Relation Extraction from Biomedical Literature with Minimal Supervision and Grouping Strategy

Mengwen Liu, Yuan Ling, Yuan An and Xiaohua Hu
College of Computing and Informatics
Drexel University
{ml943,yl638,ya45,xh29@drexel.edu}

Alan Yagoda and Rick Misra
Elsevier Inc.
{a.yagoda, r.misra}@elsevier.com

***Abstract*—We develop a novel distant supervised model that integrates the results from open information extraction techniques to perform relation extraction task from biomedical literature. Unlike state-of-the-art models for relation extraction in biomedical domain which are mainly based on supervised methods, our approach does not require manually-labeled instances. In addition, our model incorporates a grouping strategy to take into consideration the coordinating structure among entities co-occurred in one sentence. We apply our approach to extract gene expression relationship between genes and brain regions from literature. Results show that our methods can achieve promising performance over baselines of Transductive Support Vector Machine and with non-grouping strategy.**

***Keywords*—Relation Extraction; Distant Supervision; Grouping Strategy**

I. Introduction

The biomedical community has made extensive use of scientific literature to discover facts about various types of biomedical entities such as genes, proteins, drugs, etc. Semantic relation extraction between biological entities is a fundamental task for biological knowledge graph construction, which supports automated hypothesis generation and knowledge discovery. It also benefits many biomedical studies, such as gene-disease interactions, protein-protein interactions, etc.

Supervised approaches for semantic relation extraction can achieve high precision and recall since they rely on a sufficiently large amount of training data. However, their generalizability is limited when manually labeled examples are expensive to obtain. Therefore, recent work has been focused on training relation extractors with minimal human supervision. Among them, approaches based on distant supervision have started to attract attention. Those methods use a relation database to generate a number of sentences containing entity pairs as noisy training examples, which consequently do not require expensive human labeled examples.

It is generally believed that distant supervision approaches would benefit relation extraction in general domain, however, such approaches in biomedical domain are not fully explored yet. There are two possible reasons. One is that the main source of knowledge of distant supervision approaches for general domain is Freebase[1], which does not contain much specialized biomedical knowledge. Second, the distant learning models developed so far assume that each entity instance is independent. However, this assumption is often violated in biomedical domain. Take the following sentence annotated with two genes and one brain region as an example: *In the cells from the brain/BRAIN of 11.5 days fetal mice, Nestin/GENE and MAP-2/GENE were strongly expressed.* In the traditional view of relation extraction, this sentence contains two entity pairs, (brain, Nestin) and (brain, MAP-2), each of which is represented with independent types of features. Those two genes, Nestin and MAP-2, however, are parallel with each other; so both of them have the *geneExpression* relationship with brain. Ignoring such coordinating structure will undermine the performance of the distant supervised models.

In dealing with these challenges, this paper presents a novel relation extraction approach that makes the following contributions:

1) We develop a distant supervised model that incorporates the results from existing open information extraction techniques to perform relation extraction task in biomedical domain without using any hand-labeled examples;

2) We design a novel grouping strategy to capture the coordinating structure among entities;

3) We apply our approach to extract the gene expression relationship between genes and brain regions from biomedical literature. The results achieved by our model are more accurate than that by semi-supervised and non-grouping baselines.

The paper is organized as follows: Section II summarizes the related work of relation extraction. Section III presents the details of our proposed method. Section IV describes the experimental settings. Section V discusses the results, followed by conclusions and future work in Section VI.

II. Related Work

Relation extraction problem can be regarded as a multi-class classification issue with each type of relation as one class. The state-of-the-art supervised approaches for tackling relation extraction fall into two categories: feature-based approach [1] [2] and kernel-based approach [3] [4]. Due to the cost of obtaining training examples in supervised approaches, recent studies have shifted to focus on reducing the amount of human annotations to perform relation extraction task. Two lines of approaches are proposed under this context: semi-supervised approach and distant supervised approach. An example of relation extraction in semi-supervised setting is from Erkan [5]. The author trained transductive Support Vector Machines with

[1] http://www.freebase.com

978-1-4799-5669-2/14/$31.00 ©2014 IEEE

limited amount of labeled data, which achieved better results in comparison with a fully supervised way. As for the distant supervision setting, crowdsourcing has played an important role because of the blossom of the social web. Large knowledge databases have been emerging such as Wikipedia, Freebase, etc. With such freely available knowledge, it becomes possible to use a large set of known entity pairs to generate training data with target relation types. Examples of relation extraction using distant supervision approach include [6]–[8].

