BIOMEDICAL RELATION EXTRACTION WITH REDUCED MANUAL EFFORT

by

Gang Li

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science

Spring 2018

BIOMEDICAL RELATION EXTRACTION WITH REDUCED MANUAL EFFORT

by

Gang Li

Approved: ______________________________
Kathleen F. McCoy, Ph.D.
Chair of the Department of Computer and Information Sciences

Approved: ______________________________
Babatunde A. Ogunnaike, Ph.D.
Dean of the College of Engineering

Approved: ______________________________
Ann L. Ardis, Ph.D.
Senior Vice Provost for Graduate and Professional Education

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: ____

Vijay K. Shanker, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: ____

Cathy H. Wu, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: ____

Li Liao, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: ____

Zhiyong Lu, Ph.D.
Member of dissertation committee

ACKNOWLEDGEMENTS

First, I would like to express my gratitude to my advisors Profs. Cathy H. Wu and Vijay K. Shanker. This dissertation would not have been possible without their continuous support and guidance through my Ph.D. study. I learned many things from them through the countless meetings and discussions we had, such as how to conduct a research project, how to present ideas effectively and the attention to details. These skills have helped me in many ways in my research and I believe I will continue to benefit from them in my future career.

I would like to also thank the rest of my dissertation committee, Prof. Liao Li and Dr. Zhiyong Lu. Their insightful comments and suggestions have enabled me to look at my research at different angles and improve my dissertation.

In the past six years, I have spent a lot of time in the BioNLP lab with other Ph.D. students and postdoctoral researchers. I would like to thank them for the days we worked and had fun together. Special thanks go to Ruoyao Ding, Samir Gupta, Ashique Mahmood, Yifan Peng, Manabu Torii and Catalina Oana Tudor.

Many thanks to my colleagues in the Center for Bioinformatics and Computational Biology: Drs. Hongzhan Huang, Cecilia Arighi and Karen Ross. This is the place I first started as a web developer before I stepped into the field of Biomedical Text Mining. We together did many biomedical-related projects and they were always happy to answer my silly questions in the biological field.

Finally and most importantly, I wish to thank my parents, for their unconditional love and support. I would like to also thank Jia Ren for her love and longtime accompany.

TABLE OF CONTENTS

LIST OF TABLES . . . . **ix**
LIST OF FIGURES . . . . **xi**
ABSTRACT . . . . **xii**

Chapter

1 INTRODUCTION **1**

1.1 Biomedical Relation Extraction 1
1.2 Thesis Contributions 3
1.3 Outline 4

2 RELATED WORK **6**

2.1 Biomedical Relation Extraction 6

2.1.1 Rule-based System Development with Reduced Manual Effort 6
2.1.2 Automatic Training Set Construction with Distant Supervision 7
2.1.3 Deep Neural Networks 9

2.2 Relation Extraction Tasks and Corpora 9

2.2.1 miRNA-Gene Relation 9
2.2.2 Protein Post-Translational Modification 11
2.2.3 Protein-Protein Interaction 12
2.2.4 Protein Subcellular Localization 13

3 MIRNA-GENE RELATION EXTRACTION **14**

3.1 miRNA-Gene Relation 14
3.2 miRTex System 16

3.2.1 Pre-processing 17

3.2.2 Lexico-Syntactic Rules . . . 17
3.2.3 Null-Argument Rules . . . 19
3.2.4 Linking Relation . . . 20
3.2.5 Relaxing the Same-Trigger Constraint . . . 22
3.2.6 Detection of miRNA-Target Relations . . . 23

3.3 Evaluation . . . 24

3.3.1 Evaluation on In-house Developed Corpus . . . 24
3.3.2 Evaluation on Bagewadi et al. Corpus . . . 26

3.4 Full-Scale Processing . . . 26
3.5 Conclusion . . . 27

4 A GENERAL RULE-BASED RELATION EXTRACTION SYSTEM . . . 28

4.1 Extended Dependency Graph . . . 29
4.2 Lexicalized Tree-Adjoining Grammar . . . 30
4.3 Linguistic-based Pattern Generation . . . 31

4.3.1 Trigger Word Types . . . 32
4.3.2 Pattern Generation . . . 32
4.3.3 Entity Typing Incorporation . . . 35
4.3.4 System Overview . . . 36

4.4 Post-Translational Modification Information Extraction . . . 37

4.4.1 Post-Translational Modification Relation . . . 37
4.4.2 Named Entity Recognition . . . 38
4.4.3 Post-translational Modification Trigger Words . . . 38
4.4.4 PTM-specific Rules . . . 39
4.4.5 Extraction of Additional Arguments for BioNLP Corpus . . . 42

4.5 Experiment Results . . . 43

4.5.1 Evaluation on In-house Developed Corpus . . . 43
4.5.2 Evaluation on BioNLP 2011 EPI Test Corpus . . . 44
4.5.3 Effects of Referential Relations and Merging Rules . . . 46

4.6 Error Analysis . . . 47

4.6.1 Errors in In-house Developed Corpus . . . 47

4.6.2 Errors in BioNLP 2011 EPI Training Corpus 50

4.7 Conclusion 50

5 DISTANT SUPERVISION FOR BIOMEDICAL RELATION EXTRACTION 52

5.1 Distant Supervision 52

5.1.1 Heuristical Labeling Process 53
5.1.2 Noise in Positively-Labeled Data 53
5.1.3 Noise in Negatively-Labeled Data 54

5.2 Convolutional Neural Networks 56
5.3 Combining Distantly-labeled and Human-labeled Data 58
5.4 Experiment Tasks 59
5.5 Training Set Construction 59
5.6 Test Sets 60
5.7 Methods 60

5.7.1 Learning Algorithm and Feature Set 61
5.7.2 Baselines 62
5.7.3 Noise Reduction Methods 62
5.7.4 Application of Convolutional Neural Networks 66
5.7.5 Parameters Tuning 68

5.8 Results 69

5.8.1 Results for Noise Reduction Methods 69
5.8.2 Results for Convolutional Neural Networks 72

5.9 Combining Distantly-labeled and Human-labeled Data 73

5.9.1 Methods 73
5.9.2 Results 74

5.10 Conclusion 75

6 CONCLUSIONS 77

6.1 Thesis Summary and Contributions 77
6.2 Future work 80

BIBLIOGRAPHY **82**

Appendix

A PTM TRIGGER WORDS **92**
B AUTOMATICALLY MINED TRIGGER STEMS **93**
C PARAMETERS OF CONVOLUTIONAL MODELS **95**

C.1 CNN Models Training Solely on Distantly Labeled Data 95
C.2 PCNN Models Trained on Human-labeled Data (10-fold instance-level CV) 96
C.3 PCNN Models Trained on Both Distantly-labeled and Human-labeled Data (10-fold instance-level CV) 96
C.4 PCNN Models Trained on Human-labeled Data (10-fold document-level CV) 96
C.5 PCNN Models Trained on Both Distantly-labeled and Human-labeled Data (10-fold document-level CV) 97

LIST OF TABLES

2.1 Characteristics of Bagewadi et al. Corpus. 10

2.2 Characteristics of BioNLP 2011 EPI Corpus. 12

2.3 Characteristics of AIMed and BioInfer Corpus. 12

2.4 Protein-Localization Events in BioNLP 2011 Genia Corpus. 13

3.1 miRTex Development and Test Set. 25

3.2 Performance for Extraction of Three miRNA-Gene Relations. . . . 26

3.3 Result Statistics of Full-Scale Processing 27

4.1 Agent Patterns 33

4.2 Theme Patterns 34

4.3 Trigger Words for PTM Relations. 39

4.4 Characteristics of In-house Developed Corpus 43

4.5 Evaluation Results on In-house Developed Corpus 44

4.6 Evaluation Results for Full Relation Extraction on BioNLP 2011 EPI Test Corpus 45

4.7 Evaluation Results for Substrate Extraction on BioNLP 2011 EPI Test Corpus 45

5.1 Text Corpora and Databases for Three Tasks 59

5.2 Test Sets for Three Tasks. 60

5.3 Features Extracted from Example Eentence. 61

5.4 Top 10 Verb Stems and Top Pattern and Example Sentence for Three Tasks. 64

5.5 Development and Test Set for Three Relations. 68

5.6 Precisions, Recalls and F-scores for Three Extraction Tasks. 70

5.7 Results for Convolutional Neural Networks. 73

5.8 Document-level Cross Validation Results for Three Relations 75

5.9 Instance-level Cross Validation Results for Three Relations 75

LIST OF FIGURES

3.1	Overview of miRTex System	16
4.1	An Example Sentence for Extended Dependency Graph.	29
4.2	Parse Trees from the Transitive Verb Family.	30
4.3	Invalid Parse Tree for "phosphorylate".	31
4.4	Two Syntactically Transformable Dependency Parses.	33
4.5	General Relation Extraction System Architecture.	36
4.6	System Comparison for Acetylation Extraction	46
4.7	System Comparison for Glycosylation Extraction	47
4.8	System Comparison for Ubi-like Modification Extraction	47
5.1	Typical Structure of Convolutional Neural Network Model.	56
5.3	Example Sentence for Feature Extraction.	61
5.4	Model Structure of Multi-channel CNN and Piecewise CNN.	67
5.5	Results of Using Different Numbers of Trigger Stems and Patterns.	71
5.7	Precision-Recall Curves for Three Tasks.	72
5.9	Pipeline for Combining Distantly and Human-labeled Data	74

ABSTRACT

Biomedical relation extraction is an critical text-mining task that concerns automatic extraction of related bio-entities in text. Rule-based and machine learning methods are two main approaches for relation extraction. While these two methods can be used to develop high-performance relation extraction system, considerable amount of manual effort is required by both methods in different phases of the system development. This hinders fast application of both methods to extract new types of relations.

This dissertation focuses on developing techniques to assist the development of biomedical relation extraction systems using rule-based and machine learning methods. For rule-based systems, one main component requiring manual effort is pattern design. Domain experts often need to examine considerable amount of documents and exhaustively collect every pattern that can be used to extract relations. We leverage various linguistic knowledge to automatically generate a comprehensive set of patterns. Our first approach is instantiated to develop miRTex, a system that extracts three kinds of miRNA-gene relations that regulate a wide range of biological processes and are involved with diseases. Only a small number of triggers and rules are needed to achieve the state-of-the-art performance. Our second approach is to translate the ideas in Lexicalized Tree Adjoining Grammar to dependency graph for pattern generation, and adopt Extended Dependency Graph as an abstract sentence representation. This approach is applied to extract five type of post-translational modifications, a class of relations that plays an important role in cellular functions. Evaluations on BioNLP 2011 EPI task show that the resulting system achieves state-of-the-art performance.

For machine learning systems, a sizable training corpus is needed to train the extraction model, while the annotation of the corpus is time- and labor-intensive. We

adopt distant supervision in two ways. Our first contribution is to develop noise reduction techniques to improve the data quality of the automatically generated large training set, leading to improvement over existing results for distantly supervised biomedical relation extraction. Secondly, we employ distant supervision in conjunction with human-labeled data and deep neural networks to achieve state-of-the-art performance on some benchmark relation extraction tasks.

Chapter 1

INTRODUCTION

As the biomedical literature grows rapidly, it has become more and more difficult for researchers to keep up with new findings reported in the publications. With the ability to automatically extract structural information from large amount of text, relation extraction has attracted increasing interest and assisted many biological tasks, including database curation, biological hypothesis generation and knowledge discovery.

In general, two kinds of methods are commonly used for relation extraction: rule-based and machine learning approaches. Considerable amount of human effort is required by both methods and hinders the fast application to extract new types of relations. To mitigate this problem, this dissertation proposes different methods to extract biomedical relation with reduced human effort. Specifically, we incorporate various linguistic knowledge to assist automatic pattern generation for rule-based system, and adopt distant supervision to generate training set without human labeling for machine learning system. Experiments show that our methods achieve state-of-the-art results compared to previous work. We hope that the research work reported in this dissertation will facilitate development of better biomedical relation extraction systems with less human effort involved.

In Section 1.1, we will briefly discuss the task of biomedical relation extraction and existing methods. In Section 1.2, we present the contributions of this dissertation. Section 1.3 will give an outline of the rest part of the dissertation.

1.1 Biomedical Relation Extraction

Biomedical relation extraction is a task that concerns extraction of related bio-entities in text. Typical examples include protein-protein interaction [1], gene-disease association [49] and drug-drug interaction [84]. While most of the text-mining tasks focus on extraction of binary relations between two bio-entities, there has been effort on extraction of n-ary relations which involve more than two entities [38, 89, 90]. Competition events organized by the research community, such as BioCreative [4, 97] and BioNLP 2009-2013 [38–40], offer comprehensive evaluation for a wide range of biomedical relation extraction tasks.

Rule-based and machine learning approaches are the two main categories of methods used in biomedical relation extraction. In rule-based methods, linguistic and/or biological knowledge is encoded in rules that can be used to extract relations from text. Such knowledge includes plain-text/syntactic patterns, specific words that indicate biological relations and post-processing heuristics. High-performance rule-based systems have been developed for various biomedical relations, such as OpenDMAP [19], RLIMS-P [89], BioSem [12] and eFIP [90]. However, the success of rule-based systems depends on a high-quality and comprehensive set of patterns. There are two main challenges for creating such patterns. First, they are usually manually designed by domain experts by examining a development corpus and thus cost a lot of time and effort. Secondly, text variants not covered in the development corpus might not be included in the patterns, resulting in lower recall. In this dissertation, we show that with linguistic knowledge and a suitable abstract level of sentence representation we can automatically generate high-quality and comprehensive pattern set to build a high-performance relation extraction system.

Machine learning methods view relation extraction as a classification problem and train extraction models using various algorithms, e.g., Support Vector Machines [21] or Naive Bayes classifier [18], to extract relations. In many well-studied tasks, such as protein-protein interaction extraction, or Genia Event Extraction in BioNLP, machine learning methods have been widely applied [1,13,55,75]. Except tuning some model parameters on a development set, the training of the extraction model needs little human intervention. However, the training corpus is manually annotated.

The annotation process is time- and labor-intensive, and often involves multiple annotators for calculating annotation agreement. Recently, distant supervision has been applied to automatically generate labeled data using related entity pairs from database for biomedical relation extraction [88, 104]. Although it can create a training set without human effort, the generated data is noisy due to the aggressive labeling process and incompleteness of database. In this dissertation, we propose three noise reduction heuristics and achieve improved results compared to previous effort for distantly supervised biomedical relation extraction. Furthermore, since distant supervision can be used to build large data sets, we can better train deep neural networks by utilizing the large data. Finally, we combine the large distantly labeled data set and small amount of human-labeled data for training deep neural network models and obtain better performance than models solely trained on the small human-labeled set. Evaluations show that using this method we can achieve state-of-the-art result for supervised protein-protein interaction extraction.

1.2 Thesis Contributions

In this section, we briefly summarize the contributions of this thesis.

1. miRTex, a relation extraction system for miRNA-gene relations [45]**:** we developed a rule-based system that extracts three kinds of miRNA-gene relations from text. The three relations, which are miRNA-target, miRNA-gene regulation and gene-miRNA regulation, are important for understanding miRNA's role in various biological processes and disease contexts. Following the principles underlying the iXtractR system [69], the miRTex system exploits lexico-syntactic patterns and general linguistic relations and is able to achieve both high precision and recall using a small set of trigger words and rules.

2. A general relation extraction framework and application to the development of a PTM extraction system: the framework generalizes ideas from Lexicalized Tree-Adjoining Grammar [83] to dependency graph and utilizes Extended Dependency Graph (EDG) as the sentence representation [67]. It supports automatic generation of comprehensive rigorous patterns based on different verb types. This ensures high precision of the extracted results. The use of EDG abstracts away syntactic details and propagates extracted predicate arguments using is-a, member-collection and part-of relations, to increase recall. Experiments on extraction of five post-translational modification relations, acetylation, glycosylation, ubiquitination, sumoylation and neddylation, show that the framework can achieve high precision and recall with a small set of trigger words on an in-house developed corpus, and state-of-the-art performance on BioNLP 2011 EPI task for three kinds of PTM relations.

3. Improve distant supervision for biomedical relation extraction [46]**:** First, we propose three heuristics based on the notion of proximity, trigger word and confidence of patterns to leverage lexical and syntactic information to reduce noise in the distantly labeled data. Experiments on three relation extraction tasks, protein-protein interaction, miRNA-gene regulation and protein-localization event show that the proposed methods can improve the F-scores by a large margin over the baselines, and achieve better results than previous work. To better utilize the large dataset, we then train Convolutional Neural Networks (CNN) on the generated data, and achieve better performance than logistic regression methods. Finally, we further train the CNN models with a small amount of human-labeled data to improve performance. Compared to the CNN models solely trained on small human-labeled data, the models trained using both kinds of data achieve significantly better F-scores on all the three tasks. Specifically, for protein-protein interaction, a well-studied biomedical relation extraction task, we are able to surpass previous state-of-the-art with supervised learning and achieve 68.7 F-score with 69.4 precision and 68.6 recall. To the best of our knowledge, this is the first work that effectively combines distant supervision with human-labeled data using a two-step training process for biomedical relation extraction.

1.3 Outline

The rest of the dissertation is organized as follows. In Chapter 2 we will review the related works that provide the context and are useful for understanding the work in this dissertation. In Chapter 3, we present the miRTex system with the methods to reduce the rule design effort. We will discuss a general relation extraction framework and its application to extract five types of PTMs in Chapter 4. In Chapter 5, we report detailed experiment results on three relation extraction tasks using distant supervision and Convolutional Neural Networks. We conclude with contributions and future work in Chapter 6.

Chapter 2

RELATED WORK

In the first part of this chapter, we will review the related work for general biomedical relation extraction. In the second part, we introduce existing systems and corpora for the relation extraction tasks that we use for experiments in this dissertation.

2.1 Biomedical Relation Extraction

There are numerous important biomedical relations that can be mined from biomedical literature. Typical examples include protein-protein interaction [13], drug-drug interaction [84], post-translational modification [62] and disease-gene relation [49]. For each relation, considerable human effort is required from both computational as well as domain experts to develop a reasonable relation extraction system. To alleviate this situation, we propose different methods for developing rule-based and method learning systems that requires less human effort for biomedical relation extraction. First, we discuss previous effort for fast rule-based system development with reduced human effort. We then discuss the idea of distant supervision and its advantage and shortcomings. Several recent works applying deep neural networks for biomedical relation extraction will also be included.

2.1.1 Rule-based System Development with Reduced Manual Effort

Rule-based systems depend on lexical and/or syntactic patterns to extract relations. The development of patterns is a time- and labor-intensive task as it requires domain experts to examine considerable amount of documents. Furthermore, the number of patterns describing a relation can be very large. The development corpus may not be big enough to cover every possible pattern.

To enable fast development of patterns, a previous system iXtractR [69] has used a set of syntactic templates and trigger words to automatically generate rigorous patterns. The templates encode general syntactic pattern that can be found in many predicate-argument structures, e.g., subject-object structure. Then, these templates can be instantiated to produce a variety of rules when provided a list of trigger words, e.g., "bind" for protein-protein interaction. In this approach, as the syntactic templates are already defined, users only need to provide a list of trigger words for the relation to be extracted, which is a much easier task than finding all the possible patterns expressing the relation. iXtractR will combine a trigger and its corresponding syntactic templates to generate the rigorous patterns. It has been applied on BioNLP 2011 Genia corpus [39] for extraction of five different types of events and achieved comparable results to the state-of-the-art.

To increase extraction coverage, iXtractR and another rule-based system RLIMS-P [89] also leverage several linguistic notions to generalize the patterns. They exploit referential relations between two nominals, such as is-a, part-of, and member-collection relation, to find the actual target entity referred by an initially extracted argument. Combinations between the template-generated patterns and the referential relations dramatically increase the number of final rigorous patterns, thus expand the coverage of the extraction system without compromising precision. We use these referential relations for miRNA-gene relation extraction and will discuss them in detail in Chapter 3.

2.1.2 Automatic Training Set Construction with Distant Supervision

As the human-annotated corpus required by supervised machine learning algorithms needs to be manually created and often involves multiple annotators, we adopt the idea of distant supervision to construct the training set automatically. The first application of distant supervision to relation extraction was proposed by Craven et al. [20]. Craven et al. used 33 protein-localization pairs from Yeast Protein Database [29] to annotate 2889 Medline abstracts. They found the model trained on the distantly labeled set gained a better performance than a simple co-occurrence method. Mintz et al. [54] used Freebase [10] relations to annotate articles in Wikipedia and trained a logistic regression model to extract 102 different types of relations. The feature set they used contains lexical and syntactic features (e.g., dependency path) of the two entities in consideration. Evaluation showed that new relation instances which are absent in Freebase can be discovered using the trained model, and that the top-ranked 1000 instances for 10 relations achieved 0.68 precision. This work laid down a basis for an upcoming stream of research focusing on applying distant supervision with Freebase relations and Wikipedia articles. Along this direction, Riedel et al. [77] proposed to use multi-instance learning under the at-least-one assumption to mitigate noise in the positive training data. Hoffmann et al. [30] and Surdeanu et al. [86] continued to augment the multi-instance model with a multi-label classifier for each entity pair, to exploit correlations and conflicts among different relations to improve performance. In another direction, researchers focused on explicit noise reduction for the training data. Intxaurrondo et al. [36] proposed three simple heuristics to remove noise in positively-labeled data. They tried to filter out relation mentions that appear too often or have a large distance from their cluster centroid, or entity pairs that have a low partial mutual information, and found that combination of these heuristics improved the results significantly. Takamatsu et al. [87] proposed a statistical model to estimate the probability of a pattern expressing a relation, and removed low-probability patterns in the positive set. Meanwhile, Xu et al. [99] used pseudo-relevance feedback to discover high-confidence related entity pairs which do not exist in the database, and removed them from negative training data. Roller et al. [78] also tried to remove noise from the negative data by inferring new relations of the knowledge graph using a random-walk algorithm. Roth et al. [80] gave a nice review of some of the above methods.

