Relationship extraction for knowledge graph creation from biomedical literature

Relationship extraction for knowledge graph creation from biomedical literature

Nikola Milosevic[a,b], Wolfgang Thielemann[c]

[a]*University of Manchester, Faculty of Science and Engineering, Oxford road, Manchester, M13 9PL, United Kingdom*
[b]*Bayer Pharmaceuticals R&D, Mullerstrasse 178, Berlin, 13353, , Germany*
[c]*Bayer Pharmaceuticals R&D, Friedrich-Ebert-Str.475, Wuppertal, 42117, , Germany*

Abstract

Biomedical research is growing at such an exponential pace that scientists, researchers, and practitioners are no more able to cope with the amount of published literature in the domain. The knowledge presented in the literature needs to be systematized in such a way that claims and hypotheses can be easily found, accessed, and validated. Knowledge graphs can provide such a framework for semantic knowledge representation from literature. However, in order to build a knowledge graph, it is necessary to extract knowledge as relationships between biomedical entities and normalize both entities and relationship types. In this paper, we present and compare a few rule-based and machine learning-based (Naive Bayes, Random Forests as examples of traditional machine learning methods and DistilBERT and T5-based models as examples of modern deep learning transformers) methods for scalable relationship extraction from biomedical literature, and for the integration into the knowledge graphs. We examine how resilient are these various methods to unbalanced and fairly small datasets. Our experiments show that transformer-based models handle well both small (due to pre-training on a large dataset) and unbalanced datasets. The best performing model was the DistilBERT-based model fine-tuned on balanced data, with a reported F1-score of 0.89.

Keywords:
knowledge graphs, information extraction, machine learning, natural language processing, text mining, text-to-text model, linked data

1. Introduction

The amount of published scientific, especially biomedical literature is growing exponentially. In 2020, over 950,000 articles were added to Medline (National Library of Medicine, 2020), a repository of biomedical literature, meaning that on average, over 2600 biomedical articles were published daily. Scientists, researchers and professionals are not able to cope with the amount of published research in their area and are in need of tools that would help them find relevant articles and review and validate claims and hypothesis.

Finding relevant articles is the task addressed by the information retrieval sub-field of natural language processing. A number of information retrieval approaches have been

examined and several domain-specific information retrieval applications for bio-medicine have been built, such as PubMed, PubMed-Central, Quertle, Embase, etc. (Canese and Weis, 2013; Roberts, 2001; Coppernoll-Blach, 2011). Information retrieval engines can use named entity recognition algorithms and entity normalization techniques (often using dictionaries or terminologies) to improve the search results by returning semantically the most relevant articles for the searched concept regardless of the form or a synonym used for the searched entity (Jonnalagadda and Topham, 2010; Jonnagaddala et al., 2015; Hakala et al., 2016). However, information retrieval only offers a list of relevant articles for searched terms or concepts. In order to validate a hypothesis or claim, a researcher still needs to read through a significant amount of literature, which may be time-consuming.

The hypothesis and claims, that researchers often would like to validate, can be summarized in a simple sentence with two interacting concepts and a predicate describing their interaction (e.g. Aspirin *treats* pain). Hypothesis and claims are named relationships between concepts. These named relationships can be extracted with evidence (sentences from articles, stating them) from biomedical literature. Also, entities may be connected with many other entities in relationships, finally generating a large knowledge graph. This knowledge graph can be later utilized to infer knowledge by following connections ($A \rightarrow B, B \rightarrow C, therefore A \rightarrow C$), and even applications of graph machine learning to find potentially missing edges (relationships), or discover potential leads and targets in the drug discovery process.

The ways of validating claims, and inferring new knowledge from the statements in the knowledge graphs have been examined in areas of knowledge graph databases (Messina et al., 2017; Miller, 2013), semantic web (McGuinness et al., 2004; Parsia and Sirin, 2004; Sirin et al., 2007; Shearer et al., 2008) and graph machine learning (Scarselli et al., 2008; Veličković et al., 2017; Qu et al., 2019). However, in order to perform inference and validation over a knowledge graph, the knowledge needs to be extracted from the text and normalized. Most of the normalization research in the biomedical domain considers the normalization of named entities, such as diseases, genes, and compounds (Cho et al., 2017; Ji et al., 2020; Zhou et al., 2020). On the other hand, it was not given much attention to the normalization of biomedical relationships. Relationship extraction research in biomedical domain is often limited on a certain domain (e.g. cancer or cardiovascular domain), and considers limited set of possible relationship entity pairs and relationship types (Rindflesch et al., 1999; Yang et al., 2021). Normalized relationships (graph edges) are the pillar of successful systematization of knowledge in knowledge graphs.

As part of the R&D organization within Bayer pharmaceuticals, we focus on generating knowledge graphs relevant to drug discovery, target identification, and indication expansion. As well, most of the use-cases we are dealing with are related to humans. Therefore, the defined relationship model and methods for relationship extraction, described in this paper, will have a focus on the stated use-cases.

In this paper, we propose a data model for relationship normalization between drugs, targets, and diseases. We also examine and compare several rule-based and machine learning-based approaches. Using proposed methods, we generated a knowledge graph with links to the evidence sentences, based on the extracted and normalized relationships from PubMed. At the end, we discuss the results of our

knowledge graph creation method.

2. Background

Knowledge graphs have a long history spanning to the 1970s (Schneider, 1973). Knowledge graphs are a flexible knowledge representation framework, where knowledge is represented as a graph of inter-related concepts. Representing knowledge in a graph has a number of practical benefits in scenarios that involve integrating, managing, and extracting value from diverse and heterogeneous data sources. The idea of representing knowledge in graphs, particularly gained influence with Semantic Web, and lately with the development of knowledge graph announced in 2012 by Google, followed by other major tech industry players (Hogan et al., 2021). Lately, we could see applications of knowledge graphs in question answering products in wider use, such as Alexa, Google Assistant, or Siri (Zhang et al., 2018). Likewise, the pharmaceutical industry identified potential benefits knowledge graphs can bring in accelerating drug discovery, drug development, indication expansion of existing drugs, and pharmacovigilance.

In order to extract information and structure them for entry into the knowledge graph, it is necessary to perform named entity recognition of relevant entities (for bio-medicine these could be genes/targets, compounds, diseases, cell lines, pathways, organs, etc.), normalize all the possible synonyms to agreed terminology and at the end find relationships between co-mentioned entities and normalize the relationships to the agreed data model or ontology.

Biomedical named entity recognition and named entity normalization have had a long tradition of research since the late 1990s (Fukuda et al., 1998; Collier et al., 2000). A number of approaches were developed that can

be classified into dictionary-based, machine learning-based using usually hidden Markov models or Conditional Random Fields and Deep Learning-based, often using language models, such as word2vec, ELMo (Milosevic et al., 2020), BERT, and others, with transformers (Khan et al., 2020) or recurrent neural networks (Belousov et al., 2019).

In order to systematize extracted entities and input them into the knowledge graph, they need to be normalized. Normalization is a process of mapping all possible terms and variants that represent one concept to one unique entity id or preferred term (for example the concept of cancer can be stated using various expressions, such as neoplasms, tumor, cancer, malignity, etc.). For a long time, named entity normalization relied on good dictionaries and rule-based approaches (Leaman et al., 2015; Cohen, 2005). However, in recent years, there have been several deep learning-based approaches for ranking candidate entities for normalization using convolutional neural networks (Li et al., 2017; Deng et al., 2019) or language models such as Word2Vec (Cho et al., 2017) or BERT (Ji et al., 2020).