However, while approaches with minimal supervision have been widely applied in general domain, it has been rarely studied in biomedical domain. The most similar work was conducted by Thomas et al. [9]. This study utilized distant supervision approach to perform protein-protein interaction extraction from biomedical literature. However, the authors used heuristic rules to reason the relationships between an entity pair at sentence level; whereas our work concerns a statistical model to infer the relation types.

III. Methods

This section provides an overview of the major steps in our method, which is summarized in Fig. 1. Provided with a large text corpus, we first annotate sentences with entities of interest. Then, parse trees of those sentences are generated from which our proposed rules are applied to identify entities with coordinating structure. After using features to represent each pair of entities, we train a distant supervised model to decode the relationships between each entity pair. Implementation details are presented from Section III.A to Section III.E.

A. Entity Annotation

The first stage of the architecture is to annotate sentences with entities of interests. As our goal is to extract the *gene expression* relationships between genes and brain regions, we utilize the following resources and toolkit to find those entities: 1) the Brain Regions Hierarchy from Neuroscience Information Framework (NIF)[2]; 2) a comprehensive brain dictionary provided by Elsevier[3]; and 3) the Penn BioTagger[4].

We first select sentences containing brain regions and genes individually. We use string matching method to match sentences with brain mentions, and use the Penn BioTagger to annotate sentences containing genes. After that, sentences containing both brain regions and genes are retained. It is noted that there exist multiple entities occurred in one sentence. If a sentence contains m different brains and n different genes, there would be $m \times n$ different pairs of entities. We will show how our proposed grouping strategy will be applied to find entities with coordinating structure and generate corresponding features in section III.B and III.C, respectively.

B. Grouping Strategy

Once a set of sentences containing entity pairs are gathered, features representing those pairs of entities will be extracted for training a distant supervised model. Current approach of transforming a set of sentences into a set of feature representations for each entity pair is based on the assumption that entity pairs are independent. However, this is often violated in biomedical domain. Take the following sentence as an example:

> *In the AD brain, decreased BDNF protein levels were reported in hippocampus, entorhinal cortex, and temporal neocortex, while no changes were observed in areas less affected by the disease, such as the frontal, parietal, and cerebellar cortices.*

In this sentence, *BDNF* is tagged as a gene, and *hippocampus, entorhinal cortex, and temporal neocortex, frontal, parietal, and cerebellar cortices* are tagged as brain regions. Therefore, the sentence has six pairs of entities, (BDNF, hippocampus), (BDNF, entorhinal cortex), (BDNF, temporal neocortex), (BDNF, frontal cortices), (BDNF, parietal cortices), and (BDNF, cerebellar cortices). Then, six feature vectors are generated independently for each of the entity pairs (see Section III.C for details of feature representation).

However, from a linguistic perspective, the first and second three brain regions are parallel with each other. Therefore, the relationships between the two sets of three brain regions and BDNF should be the same. Indeed, entities consecutively occurred in one sentence are quite common in biomedical literature — almost one third of the sentences in our experimental dataset contain interdependent brain regions and genes.

In tackling this problem, we first utilize the Stanford parser[5] to generate parse trees for all the sentences. We then develop heuristics rules, which are summarized in Table I, to find parallel entities from each sentence. In this study, we only search for consecutive entities containing the conjunction "and", because over 90% of the entities with coordinating structure in our experimental dataset are connected by "and". Examples of rules include consecutive nouns or noun phrases. We infer that if any of the noun phrases in the chunks is annotated with a brain region or gene, any of the nouns in the chunks is a brain region or a gene. Therefore, those rules are also beneficial for finding entities that are not either included in the dictionary or identified by the tagger. Fig. 2 and 3 show the entity pair instances before and after grouping consecutive entities, respectively. Before identifying the coordinating structure, we need to deal with six independent entity pairs. After grouping entities, the six brain regions are categorized into two groups. Those two "brain combinations" are paired with the gene to construct two entity pairs.

C. Feature Generation

We used standard lexical and syntactic features introduced in the literature [6] to represent how two entities are related in a sentence. Lexical features describe specific tokens between and surrounding the two entities in the sentence in which they appear. Examples of lexical features include: 1) Bags-of-words of the two entities; 2) Bags-of-words of the sequence of words between the two entities; 3) A window of three words to the left of Entity 1; and 4) A window of three words to the right of Entity 2. Syntactic features are generated based on syntax of context surrounding the two entities. An example of syntactic features is the shortest dependency path between those two entities.