A few previous works have tried applying distant supervision for biomedical relation extraction. Zheng et al. [104] used a method similar to [54] with a few noise-filtering heuristics for extraction of subcellular-localization relations. Thomas et al. [88]

used a list of words which are frequently employed to indicate protein interaction to filter out noise in positively and negatively-labeled data for protein-protein interaction extraction. Roller et al. [79] tried to combine existing hand-labeled data with distantly labeled data to improve the performance for drug-condition relations. Roller et al. [78] used the multi-instance learning algorithm to extract two subsets of relations in UMLS database with reduced noise by a path ranking algorithm.

2.1.3 Deep Neural Networks

Recently, deep neural networks have been applied for biomedical relation extraction [5, 35, 68, 75]. The two main network structures used are convolutional and recurrent neural networks. In these models, the words in the sentences are usually represented by distributed embedding vectors [53]. Other features such as POS tags, named entity types and syntactic information can be concatenated to the word embeddings to form an extended vector for each word. Convolutional neural networks (CNN) adopt filters with fixed size [44] to learn higher-level features of the inputs. Recurrent neural networks, such as Long Short-Term Memory [28] exploits the structure of the inputs (e.g., sentences) and learns to predict future outputs depending on history. Both kinds of model structures have been applied for biomedical relation extraction and achieved state-of-the-art results [31, 35, 81].

2.2 Relation Extraction Tasks and Corpora

In this section, we will discuss existing systems and publicly available corpora for the relation extraction tasks used for experiments in this dissertation.

2.2.1 miRNA-Gene Relation

In Chapter 3, we present a rule-based system miRTex [45] for three kinds of miRNA-gene relations, i.e., miRNA-target, miRNA-gene regulation and gene-miRNA regulation relations. We will give detailed definition of these relations in that chapter. In this section, we briefly discuss existing systems and corpora for general miRNA-gene relation extraction.

Extraction Systems: Prior to miRTex, there were two previous works concerning the extraction of miRNA-gene associated pairs from text. miRSel [58] detects sentences that contain both a miRNA and a gene and classifies them into one of five categories of miRNA-gene associations: physical target, repression, co-expression, induction and cleavage. This classification is based on detecting presence of words (e.g., "target" or "repress") anywhere in the sentence. An evaluation on a set of 50 abstracts containing 103 miRNA-gene associations showed a 0.78 F-score with 0.89 recall and 0.70 precision for detecting miRNA-gene associations, and a 0.73 F-score with 0.87 recall and 0.62 precision for classifying association categories. Another text mining system [6] compared co-occurrence-based methods similar to miRSel with machine learning approaches to extract miRNA-gene associated pairs. Linear classification [24] and linear SVM [14] algorithms are applied to train extraction models with lexical features and shortest-path features from the dependency graph. Compared to the co-occurrence-based approach, the machine learning approaches performed better and achieved a 0.76 F-score with 0.67 precision and 0.87 recall on a test set of 100 abstracts with 123 miRNA-gene associated pairs. However, neither this system nor miRSel captures the direction of the extracted miRNA-gene relations, i.e., whether the miRNA regulates the gene or the gene regulates the miRNA.

Corpora: Prior to miRTex, the only publicly available corpus for miRNA-gene relation extraction is created by Bagewadi et al. [6], which consists of 301 Medline abstracts. They are divided into a training set and a test set. The relevant statistics of this corpus are listed in Table 2.1.

Table 2.1: Characteristics of Bagewadi et al. Corpus.

	Training	Test
Abstracts	201	100
Specific miRNA mentions	529	376
Gene/protein mentions	734	324
Relation triggers	1,335	625
Relations between specific miRNAs and genes	195	123

We use its test set to evaluate the miRTex system described in Chapter 3. In the test set, the curators manually annotated specific miRNA and gene mentions, as well as 123 related miRNA-gene pairs that co-occur in a sentence in the test corpus. The annotated relations include both cases where the miRNA regulates the gene, or the gene regulates the miRNA.

2.2.2 Protein Post-Translational Modification

In Chapter 4, we apply the general relation extraction framework to extract five types of protein post-translational modification (PTM) from text, i.e., acetylation, glycosylation, ubiquitination, sumoylation and neddylation. They play an important role in cellular functions and disruptions in the modification pattern have been linked to diseases.

Extraction Systems: Despite of abundant research efforts concentrated on phosphorylation extraction, only a few previous work exists for extraction of other types of PTMs. A system developed by Ohta et al. [61] extracted acetylation, glycosylation, methylation and hydroxylation information from text. Ohta et al. used SVM to train three detectors for trigger words, event edges and events, based on a rich set of lexical and syntactic features proposed by Miwa et al. [55]. Overall, they achieved 0.424 F-score with 0.521 precision and 0.357 recall for the extraction of the four PTMs. The BioNLP 2011 Epigenetics and Post-translational Modifications Task [62] focused on extraction of seven different types of PTM events. The best performing system [7] achieved an overall performance with 53.33 F-score with 53.98 precision, 52.69 recall. They applied a machine learning approach with SVM model and various lexical and syntactic features. A two-step approach was used. They first recognized trigger words, and then event arguments were extracted for each trigger word in the same sentence.

Corpora: One early publicly available corpus for post-translational modifications was created by Protein Information Resource (PIR) [34]. Pyysalo et al. [74] developed a new corpus based on the PIR corpus to include more accurate entity offset information. In this dissertation, we used the corpus from BioNLP 2011 Epigenetics and Post-translational Modifications Task [62] for evaluation of four PTMs. The training, development and test set contain 600, 200 and 440 Medline abstracts, respectively. The number of PTM events in this corpus are listed in Table 2.2.

Table 2.2: Characteristics of BioNLP 2011 EPI Corpus.

	Training Events	Development Events	Test Events
Phosphorylation	98	32	69
Acetylation	262	71	175
Glycosylation	273	73	188
Ubiquitination	263	77	181

2.2.3 Protein-Protein Interaction

In Chapter 5, we use protein-protein interaction extraction to evaluate our methods for improving distantly supervised biomedical relation extraction.

Extraction Systems: Extraction of protein-protein interaction (PPI) from text is a well-studied biomedical text-mining task. Various methods have been proposed for it, including both rule-based [66] and machine learning-based methods [1,31,56,72]. On the corpus AIMed [13], the state-of-the-art system [57] achieves 64.2 F-score using 10-fold document-level cross validation.

Corpora: There are five public corpora that are commonly used by the literature for evaluation of PPI extraction. They are AIMed [13], BioInfer [73], HPRD50 [25], IEPA [23] and LLL [27]. Among them, AIMed and BioInfer are the largest in terms of the number of annotations, and their characteristics are listed in Table 2.3. We refer readers to the publications for details of other corpora.

Table 2.3: Characteristics of AIMed and BioInfer Corpus.

Corpus	Document	Positive Pairs	Negative Pairs
AIMed	225	1,000	4,834
BioInfer	836	2,534	7,132

2.2.4 Protein Subcellular Localization

In Chapter 5, we use the task of extracting protein subcellular localization events to evaluate our methods for improving distantly supervised biomedical relation extraction.

Extraction Systems: Protein subcellular localization event extraction has been a subtask in BioNLP shared task from 2009 to 2013 in the Genia track [38, 39]. The top 4 ranked systems in BioNLP 2013 event for protein localization task are EVEX [92], TEES-2.1 [8], BioSEM [12] and HDS4NLP [48]. BioSEM, a rule-based system, achieved 57.53 F-score with 89.36 precision and 42.42 recall on the test set for protein subcellular localization task. The other three systems used SVM algorithm with various lexical and syntactic features and achieved 60.61-63.29 F-scores.

Corpora: The BioNLP Genia corpus is the main corpus used in previous work for protein localization event extraction. In this corpus, localization events are annotated with the involved protein and subcellular location, and also a trigger word indicating the occurrence of the event. In this dissertation we use the 2011 version of Genia corpus to compare with a previous work [104]. The characteristics of the Genia 2011 corpus for protein subcellular localization events are listed in Table 2.4. For Genia 2009 and 2013 corpora, readers are referred to Kim et al. [38, 40] for more details.

Table 2.4: Protein-Localization Events in BioNLP 2011 Genia Corpus.

	Training Set	Development Set	Test Set
Text source	800 abstracts 5 PMC articles	150 abstracts 5 PMC articles	260 abstracts 4 PMC articles
Events	281	67	191

Chapter 3

MIRNA-GENE RELATION EXTRACTION

In this chapter we present a rule-based system, miRTex, for extracting three kinds of miRNA-gene relations. Rule-based systems can usually achieve high precision but have limited recall due to the difficulty to manually collect all the possible patterns. To mitigate this issue, Peng et al. [69] proposed a general relation extraction framework iXtractR, which used a small set of trigger words and lexico-syntactic templates to automatically generate rigorous patterns, and then used referential relations to improve recall. Other than the experiments conducted in [69], there is no existing effort that adopts iXtractR to develop an actual relation extraction system. In this chapter, we develop miRTex based on the principles in iXtractR. First we will give a definition of the three relations, and then describe the design of the extraction system. We evaluate the performance on a new corpus annotated by curators in our group and a corpus used by previous work. Finally, we describe the application of the miRTex system to process all Medline abstracts and PMC full-length articles and report the results.

3.1 miRNA-Gene Relation

The three kinds of miRNA-gene relations are miRNA-target, miRNA-gene regulation and gene-miRNA regulation relation. miRNA-target relation is a subset of miRNA-gene regulation relation, consisting of cases when the miRNA directly impacts a protein's expression level, e.g., by binding the mRNA's 3' UTR. In miRNA-gene regulation relation, a miRNA directly or indirectly impacts the protein's expression level. On the other hand, gene-miRNA regulation relations are the cases where a gene, usually encoding a transcription factor, is involved in the regulation of a miRNA's expression. The following examples elaborate the text expressing different kinds of relations.

> **Example 1.** miRNA-gene regulation (PMID-22569260): *MicroRNA-223* **regulates** *FOXO1* expression and cell proliferation.

Here the miRNA "MicroRNA-223" regulates the gene "FOXO1". As it is unclear whether the regulation is direct and hence we simply call this case an example of a miRNA-gene regulation relation.

> **Example 2.** miRNA-target (PMID-23579289): The direct **binding** of *miR-214* to the *Osx* 3' untranslated region (3' UTR) was demonstrated by a luciferase reporter assay.

Example 2 illustrates the case of a direct miRNA-gene regulation relation, i.e., a miRNA-target relation. In this example, the miRNA "miR-214" directly binds to the gene "Osx" 3' UTR. Some other examples of direct miRNA-gene regulation relations include sentences where the gene is mentioned explicitly to be a target of the miRNA, as seen in Example 3.

> **Example 3.** miRNA-target (PMID-23041385): *TGF-betaR2* mRNA, a validated *miR-21* **target**, showed the highest expression in the leukocytes from a subset of the octogenarians.

Example 4 is a case of gene-miRNA regulation relation, where the gene "Sp1" is a regulator of the miRNA "miR-29b".

> **Example 4.** gene-miRNA regulation (PMID-23190608): We identified *Sp1*, a transcription factor endowed with oncogenic activity, as a negative **regulator** of *miR-29b* expression in MM cells.

These four examples demonstrate the kinds of relations we want to extract and illustrate some of the variety of ways in which they may be described in text. We developed an extraction system for these relations after examination of a development literature corpus.

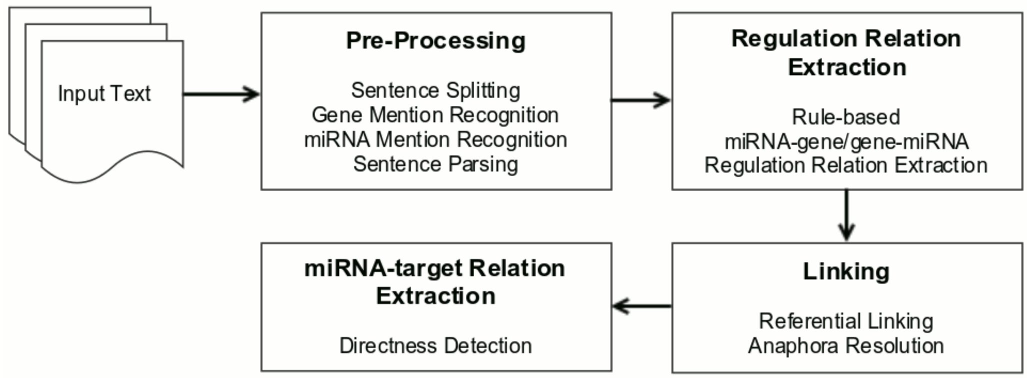


Figure 3.1: Overview of miRTex System

3.2 miRTex System

For extraction of the three kinds of relations, our discussions will focus on the extraction of miRNA-gene regulation and miRNA-target relations as their mentions are far more common than the gene-miRNA regulation relations in the literature. It only requires minor modification in the system to extract gene-miRNA regulation relations.

Through the examples presented in Section 3.1, we can see that the miRNA-gene relations are usually expressed by a trigger word, such as “regulate” and “target”. In miRNA-gene regulation relation, the miRNA is the “agent” of the trigger word, i.e., the object that performs the regulation action, while the gene is the “theme”, i.e., the object that is regulated. Although the agent and theme are semantic but not syntactic arguments, they are still usually syntactic arguments of the trigger words in the sentences. Therefore, we can use lexico-syntactic rules to extract the arguments of a predefined set of trigger words.

Another aspect of miRTex is the adoption of ideas from iXtractR and RLIMS-P to improve recall of the system by using several linguistic generalizations, including referential relations and anaphora resolution. Combining with the lexico-syntactic patterns, we apply a linking component to improve the recall of the system. The overview of the pipeline is shown in Fig. 3.1

3.2.1 Pre-processing

In pre-processing, we first split text into individual sentences and recognize named entities, i.e., miRNA and gene mentions. We use the Stanford sentence splitter [50], and a well-known name detector, BANNER [43], for gene mention. Unlike gene mention detection, miRNA mention detection is an easier task because miRNA nomenclature is well-defined [2, 26] and closely followed in the literature. The basic pattern of miRNA mentions is one prefix "miR" (or "miRNA", "microRNA") followed by a unique identifying number, which is assigned based on sequence similarity (e.g., "miR-1"). Additionally, the name might include one suffix such as "-a", "-1", "-3p" or "-5p", and/or include a prefix that denotes the species. Most systems, including miRTex, use simple regular expressions to identify miRNA mentions.

Following name detection, parse trees of the sentences are generated to match with our lexico-syntactic rules. We use the Charniak-Johnson parser with David McClosky's biomedical model [15, 16].

3.2.2 Lexico-Syntactic Rules

We use lexico-syntactic rules to extract miRNA-gene regulation relation from parse tree. The rules consist of three components: trigger words, syntactic patterns and entity typing constraints. The trigger words are a set of words that indicate the regulation of a gene by a miRNA. We formed the trigger words by the verbs and their nominal and adjective forms from the most common trigger words of regulation events in the BioNLP 2013 Genia corpus, and the trigger words seen in our development corpus. The verbs are either words that indicate regulation, such as "regulate" and "suppress", or specific action verbs that indicate how a miRNA interacts with a gene product, such as "target" and "bind".

The syntactic pattern expresses the syntactic constraints between the trigger word and its agent or theme argument. Additionally, since we are extracting miRNA-gene regulation relation, typing constraints require that the agent must be a miRNA and the theme must be a gene. To extract gene-miRNA regulation relations, the only necessary modification is to switch the typing constraints of the agent and theme.

Instead of a rule extracting the two arguments simultaneously, our rules only capture one argument of a trigger word at a time. Thus two rules will be needed to capture both agent and theme argument with the requirement that the trigger words of both rules are the same. This one-argument-per-rule design significantly simplifies the rules and reduces the number of rules that is required.

We demonstrate the extraction process using lexico-syntactic rules with an example below. In all the examples in Chapter 3, trigger words are bolded and entities are underlined.

> **Example 6.** (PMID-19487542): Taken together, these results suggested that *hsa-miR-222* **regulates** the *MMP1* expression.

In this sentence, the trigger word is "regulates" and has two syntactic arguments, the subject "hsa-miR-222" and the object "the MMP1 expression". Since the verb is in active form, these two syntactic arguments, subject and object, would correspond to the semantic arguments, agent and theme, respectively. A lexico-syntactic rule "agent: NP-VP" [45] contains the syntactic pattern for matching the verb phrase headed by an active verb and its noun phrase subject. The rule also requires that this noun phrase be of type miRNA and the head verb be "regulate" (in any tense). Matching the parse tree for the given sentence against the pattern and the constraints will allow us to extract "hsa-miR-222" as the agent of regulation. Similarly, another rule "theme: NP-VP" [45] extract the theme "the MMP1 expression". Because in regulation relations the argument could be the gene itself or its expression, we actually extract "MMP1" as the theme. Since the trigger word for both rules applied is the same, the system extracts the regulation relationship between "hsa-miR-222" and "MMP1".

While the above rules capture the relation using a verb trigger "regulate", the relation can be specified with other forms of the same trigger word. Their lexico-syntactic rules are closely related to the one for the verb and is predictable as in our previous work iXtractR [69] and RLIMS-P [89]. An example of a noun trigger word "regulation" is shown below.

> **Example 7.** (PMID-23303397): **regulation** of *RhoG* expression by the microRNA *miR-124*

"RhoG" can be extracted as the theme of the trigger word "regulation" using a rule "theme: NP-of-NP-1" [45]. The agent, "miR-124", can be extracted by another rule "agent: NP-by-NP" [45].

We also consider trigger words that are of adjective form. Similar to the noun form case, the rule for the adjective form can be easily predicted given the rule for the original verb form. The lexico-syntactic rules can be used to extract miRNA-gene regulation relations described by other trigger words, such as "suppress" and "decrease". All we need to do is replace the trigger word when matching against the parse tree.

3.2.3 Null-Argument Rules

We now discuss a few additional rules [69] where one of the two arguments is not directly syntactically related to the trigger word.

> **Example 9.** (PMID-24099915): *MicroRNA-31* inhibits cisplatin-induced apoptosis in non-small cell lung cancer cells by **regulating** the drug transporter *ABCB9*.

In Example 9, we will extract the agent "MicroRNA-31" and the theme "ABCB9" of the trigger word "regulating". The subject of "regulating" is omitted because it is the same as the one for the preceding predicate "inhibits". It is not necessary to repeat the same subject for clarity. We believe that this is a very general kind of linguistic expression. It applies to any trigger words following prepositions such as "via", "by", "through", etc. We have a set of rules inspired by RLIMS-P and iXtractR that are called null-argument rules capable to capture such cases. A null-argument rule "agent:

null-arg-VP" [45] can match this example and extract the agent "microRNA-31" for the trigger word "regulating". The theme of the same trigger word can be extracted using the rule "theme: NP-VP" [45] and thus we obtain the regulation relation between "MicroRNA-31" and "ABCB9".

3.2.4 Linking Relation

In the previous subsection we described how to extract the agent and the theme argument of a trigger word. Consider the fragment "a target for miR-142-3p" in Example 10 below. Two rules apply to the phrase, one of which will extract "miR-142-3p" as the agent of the target relation, and the other will extract "a target" as the theme. We call such phrases as referential phrases since they do not provide the actual argument but instead refer to the actual argument.

> **Example 10.** (PMID-22870299): We show that the *proinflammatory cytokine interleukin 1 alpha (IL1A)* is a **target** for *miR-142-3p*.

In Example 10, after exploiting the "is-a" relation in the sentence, we can link the phrase "a target" to the actual referred entity "proinflammatory cytokine interleukin 1 alpha (IL1A)".

Referential Linking: We call the process of linking the referential phrase to the actual argument "referential linking". The detection of an explicit is-a construction is one form of referential linking. As in iXtractR and RLIMS-P, we also use appositives (Example 11) and member-collection relation (Example 12) for referential linking.

> **Example 11.** (PMID-23041385): *TGF-betaR2* mRNA, a validated *miR-21* **target**, showed the highest expression in the leukocytes from a subset of the octogenarians.

The phrase containing the theme "a validated miR-21 target" is a description of the entity specified by "TGF-betaR2 mRNA". To detect appositives, we use the patterns from the sentence simplifier iSimp [65].

Example 12. (PMID-22465011): Over-expression of *miR-499* in rat BM-MSCs **increased** the expression of cardiac-specific genes, such as *NKx2.5*, *GATA4*, *MEF2C*, and *cTnI*.

The theme of the trigger word "increased" is "the expression of cardiac-specific genes", but the actual themes are the genes listed in the end of the sentence. Each of these genes can be considered to be in a member-collection relation with the phrase "cardiac-specific genes". By using member-collection relations miRTex will find the actual themes to be the list of genes.

Anaphora Resolution: An anaphor, e.g., a pronoun, is an expression that refers to an entity that is previously mentioned in the text. Besides referential linking, anaphora also needs to be resolved to their antecedents. The antecedent must be chosen from the recognized mentions of miRNA or gene. While in resolution, the number and typing agreement between the anaphora and its antecedent are required, and different search scopes are defined for different anaphora. For the most common anaphora "it" or "its" we look for single entity in the same sentence (typically the subject) whose type (miRNA or gene) is based on the anaphora's semantic role. The resolution of the anaphora "they", "them" and "their" is similar except that the referred phrase must contain more than one entity. For noun phrases that begin with "these", "those", we look for antecedent phrases that mention multiple entities previously in the paragraph, and for the phrases that start with "one", "two", "three", etc., we look for antecedents that contain the same number of entities previously in the paragraph.

