A Biomedical Information Extraction Primer for NLP Researchers

Surag Nair
Indian Institute of Technology Delhi
ee1130504@iitd.ac.in

Abstract

Biomedical Information Extraction is an exciting field at the crossroads of Natural Language Processing, Biology and Medicine. It encompasses a variety of different tasks that require application of state-of-the-art NLP techniques, such as NER and Relation Extraction. This paper provides an overview of the problems in the field and discusses some of the techniques used for solving them.

1 Introduction

The explosion of available scientific articles in the Biomedical domain has led to the rise of Biomedical Information Extraction (BioIE). BioIE systems aim to extract information from a wide spectrum of articles including medical literature, biological literature, electronic health records, etc. that can be used by clinicians and researchers in the field. Often the outputs of BioIE systems are used to assist in the creation of databases, or to suggest new paths for research. For example, a ranked list of interacting proteins that are extracted from biomedical literature, but are not present in existing databases, can allow researchers to make informed decisions about which protein/gene to study further. Interactions between drugs are necessary for clinicians who simultaneously administer multiple drugs to their patients. A database of diseases, treatments and tests is beneficial for doctors consulting in complicated medical cases.

The main problems in BioIE are similar to those in Information Extraction:

1. Named Entity Recognition
2. Relation Extraction
3. Event Extraction

This paper discusses, in each section, various methods that have been adopted to solve the listed problems. Each section also highlights the difficulty of Information Extraction tasks in the biomedical domain.

This paper is intended as a primer to Biomedical Information Extraction for current NLP researchers. It aims to highlight the diversity of the various techniques from Information Extraction that have been applied in the Biomedical domain. The state of biomedical text mining is reviewed regularly. For more extensive surveys, consult (Liu et al., 2016), (Aggarwal and Zhai, 2012), (Zweigenbaum et al., 2007).

2 Named Entity Recognition and Fact Extraction

Named Entity Recognition (NER) in the Biomedical domain usually includes recognition of entities such as proteins, genes, diseases, treatments, drugs, etc. Fact extraction involves extraction of Named Entities from a corpus, usually given a certain ontology. When compared to NER in the domain of general text, the biomedical domain has some characteristic challenges:

1. Synonymy: the same biomedical entity is often known by different names. E.g. “cyclin-dependent kinase inhibitor p27” and “p27kip1” are the same proteins, “heart attack” and “myocardial infarcation” refer to the same medical problem.
2. Abbreviations: The literature is rich with ambiguous abbreviations: “RA” can refer to “right atrium”, “rheumatoid arthritis”, “renal artery” or several other concepts (Pakhomov, 2002)
3. Entity names are subject to many variants, and also change over time

Some of the earliest systems were heavily dependent on hand-crafted features. The method proposed in (Fukuda et al., 1998) for recognition of protein names in text does not require any prepared dictionary. The work gives examples of diversity in protein names and lists multiple rules depending on simple word features as well as POS tags.

(de Bruijn et al., 2011) adopt a machine learning approach for NER. Their NER system extracts medical problems, tests and treatments from discharge summaries and progress notes. They use a semi-Conditional Random Field (semi-CRF) (Sarawagi et al., 2004) to output labels over all tokens in the sentence. They use a variety of token, context and sentence level features. They also use some concept mapping features using existing annotation tools, as well as Brown clustering to form 128 clusters over the unlabelled data. The dataset used is the i2b2 2010 challenge dataset. Their system achieves an F-Score of 0.85. (Tang et al., 2014) is an incremental paper on NER taggers. It uses 3 types of word-representation techniques (Brown clustering, distributional clustering, word vectors) to improve performance of the NER Conditional Random Field tagger, and achieves marginal F-Score improvements.