The extraction of the actual relationship comes as the final step of structuring information from a sentence. One of the first systems to attempt relationship extraction in biomedical domain was EDGAR (Rindflesch et al., 1999), that was extracting relationships between drugs and genes in cancer domain, using a set of rules based on syntactic analysis. Since then, several approaches to structuring relationships were explored:

- **Existence of relationship between entities** - classifies whether there is an actual semantic relationship between two entities, or the entities are co-mentioned, but there is no actual named relationship between them. This approach of extracting

general existence of the relationship is often applied for protein-protein interaction extraction (Zitnik et al., 2018; Szklarczyk et al., 2019) or gene-disease interactions (Becker et al., 2004).

- **Extracting predicate verb as relationship type** - predicate is not a closed set of possible classes. Rather any predicate verb appearing in a sentence indicating a relationship between entities is taken as the relationship type. Normalization of relationship types is left for further processing or analysis. Some of the research databases, such as Open Targets use this approach (Carvalho-Silva et al., 2019), as well as some of the commercial tools that allow relationship extraction (e.g. Linguamatics I2E).

- **Normalizing relationship types** - predicate is normalized into the set of well-defined types. This approach needs a carefully crafted data model of possible relationships, as well as a carefully developed dataset for machine learning or extraction rules. In semantic web community, there have been research on normalizing a basic set of relationships in general domain, such as is-a, part-of, equal (Arnold and Rahm, 2015; Speer and Havasi, 2013; Speer et al., 2017). Domain specific and more granular data models and datasets for this approach are rare. BioCreative VI and BioCreative VII provided data and organized shared tasks on chemical-protein interactions (Krallinger et al., 2017, 2020)

From the methodological perspective, relationship extraction can be performed using machine learning or rule-based approaches. Rule-based approaches range from using lists of relationship-related keywords and distances between concepts and keywords (Abacha and Zweigenbaum, 2011; Ravikumar et al., 2017) to using dependency parsers and evaluating whether concepts are related in grammatical sense (Erkan et al., 2007; Goertzel et al., 2006). On the other hand, machine learning approaches can be classified into two groups: (1) supervised learning, using crafted datasets (Peng et al., 2018; Liu et al., 2017) and (2) semi-supervised or distant learning approaches, where a dataset is expanded based on known relationships assuming that mentions of the same entities would entail the same relationship (Mintz et al., 2009). Since 2009, distant learning approaches have gained popularity and proved to be effective in relationship extraction, despite the assumption that all co-mentions of the same entities would entail the same relationship is not always correct and adds noise. In addition to these two general approaches, hybrid approaches, combining rules, dependency trees, and machine learning approaches have been popular for relationship extraction (Erkan et al., 2007; Muzaffar et al., 2015).

3. Method

In this paper, we compare methods for relationship extraction. All of our methods use in-house modified Linnaeus (Gerner et al., 2010) tool for named entity recognition and normalization. For relationship extraction, we present and compare three methods: (1) a rule-based method, based on sentence patterns and dictionaries of trigger verbs and phrases, (2) a machine learning method, based on traditional machine learning models (i.e. Naive Bayes, Random Forests), and (3) a deep learning method based on transformer architectures, such as DistilBERT (Sanh et al., 2019) and text to text T5 transformer (Raffel et al., 2020).

4

3.1. Named entity recognition and normalization (modification of Linnaeus)

Named entity recognition was done by an internally modified version of the Linnaeus tool (Gerner et al., 2010). We have added a number of features that would allow us to perform more flexible entity matching while relying on the Linnaeus algorithm which is fast and reliable with good dictionaries. The added features include:

- Handling defined a set of characters like white spaces and treating multiple sequential white space characters as a single white space character.

- Implementing a global flag that allows ignoring cases of letters in matches if needed

- A flag for automatic pluralization (adding *-s* and *-es* suffixes) of dictionary terms

- A functionality that can handle and transliterate Greek characters (e.g. beta - β) as well as functionality that can handle the variation of the position of Greek characters (e.g. Interferon-α vs α-interferon).

- Removing or ignoring diacritics

- Synonym level exact, case sensitive, and regular expression matching

- Flag to match only the longest match

The dictionaries used for named entity recognition and normalization into entities have been carefully internally developed, expanded, and refined over the past fifteen years by our internal information scientists. We have used dictionaries for human genes, diseases, and approved drugs.

3.2. Relationship data model

Relationships of interest are relationships between drug, gene, and disease entities. The three pairs of relationships of interest are: (1) Drug-Gene, (2) Drug-Disease, and (3) Gene-Disease. Each of these pairs may have several distinct relationship types. In order to develop the relationship model, we have organized two workshops guided by the authors with the internal experts from the Bayer R&D department. Eighteen people participated in these workshops and in the effort to create the data model. They are members of following teams within Bayer R&D department: scientific and competitive intelligence (12 people), semantics and knowledge graph technologies (3 people, including authors), research and early development, kidney disease (2 people), bioinformatics (1 person). Experts from these teams have advanced degrees in pharmacology, biology, or medicine and often substantial working experience in academia and within the pharmaceutical industry. They have helped us identify the possible relationship types for the entities we focused on and validate our model. We have also organized a meeting with one bioinformatics expert, who helped us additionally expand and validate Gene-Disease relationships and possible modes of action. Furthermore, we scouted available commercial solutions that provide knowledge graph solutions with the entities we were looking for. We have identified two companies which provide data that is close to our needs - Causally[1] and Biorelate[2]. The model that we have created contained more comprehensive and more detailed relationship types (more relationship types, relationship attributes, such as modes of action for genetic relationships) for the rele-

---

[1] https://www.causaly.com/
[2] https://www.biorelate.com/

5

vant entity pairs, at the time of writing this paper.

3.2.1. Drug-Gene relationships

CPR GROUP	TYPE
CPR:1	Part of
CPR:2	Regulator
CPR:3	Up-regulator, Activator
CPR:4	Down-regulator, Inhibitor
CPR:5	Agonist
CPR:6	Antagonist
CPR:7	Modulator
CPR:8	Co-factor
CPR:9	Substrate, Product of
CPR:10	Not

Table 1: Chemical-Protein relationship as defined by BioCreative shared task

The relationships between chemical and proteins were subject of last two BioCreative shared tasks in 2017 [3] and 2020 [4]. Both of the tasks defined the same interaction types. These can be seen in Table 1. However, for the majority of purposes, some of the defined types may be redundant. Therefore, we have simplified the data model by merging some of the classes and excluding ones that are rarely mentioned in the text. In the end, our Drug-Gene model had the following relationship classes:

- Up-regulator/activator

- Down-regulator/inhibitor

- Regulator

- Part of

- Modulator

- Co-factor

- Substrate or product of

Note that *Regulator* is a type of relationship in which it is not possible to determine the direction of regulation from the sentence.

3.2.2. Drug-Disease relationships

Relationships between drugs and diseases do not have any gold standard data model that was used in previous shared tasks. Therefore a new model was proposed containing the following relationship classes:

- Therapeutic use/Treatment

- Cause/Adverse event

- Decrease Disease

- Increase Disease

- Effect on

- Biomarker

It may seem that there is redundancy between *Therapeutic use* and *Decrease Disease* classes, or between *Cause* and *Increase Disease* classes. However, *Therapeutic use* and *Cause* indicate relationships where disease is indication or counter indication for a given drug. Increase and Decrease disease may refer to any finding that a given drug improved or worsened the state of disease, and therefore is a weaker relationship. The weakest relationship is *Effect on*, because, in this case, only the fact that there is some effect of a drug on disease is known, without any additional details (e.g. whether it improves disease or makes it worse).