Example 13. (PMID-20545570): the *miR-21* was upregulated in 0.5 Gy-treated TK6 cells and its **target** genes *programmed cell death factor 4 (hPDCD4)*, *phosphatase and tensin homolog (hPTEN)*, and *sprouty homolog 2 (hSPRY2)* were found to be downregulated in these cells.

The anaphora "its" is resolved to the miRNA "miR-21". They agree on the number and type and they are in the same sentence, satisfying the requirement on search scope.

Another kind of anaphora the system resolves is relative pronouns, which appear in relative clauses. A relative clause is a kind of subordinate clause marked by a relative pronoun, which co-refers to a phrase in the main clause. As for anaphora resolution above, if the relative pronoun acts as a semantic argument of a trigger word, we link it to the nearest noun phrase of the appropriate type by token in the same constituent with the relative clause in the parse tree. Consider the following sentence,

> **Example 14.** (PMID-23104180): Two miRNAs, *miR-15a* and *miR-16*, which act as putative tumor suppressor by **targeting** the oncogene *BCL2*, have been implicated in cell cycle, apoptosis and proliferation.

In this example, the relative pronoun "which" is first detected as the semantic argument of the trigger word "targeting" by a null-argument rule. Since it is detected as the agent, we attempt to link it to the nearest noun phrase of type miRNA, and thus we capture the regulation relation between the miRNAs "miR-15a" and "miR-16" and the gene "BCL2".

3.2.5 Relaxing the Same-Trigger Constraint

We also observed miRNA-gene regulation relations are not always described by a single trigger. Consider the following example.

> **Example 15.** (PMID-17891175): we have identified additional potential mRNA **targets** of one of the miRNAs, (*miR-125b*) that are upregulated in prostate cancer and confirmed increased expression of one of these **targets**, *EIF4EBP1*.

The miRNA "miR-125b" and the gene "EIF4EBP1" are in a regulation relation, but their trigger words are not the same. Recall that our rules require the arguments to have the same trigger word; thus, this example will be neglected by the system. Moreover, even if two arguments share the same trigger word, unseen pattern or parsing errors may sometimes prevent the rule-based system from extracting both arguments for the trigger word. We relax the same-trigger constraint to solve this problem.

The intuition is that if a miRNA is the agent of regulation and a gene is the theme of regulation but described by a different trigger word in the same sentence, and they are not extracted by any of the previous rules, it is highly likely that they are in regulation relation. Thus, if two entities are not considered for extraction using the same trigger word in a sentence, we relax the same-trigger constraint so that a miRNA-gene pair can be extracted even they have different trigger words. Note that the original rules with same-trigger constraint are still used to extract miRNA-gene regulation relations with high confidence, while the relaxed rules only operate on the entities that are not in relations extracted by the same-trigger rules.

In our experiments, we evaluate the same-trigger rules alone and together with rules with the relaxed constraint. We prove that the system using both sets of rules obtains a higher recall without significantly compromising precision.

3.2.6 Detection of miRNA-Target Relations

When we extract miRNA-gene regulation relations, we further detect if the relation is direct, thus considered as a miRNA-target relation. Usually we can find evidence indicating the directness of the relation in the sentence from which it is extracted. Below is a list of evidence that we consider for directness of the relation.

1. The trigger word is “target”, “bind”, “interact” or one of their nominal or adjective forms.

2. The trigger word is modified by the adjectives “direct” or “immediate” or the adverb “directly” or “immediately”.

3. The relation is extracted by a null-argument rule and the gene is down-regulated. The intuition is that the null-argument structure is often used to describe direct impact on an object by another object, and since the gene is down-regulated, it is very likely that the regulation is direct.

4. The following words co-occur in the sentence from which the miRNA-gene regulation relation is extracted: “translation”, “3’UTR” or one of its variances, e.g., “3’untranslated region”. (The 3’UTR is a region of a target mRNA to which a miRNA binds.)

If one of such evidence can be detected for an extracted miRNA-gene regulation pair, the pair is assumed to be in a direct (miRNA-target) relation.

3.3 Evaluation

We evaluated miRTex on two different corpora. First, we evaluated the extraction of the three kinds of relations on a test set containing 150 abstracts of an in-house developed corpus. Second, we evaluated miRTex on Bagewadi et al.’s [6] corpus.

3.3.1 Evaluation on In-house Developed Corpus

As there is no public corpus available for comparative evaluation of the three types of miRNA-gene relation extraction, we developed an in-house corpus consisting of 350 abstracts for both system development and evaluation. First, the query “microRNA[TIAB] OR miRNA[TIAB] OR miR[TIAB]” was searched in PubMed to obtain potential abstracts containing miRNA information (performed in May 2014). This gave us about 30,000 abstracts. Second, we used our miRNA mention recognizer to extract miRNA mentions in the abstracts, and randomly selected 350 abstracts from those that have at least one miRNA mention. These abstracts were further processed by BANNER to annotate gene mentions. The automatically annotated miRNA and gene mentions were manually checked and corrected by an annotator. Finally, these 350 abstracts were divided into a development set of 200 abstracts and a test set of 150 abstracts and annotated by two other annotators. The major characteristics of the corpus are listed in Table 3.1.

For example, 229 miRNA-gene regulation pairs were annotated in 86 of the 150 abstracts from the test set; of these, 128 were marked as direct regulation. We calculated the inter-annotator agreement in terms of F-score as was done in [22], where

Table 3.1: miRTex Development and Test Set.

	Development Set	Test Set
Abstracts	200	150
Abstracts with at least one related pair	107	86
Abstracts with no related pairs	93	64
miRNA-gene regulation related pairs	259	229
miRNA-target pairs	155	128
Gene-miRNA regulation related pairs	24	36

one of the annotators' annotation is taken as gold standard, and the other's annotation is contrasted with it. The inter-annotator agreement for the relations annotated on a common set of 20 abstracts is 86%.

After developing miRTex using the development set, we evaluated it by running the system on the test set. The evaluation was at the abstract level meaning that only unique miRNA-gene pairs in an abstract are counted. Three different settings were used to run the system: 1) the same-trigger rules only, 2) the same-trigger rules with the linking component, and 3) the full system with the same-trigger rules, the linking component and the relaxed rules. Comparison among the results produced by these three settings allows us to verify the extent of the improvement in recall by usage of the linking process as well as the relaxed rules and measure if there is any significant drop in precision. Additionally, in order to compare our system with previous co-occurrence-based methods, we also implemented a system that closely follows the description of the method described in the miRSel system [58]. The results of our evaluations are listed in Table 3.2.

Overall, the full system performed the best for extraction of the three kinds of relations in terms of F-scores with high precisions. The usage of linking component and relaxed rules can improve recall significantly without decrease in precision. Simple co-occurrence method resulted in low precisions, especially for extraction of miRNA-target and gene-miRNA regulation relations, although it can achieve high recalls.

Table 3.2: Performance for Extraction of Three miRNA-Gene Relations.

	miR-Gene Regulation			Gene-miR Regulation			miRNA-Target		
System	P	R	F	P	R	F	P	R	F
co-occurrence	0.61	0.94	0.74	0.09	0.97	0.17	0.35	0.97	0.52
miRTex−LK−RR	0.96	0.57	0.72	1.00	0.53	0.69	0.95	0.46	0.62
miRTex−RR	0.96	0.79	0.87	0.94	0.83	0.88	0.96	0.80	0.88
miRTex	0.96	0.91	0.94	0.94	0.83	0.88	0.96	0.81	0.88

P, R, F stand for precision, recall and F-score, respectively; co-occurrence: our implementation; miRTex−LK−RR: using same-trigger rules only; miRTex−RR: using same-trigger rules and the linking component; miRTex: miRTex full system using both same-trigger and relaxed rules and the linking component.

3.3.2 Evaluation on Bagewadi et al. Corpus

The test set of the Bagewadi et al. corpus contains 100 abstracts and was used for evaluation. The evaluation was at the mention level, as in Bagewadi et al. [6], because this corpus annotates all occurrences of miRNA-gene associations even if they are mentioned multiple times in an abstract.

We first evaluated our miRNA mention recognizer on these 100 abstracts and obtained an F-score of 0.965 with precision 0.989 and recall 0.941 for the 376 specific miRNA mentions. System discussed in Bagewadi et al. [6] reported an F-score of 0.935 with precision 0.936 and recall 0.934 for the recognition of the 376 specific miRNA mentions. Evaluation of miRTex on this test set at the mention level gave performance of 0.87 F-score with 0.92 precision and 0.82 recall for the 123 miRNA-gene associated pairs. We also evaluated the co-occurrence method implemented based on miRSel [58] on this test set and it showed a performance of 0.71 F-score with 0.55 precision and 1.00 recall. System discussed in Bagewadi et al. reported 0.76 F-score with 0.68 precision and 0.87 recall for these 123 miRNA-gene relations.

3.4 Full-Scale Processing

To unveil the rich knowledge about miRNA and gene in the text, we conducted full-scale processing of all the Medline abstracts and PMC full-length articles. We used results from PubTator [95] to normalize the gene mentions in miRNA-gene relations.

The processed results were integrated into the iTextMine system [1], which supports keyword query and network visualization of the extracted relations. Table 3.3 lists the result statistics of the full-scale processing. The statistics show that we can extract much more information from full-length articles than abstracts, highlighting the importance of the processing of full-length articles when applying full-scale processing of text-mining tools.

Table 3.3: Result Statistics of Full-Scale Processing

	Medline Abstracts	PMC Articles
Document	21,966	19,479
miRNA-Target	33,559	255,308
miRNA-Gene Regulation	44,565	377,796
Gene-miRNA Regulation	8,426	65,492

PMC statistics do not include relations extracted from abstract sections.

3.5 Conclusion

In this chapter, we described the text mining system miRTex for extraction of miRNA-target relations as well as miRNA-gene and gene-miRNA regulation relations. We combined rule-based extraction with the linking process and a relaxed same-trigger constraint. With small number of trigger words and lexico-syntactic rules, the system achieved the state-of-the-art performance on a test set of 150 abstracts. The evaluation results showed that the precision of our system greatly outperforms the co-occurrence-based method with a comparable recall.

The scalability and robustness of miRTex were validated by applying it to extract miRNA-gene relations from all the abstracts in Medline and all the full-length articles from the PMC Open Access Subset. The text mining results for all the Medline abstracts are stored in a database that can be searched through the iTextMine system website.

[1] https://research.bioinformatics.udel.edu/itextmine

Chapter 4

A GENERAL RULE-BASED RELATION EXTRACTION SYSTEM

Although the iXtractR [69] and miRTex system [45] can achieve high performance using a set of trigger words and predicate-argument rules, user still need to manually define all the possible variants of predicate-argument patterns to extract relations. In this chapter, we describe a general rule-based framework for biomedical relation extraction. The framework is based on the generalization of ideas from Lexicalized Tree-Adjoining Grammar (LTAG) [83] and the Extended Dependency Graph (EDG) [67]. It supports automatic generation of comprehensive patterns, which usually requires significant amount of time if they were created manually by domain experts. The adoption of EDG abstracts away syntactic details and enables us to extract related entities using numbered arguments, which can be propagated using a set of semantic relations, including is-a, member-collection and part-of, to improve recall. We apply the framework to extract five kinds of post-translational modification relations and experiments show that we achieve good performance on an in-house developed corpus and state-of-the-art performance on BioNLP 2011 EPI corpus [62].

We will first introduce Extended Dependency Graph, its advantages and shortcomings. We then present the linguistic-based pattern generation component and how it can be combined with EDG to form the relation extraction framework. Next we discuss the application of the framework to extract five kinds of PTM relations, with the generated patterns and a few PTM-specific rules. Finally we present the detailed experiment results.

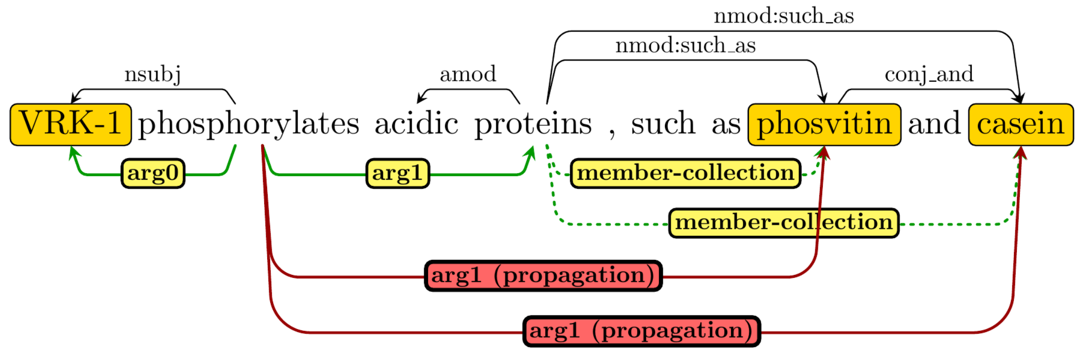


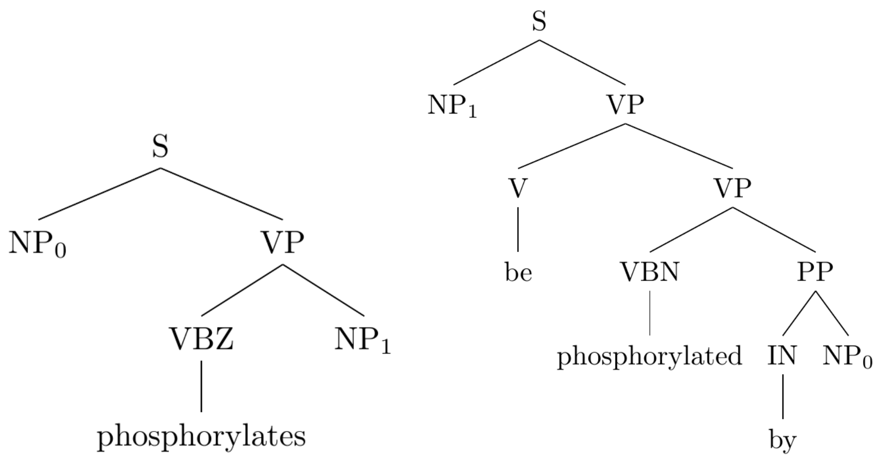
Figure 4.1: An Example Sentence for Extended Dependency Graph.

4.1 Extended Dependency Graph

Extended Dependency Graph (EDG) was proposed by Peng et al. [67]. The main idea is to abstract away syntactic details by adding new numbered argument edges to the dependency graph, thus simplify the extraction patterns for rule-based system and increase the generalizability of machine learning system. The added edges were primarily for the semantic role *agent* and *theme*. An example of EDG is shown in Fig. 4.1.

In this example sentence, two phosphorylation events can be extracted, i.e., "VRK-1" phosphorylates "phosvitin" and "casein". The traditional dependency edges are shown above the sentence, while the EDG edges are shown below. First, the phrase "acidic proteins" will be recognized as the substrate (arg1) as it is in the object position of the trigger word "phosphorylates". Next, by the enumeration (called member-collection in EDG) of the object phrase, "acidic proteins" can be linked to "phosvitin" and "casein". Based on the information extracted in these two steps, the two phosphorylation substrates of VRK-1 can be inferred.

Using EDG can avoid such a complex procedure involving multiple rules and steps, as it abstracts away the syntactic details and unifies the representation of the predicate-argument relations using numbered argument edge. EDG uses *arg0* and *arg1* edge to indicate the predicate-agent and predicate-theme relation, respectively. By using the referential relations (such as the member-collection relation mentioned above and the is-a relation), it can propagate the numbered argument to its actual



(a) Base Form (b) Passivised Form

Figure 4.2: Parse Trees from the Transitive Verb Family.

referred entities (red edges in Fig. 4.1). Thus, a simple rule "phosphorylates → arg1 → PROTEIN" is sufficient to extract the two substrates.

Although EDG has the advantages discussed above, it was mainly implemented for protein-protein interaction extraction. The predicate-argument rules used to extract the numbered argument in the first place need to be designed manually for different relations and are tightly coupled with the relation types.

We propose to develop a new component that can automatically generate predicate-argument patterns for different trigger words of different relations, by using linguistic knowledge from Lexicalized Tree-Adjoining Grammar [83]. The generated patterns then can be used to extract numbered arguments, which can be propagated by EDG using semantic relations.

4.2 Lexicalized Tree-Adjoining Grammar

We give a brief introduction of Lexicalized Tree-Adjoining Grammar (LTAG) in this section. In LTAG, constituent syntax trees are grouped into families. Each lexical anchor, usually a verb, is assigned to a tree family based on its subcategorization. For example, for the transitive verb "phosphorylate", it will be assigned to the transitive tree family. The two trees in Fig. 4.2 are from this family.

NP_0 and NP_1 represent the semantic node agent and theme, respectively. The right tree is the passivized form of the left base tree. LTAG defines a set of syntactic transformations from the base tree to other forms of trees in the same family, while all of them express the same predicate-argument relations, i.e., NP_0 is always the agent and NP_1 is always the theme. Using these transformations, we can easily generate a variety of parse trees from the base tree for a lexical anchor that can select the family, in this case, “phosphorylate”.

In LTAG, there is another aspect in the verb-tree relationship. A verb can only select a subset of trees, and the other trees are not applicable to the verb. For example, the parse tree in Fig. 4.3 is invalid for “phosphorylate”.

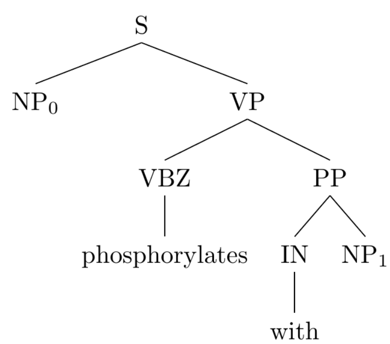


Figure 4.3: Invalid Parse Tree for “phosphorylate”.

As “phosphorylate” is a transitive verb, there can not be a preposition between the verb and its object. Thus, this tree is not in the transitive tree family and excluded from the valid set of parse trees for “phosphorylate”.

4.3 Linguistic-based Pattern Generation

In our system, we generalize the linguistic knowledge in LTAG to dependency parse, and use them to generate a comprehensive set of patterns for trigger words. We focus on extraction of three arguments that are commonly used in biomedical relations, i.e., agent, theme and location. The primary goal for this pattern generation component is then to generate predicate-argument patterns for *arg0*, *arg1* and *arg_loc*, which can be used by EDG for argument propagation.

4.3.1 Trigger Word Types

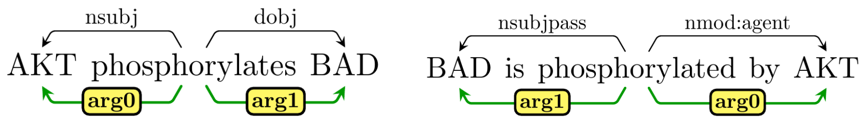
First, we adopt a simpler way to classify the trigger words. Instead of using type transitive or intransitive, we have two different trigger types, *V-dobj* and *V-pobj*. The first class corresponds to the transitive verbs, which can take a direct object, while the second class corresponds to intransitive and ditransitive verbs, which need a proposition to take an object. Example trigger words and sentences are shown below for the two types.

- ***V-dobj***
  - Example trigger: phosphorylate, regulate, promote
  - Example sentence: AKT phosphorylates BAD.
- ***V-pobj***
  - Example trigger: interact, conjugate, attach
  - Example sentence: p53 interacts with hRAD51.

4.3.2 Pattern Generation

We then define the patterns applicable to the two trigger types. A pattern is a specific dependency relation between the trigger and another word. It can be used to match the dependency graph of a sentence and extract arguments of a trigger word. Using the syntactic transformations from LTAG, we can generate new dependency patterns based on the base pattern. Below we show an example with the same sentence in Fig. 4.4 to depicts the syntactic transformation between the base pattern and a derived pattern.

In Fig. 4.4, the base pattern (left) shows a dependency parse for the sentence. The proteins AKT and BAD are subject and object of the trigger word "phosphorylates", respectively. Similar to the syntactic transformation in LTAG for constituency parse, we can derive a passivized form (right) of the base pattern. In this derived pattern, BAD is a passive subject of the trigger word, while AKT modifies the trigger



(a) Base Form (b) Passivised Form

Figure 4.4: Two Syntactically Transformable Dependency Parses.

Table 4.1: Agent Patterns

Pattern Name	Trigger Type	Pattern
base	*V-dobj*, *V-pobj*	*base_trigger* → nsubj → N
passive	*V-dobj*	*passive_trigger* → nmod:agent → N
adjectival	*V-dobj*, *V-pobj*	N-*gerund_trigger* → amod → N
adjectival-self	*V-dobj*	N-*passive_trigger*
nominal	*V-dobj*	*tion_trigger* → nmod:by → N
nominal-compound	*V-dobj*, *V-pobj*	*tion_trigger* → compound → N

word through a nmod:agent relation. Note that in both patterns, AKT is always the *arg0* (agent) and BAD is the *arg1* (theme) of the trigger. We use *arg0/arg1* labeling, following the convention in EDG [67]. Other derived forms from the base patterns include patterns with nominal or adjectival triggers.

By using the base pattern and its derived forms, we then can generate a comprehensive set of patterns for different trigger words. In our approach, we adopt the one-argument-per-pattern design as in miRTex [45] and RLIMS-P [89], and use two different patterns to extract the agent and theme argument independently. This design can then automatically enumerate the combinations between the simpler agent and theme patterns to cover all the cases that can be matched with a two-argument pattern. We first discuss the patterns for agent/theme argument. The base pattern and its derived forms for extracting *agent* and *theme* are listed in Table 4.1 and 4.2.