(Movshovitz-Attias and Cohen, 2012) propose a boostrapping mechanism to bootstrap biomedical ontologies using NELL (Carlson et al., 2010), which uses a coupled semi-supervised bootstrapping approach to extract facts from text, given an ontology and a small number of "seed" examples for each category. This interesting approach (called BioNELL) uses an ontology of over 100 categories. In contrast to NELL, BioNELL does not contain any relations in the ontology. BioNELL is motivated by the fact that a lot of scientific literature available online is highly reliable due to peer-review. The authors note that the algorithm used by NELL to bootstrap fails in BioNELL due to ambiguities in biomedical literature, and heavy semantic drift. One of the causes for this is that often common words such as "white", "dad", "arm" are used as names of genes- this can easily result in semantic drift in one iteration of the bootstrapping. In order to mitigate this, they use Pointwise Mutual Information scores for corpus level statistics, which attributes a small score to common words. In addition, in contrast to NELL, BioNELL only uses high instances as seeds in the next iteration, but adds low ranking instances to the knowledge base. Since evaluation is not possible using Mechanical Turk or a small number of experts (due to the complexity of the task), they use Freebase (Bollacker et al., 2008), a knowledge base that has some biomedical concepts as well. The lexicon learned using BioNELL is used to train an NER system. The system shows a very high precision, thereby showing that BioNELL learns very few ambiguous terms.

More recently, deep learning techniques have been developed to further enhance the performance of NER systems. (Li et al., 2015) explore recurrent neural networks for the problem of NER in biomedical text.

3 Relation Extraction

In Biomedical Information Extraction, Relation Extraction involves finding related entities of many different kinds. Some of these include protein-protein interactions, disease-gene relations and drug-drug interactions. Due to the explosion of available biomedical literature, it is impossible for one person to extract relevant relations from published material. Automatic extraction of relations assists in the process of database creation, by suggesting potentially related entities with links to the source article. For example, a database of drug-drug interactions is important for clinicians who administer multiple drugs simultaneously to their patients- it is imperative to know if one drug will have an adverse effect on the other. A variety of methods have been developed for relation extractions, and are often inspired by Relation Extraction in NLP tasks. These include rule-based approaches, hand-crafted patterns, feature-based and kernel machine learning methods, and more recently deep learning architectures. Relation Extraction systems over Biomedical Corpora are often affected by noisy extraction of entities, due to ambiguities in names of proteins, genes, drugs etc.

(Blaschke and Valencia, 2001) was one of the first large scale Information Extraction efforts to study the feasibility of extraction of protein-protein interactions (such as "protein A activates protein B") from Biomedical text. Using 8 hand-crafted regular expressions over a fixed vocabulary, the authors were able to achieve a recall of 30% for interactions present in The Dictionary of

Interacting Proteins (DIP) from abstracts in Medline. The method did not differentiate between the type of relation. The reasons for the low recall were the inconsistency in protein nomenclature, information not present in the abstract, and due to specificity of the hand-crafted patterns. On a small subset of extracted relations, they found that about 60% were true interactions between proteins not present in DIP.

(Bunescu et al., 2006) combine sentence level relation extraction for protein interactions with corpus level statistics. Similar to (Blaschke and Valencia, 2001), they do not consider the type of interaction between proteins- only whether they interact in the general sense of the word. They also do not differentiate between genes and their protein products (which may share the same name). They use Pointwise Mutual Information (PMI) for corpus level statistics to determine whether a pair of proteins occur together by chance or because they interact. They combine this with a confidence aggregator that takes the maximum of the confidence of the extractor over all extractions for the same protein-pair. The extraction uses a subsequence kernel based on (Bunescu and Mooney, 2005). The integrated model, that combines PMI with aggregate confidence, gives the best performance. Kernel methods have widely been studied for Relation Extraction in Biomedical Literature. Common kernels used usually exploit linguistic information by utilising kernels based on the dependency tree (Liu et al., 2013), (Zhang et al., 2012), (Patra and Saha, 2013).

(Chun et al., 2006) look at the extraction of diseases and their relevant genes. They use a dictionary from six public databases to annotate genes and diseases in Medline abstracts. In their work, the authors note that when both genes and diseases are correctly identified, they are related in 94% of the cases. The problem then reduces to filtering incorrect matches using the dictionary, which occurs due to false positives resulting from ambiguities in the names as well as ambiguities in abbreviations. To this end, they train a Max-Ent based NER classifier for the task, and get a 26% gain in precision over the unfiltered baseline, with a slight hit in recall. They use POS tags, expanded forms of abbreviations, indicators for Greek letters as well as suffixes and prefixes commonly used in biomedical terms.