The chemical compound can be a biomarker for some diseases. In medicine and drug discovery, it is important to have a picture of biomarkers, and therefore it is included in the model.

3.2.3. Gene-Disease relationships

The relationship between genes and diseases is the most complex one among the three types in scope. This is because a single gene can improve, worsen or even cause a certain disease. Therefore, it is often not enough to classify the type of the relationship, the algorithm needs to extract also a mode of action on the gene. In terms of possible relationship types, we have identified the following ones:

- Plays a role – From the sentence can be concluded clearly that there is a connection between the gene and the disease, however, it is not clear what kind of role the gene plays in the disease, only that it plays some role.

- Target

  - General – The gene or protein can be considered a target for the given disease, with no more specific information on the modulation of the disease.

  - Cause – The sentence indicates that activation, mutation or inhibition, or any other action over a gene is causing a given disease.

  - Modulator

    * Decrease disease – There is a clear indication that gene is responsible for decreasing and alleviating the disease.

    * Increase Disease - There is a clear indication that gene is responsible for increasing and worsening the disease.

- Biomarker – The presence or lack of a given gene/protein is an indicator for the diagnosis of disease or pathology.

3.2.4. Mode of action

Together with the relationship classes, if available, mode of action is an important modifier for Gene-Disease relationships. It establishes the action taken on a gene in order for the relationship to take place. For example, a gene may both decrease and cause disease, depending on whether the gene was activated or inhibited. Possible modes of action are (1) inhibition or down-regulation, (2) activation or up-regulation, (3) mutation or modification.

3.2.5. Negation and speculation

The relationship between entities in a sentence may be negated, which reverses the semantics of the relationship. Therefore, it is important to detect whether the relationship is negated.

Likewise, statements in the text can be factual, stated as well-known facts, or speculative. Speculative claims need to be taken with more caution and therefore, speculation detection is included in our model.

3.3. Relationship extraction using rule-based method

Based on the previously described relationship model, we have developed a rule-based method for relationship extraction. The method relies on vocabularies for relationship trigger words, negation cues, speculation cues, mode of action (MoA) cues, and grammar pattern rule set. An example of vocabularies and patterns in the ruleset with an example sentence from which a relationship is extracted using given rules and vocabularies is presented in Figure 1.

The trigger word vocabulary contains a list of relationship trigger words and phrases, with metadata, such as to which relationship given word or phrase maps, between which entities, whether entities have to be in a given order of mentioning or can be in reverse order (e.g.

Figure 1: Example of dictionaries, rules set and an example of sentence annotations in order to match relationship in a sentence

MoA	Negation	Speculation list	Trigger words
mutation-> Mutation blocking-> Inhibition Inhibition->Inhibition Underexpression->Inhibition …	no not no one neither does not …	may can could suggest Hypothesise …	is indication;Treat;Drug->Disease; efficacy in the treatment;Target;Gene->Disease involved in;Plays a role;Gene->Disease alleviating;Decrease Disease;Gene->Disease …

Rule set
Subject W2 Attribute W5 Trigger W3 Predicate
Subject W5 Negation W5 Trigger W5 Predicate
…

W5   W1

Subject | MoA   Trigger   Predicate

IL 17   blocking with Secukinumab (SEK) has proved   efficacy in the treatment   of   psoriasis.

for Drug-Disease relationship whether it is allowed for the drug to be after disease in the sentence) and whether the phrase should be interpreted as a regular expression.

Mode of action cues has a mapping to the mode of action type (i.e. Inhibition, Activation, Mutation). The vocabularies for negation and speculation cues are simple lists of words (e.g. *no, not* for negation; *hypothesise, may* for speculative).

Grammar patterns define sequences that need to be matched in order to extract relationships. This grammar has several keywords, such as *Subject*, which refers to the subject entity, *Predicate*, which refers to predicate entity, *Trigger*, referring to trigger cue, *Speculation*, referring to Speculation, *Negation*, referring to negation cue, *Subject_type*, referring to entity that is not subject in current evaluation pair of entities, but has same type as Subject entity (i.e. Drug or Gene), *Predicate_type*, referring to entity having same type as predicate, but not evaluated in current pair. Ad-

ditionally, there may be defined a number of words that are between labeled entities, trigger words, negations, and speculative phrases. For example the following pattern:

Speculation W3 Subject W3 Trigger W3 Predicate

would match sequences where the speculative cue is followed by up to three tokens, followed by Subject, followed by up to three tokens, followed by trigger phrase, followed by another three tokens and predicate. This means it would match sentences such as "We hypothesize that aspirin can alleviate most cases of headache", if the token "hypothesize" is in the list of speculative cues, "aspirin" is marked as a drug and is subject, "headache" is a disease and predicate, and "alleviate" is a trigger word.

The matching algorithm iterates over labeled entities in each sentence and finds all pairs that may constitute relationships. It annotates sentences with potential trigger

8

phrases, speculative cues, and negations. Finally, the algorithm tries to match any pattern from the grammar to the sequence in a sentence. If the matching is successful, the relationship is extracted and mapped to the relationship type and metadata, such as mode of action, negation, and speculation cues are extracted.

The confidence score is calculated as a proportion of words in sentences that exactly match named phrases (Subject, Predicate, Trigger, Speculation, Negation), divided by all the words in the pattern (this includes named phrases and tokens that were matched as part of allowed distance tokens, e.g. up to three words for each W3 statement in grammar). The rationale for this calculation is based on the assumption that additional words may change the semantics of the sentence and therefore confidence about the existence of the relationship should be lower.

In addition, the method also extracts co-occurrences, giving them fixed confidence of 0.0001, and labeling them with a "Co-occurrence" label.

Each extracted relationship contains information about entities (entity string, type, preferred term, internal ID), relationship type, whether it is negated, whether it is speculative, and confidence score. Also, the evidence sentence and Medline ID of the article where the evidence was found are recorded.

3.4. Machine learning

We have developed machine learning methods for classifying relationship types. The task was modeled as a sentence-level classification task. The initial method is based on sentence classification using traditional machine learning algorithms, such as Random Forests, Naive Bayes. We have then advanced the method by using fine-tuned transformer-based architectures for sentence classification for relation-

ship type, such as DistilBERT (Sanh et al., 2019), and a text-to-text transformer called T5 (Raffel et al., 2020). We report here results from all of the mentioned experiments.

3.4.1. Training and testing data

The data are collected by using a rule-based relationship extractor previously described for the Gene-Disease relationship. We are evaluating our approach to Gene-Disease data, as it proved to have the most complex data model and therefore is the most complex to correctly extract relationships. Also, this relationship type is important from a biomedical perspective, as it may give insights on potential targets for treating respective diseases. Out of this dataset, 2000 sentences were reviewed and corrected by human annotators. For this task, a company called Molecular Connections was contracted. Other 10 000 sentences were obtained from the rule-based method, with confidence 1. These sentences would match correct sentences, as they do not allow for any tokens that may change context, apart from named phrases. Therefore, the dataset contained about 12 000 sentences. The data was split as 90% training and 10% testing data for training and testing of machine learning approaches.

In order to create a more balanced dataset, we have generated the second dataset by taking 2000 manually annotated sentences, but then adding sentences from the rule-based method with confidence 1 in such a way that each relationship class had at least 1400 sentences (for biomarkers, we could obtain 1243 sentences with confidence 1 from a processed portion of the data we had at the time of building the dataset). The statistics about the number of sentences per relationship class in our datasets are presented in Table 2.