The different trigger notions in the pattern column indicates the inflection/part-of-speech of the trigger words. For example, base trigger corresponds to base verb form,

Table 4.2: Theme Patterns

Pattern Name	Trigger Type	Pattern
base	*V-dobj*, *V-pobj*	*base_trigger* $\rightarrow$ dobj $\rightarrow$ N
passive	*V-dobj*	*passive_trigger* $\rightarrow$ nsubjpass $\rightarrow$ N
adjectival	*V-dobj*, *V-pobj*	*passive_trigger* $\rightarrow$ amod $\rightarrow$ N
adjectival-self	*V-dobj*	N-*gerund_trigger*
nominal	*V-dobj*, *V-pobj*	*tion_trigger* $\rightarrow$ nmod:of $\rightarrow$ N
nominal-compound	*V-dobj*, *V-pobj*	*tion_trigger* $\rightarrow$ compound $\rightarrow$ N
preposition	*V-pobj*	*any_trigger* $\rightarrow$ nmod:prep $\rightarrow$ N

e.g., “phosphorylate”, while tion trigger corresponds to nominal trigger, e.g., “phosphorylation”. We use a linguistic Python tool [1] to generate different inflections/derivations of a verb trigger. User can also add more trigger variants manually. The “N” notion in the pattern column generally corresponds to a noun node in the dependency graph.

The trigger column specifies the trigger word types that can select the pattern. For example, the base agent pattern can be selected by both *V-dobj* and *V-pobj*, but the passive agent pattern can only be selected by *V-dobj*, as *V-pobj* can not be passivized. On the contrast, the preposition theme pattern can be only selected by *V-pobj* triggers, as they can use a preposition to take an object.

There is one special pattern that is included in both agent and theme patterns, which is the nominal-compound pattern. This is because the N node in this pattern can be either agent or theme, depending on the context. Consider the following two examples.

1) BAD phosphorylation by AKT

2) AKT phosphorylation of BAD

AKT and BAD are agent and theme in thse two sentences. Although their semantic roles are different, both BAD in the first sentence and AKT in the second sentence match the nominal-compound pattern. We can distinguish between their

[1] https://www.clips.uantwerpen.be/pages/pattern-en

roles by using the context. From the first sentence we can extract AKT as the agent using the nominal agent pattern, so we can determine that BAD must be the theme. Similarly, we can extract the correct agent and theme in the second sentence.

In addition, we can have patterns for additional arguments other than *arg0* and *arg1*. For example, in phosphorylation relation, the *Site* argument gives the information where the kinase attaches the phosphate-group to the substrate, and is of great interest to biologists. To extract the location argument, we can use one simple pattern. This pattern can be selected by any class of triggers, as it is optional to all the triggers.

$$\textit{any\_trigger} \rightarrow \text{nmod:(on|at|in)} \rightarrow \text{N}$$

With this method, we can generate a comprehensive set of patterns for a trigger word, given its base form and type. Using its base form, we can first generate different inflections/variants of the trigger, e.g., passive and nominal form. Then, we can replace the lexical anchor in the dependency patterns which can be selected by the trigger, and generate the final rigorous pattern that can be used by EDG.

4.3.3 Entity Typing Incorporation

We incorporate entity typing information into the system for both pattern matching and relation type classification. The entity typing information comes from external named entity recognition tools. Typical entities involved in biomedical relations include proteins/genes, miRNAs, diseases and drugs.

After a numbered argument of a trigger word are extracted, entity typing information can be used to decide the role of an entity in the relation. Consider the following example for phosphorylation relation.

- AKT phosphorylates BAD.
- AKT phosphorylates Ser-100 in BAD.

For the trigger word "phosphorylates", EDG will extract BAD and Ser-100 as *arg1* for the first and second sentence, respectively. As Ser-100 has the entity type

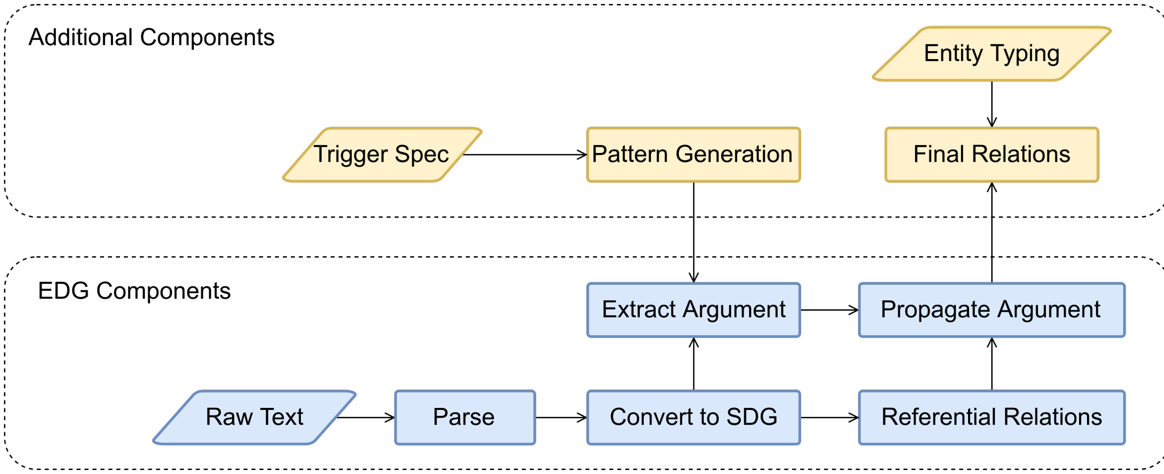


Figure 4.5: General Relation Extraction System Architecture.

SITE, it would be extracted as the phosphorylation site in the relation. Similarly, as BAD has the entity type *PROTEIN*, it would be extracted as the phosphorylation substrate. Altough they both are extracted as *arg1* initially, using entity typing we can correctly determine their roles in the relation.

Another usage of entity typing information is for relation type classification. In post-translational modification, a sentence may state "polysaccharide attached to Kv3.1". We can extract an attachment relation between "polysaccharide" and the protein "Kv3.1" using the generated predicate-argument patterns. Next, as polysaccharide has the type SUGAR, and glycosylation is the type of PTM that attaches sugar to a protein, we can determine that this is a glycosylation relation. As the entity type SUGAR contains different names, it's impossible to encode all the different sugar names into the patterns. Similar to the above example, we can specify the applicable entity types for determining the relation types and the system will exploit this information to assign accurate types to the extracted relations.

4.3.4 System Overview

By combining EDG and the pattern generation component, we can develop a general relation extraction system that only requires a set of trigger specifications as input. The overall system architecture is demonstrated in Fig. 4.5.

The blocks in blue represent the pattern generation component, while the yellow ones represent the EDG system. The system no longer needs user to define a set of predicate-argument patterns. Instead, it uses trigger specification and entity typing constraints to automatically generate the rigorous predicate-argument patterns. This can save a lot of time as trigger specification and typing constraints are much easier to define compared to finding all the pattern variants using a development corpus.

For the input sentence, we first use Charniak-Johnson parser with David McClosky's biomedical model [16, 51] to obtain the constituency parse, and then convert it to collapsed dependency parse with Stanford CoreNLP [50] with *ccprocessed* setting. EDG will then use the generated patterns to extract arguments and propagate them using semantic relations, such as is-a, member-collection and part-of relations.

4.4 Post-Translational Modification Information Extraction

In this section, to illustrate the application of the framework, we use it to extract different post-translational modification (PTM) information from text. We first briefly introduce these PTM relations, and then discuss the the application of the general relation extraction system. We add a few PTM-specific rules to cover some cases that are not extracted on the basis of simple predicate-argument structure. Finally, we present the experiment results.

4.4.1 Post-Translational Modification Relation

We extract PTM relations for acetylation, glycosylation, and three ubiquitination-like modification, namely, ubiquitination, sumoylation and neddylation. Post-translational modifications are chemical modifications of amino acid residues (site) of proteins (substrate) by an catalyst protein (enzyme). These modifications have a major implication for cell biology and disruptions in modification patterns have been linked to diseases. Phosphorylation is the most common kind of PTM and its relation extraction has been widely studied [38, 89, 94, 100]. Other types of PTMs receive less attention from the text-mining community. To the best of our knowledge, BioNLP 2011 EPI task [62] is the only shared task that focuses on extraction of multiple PTM relations. Below we show a few example sentences for different types of PTMs that we want to extract.

- Acetylation (PMID 11073948): Consistent with this observation, **CBP** *acetylated* **c-Myb** in vitro at **Lys(438)** and **Lys(441)** within the NRD.
- Glycosylation (PMID 1730617): **GlcNAc** is *attached* to the **alpha A subunits** at **serine 162**.
- Ubiquitination (PMID 18627766): **TRIP12** functions as an *E3 ubiquitin ligase* of **APP-BP1**.

We will extract three types of arguments for post-translational modifications, namely, enzyme (green), substrate (blue) and site (red). In the glycosylation sentence, GlcNAc in bold is a sugar group that attaches to the substrate. We will use it for extracting glycosylation relation when using more general trigger words, e.g., attach and conjugate. Similarly, ubiquitin, sumo proteins and Nedd8 protein in conjunction with these general triggers indicate ubiquitination, sumoylation and neddylation, respectively.

4.4.2 Named Entity Recognition

We need to first recognize protein and site mentions in the text for relation extraction. We use the gene/protein mentions in PubTator [95] as candidate proteins. We use BioNex [60] to recognize the site entity mentions. For detecting sugar group we use a simple regular expression-based method.

4.4.3 Post-translational Modification Trigger Words

To use the system for PTM extraction, we only need to specify the trigger words and their types for the system to generate the patterns. Table 4.3 below lists example trigger words for each PTM. We include the comprehensive trigger word list in Appendix A. We will explain the generic PTM relation in next subsection.

Table 4.3: Trigger Words for PTM Relations.

PTM	Base Form of Trigger	Trigger Type
Acetylation	acetylate	*V-dobj*
Glycosylation	glycosylate	*V-dobj*
Ubiquitination	ubiquitinate	*V-dobj*
Sumoylation	sumoylate	*V-dobj*
Neddylation	neddylate	*V-dobj*
Generic PTM	conjugate, attach, modify	*V-dobj*, *V-pobj*

4.4.4 PTM-specific Rules

Given the triggers in the base form above, our method will automatically generate patterns based on predicate-argument structure. However, there are several cases where simple predicate-argument structure does not appear to suffice for defining patterns for extraction. This had motivated the development of some additional rules during the system development. These rules are applicable to multiple PTM types and thus have been implemented as a general component in the PTM extraction system.

Part-of relations: sentences describing PTM relations often mention the PTM site is contained in the substrate. We can extract this kind of part-of relation, and use the EDG system to propagate the numbered argument. Consider the following example:

> Example (PMID-18235976): The **corticotropin-releasing factor (CRF) receptor type 1 (CRFR1)** contains five potential N-glycosylation sites: **N38, N45, N78, N90, and N98**.

The relation extraction system will first extract the *arg1* (site) argument of glycosylation, “N38, N45, N78, N90 and N98”, using the generated predicate-argument patterns and is-a relation. We can extract a part-of relation from the glycosylation site to the protein “CRFR1”. Then the EDG system can propagate the *arg1* argument from site to the protein, thus extract the substrate argument.

Generic modification relation: the general PTM includes relations for which we can not determine the PTM type solely based on the syntactic pattern and numbered argument. Consider the following two examples.

1) **NEDD8** is covalently *conjugated* to **H2A**

2) **NAC** is *conjugated* to **SUMO1**

Numbered arguments can be extracted for the trigger word "conjugate" in both sentences with the same syntactic pattern. The first sentence describes a neddylation relation, because *arg1* is NEDD8, a specific kind of protein that is attached to the substrate in neddylation. Similarly, we know the second sentence describes a sumoylation relation, as SUMO1 is a specific kind of protein for sumoylation. We use a simple post-processing step to determine the type of PTM for generic modification relations. Specifically, we check all the arguments of an extracted generic modification relation, as well as the words in the same noun phrase with the trigger. If a indicative term or protein (such as Nedd8 and Sumo-1) is found, we assign the corresponding PTM type to the general PTM relation.

Self-agent and self-theme: there are a set of words that describe their role in the PTM relation by themselves. For example, kinase, substrate and ligase. We add a few simple rules to extract PTM relations described in this way. Consider the following example,

> Example (PMID-18627766): **TRIP12** functions as an E3 ubiquitin ligase of **APP-BP1.**

For example, here "ligase of" is used as a trigger from which we can identify the substrate (i.e., *arg1*). Thus, in the above sentence, *arg1* is assigned to APP-BP1. Further, as mentioned above, the indicative term "ubiquitin" can be used to change the generic relation to ubiquitination. However, this pattern assigns arg1 (agent) role only to itself. Note the EDG framework uses is-a relation and propagation of arg edges.

Thus in this case, we can extract TRIP12 as the ligase through a is-a relation, and APP-BP1 as the substrate through the "ligase of protein" pattern.

Event-argument rules: authors often mention how a PTM affects the functionality of a protein, indicating the protein is the substrate of the PTM relation. Thus, following a heuristic in RLIMS-P, if a trigger does not lead to the extraction of a theme (such as the occurrence of "glycosylation" in the sentence below), and the trigger word is also the agent of a regulation relation then the theme of the PTM can be extracted from the theme of the regulation event.

Example (PMID-8402259): This result indicates that glycosylation affects the thermostability of **renin**.

Thus, we can extract the glycosylation relation with the substrate renin for the above sentence.

Anaphora resolution: if an extracted argument is an anaphor, e.g., "it", "its" or "itself", we resolve it to the protein in the subject position in the sentence and use that protein as the argument.

Post-processing rules: after all the PTM relations are extracted from the document, we try to merge some of them or add new low-confidence relations, inspired by RLIMS-P [59, 89]. We define two classes of rules which will merge relations in an intra- and inter-sentence manner. Both classes of rules are used to improve recall for information-centric extraction, and only intra-sentence rules are used for mention-level extraction (see Subsection 4.5.2). The post-processing rules are listed below.

- Rules for intra-sentence post-processing:
  - If there are two relations extracted with the same PTM type, one of them only has substrate and the other only has site, then merge them into one relation.
  - If there is only one PTM relation extracted and it only has site argument, and there is only one protein, and the protein is extracted as substrate in other sentences, then use that protein as substrate.

- If in a sentence there is only one PTM trigger and only one protein, and the protein has been extracted as substrate in other sentences, then use that protein as substrate for the trigger.

- Rules for inter-sentence post-processing:

  - If a PTM relation only has site arguments, and the same site arguments are also extracted in another PTM relation with a substrate, use this substrate for the first relation.

  - If a PTM relation only has site arguments, we try to find the focus entity as in RLIMS-P [89] and use it as the substrate.

  - If a PTM relation only has enzyme and substrate argument, and another PTM relation only has substrate and site argument, merge the two relations if the substrate is the same.

4.4.5 Extraction of Additional Arguments for BioNLP Corpus

As the acetylation relations in BioNLP corpus involve an additional argument *Contextgene*, the gene affected by the acetylation. We apply a simple heuristic to extract the *Contextgene* argument after extracting the other arguments. Specifically, we look for gene names in a noun phrase with the head word "gene", "promoter", "transcript", etc., in the current and one sentence before. We then use the closest gene as the *Contextgene* argument for the acetylation relation.

For glycosylation relations, the BioNLP corpus involves an additional *Sidechain* argument, which is the sugar attached to the modification site of the protein. We develop a simple regular expression-based recognizer of sugar groups, and assign the *Sidechain* to be the sugars extracted as an argument for glycosylation.

4.5 Experiment Results

In this section, we report the evaluation results of the PTM extraction system on an in-house developed corpus and the BioNLP 2011 EPI task test corpus [62]. We then analyze the errors for future improvement of the system.

4.5.1 Evaluation on In-house Developed Corpus

We have annotated in-house corpus for the five kinds of PTM relations using Medline abstracts. Table 4.4 below shows the characteristic of the corpus.

Table 4.4: Characteristics of In-house Developed Corpus

	Development Set		Test Set	
Relation	Abstract	Relation Annotation	Abstract	Relation Annotation
Acetylation	50	61	101	130
Glycosylation	50	65	107	127
Ubi-like relations	50	65	159	153

The statistics of ubi-like relations include ubiquitination, sumoylation and neddylation, as they are similar post-translational modifications and have smaller number of annotated relations.

The annotation is conducted in a information-centric manner by two biologists in our group. For each abstract, only unique relations are annotated. If there are multiple mentions of the same relation stated in the abstract, only one relation instance will be annotated. The evaluation is thus on document-level. We will discuss the evaluation on a relation mention level in the experiment with BioNLP 2011 EPI test corpus.

We use the annotated proteins in the corpus with proteins recognized by PubTator [95] as protein candidates for relation extraction. Table 4.5 lists the evaluation results on the in-house developed test corpus. As the PTM relation involves three kinds of arguments, i.e., enzyme, substrate and site, an abstract may not include all the arguments and only discuss all of them in the full-length article [33]. Thus, we report evaluation results on four different levels of extraction, which are extraction of the full relation (enzyme-substrate-site), enzyme-substrate pairs, substrate-site pairs, and the substrates.

Table 4.5: Evaluation Results on In-house Developed Corpus

Relation	Precision	Recall	F-score
	Extraction of Enzyme-Substrate-Site		
Acetylation	83.95	69.39	75.98
Glycosylation [a]	100.00	100.00	100.00
Ubi-like modification	100.00	48.39	65.22
	Extraction of Enzyme-Substrate		
Acetylation	81.48	70.97	75.86
Glycosylation	100.00	62.50	76.92
Ubi-like modification	100.00	52.83	69.14
	Extraction of Substrate-Site		
Acetylation	84.07	84.82	84.44
Glycosylation	88.31	75.56	81.44
Ubi-like modification	97.26	64.55	77.60
	Extraction of Substrate		
Acetylation	85.16	94.78	89.71
Glycosylation	98.06	86.32	91.82
Ubi-like modification	92.24	85.60	88.80

[a] Glycosylation only has 3 enzyme-substrate-site relations.

Generally, the system achieves high precision for all the relations. For substrate-site pairs and substrates the system also obtains good recall. The two sub-relations involving enzymes have lower recall, partially due to that the enzymes are often stated in a different sentence or complex pattern. We will analyze these errors in Section 4.6.

4.5.2 Evaluation on BioNLP 2011 EPI Test Corpus

To evaluate the system for mention-level relation extraction, we use the BioNLP 2011 EPI test corpus. The statistics of the corpus are listed in Table 2.2. Three of the five kinds of PTMs are annotated in this corpus, which are acetylation, glycosylation and ubiquitination. We report the evaluation results of these three relations as well as

Table 4.6: Evaluation Results for Full Relation Extraction on BioNLP 2011 EPI Test Corpus

	Our System			Top System in BioNLP 2011 [7]		
Relation	P	R	F	P	R	F
Acetylation	70.75	59.43	**64.60**	53.77	61.14	57.22
Glycosylation	81.91	40.96	**54.61**	53.05	46.28	49.43
Ubiquitination	87.67	70.72	**78.29**	73.94	76.80	75.34
Phosphorylation	79.59	56.52	66.10	63.64	71.01	**67.12**

Table 4.7: Evaluation Results for Substrate Extraction on BioNLP 2011 EPI Test Corpus

	Our System			Top System in BioNLP 2011 [7]		
Relation	P	R	F	P	R	F
Acetylation	94.96	83.02	**88.59**	78.53	78.53	82.74
Glycosylation	87.06	43.79	58.27	64.10	59.17	**61.54**
Ubiquitination	92.47	75.00	**82.82**	81.62	83.89	82.74
Phosphorylation	93.88	70.77	**80.70**	74.67	86.15	80.00

phosphorylation, a well-studied PTM for biomedical relation extraction. To compare with previous systems, we also include the results of the top system on BioNLP 2011 EPI test corpus [7].

Compared to the top system in the task, we obtain better precision for all the four PTM relations, and better F-score for the three PTM relations we are interested in. For phosphorylation we obtain comparable F-score, although we just used the general approach for this PTM and did not consider any special aspects since RLIMS-P had already been developed. This demonstrates the generalizability of the system for extract different kinds of PTM relations with the automatically generated patterns.

In addition to the results for full relation extraction given above, we also report the performance for extracting the substrates in Table 4.7.

Compared to the top system in BioNLP 2011 EPI task, the system achieves better precision for all the four relations. For acetylation, ubiquitination and phosphorylation we obtain better F-score. Due to the lower recall for glycosylation, the

F-score of our system did not surpass the top system, which has more balanced scores of precision and recall.

4.5.3 Effects of Referential Relations and Merging Rules

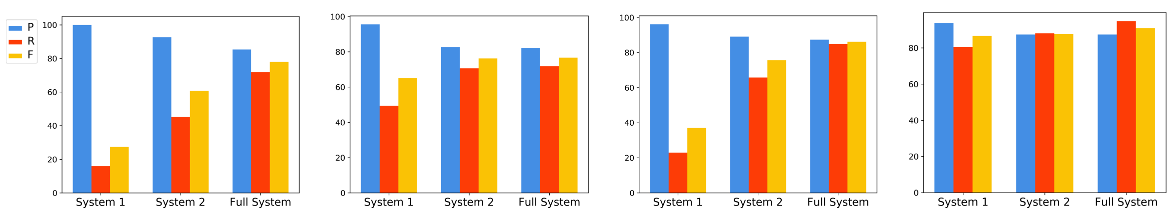
As the PTM extraction system uses referential relations (including anaphora resolution and part-of relation) and merging rules to increase recall, we investigate how the usage of these information affects the results (especially the recall) quantatively. For the extraction of acetylation, glycosylation and ubi-like modification, we evaluated three different system settings.