(Bui et al., 2014) adopt a supervised feature-based approach for the extraction of drug-drug interaction (DDI) for the DDI-2013 dataset (Herrero-Zazo et al., 2013). They partition the data in subsets depending on the syntactic features, and train a different model for each. They use lexical, syntactic and verb based features on top of shallow parse features, in addition to a hand-crafted list of trigger words to define their features. An SVM classifier is then trained on the feature vectors, with a positive label if the drug pair interacts, and negative otherwise. Their method beats other systems on the DDI-2013 dataset. Some other feature-based approaches are described in (Leaman et al., 2015), (Bui et al., 2011).

Distant supervision methods have also been applied to relation extraction over biomedical corpora. In (Liu et al., 2014), 10,000 neuroscience articles are distantly supervised using information from UMLS Semantic Network to classify brain-gene relations into geneExpression and otherRelation. They use lexical (bag of words, contextual) features as well as syntactic (dependency parse features). They make the "at-least one" assumption, i.e. at least one of the sentences extracted for a given entity-pair contains the relation in database. They model it as a multi-instance learning problem and adopt a graphical model similar to (Hoffmann et al., 2011). They test using manually annotated examples. They note that the F-score achieved are much lesser than that achieved in the general domain in (Hoffmann et al., 2011), and attribute to generally poorer performance of NER tools in the biomedical domain, as well as less training examples. (Thomas et al., 2011) explore distant supervision methods for protein-protein interaction extraction.

More recently, deep learning methods have been applied to relation extraction in the biomedical domain. One of the main advantages of such methods over traditional feature or kernel based learning methods is that they require minimal feature engineering. In (Jiang et al., 2016), skip-gram vectors (Mikolov et al., 2013) are trained over 5.6Gb of unlabelled text. They use these vectors to extract protein-protein interactions by converting them into features for entities, context and the entire sentence. Using an SVM for classification, their method is able to outperform many kernel and feature based methods over a variety of datasets.

(Sahu et al., 2016) follow a similar method by using word vectors trained on PubMed articles. They use it for the task of relation extraction from clinical text for entities that include problem, treatment and medical test. For a given sentence, given labelled entities, they predict the type of relation exhibited (or None) by the entity pair. These types include "treatment caused medical problem", "test conducted to investigate medical problem", "medical problem indicates medical problems", etc. They use a Convolutional Neural Network (CNN) followed by feedforward neural network architecture for prediction. In addition to pre-trained word vectors as features, for each token they also add features for POS tags, distance from both the entities in the sentence, as well BIO tags for the entities. Their model performs better than a feature based SVM baseline that they train themselves.

The BioNLP'16 Shared Tasks has also introduced some Relation Extraction tasks, in particular the BB3-event subtask that involves predicting whether a "lives-in" relation holds for a Bacteria in a location. Some of the top performing models for this task are deep learning models. (Mehryary et al., 2016) train word embeddings with six billions words of scientific texts from PubMed. They then consider the shortest dependency path between the two entities (Bacteria and location). For each token in the path, they use word embedding features, POS type embeddings and dependency type embeddings. They train a unidirectional LSTM (Hochreiter and Schmidhuber, 1997) over the dependency path, that achieves an F-Score of 52.1% on the test set.

(Li et al., 2016) improve the performance by making modifications to the above model. Instead of using the shortest dependency path, they modify the parse tree based on some pruning strategies. They also add feature embeddings for each token to represent the distance from the entities in the shortest path. They then train a Bidirectional LSTM on the path, and obtain an F-Score of 57.1%.

The recent success of deep learning models in Biomedical Relation Extraction that require minimal feature engineering is promising. This also suggests new avenues of research in the field. An approach as in (Zeng et al., 2015) can be used to combine multi-instance learning and distant supervision with a neural architecture.