We have also created a dataset for the classification of a mode of action. We created again

9

Relationship type	Unbalanced dataset	Balanced dataset
Biomarker	198	1243
No Explicit Relationship	446	446
Plays a role	7393	1532
Target->Causative	1460	1508
Target->General	656	1414
Target->Modulator->Decrease Disease	1108	1450
Target->Modulator->Increase Disease	720	1422

Table 2: Number of sentences for each relationship type in our balanced and unbalanced data sets

one unbalanced and balanced dataset. Since for the mutation class we had only 140 examples, we initially balanced the dataset by taking 140 examples from each class. This is a fairly small dataset and may be improved by adding examples. We have created an additional dataset taking 300 data samples from each class, allowing duplication for classes that did not have enough samples (e.g. mutation). The access to the generated datasets can be requested at `https://zenodo.org/record/6466316#.Ylw3T-dS9Ea`.

3.4.2. Initial experiments using Random forest and Naive Bayes classifiers

Initially, we attempted to use traditional machine learning algorithms, such as Naive Bayes and Random Forests. For both of them, sentences were tokenized and stemmed using Porter Stemmer (Porter, 1980). Since for relationship extraction, it is important to examine the sequences of words, the features for our classifiers were uni-grams, bi-grams, tri-grams, and four-grams. Finally, data was trained using Random Forest and Naive Bayes Classifier.

3.4.3. Transformer based architectures: DistilBERT and T5 Text-to-text model

Transformer-based models are currently the state-of-the-art machine learning methods that perform well on a variety of tasks, ranging from classification to summarization and question answering (Devlin et al., 2018; Raffel et al., 2020). In the past few years, a number of language models were developed and pre-trained on datasets such as common crawl. Many of these models are based on the BERT model, with various modifications to reduce the size of the model or increase speed (Sanh et al., 2019; Liu et al., 2019; Lan et al., 2019). These models can be used for classification by using and training head - a feed-forward neural network on top of the transformer network that outputs word embeddings. We will use a BERT-based model that was optimized for size and speed, called DistilBERT (Sanh et al., 2019), whose authors claimed that has 40% fewer parameters, runs 60% faster while preserving over 97% of BERT's performances as measured on the GLUE language understanding benchmark. In 2020, Google released a text-to-text transformer called T5. This model is generating textual output and a single model can be trained to perform multiple tasks (specifying task in the prefix of the input). In the original paper, the authors of T5 claimed that the model exhibits state-of-the-art performance and on most of the tasks it outperformed BERT. In this paper, we will evaluate that claim on the sentence-level classification of biomedical relationships (gene-disease).

The learning task was defined as a sentence classification task. For a given sentence, con-

taining entities, the model is supposed to provide a normalized relationship type from our data model. In the training and testing sentences, the text was pre-prepared in such a way that subject of the relationship (e.g. gene) was masked with the keyword **SUBJECT** and the predicated of the relationship (e.g. disease) was masked with the keyword **PREDICATE**. In this way, we hypothesized that the internal attention mechanism of the model would be able to learn how to treat the vicinity of subjects and the predicates of the relationships.

The DistilBERT model was based on the DistilBERT base uncased model available on HuggingFace[5]. This model was fine-tuned for the classification task, and trained on our unbalanced and balanced datasets for 5 epochs (learning rate=0.00002). DistilBERT is an encoder model, to which a decoder can be created using a pooling and feed-forward network whose output layers dimension is equal to the number of classes (in our case 8).

On the other hand, the T5 model has encoder-decoder architecture, and therefore we do not define additional layers. We have fine-tuned the T5 model that is readily available on HuggingFace[6]. T5 is a multi-task model that can be fine-tuned and new tasks can be added during the fine-tuning of the model. The multi-tasking nature of T5 is convenient, since the same model can be deployed once performing multiple tasks (e.g. question-answering, summarization, translation, and relationship extraction within the same API). During the fine-tuning of the model, we have added new prefixes for gene-disease relationship classification and gene mode-of-action classification (with four classes - activation, inhibition, mutation, and not reported). We have fine-tuned the model on our dataset us-

ing Adafactor optimizer (Shazeer and Stern, 2018). The model was trained for 5 epochs (learning rate=0.00002).

Encoder layers of T5 have a size of 512 tokens, while DistilBERT has an encoder size of 728 tokens. Both of the sizes are longer than the longest sentence in our dataset, therefore the size difference should not affect the training and we used padding to fill the sequence with special padding tokens.

4. Results

### 4.1. Rule-based relationship extraction

We have processed base Medline data until January 2021, containing about 35 million abstracts. The processing with both Linnaeus and the relationship extraction engine took about 7 days on a single machine. We managed to extract in total 4,784,985 relationships (with co-occurrences 35,900,521). There were 631,573 named relationships found between Drug-Genes (6,885,810 including co-occurrences), 1,468,639 relationships between Drug-Diseases (8,378,599 including co-occurrences), and 2,684,742 relationships between Genes and Diseases (20,065,385 including co-occurrences).

Extracted relationships can be loaded into a graph or relational database, where these relationships can answer complex medical questions with evidence. By summing confidence scores, it is possible to retrieve genes interacting with a given drug (e.g. top results for drug *Tolvaptan* was inhibition of AVPR2, while the second one was inhibition of vasopressin receptor family), drugs that have an effect on certain disease (e.g. for *autosomal dominant polycystic kidney disease* retrieved Tolvaptan, which is approved for autosomal dominant polycystic kidney disease, Sirolimus, which inhibits mTOR and as well have been often