1. System 1: the system only uses generated patterns, but not referential relations/anaphora resolution and merging rules.

2. System 2: the system uses both generated patterns and referential relations/anaphora resolution, but not merging rules.

3. Full system: the system uses all the three components: the generated patterns, referential relations/anaphora resolution and the merging rules.

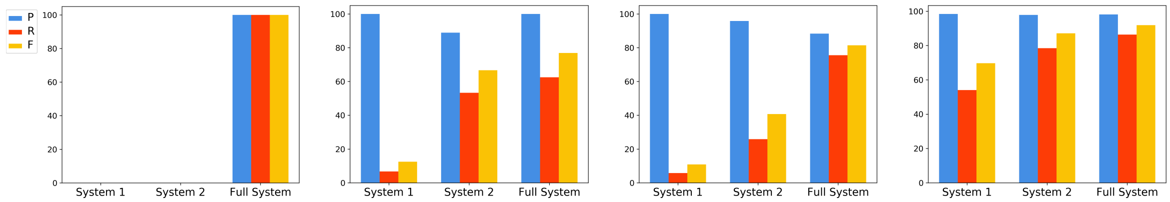
In Figure 4.6-4.8, we show the results using the three system settings for extraction of acetylation, glycosylation and ubi-like modification.



(a) Enzyme-Substrate-Site (b) Enzyme-Substrate (c) Substrate-Site (d) Subtrate

Figure 4.6: System Comparison for Acetylation Extraction

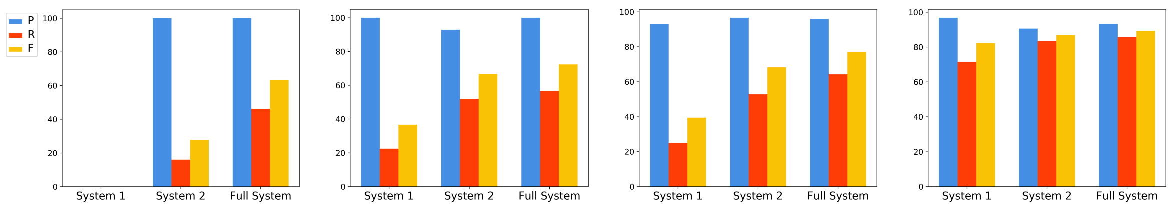
Figure 4.6-4.8 demonstrate a common pattern of increasing recall by using referential relations and merging rules. The results indicate that the usage of referential relations and merging rules are crucial for extracting comprehensive PTM information from text. For example, consider the extraction of glycosylation substrate-site in



(a) Enzyme-Substrate-Site (b) Enzyme-Substrate (c) Substrate-Site (d) Subtrate

Figure 4.7: System Comparison for Glycosylation Extraction

There are only 3 enzyme-substrate-site relations for glycosylation. System 1 and 2 extracts none of them.



(a) Enzyme-Substrate-Site (b) Enzyme-Substrate (c) Substrate-Site (d) Subtrate

Figure 4.8: System Comparison for Ubi-like Modification Extraction

System 1 extracts none of the enzyme-substrate-site relations for ubi-like modification.

Figure 4.7 (c), using generated patterns only gives 5.75 recall, and using both referential relations and merging rules increases the recall to 75.56 with small decrease in precision.

4.6 Error Analysis

In this section we analyze the error cases in the in-house developed corpus. For BioNLP 2011 EPI task, as we do not have direct access to the gold annotations in the test corpus, we use the training corpus for error analysis.

4.6.1 Errors in In-house Developed Corpus

We specifically check the false negatives in enzyme-involved sub-relations for acetylation and ubi-like relations, as their mentions are much more common than enzyme-involved glycosylation. Below are the main causes we find for the error cases.

Lack of domain knowledge: as the annotation is on document level, biologists can leverage domain knowledge and infer from the context to annotate the relation. An example is shown below.

> Example (PMID 11509661): In this study, we show that the interaction between CtBP and the nuclear hormone receptor corepressor RIP140 is regulated similarly, in this case by **p300/CBP** itself.

p300/CBP is annotated as the enzyme of an acetylation relation even though the sentence does not include an acetylation trigger word. p300/CBP is a well-known acetyltransferase and the first sentence in the abstract describes an acetylation relation with RIP140 as a substrate. So, presumably the annotators have combined these two sentences and inferred that the interaction of RIP140 is the cause of its acetylation and p300/CBP is the enzyme. For Ubi-like relations, this is a main cause for the false negatives in enzyme-involved PTM relations. For example, interaction between an SUMO-protein ligase and a sumoylation substrate implies the ligase is the sumoylation enzyme in the literature. We can use SUMO-ligase knowledge in the system and use interaction patterns to extract more enzyme-substrate relations for ubi-like PTM relations.

Cross-sentence relation: the enzyme is described in a different sentence than the one for the substrate-pair. As the system currently mainly extracts relations in the same sentence, these constitute false negatives in the evaluation. An example is shown below.

> Example (PMID 24051374): Using mouse models and in vitro assays, we show that **histone H3** *acetylation* at **Lys 9** (H3K9Ac) was elevated over gluconeogenic genes and contributed to increased hepatic glucose production during fasting and in diabetes. Dephosphorylation of CRTC2 promoted increased H3K9Ac through recruitment of the **lysine acetyltransferase 2B (KAT2B)** and WD repeat-containing protein 5 (WDR5), a core subunit of histone methyltransferase (HMT) complexes.

Note that the enzyme KAT2B and the substrate histone H3 appear in two different sentences and this information is combined by the annotators.

Although we used merging rules in Subsection 4.4.4 to combine some relations extracted from different sentences, they are primarily focused on extracting missing substrate in a PTM relation. To increase recall for enzyme-involved PTM relations, we can try to use and improve additional merging rules for enzyme extraction.

Negated relation: our system currently extracts negated relations as the predicate-argument patterns can be applied for these cases. However, biologists do not annotate these negated relations and thus they constitute false positives in the evaluation.

> Example (PMID 11739381): Here, we show that the p50 subunit as well as the p50/p65 of NF-kappaB, and not other factors such as **SP1**, **TFIIB**, **polymerase II**, **TFIIA**, or **p65**, can be *acetylated* by CBP/p300 HAT domain.

Since the system currently does not consider negation, it extracts five acetylation relations for the five proteins in bold, generating five false positives.

Complex expression for sites: there is a noticeable drop in recall for ubi-like relations for the substrate-site evaluation. The drop is mainly due to that sites are often not connected to the trigger in a straightforward way, as exemplified by the sentence below.

> Example (PMID 15824120): **Lys(16)**, **Lys(194)** preceding the polyglutamine tract, **Lys(610)/Lys(697)** in the C-terminal ataxin high mobility group domain, and **Lys(746)** all contribute to **ataxin-1** *SUMOylation*.

For this reason, the previous version of RLIMS-P [59] uses a post-processing rule to extract site argument for relations that only have substrate. If a PTM relation only has substrate arguments and there is a site entity in the same sentence, then it will be extracted as the site argument for the PTM relation. As we focus on the theme extraction, and introduction of this rule causes many false positives, we did not use this rule in the system. A more restricted version of the rule might work better and can be added to improve the recall.

4.6.2 Errors in BioNLP 2011 EPI Training Corpus

We obtained low recall for glycosylation in the test corpus. It turns out 54 (19.7%) of the 273 glycosylation relations in the training corpus involve entities in different sentences. This is a main cause for the false negatives. Another reason is that some glycosylation relations annotated in the BioNLP corpus are not expressed using predicate-argument structure of a trigger. An example is shown below.

> Example (PMID 16920952): We report that the **UL16-MICB** interaction is independent of *glycosylation*.

For all the relations, the BioNLP 2011 EPI task also requires the correct occurrence of the trigger words to be annotated. In several instances, we obtain the right relation but for a trigger we might find the same protein but a different occurrence of its mention than the one that was annotated in the corpus. In such cases, while the relation (e.g., the substrate and site) is correctly extracted, if will be counted as both a false positive and a false negative. We try to match with the annotation standard of the corpus but it is possible that in the test corpus we might not use the correct trigger for some cases. This error cause affects all the PTM relations that we extract for BioNLP 2011 EPI task.

4.7 Conclusion

In this chapter, we describe a general rule-based framework for biomedical relation extraction based on generalization of Lexicalized Tree-Adjoining Grammar (LTAG) and the usage of Extended Dependency Graph (EDG). We leverage the linguistic knowledge in LTAG to support automatic pattern generation and EDG as the sentence representation to enable extraction of the numbered argument. The system only needs a set of trigger specification as input and mitigate the need for manual pattern design. To illustrate the application, we use it to develop a PTM extraction system with some PTM-specific rules. Experiments on BioNLP 2011 EPI task show that the system achieves state-of-the-art performance for acetylation, glycosylation and ubiquitination relations. We also show that the system obtains good results on an in-house developed corpus for information-centric extraction. The framework can be adopted to extract other biomedical relations described using predicate-argument structure with little effort.

Chapter 5

DISTANT SUPERVISION FOR BIOMEDICAL RELATION EXTRACTION

In this chapter, we present our work to improve distantly supervised biomedical relation extraction. As the training data generated by distant supervision contains much noise, we first apply noise reduction techniques to improve the data quality, thus increase the performance of supervised learning algorithms for biomedical relation extraction. Next, as the size of the generated data set is large, we can better train Convolutional Neural Networks (CNN) and improve the results over simpler models. Last, we show that we can further improve the extraction performance by combining both distantly-labeled and human-labeled data. The experiment results show that we can obtain state-of-the-art performance for protein-protein interaction, a well-studied biomedical relation extraction task.

In the rest of the chapter, we first introduce the idea of distant supervision, Convolutional Neural Networks and combination of distantly-labeled and human-labeled data. We then present detailed experiment results to show that we can achieve state-of-the-art performance with the proposed methods for distantly supervised biomedical relation extraction.

5.1 Distant Supervision

Distant supervision has been used to generate training sets for machine learning algorithms by heuristically labeling training instances with a database. The heuristical labeling process does not require human effort, which makes it easier to apply it to extract new relations without hand-labeled training set. However, the heuristical labeling process does not produce a perfect training set. Noise (wrongly labeled instances) exists in both positively- and negatively-labeled data. In this section, we first introduce the basic idea of heuristic labeling, followed by an analysis of noise sources and a review of existing techniques to handle them.

5.1.1 Heuristical Labeling Process

The main component of distant supervision is a heuristical labeling process used to label text corpus automatically. When applying it to binary relation extraction, in its simplest form, a database containing related entity pairs is used. Then text (appropriately defined) that contains a pair of related entities mentioned in the database is used as basis of a positive instance. We currently only consider relation extraction from individual sentences. To illustrate the labeling process, for the protein-protein interaction (PPI) extraction, we consider the database IntAct [63] and Medline abstract sentences as the text corpus. Consider the pair <NgR, p75> found in IntAct database and the sentence: **NgR** interacted with **p75** in lipid rafts. The sentence is labeled as a positive instance and thereby assumed to express protein-protein interaction relation between the protein pair. When a sentence contains a pair of proteins which is not recorded by IntAct, it is then labeled as a negative instance. As the heuristic labeling process has no additional guidance, incorrectly labeled instances can exist in both positively- and negatively-labeled data.

5.1.2 Noise in Positively-Labeled Data

The main cause of the noise in the positively-labeled data is that the text that contains a related entity pair does not necessarily express any relation between the two entity mentions, or possibly describes a relation different from the one of interest. For example, according to IntAct database, protein “CycC” interacts with protein “Med12”. However, the following sentence does not express an interaction relation between the two proteins but would be labeled as a positive example by the heuristic labeling process, thus constituting noise in the positively-labeled data.

Example (PMID-17290221): We analyzed the in vivo functions of this module using null mutants for Cdk8, **CycC**, **Med12** and Med13.

Researchers have proposed various methods to mitigate the noise in the positively-labeled data. Multi-instance learning is used by [77] to extract related entity pairs directly using supervision from curated entity pairs in database, with the hope that it can tolerate noise at sentence level to some extent. In detail, the multi-instance model is trained using the entire set of sentences for an entity pair, such that it can combine evidences from multiple sentences. Even though noise could exist in some of the sentences, multi-instance learning assumes that at least one sentence of an entity pair expresses the target relation (at-least-one assumption). Multi-instance multi-label learning [30, 86] was built upon the multi-instance learning method to exploit correlations and/or conflicts among relations.

Researchers have also used statistical models to estimate the probability *P(relation|pattern)* directly, and then remove instances with low-confident patterns in positively-labeled data [87].

Distant supervision has been applied to extract biomedical relation with methods to mitigate noise in positively-labeled data. Zheng et al. [104] used a heuristic based on dependency path frequency to reduce noise in the positively-labeled data for extraction of protein-localization relations. Thomas et al. [88] used a list of words which are frequently employed to indicate protein interaction to filter out noise for protein-protein interaction extraction. Roller et al. Multi-instance learning was used by Roller et al. [78] to extract two subsets of relations in UMLS database, and by Lamurias et al. [42] to extract miRNA-gene relations.

5.1.3 Noise in Negatively-Labeled Data

The main cause of the noise in negatively-labeled data is the incompleteness of database. If a related entity pair is not recorded in the database, all of the sentences mentioning this entity pair will be labeled as negative instances. For example, UniProt database [91] does not record the protein-localization relation between protein "DAT" and subcellular location "filopodia", and the protein-location pairs in the following sentence would be labeled as a negative instance. However, the sentence expresses that the protein is enriched in "filopodia". This is an example of noise in negative data.

Example (PMID-25936602): **DAT** was enriched in **filopodia**.

To remove noise from negatively-labeled data, researchers have tried to discover novel relations that are not recorded by the database. High-confidence new relations can then be removed from the negatively-labeled data. Xu et al. [99] used pseudo-relevance feedback to identify new relations. Roller et al. [78] used a Path Ranking Algorithm to infer new relations based on an existing knowledge graph in UMLS.

The possibility of using labeled data to remove noise from negatively-labeled data has also been investigated. Pershina et al. [70] generates *training guidelines* from a labeled data set and uses them to re-label negatively-labeled data. The training guideline consists of entity types, shortest dependency path between the two entities and optional span words, and can be use as a pattern to match negatively-labeled data. Angeli et al. [3] tried to improve distant supervision performance over KBP tasks by querying human-annotated labels of ambiguous instances, which adopted the idea of active learning. In the multi-instance setting, the human-labeled instances can be used to replace the distantly labeled ones.

Besides the noise in negatively-labeled data, distantly labeled data is usually skewed in that the negative set is much larger than the positive set. Thus, negative subsampling is frequently used to build a more balanced training set.

For biomedical relation extraction, Thomas et al. [88] filtered out any negatively-labeled instances where the sentence contains an interaction word. Researchers also tried to build reliable negative data sets, which may contain less noise. Thomas et al. [88] used a database containing confirmed non-related entity pairs to label negative instances. Zheng et al. [104] built a negative training set for subcellular-localization using sentences which do not contain a pair of protein and subcellular location. They

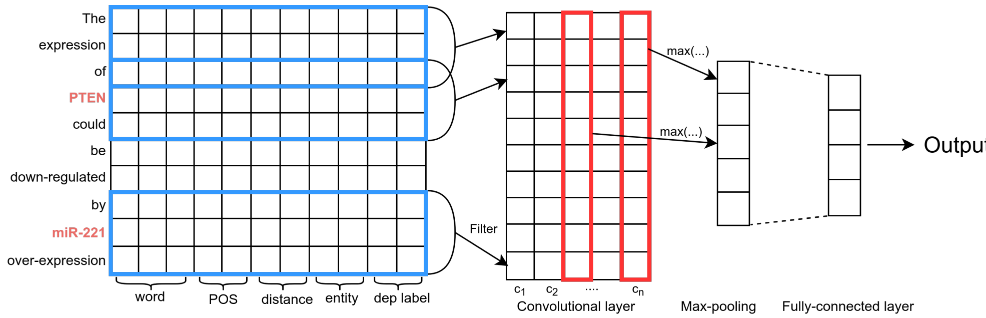


Figure 5.1: Typical Structure of Convolutional Neural Network Model.

The input sentence is "The expression of PTEN could be down-regulated by miR-221 overexpression". "PTEN" and "miR-221" are the two candidate entities for miRNA-gene regulation relation.

pair a gene or a subcellular location with another random word in a sentence to form a negative instance.

5.2 Convolutional Neural Networks

Convolutional neural networks (CNN) have been used for general relation extraction tasks and achieve state-of-the-art performance [98, 102]. Recently, they have been applied to extract biomedical relations [5, 68, 81]. Fig. 5.1 depicts a typical CNN model for relation extraction with an example sentence for miRNA-gene regulation relation as input. The inputs to CNN models are word embedding vectors and additional lexical and syntactic information, as shown in the first matrix in Fig. 5.1. A set of fixed-size filters (blue blocks) are then used to generate non-local features. The output of each filter is feed into a max-pooling layer, which only selects the element with maximum value in the pooling window (red blocks). In the model presented in Fig. 5.1, the max pooling is applied to the entire output of each filter, and the generated output vector no longer depends on the sentence length. At last, fully-connected layers are used to learn higher-level features and the final output can be generated using sigmoid function.

Convolutional layer: it consists a set of fixed-size convolutional filters, which can merge the local word-level information and generate higher-level features. Assuming the input sentence length is l, the feature dimension of each word is e, the input matrix of the sentence is $W \in \Re^{l \times e}$. A convolutional filter is a matrix of weights with the shape $F \in \Re^{w_1 \times w_2}$, where w_1 and w_2 are the convolutional window size. Usually w_2 is set to be equal to the word feature dimension e. We then multiply the filter with the input matrix convolutionally as in Equation 5.1.

$$c = f(F \odot W_{i:i+w_1} + b) \tag{5.1}$$

where $i \in [1, l - w_1 + 1]$, b is the bias term and f is the activation function, e.g., Sigmoid or ReLU. Assuming there are m independent filters, the output of the convolutional layer is then $C = [c_1, c_2, ..., c_m]$.

Max-pooling layer: Max-pooling is an operation that takes a matrix as input and outputs the max value among all the elements in a window. Using this operation, we can keep the most important features in the convolutional layer while reducing the number of features, which can significantly increase training speed. In previous works [68, 75], the max-pooling is applied on each component in the output C of convolutional layer. The max-pooling operation is defined as:

$$d_i = max(C_i) \tag{5.2}$$

where $i \in [1, F]$. After max-pooling, the output no longer depends on sentence length.

Fully-connected layer: all the nodes in this layers connect to all the nodes from the previous layer, defined as:

$$O = f(W_o D + b) \tag{5.3}$$

where W_o is the weights of fully-connected layer, b is the bias term. f is the activation function, e.g., Sigmoid or ReLU.

Although CNN models can achieve good performance without complex human-engineered features, it requires large amount of data to effectively train the numerous parameters in the model. Previous work applying CNN for biomedical relation extraction only use small annotated data set as training data. In this chapter, we use distant supervision to generate a large data set for training CNN models and achieve better performance than simple logistic regression methods.

5.3 Combining Distantly-labeled and Human-labeled Data

As many biomedical relation extraction tasks have small amount of human-labeled data available, we try to combine it with distantly-labeled data to improve results. Similar (but not exactly the same) idea has been used in the transfer learning methods, which leverage existing labeled data in a similar domain to help learning models in the target domain that does not have sufficient training data [64]. In our scenario, we have large amount of noisy labeled data generated by distant supervision, and only small amount of human-labeled data for biomedical relations. As the CNN models require large amount of data to effectively train the model parameters, we can use a trained CNN model on distantly labeled data as a better starting point and re-train it on small amount of human-labeled data to increase performance.

A few previous works investigated the possibilities to combine distantly-labeled and human-labeled data to improve results. Angeli et al. [3] tried to use human-labeled data to guide the training MIML-RE model [86], and increased the F-score by 3 points. Roller et al. [79] merged human-labeled and distantly-labeled data into one corpus and found that training on the mixture corpus achieved better F-score by about 1 point compared to training solely on the human-labeled data for extraction of adverse drug effects. Rao et al. [76] trained LSTM model using both human-labeled and distantly-labeled data for event extraction on BioNLP 2013 Genia corpus [38], and achieved better recalls for some of the event types. To the best of our knowledge, our work presented in Section 5.9 is the first to combine distantly-labeled and human-labeled data to train the models in a two-step manner.

Table 5.1: Text Corpora and Databases for Three Tasks

Task	Database	Text corpus	Positive / Negative
PPI	IntAct	14,769 abstracts curated by IntAct	67,009 / 108,016
MIRGENE	Tarbase [93] miRTarBase [32]	30,000 abstracts	75,632 / 97,118
PLOC	UniProt	30,000 abstracts	28,985 / 82,132

5.4 Experiment Tasks

We use three tasks, extraction of protein-protein interaction (PPI), miRNA-gene regulation relation (MIRGENE), and protein-localization event (PLOC) in the experiments to evaluate our methods.

To annotate a text corpus using the heuristical labeling process, we need to use named entity recognition and normalization tools name entity names to database IDs. For gene/protein, we use the output from GenNorm++ [96], which has been used to process all the Medline abstracts and the results. The miRNA recognition tool from miRTex is used to recognize miRNA mentions, and extract the miRNA number to map it to a miRNA entry in the database. For subcellular location, we use a dictionary from UniProt [91] and perform string matching to find subcellular location mentions. The dictionary contains name variants for each location, and we normalize matched variant to its standard name.

5.5 Training Set Construction

To construct training set using distant supervision for a relation, we need a database containing related entity pairs and a large text corpus which is used for heuristical labeling. We list the databases and text corpora for the three tasks in Table 5.1.