4 Event Extraction

Event Extraction in the Biomedical domain is a task that has gained more importance recently. Event Extraction goes beyond Relation Extraction. In Biomedical Event Extraction, events generally refer to a change in the state of biological molecules such as proteins and DNA. Generally, it includes detection of targeted event types such as gene expression, regulation, localisation and transcription. Each event type in addition can have multiple arguments that need to be detected. An additional layer of complexity comes from the fact that events can also be arguments of other events, giving rise to a nested structure. This helps to capture the underlying biology better (Aggarwal and Zhai, 2012). Detecting the event type often involves recognising and classifying trigger words. Often, these words are verbs such as "activates", "inhibits", "phosphorylation" that may indicate a single, or sometimes multiple event types. In this section, we will discuss some of the successful models for Event Extraction in some detail.

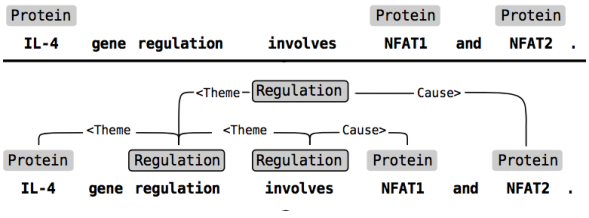

Figure 1: An example of an input sentence with annotations, and expected output of the event extraction system (Borrowed from (Björne et al., 2009))

Event Extraction gained a lot of interest with the availability of an annotated corpus with the BioNLP'09 Shared Task on Event Extraction (Kim et al., 2008). The task involves prediction of trigger words over nine event types such as expression, transcription, catabolism, binding, etc. given only annotation of named entities (proteins, genes, etc.). For each event, its class, trigger expression and arguments need to be extracted. Since the events can be arguments to other events, the final output in general is a graph representation with events and named entities as nodes, and edges that correspond to event arguments. (Björne et al., 2009) present a pipeline based method that is heavily dependent on dependency parsing. Their pipeline approach consists of three steps: trigger detection, argument detection and semantic post-

processing. While the first two components are learning based systems, the last component is a rule based system. For the BioNLP'09 corpus, only 5% of the events span multiple sentences. Hence the approach does not get affected severely by considering only single sentences. It is important to note that trigger words cannot simply be reduced to a dictionary lookup. This is because a specific word may belong to multiple classes, or may not always be a trigger word for an event. For example, "activate" is found to not be a trigger word in over 70% of the cases. A multi-class SVM is trained for trigger detection on each token, using a large feature set consisting of semantic and syntactic features. It is interesting to note that the hyperparameters of this classifier are optimised based on the performance of the entire end-to-end system.

For the second component to detect arguments, labels for edges between entities must be predicted. For the BioNLP'09 Shared Task, each directed edge from one event node to another event node, or from an event node to a named entity node are classified as "theme", "cause", or None. The second component of the pipeline makes these predictions independently. This is also trained using a multi-class SVM which involves heavy use of syntactic features, including the shortest dependency path between the nodes. The authors note that the precision-recall choice of the first component affects the performance of the second component: since the second component is only trained on Gold examples, any error by the first component will lead to a cascading of errors. The final component, which is a semantic post-processing step, consists of rules and heuristics to correct the output of the second component. Since the edge predictions are made independently, it is possible that some event nodes do not have any edges, or have an improper combination of edges. The rule based component corrects these and applies rules to break directed cycles in the graph, and some specific heuristics for different types of events. The final model gives a cumulative F-Score of 52% on the test set, and was the best model on the task.

(Poon and Vanderwende, 2010) note that previous approaches on the task suffer due to the pipeline nature and the propagation of errors. To counter this, they adopt a joint inference method based on Markov Logic Networks (Richardson and Domingos, 2006) for the same task on BioNLP'09. The Markov Logic Network jointly predicts whether each token is a trigger word, and if yes, the class it belongs to; for each dependency edge, whether it is an argument path leading to a "theme" or a "cause". By formulating the Event Extraction problem using an MLN, the approach becomes computationally feasible and only linear in the length of the sentence. They incorporate hard constraints to encode rules such as "an argument path must have an event", "a cause path must start with a regulation event", etc. In addition, they also include some domain specific soft constraints as well as some linguistically-motivated context-specific soft constraints. In order to train the MLN, stochastic gradient descent was used. Certain heuristic methods are implemented in order to deal with errors due to syntactic parsing, especially ambiguities in PP-attachment and coordination. Their final system is competitive and comes very close to the system by (Björne et al., 2009) with an average F-Score of 50%. To further improve the system, they suggest leveraging additional joint-inference opportunities and integrating the syntactic parser better. Some other more recent models for Biomedical Event Extraction include (Riedel and McCallum, 2011), (McClosky et al., 2012).