11

Figure 2: Section of knowledge graph showing nodes that are in relationship with autosomal dominant polycystic kidney disease (ADPKD). Orange entities are diseases (ADPKD), entities in blue are drugs and in green are genes/proteins. Label on edges present relationship type, number of mentions and cumulative confidence score for the given relationship between two entities.
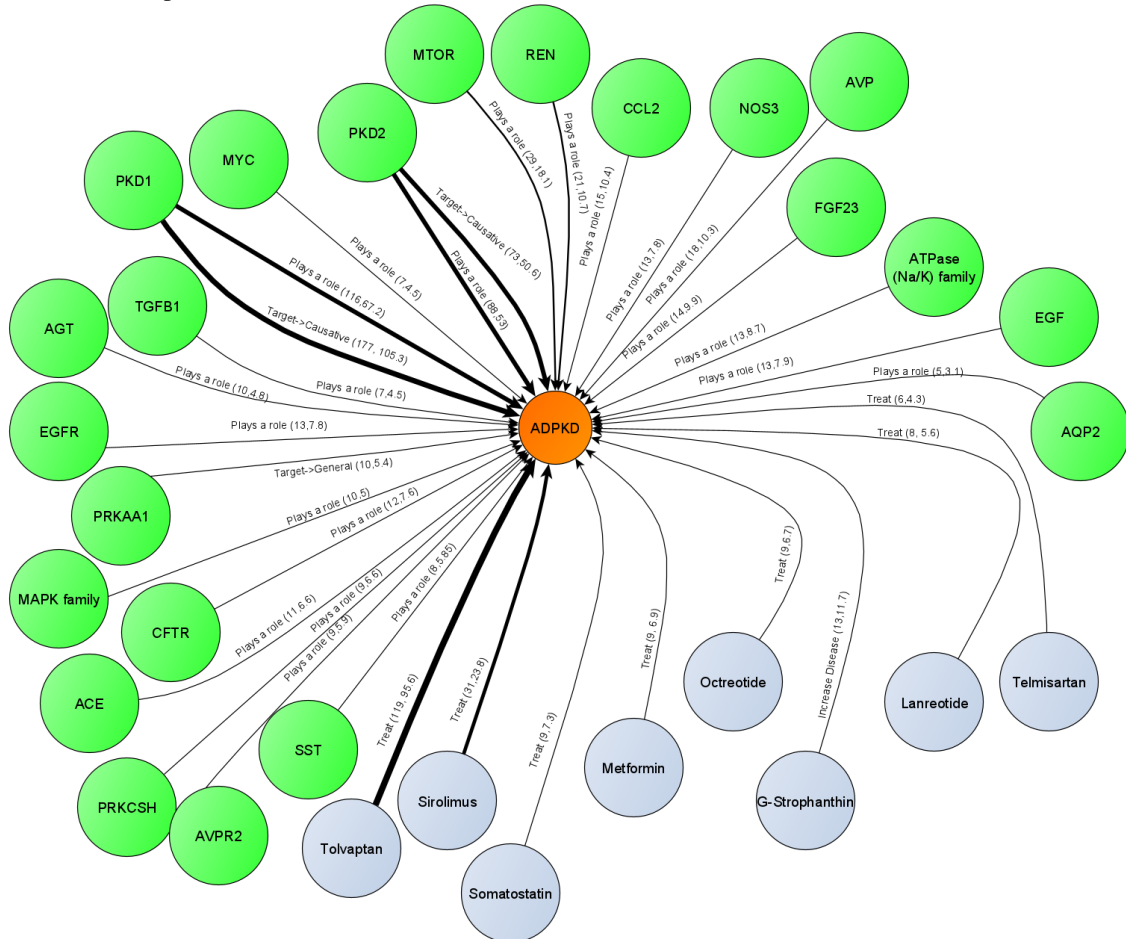
used in polycystic kidney disease, and Somatostatin, which was published as a hormone having a potential role in the treatment of autosomal dominant polycystic kidney disease (Messchendorp et al., 2020)), or what genes are important for a given disease (e.g. for autosomal dominant polycystic kidney disease retrieved PKD1 and PKD2 as targets that both play a role and have a causative relationship with disease, as well as mTOR, REN, CCL2). We evaluated a case study related to autosomal dominant polycystic kidney disease. We

created a graph whose edges end in autosomal dominant polycystic kidney disease. In order to reduce noise, we consider only edges that represent the relationship that was mentioned at least 5 times in PubMed. We have then evaluated the graph and all entities were indeed known from the literature to experts in the kidney disease team. A portion of the knowledge graph with relationships ending in ADPKD is presented in Figure 2.

We have evaluated 100 abstracts containing at least one relationship manually and calcu-

lated precision, recall, and F1-score. The evaluation is depending on the extent of trigger phrases and completeness of grammar, which is overtime improving. The measured performance was 0,88 precision, 0,74 recall and 0,80 F1-score. It is expected that the rule-based approach would have high precision and lower recall, as it would miss some of the relationships, but annotate relatively few false positives. Despite making some false positive relationships, generated data perform well in answering relevant biomedical questions between genes, drugs, and disease. The cumulative effect is that noise can be ignored by setting a threshold and manually validating results below the given threshold if necessary.

4.2. Naive Bayes and Random Forests-based Relationship extraction

The machine learning method was evaluated on two sets (2000 manually annotated relationships + 10,000 random relationships with confidence 1 - unbalanced set, and 2000 manually annotated relationships + random relationships with confidence 1, so there are at least 1,500 examples for each class - balanced set). For both data sets, 90% of data was used as training data, while 10% of data (about 1200 sentences) was used as a testing set. The results of our evaluation can be seen in the table 3.

Balancing data significantly improves precision and recall in both classifiers. With unbalanced data, Naive Bayes learned to always pick the most probable class - the class with the most results. The random forest classifier was better at learning how to recognize classes. However, balancing data, gained 26% in F1-score for Naive Bayes and 14% for overall results in Random Forests. The worst performance has a class which we were unable to balance due to the lack of annotated examples - *No Explicit relationships* (477 sentences

in the unbalanced set, that was annotated by annotators). Other classes performed with an F1-score over 70%.

4.3. Transformer-based relationship extraction

We have fine-tuned base T5 for relationship extraction by adding a new prefix (*"Relationship extraction:"*) on both unbalanced and balanced data. We have monitored the performance of the algorithm over epochs. The results can be seen in Figure 3.

Likewise, we have trained the DistilBERT model on both datasets.

Balancing data improves the T5 model, although the increase in performance is just 2% (F1-score increase from 0.86 to 0.88 after five epochs). However, certain relationship types in the unbalanced dataset had a large gap between precision and recall (e.g. *"No Explicit relationship"* in unbalanced had P=0.88, R=0.26), while in the balanced dataset precision and recall were closer (for the same class P=0.88, R=0.72).

We present the results of relationship classification after five epochs in Table 4. Overall, the DistilBERT model performed better on both data sets, even though the performance difference was just 2%. Also, the model performed better on the majority of relationship types (apart from Increase Disease and Biomarker types on both datasets). The stronger performance of DistilBERT is surprising and interesting due to its much smaller nature (66 million parameters in DistilBERT base compared to 220 million parameters in T5 base). This may be due to the multi-task and text-to-text nature of the T5 model, as a number of parameters need to be retained for other tasks and prefixes, as well as encoding to textual output.

We have added prefix into the T5 model and trained it for the classification of gene-

Class	Precision	Recall	F1-score
Unbalanced dataset			
Naive Bayes	0.39	0.62	0.48
Biomarker	0	0	0
No Explicit Relationship	0	0	0
Plays a role	0.62	1	0.77
Target->Causative	0	0	0
Target->General	0	0	0
Target->Modulator->Decrease Disease	0	0	0
Target->Modulator->Increase Disease	0	0	0
Random Forests	0.74	0.71	0.66
Biomarker	0.80	0.16	0.27
No Explicit Relationship	0.89	0.30	0.45
Plays a role	0.70	0.99	0.82
Target->Causative	0.81	0.34	0.48
Target->General	0.75	0.24	0.37
Target->Modulator->Decrease Disease	0.76	0.31	044
Target->Modulator->Increase Disease	0.81	0.17	0.28
Balanced dataset			
Naive Bayes	0.73	0.75	0.74
Biomarker	0.94	0.91	0.92
No Explicit Relationship	0	0	0
Plays a role	0.66	0.75	0.70
Target->Causative	0.66	0.89	0.76
Target->General	0.83	0.73	0.78
Target->Modulator->Decrease Disease	0.74	0.72	0.73
Target->Modulator->Increase Disease	0.84	0.76	0.80
Random Forests	0.79	0.79	0.78
Biomarker	0.97	0.85	0.91
No Explicit Relationship	0.64	0.15	0.24
Plays a role	0.65	0.80	0.72
Target->Causative	0.81	0.87	0.84
Target->General	0.76	0.84	0.80
Target->Modulator->Decrease Disease	0.75	0.81	0.78
Target->Modulator->Increase Disease	0.91	0.81	0.86