For miRNA-gene regulation relation, we first retrieved all the Medline abstracts containing sentences in which a pair of miRNA and gene is mentioned, and sampled 30,000 abstracts from it. For protein-localization event, similarly, we retrieved 30,000 from all the abstracts containing sentences in which a pair of protein and subcellular location appears. For protein-protein interaction, since the use of Medline abstracts led to a skewed labeled data set (1:7.4 positive/negative ratio), we turned to use all the abstracts that are curated by IntAct database as the text corpus. Although this may result in less noise, we will show that our proposed methods can still increase performance over the baseline in the experiments. Table 5.1 also lists the numbers of positive and negative instances labeled automatically by distant supervision.

5.6 Test Sets

To evaluate different methods, we need a test set for each relation. Note that in the context of distant supervision, we should expect little or no hand-labeled data. Hence, we can not assume the availability of a development set for purposes of parameter tuning. Thus, when a method has multiple possible choices for a parameter, we will report the evaluation results using different parameter values. The test sets for each task are listed in Table 5.2.

Table 5.2: Test Sets for Three Tasks.

Task	Test set	Size
PPI	AIMed	225 abstracts
MIRGENE	miRTex dev set	200 abstracts
PLOC	BioNLP 2011 Genia training/dev set	1,167 abstracts/full-text sections

Of all the protein localization events in BioNLP corpus, we only use those with a location argument as the test instances, same with [104]. We ensure that the abstracts used by the test sets are not included in the training sets, i.e., the training set is independent from the test set for each task.

5.7 Methods

In this section, we first describe the machine learning model and the feature set. Next, we describe the baseline systems used for comparison. Then we present the three noise reduction methods based on the notion of proximity, trigger word and confidence of patterns.

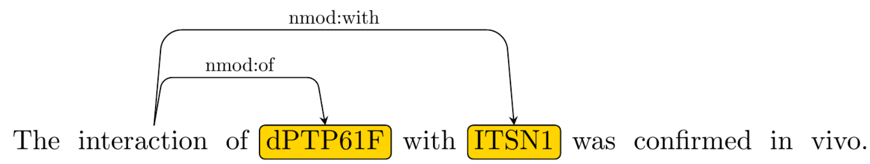


Figure 5.3: Example Sentence for Feature Extraction.

5.7.1 Learning Algorithm and Feature Set

We use logistic regression to train the extraction models in all the experiments. Used features include lexical features extracted from the surrounding words and syntactic features extracted from the shortest dependency path of the entity pair. An example sentence and its extracted features are shown in Fig. 5.3 and Table 5.3. E-walk and v-walk features are (edge, stem, edge) and (stem, edge, stem) triples including direction extracted from the shortest dependency path. They preserve partial structure information and are more generalizable than the full dependency path.

Name	Feature
Shortest dependency path	P1←nmod:of←→nmod:with→P2
E-walk	nmod:of←interact→nmod:with
V-walk	P1←nmod:of→interact interact→nmod:with→P2
Context	P1_with_P2 of_P2_with_P2_be interact_of_P1_with_P2_be_confirm
Length of shortest dependency path	2
Number of words between P1 and P2	1

Table 5.3: Features Extracted from Example Eentence.

For the lexical terms, we use the Porter's stemmer [71] to obtain their stems. Charniak parser with the biomedical model [15, 16] is used to produce constituency parse for each sentence, which is converted to collapsed dependency parse using Stanford CoreNLP converter [50] with *CCprocessed* setting. We include extended word stem sequences between the two entities by adding 1 and 2 words before the first and after the second entity (the context of the entities), thus producing two additional stem sequences. For all the experiments, we remove features that only appear once in the training data.

5.7.2 Baselines

The baseline is a logistic regression (LR) model trained on the distantly labeled set without any filtering of noise. We also implement two previous methods as baselines for comparison. First, we train a LR model on the distantly labeled set filtered by a heuristic proposed by Zheng et al. [104], which removes positively-labeled instances with a shortest dependency path that appear less than k times in the positive set. They hypothesize that rare dependency path is unlikely to express a relation. As we tried different values of k and obtained similar F-scores for the three tasks, we only report the results for $k = 5$ to save space. Note that since different features, text corpus and named entity recognition tool are used, we are not trying to reproduce the exact results reported in Zheng et al. [104]. In addition, we implement a widely-used multi-instance model described in [86] and train it on unfiltered distantly labeled data. We name these three baselines as LR, DPFreq and Multi-instance and report their performances in Table 5.6.

5.7.3 Noise Reduction Methods

We propose three novel filtering methods to remove noise from both positively and negatively labeled data. These methods are applied in a sequential manner so that each step removes more noise based on the filtered data from the previous step.

The first heuristic is concerned with multiple mentions of an entity in a sentence. If the entity is related to another entity mentioned in the sentence, all the binary combinations of their mentions will be labeled as positive by the default labeling process. This usually introduces noise, since not all combinations are likely to be in the relation. For example, consider the sentence below.

> Overexpression of **miR-193b** inhibited the expression of **CCND1**, and knock-down of **CCND1** inhibited the proliferation of GC cells, suggesting that **miR-193b** exerted its anti-tumorigenic role in GC cells through targeting **CCND1** gene.

miR-193b regulates CCND1 according to the database TarBase. The six binary combinations between miR-193b and CCND1 in the sentence will be labeled as positive instances. However, the sentence only expresses miRNA-gene regulation relation for the first and the last combination. The other four are wrongly labeled and hence constitute noise in the positively-labeled data. To remove such noise, we hypothesize that only the closest pair of the entity mentions express the relation.

The closest pair is defined as following: for a positively-labeled entity mention pair $<e_1, e_2>$, if their shortest dependency path has the smallest length among all the positively-labeled pairs that involve either e_1 or e_2, the pair $<e_1, e_2>$ is considered as a closest pair. When computing the dependency path length, we skip the appositive relation. The heuristic is described as below.

> **Heuristic of closest pairs (CP):** remove positively-labeled instances that are not closest pair, when multiple mentions of one or both entities are present in the sentence.

For the three tasks and many other biomedical text-mining tasks, the relation or event is often indicated by a small set of trigger words (e.g., interact/bind for PPI, regulate/target for MIRGENE, and localize/translocate for PLOC). Following the usage in the BioNLP Genia corpus, we can term these words as trigger words. With knowledge of a comprehensive set of trigger words, we can hypothesize that sentences without a trigger word are less likely to express the target relation or event. Since we wish to apply distant supervision on a wide variety of relation extraction tasks, we don't assume any prior knowledge of the trigger words for each relation. Instead, we propose to automatically mine the trigger words from the large distantly-labeled corpus itself, and use them to remove noise from the positively-labeled data.

Trigger words are usually verbs, or in their nominal or adjectival form. Our target is then to identify stems of verb triggers, which can also be used to match nominal or adjectival form of the verb. A simple procedure is used: first, count all the verb stems on the shortest dependency paths of the positively-labeled instances generated by the heuristic labeling process. As we want to choose triggers that are strongly associated with the relation, we only use dependency paths that contain one token, excluding the two entity mentions. These verb stems are then sorted by frequency and the high-frequency stems are chosen for the trigger list. We list the top 10 verb stems for the three tasks in Table 5.4.

Table 5.4: Top 10 Verb Stems and Top Pattern and Example Sentence for Three Tasks.

Task	Verb stems	Pattern and example sentence
PPI	interact bind associ phosphoryl recruit activ coloc coimmunoprecipit co-immunoprecipit regul	PROTEIN1←nsubj←interact→nmod:with→PROTEIN2 **mGrb10** interacts with **Nedd4**.
MIRGENE	target regul inhibit downregul suppress repress down-regul correl induc promot	GENE←dobj←target←advcl←root→nsubj→MIRNA **MiR-429** play its role in PDAC by targeting **TBK1**.
PLOC	local transloc express associ interact detect coloc find co-loc target	PROTEIN←nmod:of←transloc→amod→LOCATION Importin beta mediates **nuclear** translocation of **Smad 3**.

For each positively-labeled instance, we search for trigger stems in the tokens on its shortest dependency path or in the maximum dominating noun phrase. A maximum dominating noun phrase is defined as the maximally-spanning noun phrase that encloses the two entity mentions, with only noun or prepositional phrases as descendants. For example, in the text fragment "interaction between **FAK** and **PP1** regulates a process", the maximum dominating noun phrase is "interaction between FAK with PP1" for this protein mention pair. As sentences without a trigger word are less likely to express the target relation or event, we use the heuristic described below to remove noise.

Heuristic of trigger word (TW): remove positively-labeled instances if a trigger stem is not found on the shortest dependency path or in the maximum dominating noun phrase of the entity mention pair.

By using heuristic CP and TW, we can already filter out a substantial part of the positively-labeled data. Using heuristic CP+TW with 50 trigger stems, 65% of the positively-labeled data can be removed for PPI. For MIRGENE and PLOC, the removal ratio is 38% and 59%, respectively. We hypothesize that the remaining set will still contain a large amount of data for training and more importantly, it will be of high quality, and thus it would be possible to discover high-confidence patterns from it using pattern occurrence frequency.

Finally, we turn to the last heuristic that we introduce. Recall noisy instances in negatively-labeled data should be labeled as positive but are negatively labeled because of incompleteness of the database used for distant supervision. We try to mine some high-confidence patterns from the purified positively-labeled set after the application of heuristic CP and TW. We define a pattern as a shortest dependency path lexicalized by a trigger stem between the entity mention pair. The pattern frequencies in the positively-labeled data filtered by heuristic CP and TW are counted. The most frequent pattern and an example sentence for each task are shown in Table 5.4. Our hypothesis is that any entity mention pair connected by a high-confidence pattern is likely to be related and hence probably constitute noise in the negatively-labeled data. Therefore, we consider the next heuristic described below.

Heuristic of high-confidence patterns (HP): remove negatively-labeled instances which match a high-confidence pattern mined from positively-labeled data.

Note that heuristic DPFreq, CP and TW remove instances from the positively-labeled data, whereas HP is the only heuristic that removes instances from the negatively-labeled data. Heuristic TW depends on the number of trigger stems, while heuristic

HP depends on both the number of trigger stems and high-confidence patterns, as it needs the trigger stems to lexicalize the shortest dependency path to form a pattern.

5.7.4 Application of Convolutional Neural Networks

With the large amount of distantly labeled data, we can train convolutional neural networks for the three relation extraction tasks, and compare the results with logistics regression. The input to the CNN model is the entire sentence. We map each word to pre-trained word embeddings and include POS, entity typing and other information. For each word, we concatenate all the information into one vector as the representation of the word. Below we list the features used to generate the input vector.

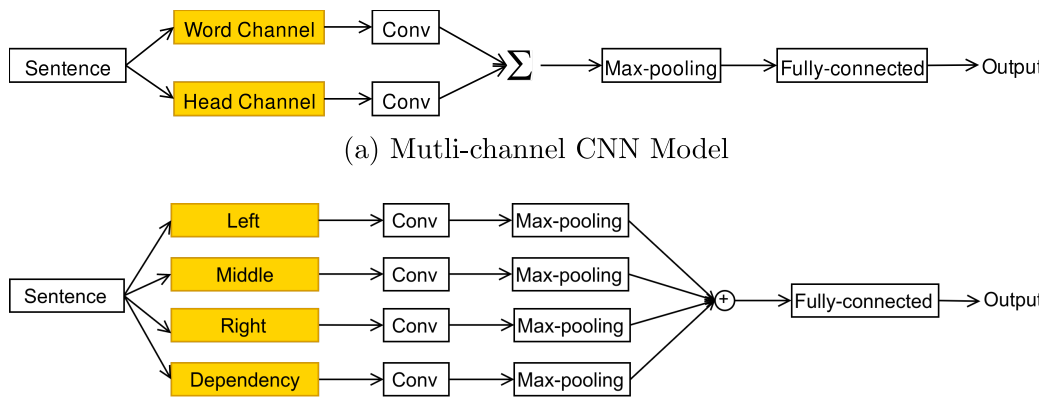
Word embedding: we use pre-trained word embeddings over the entire PubMed abstracts [17]. Each word is mapped to a 200-dimension vector.

POS tag: we map the POS tag of each word to a 10-dimension vector. We tried using dimension of 5 and 20 in the experiments and obtained similar results. The vectors are initialized to values between -0.1 and 0.1 randomly.

Distance: similar to previous works using CNN models [68, 102], this feature encodes the distance from each word to the two candidate entities. We map the two distances to two 5-dimension vectors. We tried using dimension of 10 in the experiments and obtained similar results. The vectors are initialized to values between -0.1 and 0.1 randomly.

Token entity type: we use four types: *E1* for tokens of the first candidate entity, *E2* for the second candidate entity, *E* for other entities in the sentence and *O* for other tokens. The dimension of this feature is thus 4. We use one-hot representation for this feature and make it untrainable.

Dependency relations: we use incoming dependency relations of each token and map them to a 10-dimension vector. We tried using dimension of 5 and 20 in the experiments and obtained similar results. The vectors are initialized to values between -0.1 and 0.1 randomly.



(a) Mutli-channel CNN Model

(b) Piecewise CNN Model

Figure 5.4: Model Structure of Multi-channel CNN and Piecewise CNN.

We adopted two different network structures, which are inspired by multi-channel CNN (MC-CNN) [68] and piecewise CNN (PCNN) model [101]. The main difference between the two models lies in how we organize the inputs for the model. Fig. 5.4 demonstrates the two model structures. For simplicity we omit the dropout and batch normalization layers.

The main difference between the MC-CNN and PCNN model is the input information and how they incoporate syntactic information (yellow blocks in Fig. 5.4). Specifically, multi-channel CNN tries to captures the syntactic dependency between words and their heads by combining the word and head channels. On the other hand, piecewise CNN exploits the tokens on the shortest path dependency between the two entities directly. Other methods to incorporate dependency information include [35] and [103]. In addition, piecewise CNN divides the sentence into three components and applies three individual convolutional and max-pooling operations. The three components are the words before the first entity (included), the words between the two entities and the words after the second entity (included). The hypothesis is that the three components may contain different important information for extracting the relation and thus should be separated for applying max-pooling. The max-pooled outputs for these three components and the tokens on shortest dependency path are concatenated into one vector as input to next layer. The later layers used by the two models are similar, which are typical CNN layers, e.g., convolutional layer, max-pooling layer and fully-connected layer.

After the input, convolutional and fully-connected layers, we also apply batch normalization [37] and dropout [85], mainly for faster training and avoiding overfitting.

5.7.5 Parameters Tuning

To train the CNN models, we use the distantly-labeled data filtered by the three noise reduction heuristics. We randomly split the data into training and development set, as shown in Table . The develoment set is used to tune the parameters of the models for each of the three relations.

Table 5.5: Development and Test Set for Three Relations.

Relation	Training Instances	Development Instances
PPI	121,229	5,000
MIRGENE	131,874	2,000
PLOC	92,522	500

For all the models, we use one convolutional layer with 400 filters and two fully-connected layers with 512 and 256 hidden nodes, respectively. For all the trainable variables in the model we apply L2 regularization. We use Adam Optimizer [41] with a batch size 128 to minimize the loss function. We record the learning rates and other parameters in Appendix C.

To make the model not depend on the actual entity names and can generalize to unseen entity mentions, we mask the two candidate entities in the sentence with an artificial token *ARG_ENTITY*. We also mask other entities in the same sentence with *ENTITY*. The word embeddings of these two artificial tokens are intialized randomly to small values between -0.001 and 0.001.

We use a maximum sentence length of 160 for all the tasks. Sentences shorter than 160 tokens are padded with zero values, while the ones longer than 160 are cut off.

5.8 Results

In this section we present the results for noise reduction methods and convolutional neural network models.

5.8.1 Results for Noise Reduction Methods

We use precision, recall and F-score to evaluate the baselines and proposed methods. The top 50 trigger stems were used in heuristic TW, while the top 50 trigger stems and the top 100 patterns were used in heuristic HP. The results are presented in Table 5.6. We will discuss how different numbers of trigger stems and patterns may affect the results later. Table 5.6 shows that the multi-instance model and the use of heuristic DPFreq or CP increased precision compared to the baseline for all the three tasks, indicating that they can effectively remove noise from the positively-labeled data. Using heuristic CP+TW further improved precisions over heuristic CP for the three tasks. However, using heuristic DPFreq, CP or CP+TW did not improve the F-score over the baseline for PPI and MIRGENE, due the decreased recall. By removing noise from the negatively-labeled data using heuristic HP in addition to CP and TW, the recalls can be improved with minor or no decrease in precision, resulting in higher F-scores than the baseline, the MI model and other heuristics for all the three tasks. This suggests that the proposed heuristics can effectively remove noise from both positively and negatively-labeled data, and to obtain better F-scores, it is important to filter both positive and negative set to improve precision and recall simultaneously. Although PLOC extraction did not obtain a good precision in all the experiments, we will show that high precision can be achieved for high-confidence PLOC extraction later in this section. By applying heuristic CP+TW+HP, the F-score can be improved by 10 points for PPI extraction compared to previous work of Bobic et al. [9], and 11 points for PLOC extraction compared to Zheng et al. [104].

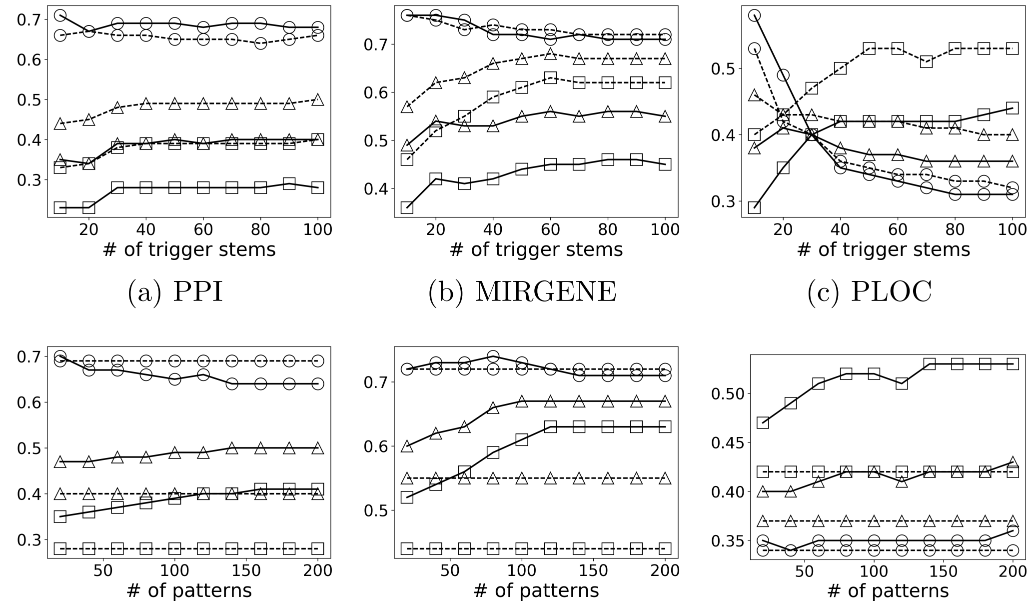
Different numbers of trigger stems: as different numbers of trigger stems can be used in heuristic TW and HP, we investigated how they affect the performance for the three tasks. In Fig. 5.5 (a)-(c), precisions, recalls and F-scores are shown for

Table 5.6: Precisions, Recalls and F-scores for Three Extraction Tasks.

	PPI			MIRGENE			PLOC		
Method	P	R	F	P	R	F	P	R	F
Bobic et al. [9]	0.26	**0.78**	0.39	-	-	-	-	-	-
Zheng et al. [104]	-	-	-	-	-	-	**0.43**	0.25	0.31
Baseline	0.37	0.52	0.43	0.56	0.58	0.57	0.18	**0.57**	0.28
Multi-instance (MI)	0.57	0.35	0.43	0.64	0.56	0.59	0.22	0.38	0.29
DPFreq	0.42	0.41	0.41	0.63	0.50	0.56	0.21	0.39	0.29
CP	0.55	0.34	0.42	0.68	0.50	0.57	0.26	0.51	0.35
CP+TW	**0.69**	0.28	0.40	0.72	0.44	0.55	0.34	0.42	0.37
CP+TW+HP	0.65	0.39	**0.49**	**0.73**	**0.61**	**0.67**	0.35	0.53	**0.42**

applying heuristic CP+TW and CP+TW+HP (using top 100 patterns) with different numbers of trigger stems. PPI and MIRGENE extraction maintained a stable precision with increasing recall when the number of trigger stem increased. For PLOC extraction the precision decreased with an increased recall when more trigger stems were used. We inspected the automatically mined trigger stems for PLOC and found that although the top trigger stems express the PLOC relation, e.g., “transloc” and “local”, many stems not clearly expressing the PLOC relation were also included. For example, “find”, “interact”, “regul”, “function” and “induce”. As a result, using more trigger stems failed to remove noise from the positively-labeled data, leading to a lower precision. Using 100 patterns to remove noise resulted in much better recalls and F-scores for all the three tasks across different numbers of trigger stems, further confirming that heuristic HP is an effective method to remove noise from the negatively-labeled data.