5 Conclusion

We have discussed some of the major problems and challenges in BioIE, and seen some of the diverse approaches adopted to solve them. Some interesting problems such as Pathway Extraction for Biological Systems (Ananiadou et al., 2010), (Rzhetsky et al., 2004) have not been discussed.

Biomedical Information Extraction is a challenging and exciting field for NLP researchers that demands application of state-of-the-art methods. Traditionally, there has been a dependence on hand-crafted features or heavily feature-engineered methods. However, with the advent of deep learning methods, a lot of BioIE tasks are seeing an improvement by adopting deep learning models such as Convolutional Neural Networks and LSTMs, which require minimal feature engineering. Rapid progress in developing better systems for BioIE will be extremely helpful for clinicians and researchers in the Biomedical domain.

References

Charu C Aggarwal and ChengXiang Zhai. 2012. *Mining text data*. Springer Science & Business Media.

Sophia Ananiadou, Sampo Pyysalo, Junichi Tsujii, and Douglas B Kell. 2010. Event extraction for systems biology by text mining the literature. *Trends in biotechnology* 28(7):381–390.

Jari Björne, Juho Heimonen, Filip Ginter, Antti Airola, Tapio Pahikkala, and Tapio Salakoski. 2009. Extracting complex biological events with rich graph-based feature sets. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*. Association for Computational Linguistics, pages 10–18.

Christian Blaschke and Alfonso Valencia. 2001. Can bibliographic pointers for known biological data be found automatically? protein interactions as a case study. *Comparative and Functional Genomics* 2(4):196–206.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. AcM, pages 1247–1250.

Quoc-Chinh Bui, Sophia Katrenko, and Peter MA Sloot. 2011. A hybrid approach to extract protein–protein interactions. *Bioinformatics* 27(2):259–265.

Quoc-Chinh Bui, Peter MA Sloot, Erik M Van Mulligen, and Jan A Kors. 2014. A novel feature-based approach to extract drug–drug interactions from biomedical text. *Bioinformatics* page btu557.

Razvan Bunescu, Raymond Mooney, Arun Ramani, and Edward Marcotte. 2006. Integrating co-occurrence statistics with information extraction for robust retrieval of protein interactions from medline. In *Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis*. Association for Computational Linguistics, pages 49–56.

Razvan Bunescu and Raymond J Mooney. 2005. Subsequence kernels for relation extraction. In *NIPS*. pages 171–178.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. 2010. Toward an architecture for never-ending language learning. In *AAAI*. volume 5, page 3.

Hong-Woo Chun, Yoshimasa Tsuruoka, Jin-Dong Kim, Rie Shiba, Naoki Nagata, Teruyoshi Hishiki, and Jun'ichi Tsujii. 2006. Extraction of gene-disease relations from medline using domain dictionaries and machine learning. In *Pacific Symposium on Biocomputing*. volume 11, pages 4–15.

Berry de Bruijn, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Xiaodan Zhu. 2011. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *Journal of the American Medical Informatics Association* 18(5):557–562.

K Fukuda, A Tamura, T Tsunoda, and T Takagi. 1998. Toward ie: Identifying protein names from biological papers. In *Proceedings of the Pacific Symposium on Biocomputing (PSB98)*.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics* 46(5):914–920.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 541–550.

Zhenchao Jiang, Lishuang Li, and Degen Huang. 2016. A general protein-protein interaction extraction architecture based on word representation and feature selection. *International Journal of Data Mining and Bioinformatics* 14(3):276–291.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC bioinformatics* 9(1):10.