Table 3: Results of Naive Bayes and Random Forests classifiers

Figure 3: F1-score by epoch in fine-tuned DistilBERT and T5 models on both unbalanced and balanced datasets
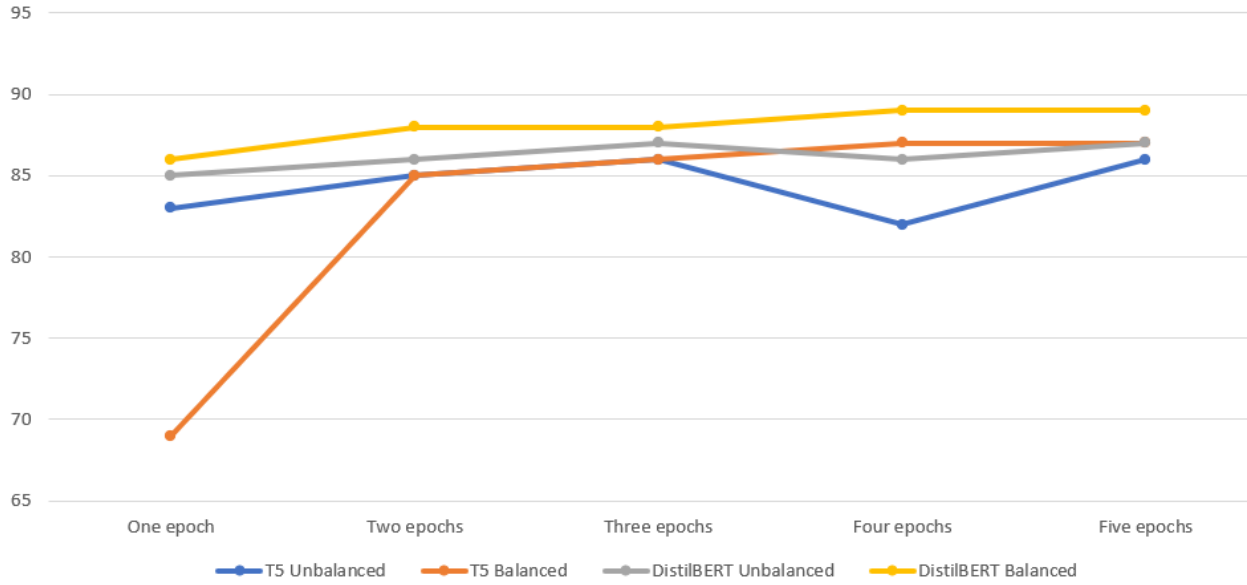
	T5			DistilBERT		
Class	Precision	Recall	F1-score	Precision	Recall	F1-score
Unbalanced dataset						
Overall (weighted average)	0.87	0.87	0.86	0.89	0.89	0.89
Biomarker	1.00	0.52	0.69	0.75	0.63	0.69
No Explicit Relationship	0.88	0.26	0.40	0.57	0.52	0.54
Plays a role	0.91	0.95	0.93	0.96	0.95	0.95
Target->Causative	0.84	0.90	0.87	0.89	0.94	0.91
Target->General	0.75	0.79	0.77	0.72	0.75	0.74
Target->Decrease Disease	0.77	0.79	0.78	0.74	0.82	0.78
Target->Increase Disease	0.85	0.82	0.83	0.78	0.79	0.79
Balanced dataset						
Overall (weighted average)	0.88	0.88	0.88	0.91	0.91	0.91
Biomarker	0.97	0.95	0.96	0.91	0.93	0.92
No Explicit Relationship	0.88	0.72	0.79	0.92	0.86	0.89
Plays a role	0.86	0.80	0.83	0.86	.0.82	0.84
Target->Causative	0.90	0.96	0.93	0.97	0.95	0.96
Target->General	0.83	0.87	0.85	0.92	0.93	0.93
Target->Decrease Disease	0.83	0.91	0.87	0.84	0.93	0.88
Target->Increase Disease	0.91	0.95	0.93	0.91	0.91	0.91

Table 4: Results of the best performing fine-tuned T5 and DistilBERT models (after 5 epochs)

15

associated modes of action into four possible classes: (1) activation, (2) inhibition, (3) mutation, and (4) not reported. The utility of the T5 model is that a single model can perform both classifications of sentences by relationship type as well as the mode of action, for which we would need separate DistilBERT-based models. The model was trained on the unbalanced and balanced dataset (each class containing 300 examples of each class). The model was trained for 5 epochs. The results are presented in Table 5.

Mode-of-action detection performs well with quite a small amount of data. This is because terms used for mode-of-action are in a relatively closed set (activation, inhibition, inhibitor, agonist, antagonist, mutation, modulation, etc.), and the language model is able to transfer and infer them from the time model was pre-trained on the C4 dataset. However, adding data helps improve it.

5. Discussion

The presented rule-based methodology is currently the base of the developed knowledge graph. With about 5 million typed relationships and over 30 million co-occurrences, it presents a powerful tool for drug discovery, target identification, indication expansion, and even pharmacovigilance. The graph structure allows for analysis over multiple hops. This will be further improved by adding protein-protein, drug-drug, and disease-disease interactions, on which we are working. It allows visualization of interaction pathways for diseases and graph learning for finding potentially missing relationships and validating hypotheses about weak relationships.

The current number of relationships in our graph is comparable with other state-of-the-art databases and graphs that consider the same or similar relationships. (Yang et al., 2021) de-

veloped a method for creating a stroke knowledge graph using PKDE4J based on 9 entity types and 30 relation types. The relationship extraction method based on BioBERT performed with an F1-score or 84.26% and they managed to extract 157 000 relationships based on stroke only papers in Pubmed (about 130 000 abstracts). (Lee et al., 2022) developed likewise a method based on PKDE4J for entity identification and SciBERT for classification of relationships between genes, diseases, and compounds. The data model had 8 relationship types based on whether the relationship is directed, undirected, positive, negative, and has an increasing or decreasing effect on the object entity. The best performance of their model is an F1-score of 81.7%. We believe that our model performed better (91% F1-score), because the data model is more granular and crafted for particular entity pair relationships, therefore easier to learn than relationship types generic for any biomedical entity pair.

Kim et al. (2017) focused on extracting gene-disease evidence sentences, using existing tools extracting genetic events, but did not classify the relationship between gene and disease. They managed to extract about 7.3 million evidence sentences from PubMed. Our system is extracting mode of action, which partially compares to biological events, on top of which it also extracts typed relationships. Our system extracted 2.7 million typed relationships and 20 million co-occurrences, therefore both wider (co-occurrences) and more detailed (typed relationships with modes of action) evidence.

Bhasuran and Natarajan (2018) used SVM on word embeddings to classify the existence of gene-disease relationships. The method was evaluated with an F1-score of about 83%-87%. RENET2 achieved 72.13% F1-score for extracting gene-disease relationships from Pub-

Class	Precision	Recall	F1-score
Unbalanced dataset			
Overall (weighted average)	0.95	0.95	0.94
Activation	1.00	1.00	1.00
Inhibition	0.87	1.00	0.93
Mutation	1.00	0.77	0.87
Not reported	0.94	1.00	0.97
Balanced dataset (300 examples per class)			
Overall (weighted average)	0.98	0.97	0.97
Activation	1.00	1.00	1.00
Inhibition	0.93	1.00	0.96
Mutation	0.97	1.00	0.99
Not reported	1.00	0.90	0.95

Table 5: Results of the best performing mode-of-action T5 model

MedCentral articles (Su et al., 2021). However, genes-disease association types were not classified, only the existence of the association was annotated. A number of methods were proposed for similar gene-disease association extraction without naming the relationship type based on DisGeNet dataset (Piñero et al., 2016; Hebbar and Xie, 2021; Parmar et al., 2020).Since the publication of DisGeNet, the research in this area accelerated. Publicly available datasets with annotated biomedical relationship types are rare. Therefore, there was a need for the creation of new gene-disease relationship data set with our described data model.

Our method based on DistilBERT is reporting better results than most methods reported in the literature. However, the model relies on the relationship types, consistency of annotators, and size of the dataset.