Different numbers of patterns: we investigated how different numbers of patterns used by heuristic HP affect the results. In Fig. 5.5 (d)-(f), precisions, recalls and F-scores are shown for applying CP+TW+HP (using top 50 trigger stems) with different number of patterns. The performances using heuristic CP+TW with 50 trigger stems are included for comparison. We can see that recalls can be consistently improved when more patterns were used, with minor or no decrease in precision. Compared to the results only using heuristic CP+TW, even using small number of patterns



(a) PPI (b) MIRGENE (c) PLOC

(d) PPI (e) MIRGENE (f) PLOC

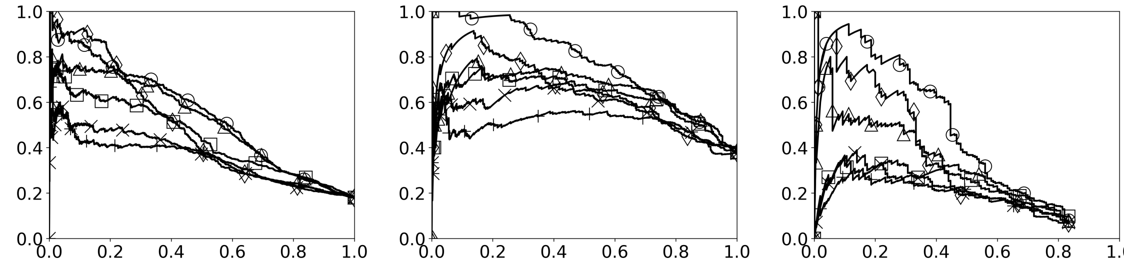
Figure 5.5: Results of Using Different Numbers of Trigger Stems and Patterns.

(a)-(c): different number of trigger stems; (d)-(f): different number of patterns. Markers: precision (circle), recall (square), F-score (triangle). (a)-(c): CP+TW (solid) and CP+TW+HP (dashed). (d)-(f): CP+TW (dashed) and CP+TW+HP (solid).

can achieve better performance.

A major use case of biomedical relation extraction is to help identify high-confidence entity pairs to facilitate manual curation for databases. Thus, a desired property of a relation extractor is to achieve high precision for such high-confidence extractions. Logistic regression model outputs a probability for each test instance, and high probability indicates high confidence to be positive.

To investigate the performance of the proposed methods for the high-confidence extractions, we draw precision-recall curves using the probability produced by the logistic regression model. By definition, logistic regression model predicts an instance as positive if the probability is greater than 0.5. By varying the threshold, we can calculate precisions at different recall levels. For example, when the threshold is set to 0.9, the model only predicts an instance with probability greater than 0.9 as positive. Ideally the model should achieve better precision when the threshold is high. For each



(a) PPI (b) MIRGENE (c) PLOC

Figure 5.7: Precision-Recall Curves for Three Tasks.

Y-axis represents precision and X-axis represents recall. Markers: baseline (+), multi-instance (diamond), DPFreq (x), CP (square), CP+TW using 50 trigger stems (triangle), CP+TW+HP using 50 trigger stems and 100 patterns (circle).

task, six curves are drawn in Fig. 5.7. We can see that using heuristic CP+TW+HP obtained higher precisions than the baselines and other heuristics on the left side of the figures, which correspond to the performance for high-confidence extractions. The multi-instance model also obtained better precisions compared to the baseline at lower recall levels. Specifically, by using heuristic CP+TW+HP, PPI, MIRGENE and PLOC extraction can achieve the highest precisions among the six curves, which are 0.71, 0.95 and 0.77, respectively, at recall level 0.30.

5.8.2 Results for Convolutional Neural Networks

The results of CNN models are listed in the Table 5.7 below. The results of logistic regression are produced by the models trained on the data filtered by all the three noise reduction techniques [46]. All the results below are obtained by evaluation on the test set in Table 5.5.

We can see that CNN models achieved better F-scores and recalls than the logistic regression models for all the three relations. One of the reasones may be the different feature representations used by CNN and logistic regression. For logistic regression, the features are encoded with the one-hot representation, which is sparse and possibily hampering generalization if a feature in test set has not been seen before. For CNN model we used distributed embeddings for words and their POS, entity

Table 5.7: Results for Convolutional Neural Networks.

Relation	Model	Precision	Recall	F-score
PPI	LR	**65.4**	38.7	48.6
	MC-CNN	63.5	42.8	51.1
	PCNN	61.2	**56.8**	**58.9**
MIRTAR	LR	**73.0**	61.2	66.6
	MC-CNN	72.1	69.6	70.8
	PCNN	73.3	**75.6**	**74.4**
PLOC	LR	34.9	53.3	42.2
	MC-CNN	**46.8**	**57.6**	**51.6**
	PCNN	46.7	56.0	50.9

typing and syntactic information. As similar words will have similar embeddings [53], the learned model may generalize better to unseen words and features if they have similar embeddings with the ones used in the training set.

5.9 Combining Distantly-labeled and Human-labeled Data

With distant supervision and convolutional neural networks, we can achieve better performance for the three relations than the previous works on distantly supervised biomedical relation extraction. In this section, we wish to see a different application for distant supervision. While impressive supervised learning results have been obtained using CNN and deep learning in general [31, 52], due to the large number of parameters in the network, their performance is likely to be hampered when there isn't sufficiently large training data. In such cases, we wish to see whether we can improve supervised learning results if we supplement small amounts of annotated data with large amounts of distantly supervised data, especially if we apply the noise reduction methods discussed in this chapter.

5.9.1 Methods

Figure 5.9 demonstrates the pipeline we use to combine both distantly-labeled and human-labeled data for training the CNN models. First, we train piecewise CNN

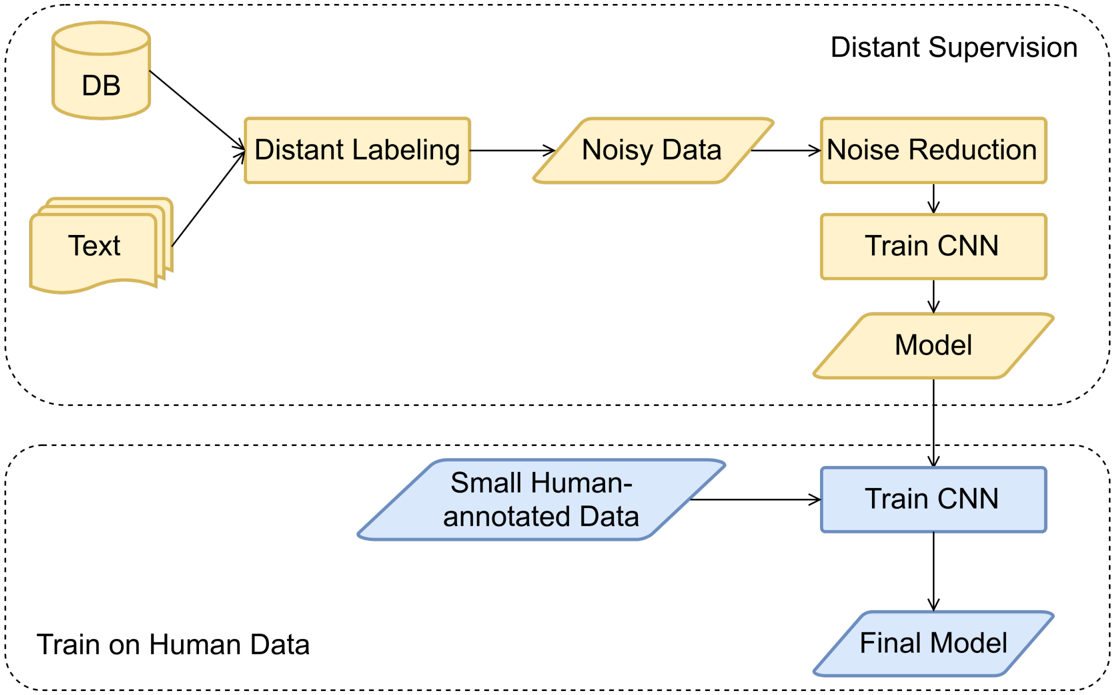


Figure 5.9: Pipeline for Combining Distantly and Human-labeled Data

models on the distantly-labeled data. Then, to further train the piecewise CNN models we use 10-fold cross validation at document level on AIMed corpus, a setting used by previous works [57, 68]. We also conduct 10-fold cross validation at instance level, as some other previous works [47] adopted this setting.

5.9.2 Results

We evaluate the pipeline for all the three relations. As baselines, we train the piecewise CNN models over the human-annotated dataset using 10-fold cross validation. The results for the three relations are listed in Table 5.8 and 5.9. The precision, recall and F-score are averaged over the 10 folds.

The results show that using both kinds of data can achieve better performance than the models solely trained on small human-annotated data. This indicates that the CNN models trained on noisy distantly-labeled data can be used as a better starting point for training on small human-annotated data. For PPI, we are able to improve the performance further by 4.5 points in F-score over previous state-of-the-art using

Table 5.8: Document-level Cross Validation Results for Three Relations

Relation	Model	Precision	Recall	F-score
PPI	PCNN-human	65.6	60.1	62.0
	PCNN-both	**69.4**	**68.6**	**68.7**
MIRTAR	PCNN-human	76.2	70.5	71.2
	PCNN-both	**79.2**	**80.5**	**79.5**
PLOC	PCNN-human	**77.7**	51.6	59.5
	PCNN-both	73.7	**67.6**	**68.5**

Table 5.9: Instance-level Cross Validation Results for Three Relations

Relation	Model	Precision	Recall	F-score
PPI	PCNN-human	76.0	71.0	73.4
	PCNN-both	**76.5**	**80.4**	**78.4**
MIRTAR	PCNN-human	**91.9**	89.1	90.3
	PCNN-both	91.5	**91.5**	**91.3**
PLOC	PCNN-human	**90.8**	68.9	77.4
	PCNN-both	87.6	**75.4**	**80.7**

10-fold cross validation at document level [57], which achieved a 64.2 F-score on AIMed corpus.

5.10 Conclusion

In this chapter, we focused on improving distant supervision for biomedical relation extraction, and used small human-annotated data set with distantly labeled data to achieve state-of-the-art performance for protein-protein interaction. We first proposed three noise reduction techniques to improve the quality of distantly labeled data, resulting in 10-11 points increase in F-score over previous works for PPI and PLOC extraction. Next, we utilized the large generated data to train Convolutional Neural Networks (CNN) and were able to increase the F-scores by 5-12 points further with significantly higher recalls than the logistic regression models. This indicates that CNN models can learn from large data set better than simpler machine learning models, and have better generalization. To investigate the possibility to combine hand-labeled data for better performance, we trained the CNN models first with distantly-labeled data and then a small human-labeled data set. Experiment results showed that we can achieve 6.7-9.0 points higher in F-scores than the strong CNN models solely trained on small human-labeled data using 10-fold document-level cross validation. For PPI extraction, we were able to surpass the previous state-of-the-art using 10-fold document-level cross validation [57] by 4.5 points in F-score. By applying the techniques reported in this chapter on three biomedical relation extraction tasks, we show that the proposed methods can generalize across different biomedical relations and advance the research in the field of biomedical relation extraction.

Chapter 6

CONCLUSIONS

Biomedical relation extraction systems are important tools for biological research. With the ability to automatically extract structure information from large amount of text, it has assisted many biological tasks, including database curation, biological hypothesis generation and knowledge discovery. As the two main methods for biomedical relation extraction, rule-based and machine learning approaches require considerable human effort for pattern collection and training corpus annotation, respectively, which hinders the fast application of these methods to extract new types of relations. To address these issues, this dissertation adopted various linguistic knowledge to enable automatic construction of patterns for rule-based systems, and used distant supervision to label a large training set without human effort for machine learning systems. Furthermore, we combine distantly-labeled and human-labeled data to further improve the results. In this chapter, we summarize the contributions of this thesis and discuss a few future directions.

6.1 Thesis Summary and Contributions

1. We developed a high-performance rule-based system, miRTex [45], to extracts three kinds of miRNA-gene relations. The three relations, which are miRNA-target, miRNA-gene regulation and gene-miRNA regulation, are important for understanding miRNA's role in various biological processes and disease contexts. As it is difficult to collect every possible pattern to extract the relations, we follow the principles underlying the iXtractR system [69] to reduce the rule design effort. miRTex system uses a small set of trigger words and lexico-syntactic templates to generate rigorous patterns, which ensure high precision of the system. Referential relations and anaphora resolution were then used to improve recall. Experiment results showed that we can obtain high precision and recall on a in-house developed corpus and outperformed previous systems on a publicly available corpus [6]. miRTex has been applied to process the entire Medline abstracts and PMC full-length articles. The results are incorporated into the iTextMine system [1].

2. We described a general relation extraction framework and applied it to build a PTM extraction system. Although the miRTex system achieved high performance using a small set of triggers and lexico-syntactic templates, it is based on constituency parse and did not exploit the advantage of Extended Dependency Graph (EDG) [67]. Users still need to define the set of lexico-syntactic templates for each new relation, which could require considerable amount of human effort. This framework generalizes ideas from Lexicalized Tree-Adjoining Grammar [83] to dependency graph and utilizes EDG as the sentence representation [67]. It supports automatic generation of comprehensive rigorous patterns based on different verb types. This ensures high precision of the extracted results. The use of EDG abstracts away syntactic details and propagates extracted predicate arguments using is-a, member-collection and part-of relations, to increase recall. The input required by the framework is only a set of trigger specifications.

To illustrate the application of the framework, we applied it to develop an extraction system for five types of post-translational modification (PTM), i.e., acetylation, glycosylation, ubiquitination, sumoylation and neddylation. PTMs play critical roles in regulating many cellular processes, and disruption of the modification patterns has been linked to diseases. Evaluation results on the BioNLP 2011 EPI task and an in-house developed corpus showed that the system achieved high performance for each PTM and outperformed the best system in BioNLP 2011 EPI task [7], while it was implemented as a general extraction system for all the five types of PTM. This demonstrates that the framework can be used to extract a wide range of biomedical

[1] https://research.bioinformatics.udel.edu/itextmine

relations which are described using predicate-argument structures.

3. We improved distantly supervised biomedical relation extraction using noise reduction techniques [46], and achieved state-of-the-art for supervised biomedical relation extraction using transfer learning. First, we proposed three heuristics based on the notion of proximity, trigger word and confidence of patterns to leverage lexical and syntactic information to reduce noise in the distantly labeled data. Experiments on three relation extraction tasks, protein-protein interaction, miRNA-gene regulation and protein-localization event showed that the proposed methods can improve the F-scores by a large margin over the baselines and previous works. To better utilize the large dataset, we then trained Convolutional Neural Networks (CNN) on the generated data, and achieved better performance than logistic regression methods.

As small amount of human-labeled data is available for many biomedical relation extraction tasks, we investigated the possibility to use the existing small amount of annotated data to improve powerful techniques which require large amount of annotated data. We trained CNN models first with distantly-labeled data, and then with a small amount of human-labeled data. Compared to the CNN models solely trained on small human-labeled data, the models trained using both data achieved significantly better F-scores on all the three tasks. Specifically, for protein-protein interaction, a well-studied biomedical relation extraction task, we were able to surpass previous state-of-the-art [57] with supervised learning and achieve 68.7 F-score with 69.4 precision and 68.6 recall. To the best of our knowledge, this is the first work that combines distant supervision and human-labeled data with a two-step training process for biomedical relation extraction. The proposed method is not specific to any relation and can be applied to improve the current state-of-the-art in supervised biomedical relation extraction.

6.2 Future work

There are a few possible directions for future work in addition to the research reported in this thesis.

Application of the general relation extraction framework to extract new types of relations. As users only need to specify a set of trigger specifications as input to the general relation extraction framework, it requires much less effort to apply the framework for other types of relations. An immediate possibility is to extract other types of PTMs, e.g., methylation and hydroxylation. Any other relations described by predicate-argument structures, e.g., protein-protein interaction and drug-drug interaction, are good candidates.

Improve noise reduction heuristics for distantly supervised biomedical relation extraction. In Chapter 5, we introduced two heuristics to reduce noise in distantly labeled data based on the notions of trigger words and high-confidence patterns. As the trigger words and patterns are automatically mined from distantly labeled data, one possible way to improve them is to manually inspect them and remove incorrectly mined triggers and patterns. Another possibility is to incorporate biological ontology for the trigger words, and compare results of using low-level specific and more general trigger words.

Improve state-of-the-art for extraction of other biomedical relations using distant supervision and human-labeled data. Compared to models solely trained on small amount of human-labeled data, transfer learning was able to improve the F-scores by a large margin for the three tasks used in Chapter 5. It is promising to apply this paradigm to improve performance for other relations where there is a database and large amount of text available for distant supervision and small amount of human-annotated data. Such relations include drug-drug interaction [82] and gene-disease relation [11].

Large-scale processing of Medline abstracts and PMC full-length articles. The ultimate goal of developing a biomedical relation extraction system is to extract structural information from large amount of text and use the information to support biological research. Medline abstracts and PMC full-length articles are two large sources of biomedical text and the extracted information from them would be of great use to biology researchers. We can apply the developed tools to process all the Medline abstracts and PMC full-length articles and provide a web interface for querying and downloading the results.

BIBLIOGRAPHY

[1] Antti Airola, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinformatics*, 9 Suppl 11:S2, November 2008.

[2] Victor Ambros, Bonnie Bartel, David P Bartel, Christopher B Burge, James C Carrington, Xuemei Chen, Gideon Dreyfuss, Sean R Eddy, Sam Griffiths-Jones, Mhairi Marshall, Marjori Matzke, Gary Ruvkun, and Thomas Tuschl. A uniform system for microRNA annotation. *RNA*, 9(3):277–279, March 2003.

[3] Gabor Angeli, Julie Tibshirani, Jean Wu, and Christopher D Manning. Combining distant and partial supervision for relation extraction. In *EMNLP*, pages 1556–1567, 2014.

[4] Cecilia N Arighi, Cathy H Wu, Kevin B Cohen, Lynette Hirschman, Martin Krallinger, Alfonso Valencia, Zhiyong Lu, John W Wilbur, and Thomas C Wiegers. BioCreative-IV virtual issue. *Database*, 2014, May 2014.

[5] Masaki Asada, Makoto Miwa, and Yutaka Sasaki. Extracting Drug-Drug interactions with attention CNNs. *BioNLP 2017*, pages 9–18, 2017.

[6] Shweta Bagewadi, Tamara Bobić, Martin Hofmann-Apitius, Juliane Fluck, and Roman Klinger. Detecting miRNA mentions and relations in biomedical literature. *F1000Res.*, 3, August 2014.

[7] Jari Björne, Filip Ginter, and Tapio Salakoski. University of turku in the BioNLP'11 shared task. *BMC Bioinformatics*, 13 Suppl 1(Suppl 11):S4, January 2012.

[8] Jari Björne and Tapio Salakoski. TEES 2.1: Automated annotation scheme learning in the BioNLP 2013 shared task. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 16–25, 2013.

[9] Tamara Bobić, Roman Klinger, Philippe Thomas, and Martin Hofmann-Apitius. Improving distantly supervised extraction of drug-drug and protein-protein interactions. In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, ROBUS-UNSUP '12, pages 35–43, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[10] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 1247–1250, New York, NY, USA, 2008. ACM.

[11] Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I Furlong. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics*, 16:55, February 2015.

[12] Quoc-Chinh Bui, David Campos, Erik van Mulligen, and Jan Kors. A fast rule-based approach for biomedical event extraction. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 104–108, 2013.

[13] Razvan Bunescu, Ruifang Ge, Rohit J Kate, Edward M Marcotte, Raymond J Mooney, Arun K Ramani, and Yuk Wah Wong. Comparative experiments on learning information extractors for proteins and their interactions. *Artif. Intell. Med.*, 33(2):139–155, February 2005.

[14] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, May 2011.

[15] Eugene Charniak. A maximum-entropy-inspired parser. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, pages 132–139, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[16] Eugene Charniak and Mark Johnson. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, volume 1 of *ACL '05*, pages 173–180, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[17] Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. How to train good word embeddings for biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 166–174, 2016.

[18] Rajesh Chowdhary, Jinfeng Zhang, and Jun S Liu. Bayesian inference of protein-protein interactions from biological literature. *Bioinformatics*, 25(12):1536–1542, June 2009.

[19] K Bretonnel Cohen, Karin Verspoor, Helen L Johnson, Chris Roeder, Philip V Ogren, William A Baumgartner, Jr, Elizabeth White, Hannah Tipney, and Lawrence Hunter. HIGH-PRECISION BIOLOGICAL EVENT EXTRACTION: EFFECTS OF SYSTEM AND OF DATA. *Comput. Intell.*, 27(4):681–701, November 2011.

[20] M Craven and J Kumlien. Constructing biological knowledge bases by extracting information from text sources. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, pages 77–86, 1999.

[21] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods.* Cambridge university press, 2000.

[22] Louise Deléger, Anne-Laure Ligozat, Cyril Grouin, Pierre Zweigenbaum, and Aurélie Névéol. Annotation of specialized corpora using a comprehensive entity and relation scheme. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC), Reykjavik: ELRA*, 2014.

[23] J Ding, D Berleant, D Nettleton, and E Wurtele. Mining MEDLINE: abstracts, sentences, or phrases? *Pac. Symp. Biocomput.*, pages 326–337, 2002.

[24] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June 2008.

[25] Katrin Fundel, Robert Küffner, and Ralf Zimmer. RelEx–relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371, February 2007.

[26] Sam Griffiths-Jones. The microRNA registry. *Nucleic Acids Res.*, 32(Database issue):D109–11, January 2004.

[27] Jörg Hakenberg, Conrad Plake, Ulf Leser, Harald Kirsch, and Dietrich Rebholz-schuhmann. LLL05 challenge: genic interaction extraction – identification . . . with alignments and finite state automata. In *IN PROC LEARNING LANGUAGE IN LOGIC WORKSHOP (LLL05) AT THE 22ND INT CONF ON MACHINE LEARNING*, 2005.

[28] S Hochreiter and J Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.

[29] P E Hodges, W E Payne, and J I Garrels. The yeast protein database (YPD): a curated proteome database for saccharomyces cerevisiae. *Nucleic Acids Res.*, 26(1):68–72, January 1998.