Robert Leaman, Chih-Hsuan Wei, and Zhiyong Lu. 2015. tmchem: a high performance approach for chemical named entity recognition and normalization. *Journal of cheminformatics* 7(1):S3.

Lishuang Li, Liuke Jin, and Degen Huang. 2015. Exploring recurrent neural networks to detect named entities from biomedical text. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, Springer, pages 279–290.

Lishuang Li, Jieqiong Zheng, Jia Wan, Degen Huang, and Xiaohui Lin. 2016. Biomedical event extraction via long short term memory networks along dynamic extended tree. In *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*. IEEE, pages 739–742.

Feifan Liu, Jinying Chen, Abhyuday Jagannatha, and Hong Yu. 2016. Learning for biomedical information extraction: Methodological review of recent advances. *arXiv preprint arXiv:1606.07993* .

Jianzhou Liu, Liang Xiao, and Xiongkai Shao. 2013. A new approach to extract biomedical events based on composite kernel. In *Information Science and Technology (ICIST), 2013 International Conference on*. IEEE, pages 39–42.

Mengwen Liu, Yuan Ling, Yuan An, and Xiaohua Hu. 2014. Relation extraction from biomedical literature with minimal supervision and grouping strategy. In *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*. IEEE, pages 444–449.

David McClosky, Sebastian Riedel, Mihai Surdeanu, Andrew McCallum, and Christopher D Manning. 2012. Combining joint models for biomedical event extraction. *BMC bioinformatics* 13(11):S9.

Farrokh Mehryary, Jari Björne, Sampo Pyysalo, Tapio Salakoski, and Filip Ginter. 2016. Deep learning with minimal training data: Turkunlp entry in the bionlp shared task 2016. *ACL 2016* page 73.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Dana Movshovitz-Attias and William W Cohen. 2012. Bootstrapping biomedical ontologies for scientific text using nell. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics, pages 11–19.

Serguei Pakhomov. 2002. Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, pages 160–167.

Rakesh Patra and Sujan Kumar Saha. 2013. A kernel-based approach for biomedical named entity recognition. *The Scientific World Journal* 2013.

Hoifung Poon and Lucy Vanderwende. 2010. Joint inference for knowledge extraction from biomedical literature. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 813–821.

Matthew Richardson and Pedro Domingos. 2006. Markov logic networks. *Machine learning* 62(1-2):107–136.

Sebastian Riedel and Andrew McCallum. 2011. Fast and robust joint models for biomedical event extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1–12.

Andrey Rzhetsky, Ivan Iossifov, Tomohiro Koike, Michael Krauthammer, Pauline Kra, Mitzi Morris, Hong Yu, Pablo Ariel Duboué, Wubin Weng, W John Wilbur, et al. 2004. Geneways: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of biomedical informatics* 37(1):43–53.

Sunil Kumar Sahu, Ashish Anand, Krishnadev Oruganty, and Mahanandeeshwar Gattu. 2016. Relation extraction from clinical texts using domain invariant convolutional neural network. *arXiv preprint arXiv:1606.09370* .

Sunita Sarawagi, William W Cohen, et al. 2004. Semi-markov conditional random fields for information extraction. In *NIPs*. volume 17, pages 1185–1192.

Buzhou Tang, Hongxin Cao, Xiaolong Wang, Qingcai Chen, and Hua Xu. 2014. Evaluating word representation features in biomedical named entity recognition tasks. *BioMed research international* 2014.

Philippe Thomas, Illés Solt, Roman Klinger, and Ulf Leser. 2011. Learning protein protein interaction extraction using distant supervision. *Robust Unsupervised and Semi-Supervised Methods in Natural Language Processing* pages 34–41.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *EMNLP*. pages 1753–1762.

Yijia Zhang, Hongfei Lin, Zhihao Yang, Jian Wang, and Yanpeng Li. 2012. Hash subgraph pairwise kernel for protein-protein interaction extraction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 9(4):1190–1202.

Pierre Zweigenbaum, Dina Demner-Fushman, Hong Yu, and Kevin B Cohen. 2007. Frontiers of biomedical text mining: current progress. *Briefings in bioinformatics* 8(5):358–375.