6. Future work

The creation of a comprehensive biomedical knowledge graph for target identification, indication expansion, and drug discovery is a long-term task. Some of the future activities on utilizing and improving our knowledge graph will involve:

- **Develop machine learning, transformer-based models for other relationship types (drug-gene, drug-disease, drug-drug, gene-gene, disease-disease)**. This may involve further annotation of data for other relationship pairs and creating a model based on these annotations.

- **Unifying relationships obtained from unstructured (literature, clinical trials, expert reports, grant proposals) and structured data sources** - Combining structured and unstructured data provides better quality of results and opens the possibility for a more detailed and comprehensive evaluation of links in the graph.

- **Developing an interface for exploring relationships and their evidence** - graphical user interface would enable a wider scientific audience to utilize the graph. This is especially important due to the fact that the majority of pharmacolo-

17

gists and biologists working in the pharmaceutical industry do not have an extensive computational background. This would allow them to be more efficient in generating and evaluating hypotheses before going to the laboratory.

- **Predicting novel target candidates using graph and temporal graph learning methods** - based on the chronology of the appearance of relationships in the graph, it may be possible to learn patterns and predict relationships between entities that would be identified in future. Therefore, it would be possible to automatically generate a hypothesis about the existence of yet undiscovered biological relationships using temporal graph neural networks (Wang et al., 2020).

7. Conclusions

In this paper, we have presented one rule-based approach, that mainly served as a starting point for obtaining biomedical relationship data. Further, we have compared traditional machine learning approaches, with modern, state-of-the-art language models and transformer approaches (DistilBERT and T5 models).

In all approaches, it was visible improvement when we used balanced datasets, however, fine-tuned DistilBERT and T5 models were more resilient and did not depend so much on balanced data sets as some older and traditional approaches would. Also, transformer-based models, due to their pre-training on larger data, are able to generalize well from a fairly small amount of data. DistilBERT performed slightly better than the T5 model, which was a surprising finding since T5 has about 4 times more parameters than DistilBERT. However, this may be due to the

multi-task nature of T5, as well as the fact that part of the model has to be used for text-to-text generation.

Developing machine learning data sets for tasks such as relationship extraction can be quite expensive. On the market, the pricing of a single annotated sentence can range between 1-3 euros, depending on the complexity of the task. However, this quickly scales, once the data set has 7 annotation classes and there is a need for over 1000 examples per annotation class in the data set. The commissioned manual annotations of our data set (around 7,000 sentences in total) cost 16,200 euros. The further cost comes from cloud infrastructure and machine learning engineering. Costs in developing relationship extraction models and approaches remain one of the main challenges.

Nevertheless, fine-tuning transformer models proved to be a promising approach. First of all, the performance of the model outperformed all other approaches, with over 91% F1-score overall, and with the majority of annotation classes breaching 85% F1-score. A review of the literature showed that the model performance is state-of-the-art for biomedical typed relationship extraction. Also, the model showed stability in terms of both precision and recall, where both were for the majority of classes similar (unlike the rule-based approach, where there is high precision but fairly low recall). On the other hand, T5 models are multi-task models, where it is possible to successfully address multiple problems with a single model, which makes valuable savings in terms of managing and updating the models. Lastly, fine-tuned T5 models, as they are text-to-text models, are easy to use and data preparation is kept simple.

In terms of limitation, T5 models are large, multi-task models, whose base model contains 220 million parameters. This is, for example, twice the size of the BERT base and over

four times the size of the DistilBERT model, and it contributes heavily to the speed of fine-tuning and execution, making the processing slow. Despite the fact that DistilBERT can be trained only on a single task and there is a need for post-processing of outputs, the model has both performance and speed advantages compared to the T5 model.

References

Abacha, A.B., Zweigenbaum, P., 2011. Automatic extraction of semantic relations between medical entities: a rule based approach. Journal of biomedical semantics 2, 1–11.

Arnold, P., Rahm, E., 2015. Semrep: A repository for semantic mapping. Datenbanksysteme für Business, Technologie und Web (BTW 2015) .

Becker, K.G., Barnes, K.C., Bright, T.J., Wang, S.A., 2004. The genetic association database. Nature genetics 36, 431–432.

Belousov, M., Milosevic, N., Dixon, W., Nenadic, G., 2019. Extracting adverse drug reactions and their context using sequence labelling ensembles in tac2017. arXiv preprint arXiv:1905.11716 .

Bhasuran, B., Natarajan, J., 2018. Automatic extraction of gene-disease associations from literature using joint ensemble learning. PloS one 13, e0200699.

Canese, K., Weis, S., 2013. Pubmed: the bibliographic database, in: The NCBI Handbook [Internet]. 2nd edition. National Center for Biotechnology Information (US).

Carvalho-Silva, D., Pierleoni, A., Pignatelli, M., Ong, C., Fumis, L., Karamanis, N., Carmona, M., Faulconbridge, A., Hercules, A., McAuley, E., et al., 2019. Open targets platform: new developments and updates two years on. Nucleic acids research 47, D1056–D1065.

Cho, H., Choi, W., Lee, H., 2017. A method for named entity normalization in biomedical articles: application to diseases and plants. BMC bioinformatics 18, 1–12.

Cohen, A., 2005. Unsupervised gene/protein named entity normalization using automatically extracted dictionaries, in: Proceedings of the acl-ismb workshop on linking biological literature, ontologies and databases: Mining biological semantics, pp. 17–24.

Collier, N., Nobata, C., Tsujii, J., 2000. Extracting the names of genes and gene products with a hidden markov model, in: COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics.

Coppernoll-Blach, P., 2011. Quertle: the conceptual relationships alternative search engine for pubmed. Journal of the Medical Library Association: JMLA 99, 176.

Deng, P., Chen, H., Huang, M., Ruan, X., Xu, L., 2019. An ensemble cnn method for biomedical entity normalization, in: Proceedings of the 5th workshop on BioNLP open shared tasks, pp. 143–149.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

Erkan, G., Özgür, A., Radev, D., 2007. Semi-supervised classification for extracting protein interaction sentences using dependency parsing, in: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 228–237.

Fukuda, K.i., Tsunoda, T., Tamura, A., Takagi, T., et al., 1998. Toward information extraction: identifying protein names from biological papers, in: Pac symp biocomput, Citeseer. pp. 707–718.

Gerner, M., Nenadic, G., Bergman, C.M., 2010. Linnaeus: a species name identification system for biomedical literature. BMC bioinformatics 11, 85.

Goertzel, B., Pinto, H., Heljakka, A., Ross, M., Pennachin, C., Goertzel, I., 2006. Using dependency parsing and probabilistic inference to extract relationships between genes, proteins and malignancies implicit among multiple biomedical research abstracts, in: Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology, pp. 104–111.

Hakala, K., Kaewphan, S., Salakoski, T., Ginter, F., 2016. Syntactic analyses and named entity recognition for pubmed and pubmed central—up-to-the-minute, in: Proceedings of the 15th Workshop on Biomedical Natural Language Processing, pp. 102–107.

Hebbar, S., Xie, Y., 2021. Covidbert-biomedical relation extraction for covid-19, in: The International FLAIRS Conference Proceedings.

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G.d., Gutierrez, C., Kirrane, S., Gayo, J.E.L., Navigli, R., Neumaier, S., et al., 2021. Knowledge graphs. Synthesis Lectures on Data, Semantics, and Knowledge 12, 1–257.

Ji, Z., Wei, Q., Xu, H., 2020. Bert-based ranking for

biomedical entity normalization. AMIA Summits on Translational Science Proceedings 2020, 269.

