedical articles. ACM SIGKDD Explorations Newsletter 4, 87–89.

Yu, W., Gwinn, M., Clyne, M., Yesupriya, A., Khoury, M.J., 2008. A navigator for human genome epidemiology. Nature genetics 40, 124–125. doi:10.1038/ng0208-124.

Zhang, C., 2015. DeepDive: a data management system for automatic knowledge base construction. Ph.D. thesis. The University of Wisconsin-Madison.

Zhu, M., Ahuja, A., Wei, W., Reddy, C.K., 2019. A hierarchical attention retrieval model for healthcare question answering, in: The World Wide Web Conference, pp. 2472–2482.

10. Author Biography

Andre Lamurias is an assistant professor at School of Science and Technology, NOVA University of Lisbon, and a researcher at NOVA LINCS. In 2020,

he worked as a research scientist at Priberam Labs, and he has previously been a post-doctoral researcher in the Department of Computer Science at Aalborg University and at LASIGE, Faculty of Sciences, University of Lisbon. He completed his PhD in 2019 at the University of Lisbon, where he developed text-mining approaches for disease network discovery and systems biology. Prior to this, he obtained his Master's degree in Bioinformatics and Computational Biology from the same institution. His research focus is on information extraction applied to biomedical data, such as scientific papers, electronic health records, and genomics.

Diana F. Sousa completed her PhD in Computer Science in 2023 at the University of Lisbon, where she focused on biomedical relation extraction with added external knowledge. Before this, she obtained a bachelor's degree in Biochemistry and a master's degree in Bioinformatics and Computational Biology from the same university in 2017 and 2019, respectively. Since 2016, she has been a researcher at LASIGE, Department of Informatics, Faculty of Sciences, University of Lisbon. Her research interests rely mainly on information extraction applied to the clinical and biomedical domains.

Francisco M. Couto is currently an associate professor with habilitation at Universidade de Lisboa (Faculty of Sciences) and a researcher at LASIGE. He graduated (2000) and has a master (2001) in Informatics and Computer Engineering from the IST. He concluded his doctorate (2006) in Informatics, specialization Bioinformatics, from the Universidade de Lisboa. He was an invited researcher at EBI, AFMB-CNRS, BioAlma during his doctoral studies. His main research contributions cover several key aspects of bioinformatics and knowledge management, namely in proposing and developing: various text mining solutions that explore the semantics encoded in ontologies; semantic similarity measures and tools using biomedical ontologies; and ontology and linked data matching systems. Until August 2022, he published 2 books; was co-author of 10 chapters, 62 journal papers (47 Q1 Scimago), and 32 conference papers (10 core A and A*); and was the supervisor of 10 PhD theses and of 51 master theses. He received the Young Engineer Innovation Prize 2004 from the Portuguese

Engineers Guild, and an honorable mention in 2017 and the prize in 2018 of the ULisboaCaixa Geral de Depósitos (CGD) Scientific Prizes.