[30] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 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*, HLT '11, pages 541–550, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[31] Yu-Lun Hsieh, Yung-Chun Chang, Nai-Wen Chang, and Wen-Lian Hsu. Identifying protein-protein interactions in biomedical literature using recurrent neural networks with long Short-Term memory. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 240–245, 2017.

[32] Sheng-Da Hsu, Yu-Ting Tseng, Sirjana Shrestha, Yu-Ling Lin, Anas Khaleel, Chih-Hung Chou, Chao-Fang Chu, Hsi-Yuan Huang, Ching-Min Lin, Shu-Yi Ho, Ting-Yan Jian, Feng-Mao Lin, Tzu-Hao Chang, Shun-Long Weng, Kuang-Wen Liao, I-En Liao, Chun-Chi Liu, and Hsien-Da Huang. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res.*, 42(Database issue):D78–85, January 2014.

[33] Z Z Hu, M Narayanaswamy, K E Ravikumar, K Vijay-Shanker, and C H Wu. Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics*, 21(11):2759–2765, June 2005.

[34] Zhang-Zhi Hu, Inderjeet Mani, Vincent Hermoso, Hongfang Liu, and Cathy H Wu. iProLINK: an integrated protein resource for literature mining. *Comput. Biol. Chem.*, 28(5-6):409–416, December 2004.

[35] Lei Hua and Chanqin Quan. A shortest dependency path based convolutional neural network for Protein-Protein relation extraction. *Biomed Res. Int.*, 2016:8479587, July 2016.

[36] Ander Intxaurrondo, Mihai Surdeanu, Oier Lopez de Lacalle, and Eneko Agirre. Removing noisy mentions for distant supervision. *Procesamiento del Lenguaje Natural*, 51(0):41–48, September 2013.

[37] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. February 2015.

[38] J D Kim, Y Wang, and Y Yasunori. The genia event extraction shared task, 2013 edition-overview. *Proceedings of the BioNLP*, 2013.

[39] Jin-Dong Kim, Ngan Nguyen, Yue Wang, Jun'ichi Tsujii, Toshihisa Takagi, and Akinori Yonezawa. The genia event and protein coreference tasks of the BioNLP shared task 2011. *BMC Bioinformatics*, 13 Suppl 1(Suppl 11):S1, January 2012.

[40] Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, BioNLP '09, pages 1–9, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[41] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. December 2014.

[42] Andre Lamurias, Luka A Clarke, and Francisco M Couto. Extracting microRNA-gene relations from biomedical literature using distant supervision. *PLoS One*, 12(3):e0171929, March 2017.

[43] Robert Leaman and Graciela Gonzalez. BANNER: an executable survey of advances in biomedical named entity recognition. *Pac. Symp. Biocomput.*, pages 652–663, 2008.

[44] Y Lecun, L Bottou, Y Bengio, and P Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, November 1998.

[45] Gang Li, Karen E Ross, Cecilia N Arighi, Yifan Peng, Cathy H Wu, and K Vijay-Shanker. miRTex: A text mining system for miRNA-Gene relation extraction. *PLoS Comput. Biol.*, 11(9):e1004391, September 2015.

[46] Gang Li, Cathy H Wu, and K Vijay-Shanker. Noise reduction methods for distantly supervised biomedical relation extraction. In *Proceedings of the BioNLP 2017 workshop*, pages 184–193, 2017.

[47] L Li, Z Jiang, and D Huang. A general instance representation architecture for protein-protein interaction extraction. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 497–500, November 2014.

[48] Xiao Liu, Antoine Bordes, and Yves Grandvalet. Biomedical event extraction by multi-class classification of pairs of text entities. In *BioNLP Shared Task 2013 Workshop*, pages 45–49, 2013.

[49] A S M Ashique Mahmood, Tsung-Jung Wu, Raja Mazumder, and K Vijay-Shanker. DiMeX: A text mining system for Mutation-Disease association extraction. *PLoS One*, 11(4):e0152725, April 2016.

[50] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60, 2014.

[51] David McClosky and Eugene Charniak. Self-Training for biomedical parsing. In *Proceedings of ACL-08: HLT, Short Papers*, pages 101–104, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[52] Farrokh Mehryary, Kai Hakala, Suwisa Kaewphan, Jari Björne, Tapio Salakoski, and Filip Ginter. End-to-End system for bacteria habitat extraction. *BioNLP 2017*, pages 80–90, 2017.

[53] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In

C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[54] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 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: Volume 2 - Volume 2*, ACL '09, pages 1003–1011, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[55] Makoto Miwa, Rune Saetre, Jin-Dong Kim, and Jun'ichi Tsujii. Event extraction with complex event classification using rich features. *J. Bioinform. Comput. Biol.*, 8(1):131–146, February 2010.

[56] Makoto Miwa, Rune Saetre, Yusuke Miyao, and Jun'ichi Tsujii. Protein-protein interaction extraction by leveraging multiple kernels and parsers. *Int. J. Med. Inform.*, 78(12):e39–46, December 2009.

[57] Makoto Miwa, Rune Sætre, Yusuke Miyao, and Jun'ichi Tsujii. A rich feature vector for protein-protein interaction extraction from multiple corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 121–130, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[58] Haroon Naeem, Robert Küffner, Gergely Csaba, and Ralf Zimmer. miRSel: automated extraction of associations between microRNAs and genes from the biomedical literature. *BMC Bioinformatics*, 11:135, March 2010.

[59] M Narayanaswamy, K E Ravikumar, and K Vijay-Shanker. Beyond the clause: extraction of phosphorylation information from medline abstracts. *Bioinformatics*, 21 Suppl 1:i319–27, June 2005.

[60] Meenakshi Narayanaswamy, K E Ravikumar, and K Vijay-Shanker. A biological named entity recognizer. *Pac. Symp. Biocomput.*, pages 427–438, 2003.

[61] Tomoko Ohta, Sampo Pyysalo, Makoto Miwa, Jin-Dong Kim, and Jun'ichi Tsujii. Event extraction for post-translational modifications. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, BioNLP '10, pages 19–27, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[62] Tomoko Ohta, Sampo Pyysalo, and Jun'ichi Tsujii. Overview of the epigenetics and post-translational modifications (EPI) task of BioNLP shared task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, BioNLP Shared Task '11, pages 16–25, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[63] Sandra Orchard, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H Campbell, Gayatri Chavali, Carol Chen, Noemi del Toro, Margaret Duesbury, Marine Dumousseau, Eugenia Galeota, Ursula Hinz, Marta Iannuccelli, Sruthi Jagannathan, Rafael Jimenez, Jyoti Khadake, Astrid Lagreid, Luana Licata, Ruth C Lovering, Birgit Meldal, Anna N Melidoni, Mila Milagros, Daniele Peluso, Livia Perfetto, Pablo Porras, Arathi Raghunath, Sylvie Ricard-Blum, Bernd Roechert, Andre Stutz, Michael Tognolli, Kim van Roey, Gianni Cesareni, and Henning Hermjakob. The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, 42(Database issue):D358–63, January 2014.

[64] S J Pan and Q Yang. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359, October 2010.

[65] Y Peng, C O Tudor, M Torii, C H Wu, and K Vijay-Shanker. isimp: A sentence simplification system for biomedicail text. In *2012 IEEE International Conference on Bioinformatics and Biomedicine*, pages 1–6, October 2012.

[66] Yifan Peng, Cecilia Arighi, Cathy H Wu, and K Vijay-Shanker. BioC-compatible full-text passage detection for protein-protein interactions using extended dependency graph. *Database*, 2016, May 2016.

[67] Yifan Peng, Samir Gupta, Cathy H Wu, and K Vijay-Shanker. An extended dependency graph for relation extraction in biomedical texts. *ACL-IJCNLP 2015*, page 21, 2015.

[68] Yifan Peng and Zhiyong Lu. Deep learning for extracting protein-protein interactions from biomedical literature. In *BioNLP 2017*, pages 29–38, Stroudsburg, PA, USA, 2017. Association for Computational Linguistics.

[69] Yifan Peng, Manabu Torii, Cathy H Wu, and K Vijay-Shanker. A generalizable NLP framework for fast development of pattern-based biomedical relation extraction systems. *BMC Bioinformatics*, 15(1):285, August 2014.

[70] Maria Pershina, Bonan Min, Wei Xu, and Ralph Grishman. Infusion of labeled data into distant supervision for relation extraction. In *Proceedings of ACL*, 2014.

[71] M F Porter. An algorithm for suffix stripping. *Programmirovanie*, 14(3):130–137, 1980.

[72] Sampo Pyysalo, Antti Airola, Juho Heimonen, Jari Björne, Filip Ginter, and Tapio Salakoski. Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, 9 Suppl 3:S6, April 2008.

[73] Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8:50, February 2007.

[74] Sampo Pyysalo, Tomoko Ohta, Makoto Miwa, and Jun'ichi Tsujii. Towards exhaustive protein modification event extraction. In *Proceedings of BioNLP 2011 Workshop*, BioNLP '11, pages 114–123, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[75] Chanqin Quan, Lei Hua, Xiao Sun, and Wenjun Bai. Multichannel convolutional neural network for biological relation extraction. *Biomed Res. Int.*, 2016:1850404, December 2016.

[76] Sudha Rao, Daniel Marcu, Kevin Knight, and Hal Daumé, III. Biomedical event extraction using abstract meaning representation. *BioNLP 2017*, pages 126–135, 2017.

[77] Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In José Luis Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 6323 of *Lecture Notes in Computer Science*, pages 148–163. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[78] Roland Roller, Eneko Agirre, Aitor Soroa, and Mark Stevenson. Improving distant supervision using inference learning. In *53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 273–278, September 2015.

[79] Roland Roller and Mark Stevenson. Making the most of limited training data using distant supervision. *ACL-IJCNLP 2015*, page 12, 2015.

[80] Benjamin Roth, Tassilo Barth, Michael Wiegand, and Dietrich Klakow. A survey of noise reduction methods for distant supervision. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 73–78. ACM, October 2013.

[81] S K Sahu, A Anand, K Oruganty, M Gattu arXiv preprint arXiv:1606.09370, and 2016. Relation extraction from clinical texts using domain invariant convolutional neural network. *arxiv.org*, 2016.

[82] Sunil Kumar Sahu and Ashish Anand. Drug-Drug interaction extraction from biomedical text using long short term memory network. January 2017.

[83] Yves Schabes. Stochastic lexicalized tree-adjoining grammars. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, volume 23 of *COLING '92*, pages 425–432, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics.

[84] Isabel Segura-Bedmar, Paloma Martınez, and Marıa Herrero-Zazo. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). *Proceedings of Semeval*, pages 341–350, 2013.

[85] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, January 2014.

[86] Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 455–465, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[87] Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. Reducing wrong labels in distant supervision for relation extraction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 721–729, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[88] P Thomas, I Solt, R Klinger, and U Leser. Learning to extract protein–protein interactions using distant supervision. *ROBUS 2011*, 2011.

[89] M Torii, C N Arighi, G Li, Q Wang, C H Wu, and K Vijay-Shanker. RLIMS-P 2.0: A generalizable Rule-Based information extraction system for literature mining of protein phosphorylation information. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, PP(99):1–1, 2014.

[90] Catalina O Tudor, Cecilia N Arighi, Qinghua Wang, Cathy H Wu, and K Vijay-Shanker. The eFIP system for text mining of protein interaction networks of phosphorylated proteins. *Database*, 2012:bas044, December 2012.

[91] UniProt Consortium. Activities at the universal protein resource (UniProt). *Nucleic Acids Res.*, 42(Database issue):D191–8, January 2014.

[92] Sofie Van Landeghem, Filip Ginter, Yves Van de Peer, and Tapio Salakoski. EVEX: A pubmed-scale resource for homology-based generalization of text mining predictions. In *Proceedings of BioNLP 2011 Workshop*, BioNLP '11, pages 28–37, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[93] Ioannis S Vlachos, Maria D Paraskevopoulou, Dimitra Karagkouni, Georgios Georgakilas, Thanasis Vergoulis, Ilias Kanellos, Ioannis-Laertis Anastasopoulos, Sofia Maniou, Konstantina Karathanou, Despina Kalfakakou, Athanasios Fevgas, Theodore Dalamagas, and Artemis G Hatzigeorgiou. DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Res.*, November 2014.

[94] Mang Wang, Hong Xia, Dongdong Sun, Zhaoxiong Chen, Minghui Wang, and Ao Li. Literature mining of protein phosphorylation using dependency parse trees. *Methods*, 67(3):386–393, June 2014.

[95] Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res.*, 41(Web Server issue):W518–22, July 2013.

[96] Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. GNormPlus: An integrative approach for tagging genes, gene families, and protein domains. *Biomed Res. Int.*, 2015:918710, August 2015.

[97] Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu. Overview of the BioCreative V chemical disease relation (CDR) task. In *Proceedings of the fifth BioCreative challenge evaluation workshop*, pages 154–166, 2015.

[98] Kun Xu, Yansong Feng, Songfang Huang, and Dongyan Zhao. Semantic relation classification via convolutional neural networks with simple negative sampling. June 2015.

[99] Wei Xu, Raphael Hoffmann, Le Zhao, and Ralph Grishman. Filling knowledge base gaps for distant supervision of relation extraction. In *ACL (2)*, pages 665–670. anthology.aclweb.org, 2013.

[100] Yun Xu, Da Teng, and Yiming Lei. MinePhos: a literature mining system for protein phoshphorylation information extraction. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 9(1):311–315, January 2012.

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

[102] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao, and Others. Relation classification via convolutional deep neural network. In *COLING*, pages 2335–2344, 2014.

[103] Zhehuan Zhao, Zhihao Yang, Ling Luo, Hongfei Lin, and Jian Wang. Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. *Bioinformatics*, 32(22):3444–3453, November 2016.

[104] Wu Zheng and Catherine Blake. Using distant supervised learning to identify protein subcellular localizations from full-text scientific articles. *J. Biomed. Inform.*, July 2015.

Appendix A
PTM TRIGGER WORDS

Here we list the PTM trigger words in their base verb form used in the PTM extraction system in Chapter 4.

Acetylation: acetylate

Glycosylation: glycosylate, galatosylate, fucosylate, galactosylate, glucosylate, glycanate, mannosylate, glcnacylate

Ubiquitination: ubiquitinate, ubiquitylate, ubiquitinylate

Neddylation: neddylate

Phosphorylation: phosphorylate

Generic modification: modify, conjugate, attach, link, transfer, target

Self-theme: substrate, target

Self-agent: ligase, enzyme, kinase

Regulation: induce, promote, mediate, enhance, catalyze, affect, modulate, activate, impair, increase, decrease, regulate

Appendix B

AUTOMATICALLY MINED TRIGGER STEMS

Here we list the top 100 trigger stems for the three relations used in experiments with distant supervision in Chapter 5.

- PPI trigger stems: interact, bind, associ, phosphoryl, recruit, activ, coloc, coimmunoprecipit, co-immunoprecipit, regul, co-loc, know, inhibit, form, link, requir, contain, cooper, stabil, function, target, deacetyl, sequest, act, dephosphoryl, call, co-precipit, complex, recogn, express, coprecipit, local, dissoci, compos, induc, have, consist, ubiquitin, includ, mediat, deloc, bridg, connect, assembl, compet, downregul, involv, degrad, cleav, releas, displac, enhanc, compris, promot, repress, identifi, direct, destabil, encod, co-purifi, find, cosedi, control, copurifi, pull, tether, combin, display, methyl, bring, inactiv, co-express, remov, liber, signal, coupl, hold, suppress, appear, stimul, conjug, transport, term, play, protect, ubiquityl, work, describ, detect, acetyl, retain, down-regul, reloc, anchor, implic, depend, maintain, attach, replac, increas

- MIRGENE trigger stems: target, regul, inhibit, downregul, suppress, repress, down-regul, correl, induc, promot, modul, upregul, control, decreas, activ, upregul, bind, mediat, involv, interact, associ, increas, play, act, have, function, affect, silenc, enhanc, stimul, attenu, link, form, identifi, dysregul, restrict, contribut, express, encod, connect, transactiv, exert, result, prevent, restor, mimic, combin, diminish, reduc, determin, remov, find, show, trigger, compos, observ, co-regul, caus, compris, antagon, use, methyl, cooper, recruit, particip, alter, parallel, exhibit, co-express, share, counterregul, overexpress, predict, keep, protect, converg, derepress, valid, improv, co-repress, requir, facilit, consist, respond, measur, quantifi, establish, reintroduc, synerg, detect, maintain, impact, sensit, degrad, down-modul, outfox, acceler, over-express, reliev, knock

- PLOC trigger stems: local, transloc, express, associ, interact, detect, coloc, find, co-loc, target, recruit, bind, shuttl, accumul, distribut, transport, regul, observ, contain, sequest, induc, function, stain, releas, localis, anchor, resid, exclud, have, increas, link, reloc, activ, inhibit, enter, form, recycl, retain, prevent, trap, dissoci, concentr, import, enrich, remain, identifi, phosphoryl, decreas, mediat, export, demonstr, act, re-loc, deliv, purifi, transfer, traffic, move, exist, redistribut, lack, occur, reveal, overexpress, migrat, maintain, involv, up-regul, reduc, cleav, recogn, carri, control, co-express, heterodimer, share, tight-junction-associ, displac, escort, play, degrad, be, deacetyl, drive, compris, appear, tether, deposit, complex, reach, direct, see, coexist, convert, solubil, label, coexpress, modul, remov, build

Appendix C

PARAMETERS OF CONVOLUTIONAL MODELS

C.1 CNN Models Training Solely on Distantly Labeled Data

1. MC-CNN Model

   (a) PPI: learning rate: 0.003, epoch: 10, l2 regularization weight: 0.0001, learning rate decay step: 200, decay rate: 0.9, conv dropout: 0.5, dense dropout: 0.2

   (b) MIRTAR: learning rate: 0.001, epoch: 10, l2 regularization weight: 0.0001, learning rate decay step: 200, decay rate: 0.9, conv dropout: 0.5, dense dropout: 0.2

   (c) PLOC: learning rate: 0.001, epoch: 10, l2 regularization weight: 0.0001, learning rate decay step: 50, decay rate: 0.9, conv dropout: 0.5, dense dropout: 0.2

2. PCNN Model

   (a) PPI: learning rate: 0.001, epoch: 10, l2 regularization weight: 0.0001, learning rate decay step: 200, decay rate: 0.9, conv dropout: 0.5, dense dropout: 0.2

   (b) MIRTAR: learning rate: 0.001, epoch: 10, l2 regularization weight: 0.0001, learning rate decay step: 200, decay rate: 0.9, conv dropout: 0.5, dense dropout: 0.2

   (c) PLOC: learning rate: 0.001, epoch: 10, l2 regularization weight: 0.0001, learning rate decay step: 200, decay rate: 0.9, conv dropout: 0.5, dense dropout: 0.2

C.2 PCNN Models Trained on Human-labeled Data (10-fold instance-level CV)

1. PPI: learning rate: 0.003, epoch: 200, l2 regularization weight: 0.0001, learning rate decay step: 200, decay rate: 0.95, conv dropout: 0.2, dense dropout: 0.2

2. MIRTAR: learning rate: 0.003, epoch: 200, l2 regularization weight: 0.0001, learning rate decay step: 200, decay rate: 0.95, conv dropout: 0.2, dense dropout: 0.2

3. PLOC: learning rate: 0.003, epoch: 200, l2 regularization weight: 0.0001, learning rate decay step: 200, decay rate: 0.9, conv dropout: 0.2, dense dropout: 0.2

C.3 PCNN Models Trained on Both Distantly-labeled and Human-labeled Data (10-fold instance-level CV)

1. PPI: learning rate: 0.003, epoch: 100, l2 regularization weight: 0.0001, learning rate decay step: 50, decay rate: 0.9, conv dropout: 0.2, dense dropout: 0.2

2. MIRTAR: learning rate: 0.003, epoch: 100, l2 regularization weight: 0.0001, learning rate decay step: 50, decay rate: 0.95, conv dropout: 0.5, dense dropout: 0.2

3. PLOC: learning rate: 0.01, epoch: 100, l2 regularization weight: 0.0001, learning rate decay step: 200, decay rate: 0.99, conv dropout: 0.2, dense dropout: 0.2

C.4 PCNN Models Trained on Human-labeled Data (10-fold document-level CV)

1. PPI: learning rate: 0.001, epoch: 200, l2 regularization weight: 0.0001, learning rate decay step: 100, decay rate: 0.95, conv dropout: 0.4, dense dropout: 0.3

2. MIRTAR: learning rate: 0.001, epoch: 200, l2 regularization weight: 0.0001, learning rate decay step: 200, decay rate: 0.95, conv dropout: 0.4, dense dropout: 0.2

3. PLOC: learning rate: 0.0001, epoch: 200, l2 regularization weight: 0.0001, learning rate decay step: 200, decay rate: 0.9, conv dropout: 0.2, dense dropout: 0.2

C.5 PCNN Models Trained on Both Distantly-labeled and Human-labeled Data (10-fold document-level CV)

1. PPI: learning rate: 0.0006, epoch: 100, l2 regularization weight: 0.0001, learning rate decay step: 100, decay rate: 0.9, conv dropout: 0.2, dense dropout: 0.2

2. MIRTAR: learning rate: 0.0003, epoch: 100, l2 regularization weight: 0.0001, learning rate decay step: 50, decay rate: 0.9, conv dropout: 0.2, dense dropout: 0.2

3. PLOC: learning rate: 0.0003, epoch: 100, l2 regularization weight: 0.0001, learning rate decay step: 50, decay rate: 0.9, conv dropout: 0.2, dense dropout: 0.2