Jonnagaddala, J., Chang, N.W., Jue, T.R., Dai, H.J., 2015. Recognition and normalization of disease mentions in pubmed abstracts, in: Proceedings of the fifth BioCreative challenge evaluation workshop, Sevilla, Spain, pp. 9–11.

Jonnalagadda, S., Topham, P., 2010. Nemo: Extraction and normalization of organization names from pubmed affiliation strings. Journal of Biomedical Discovery and Collaboration 5, 50.

Khan, M.R., Ziyadi, M., AbdelHady, M., 2020. Mt-bioner: Multi-task learning for biomedical named entity recognition using deep bidirectional transformers. arXiv preprint arXiv:2001.08904 .

Kim, J., Kim, J.j., Lee, H., 2017. An analysis of disease-gene relationship from medline abstracts by digsee. Scientific reports 7, 1–13.

Krallinger, M., Miranda, A., Mehryary, F., Luoma, J., Pyysalo, S., Valencia, A., 2020. Drugprot shared task (biocreative vii track 1-2021) text mining drug-protein/gene interactions (drugprot) shared task .

Krallinger, M., Rabal, O., Akhondi, S.A., Pérez, M.P., Santamaría, J., Rodríguez, G.P., Tsatsaronis, G., Intxaurrondo, A., 2017. Overview of the biocreative vi chemical-protein interaction track, in: Proceedings of the sixth BioCreative challenge evaluation workshop, pp. 141–146.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R., 2019. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942 .

Leaman, R., Wei, C.H., Lu, Z., 2015. tmchem: a high performance approach for chemical named entity recognition and normalization. Journal of chem-informatics 7, 1–10.

Lee, Y., Son, J., Song, M., 2022. Bertsrc: Bert-based semantic relation classification .

Li, H., Chen, Q., Tang, B., Wang, X., Xu, H., Wang, B., Huang, D., 2017. Cnn-based ranking for biomedical entity normalization. BMC bioinformatics 18, 79–86.

Liu, S., Shen, F., Wang, Y., Rastegar-Mojarad, M., Elayavilli, R.K., Chaudhary, V., Liu, H., 2017. Attention-based neural networks for chemical protein relation extraction. Training 1020, 4157.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 .

McGuinness, D.L., Van Harmelen, F., et al., 2004. Owl

web ontology language overview. W3C recommendation 10, 2004.

Messchendorp, A.L., Casteleijn, N.F., Meijer, E., Gansevoort, R.T., 2020. Somatostatin in renal physiology and autosomal dominant polycystic kidney disease. Nephrology Dialysis Transplantation 35, 1306–1316.

Messina, A., Pribadi, H., Stichbury, J., Bucci, M., Klarman, S., Urso, A., 2017. Biograkn: A knowledge graph-based semantic database for biomedical sciences, in: Conference on Complex, Intelligent, and Software Intensive Systems, Springer. pp. 299–309.

Miller, J.J., 2013. Graph database applications and concepts with neo4j, in: Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA, USA.

Milosevic, N., Kalappa, G., Dadafarin, H., Azimaee, M., Nenadic, G., 2020. Mask: A flexible framework to facilitate de-identification of clinical texts. arXiv preprint arXiv:2005.11687 .

Mintz, M., Bills, S., Snow, R., Jurafsky, D., 2009. Distant supervision for relation extraction without labeled data, in: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 1003–1011.

Muzaffar, A.W., Azam, F., Qamar, U., 2015. A relation extraction framework for biomedical text using hybrid feature set. Computational and mathematical methods in medicine 2015.

National Library of Medicine, 2020. Citations added to medline by fiscal year @online available at https://www.nlm.nih.gov/bsd/stats/cit_added.html. URL: https://www.nlm.nih.gov/bsd/stats/cit_added.html.

Parmar, J., Koehler, W., Bringmann, M., Volz, K.S., Kapicioglu, B., 2020. Biomedical information extraction for disease gene prioritization. arXiv preprint arXiv:2011.05188 .

Parsia, B., Sirin, E., 2004. Pellet: An owl dl reasoner, in: Third international semantic web conference-poster, Citeseer. p. 13.

Peng, Y., Rios, A., Kavuluru, R., Lu, Z., 2018. Chemical-protein relation extraction with ensembles of svm, cnn, and rnn models. arXiv preprint arXiv:1802.01255 .

Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., Furlong, L.I., 2016. Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. Nu-

cleic acids research , gkw943.

Porter, M.F., 1980. An algorithm for suffix stripping. Program .

Qu, M., Bengio, Y., Tang, J., 2019. Gmnn: Graph markov neural networks, in: International conference on machine learning, PMLR. pp. 5241–5250.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research 21, 1–67.

Ravikumar, K., Rastegar-Mojarad, M., Liu, H., 2017. Belminer: adapting a rule-based relation extraction system to extract biological expression language statements from bio-medical literature evidence sentences. Database 2017.

Rindflesch, T.C., Tanabe, L., Weinstein, J.N., Hunter, L., 1999. Edgar: extraction of drugs, genes and relations from the biomedical literature, in: Biocomputing 2000. World Scientific, pp. 517–528.

Roberts, R.J., 2001. Pubmed central: The genbank of the published literature.

Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 .

Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G., 2008. The graph neural network model. IEEE transactions on neural networks 20, 61–80.

Schneider, E.W., 1973. Course modularization applied: The interface system and its implications for sequence control and data analysis. .

Shazeer, N., Stern, M., 2018. Adafactor: Adaptive learning rates with sublinear memory cost, in: International Conference on Machine Learning, PMLR. pp. 4596–4604.

Shearer, R., Motik, B., Horrocks, I., 2008. Hermit: A highly-efficient owl reasoner., in: Owled, p. 91.

Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., Katz, Y., 2007. Pellet: A practical owl-dl reasoner. Journal of Web Semantics 5, 51–53.

Speer, R., Chin, J., Havasi, C., 2017. Conceptnet 5.5: An open multilingual graph of general knowledge, in: Thirty-first AAAI conference on artificial intelligence.

Speer, R., Havasi, C., 2013. Conceptnet 5: A large semantic network for relational knowledge, in: The People's Web Meets NLP. Springer, pp. 161–176.

Su, J., Wu, Y., Ting, H.F., Lam, T.W., Luo, R., 2021. Renet2: high-performance full-text gene–disease relation extraction with iterative training data expansion. NAR Genomics and Bioinformatics 3, lqab062.

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.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., 2017. Graph attention networks. arXiv preprint arXiv:1710.10903 .

Wang, X., Ma, Y., Wang, Y., Jin, W., Wang, X., Tang, J., Jia, C., Yu, J., 2020. Traffic flow prediction via spatial temporal graph neural network, in: Proceedings of The Web Conference 2020, pp. 1082–1092.

Yang, X., Wu, C., Nenadic, G., Wang, W., Lu, K., 2021. Mining a stroke knowledge graph from literature. BMC bioinformatics 22, 1–19.

Zhang, Y., Dai, H., Kozareva, Z., Smola, A.J., Song, L., 2018. Variational reasoning for question answering with knowledge graph, in: Thirty-Second AAAI Conference on Artificial Intelligence.

Zhou, H., Ning, S., Liu, Z., Lang, C., Liu, Z., Lei, B., 2020. Knowledge-enhanced biomedical named entity recognition and normalization: application to proteins and genes. BMC bioinformatics 21, 35.

Zitnik, M., Agrawal, M., Leskovec, J., 2018. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics 34, i457–i466.