[2]http://neurolex.org/wiki/Brain_Regions_Hierarchy

[3]http://www.elsevier.com/

[4]http://www.seas.upenn.edu/ strctlrn/BioTagger/BioTagger.html

[5]http://nlp.stanford.edu/software/lex-parser.shtml

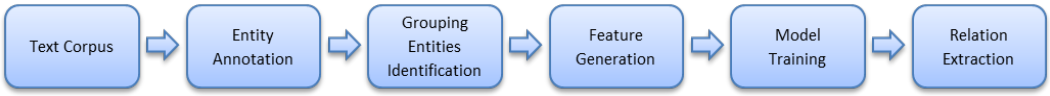

Fig. 1: Overall Architecture

TABLE I: Rules for Grouping Entities

Category	Examples of patterns	Examples of phrases
Two consecutive nouns	(NP (NN striatum) (CC and) (NN nucleus))	Striatum and nucleus
	(NP (DT the) (NN cortex) (CC and) (NN hippocampus))	The cortex and hippocampus
Two consecutive noun phrases	(NP (NP (JJ ventral)) (CC and) (NP (NN dorsal) (NN hippocampus)))	Ventral and dorsal hippocampus
	(NP (NP (NN AChE)) (CC and) (NP (JJ antioxidant) (NNS enzymes)))	AchE and antioxidant enzymes
Multiple consecutive nouns	(NP (NN cortex) (, ,) (NN hippocampus) (CC and) (NN brain))	The cerebral cortex, hippocampus and striatum
	(NP (NN GABA) (, ,) (NN aspartate) (, ,) (NN glutamate) (, ,) (NN glycine) (CC and) (NN alanine))	The GABA, aspartate, glutamate, glycine, and alanine
Multiple consecutive noun phrases	(NP (NP (DT the) (JJ medial) (JJ prefrontal) (NN cortex)) (, ,) (NP (NN nucleus) (NNS accumbens)) (, ,) (NP (JJ medial) (NN corpus) (NN striatum)) (, ,) (CC and) (NP (NN hippocampus))))	The medial prefrontal cortex, nucleus accumbens, medial corpus striatum, and hippocampus
	(NP (NP (DT the) (NN hippocampus)) (, ,) (NP (NN choroid) (NN plexus)) (, ,) (CC and) (NP (JJ frontal) (NN cortex))))	The hippocampus, choroid plexus, and frontal cortex

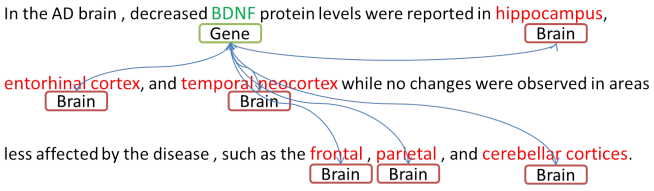

Fig. 2: Before Grouping Entities

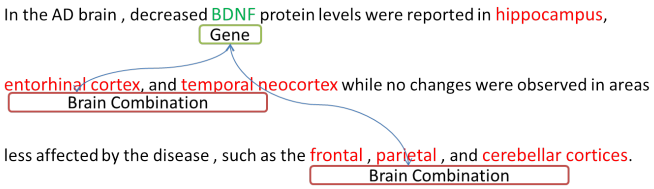

Fig. 3: After Grouping Entities

Based on our grouping strategy, a set of entity pairs in which one entity in paired with a group of entities are represented with some of the same lexical features (i.e. bags-of-words) and syntactic features. Take the three entity pairs (BDNF, hippocampus), (BDNF, entorhinal cortex), and (BDNF, temporal neocortex) extracted from the sentence mentioned in Section III.B for example. Before grouping those brain regions into one group, the bags-of-word features between those two entities are "protein levels were reported in", "protein levels were reported in hippocampus", and "protein levels were reported in hippocampus, entorhinal cortex, and", respectively; while after grouping those brain entities, the features become "protein levels were reported in", which is taken from the sequence between the gene BDNF and the "brain combination".

D. Model Training

As obtaining a lot of training examples for supervised approach is expensive, we aim at developing a distant learning model to make up the shortage of labeled examples. Our approach is motivated by Riedel et al. [7] and Hoffmann et al. [8]. The key idea is to use facts about relations between entities from a knowledgebase such as Freebase, to construct a large number of noisy training examples. Those facts instead of annotated examples are used as the source of supervision.

In our scenario, we utilize the results from SemRep [10], an existing open information extraction toolkit. SemRep incorporates the knowledge from the UMLS Semantic Network, which provides a comprehensive set of semantic relationships between biomedical concepts. Inspired by Riedel et al. [7] and Hoffmann et al. [8], we make the following assumption: as long as there is a relation type (i.e. *geneExpression*) between an entity pair stored in a knowldgebase (i.e. UMLS Semantic Network), at least one occurrence of the entity pair will express the relation.

Based on the "at-least-one" assumption, we design an undirected graphical model for every pair of biomedical entities. An example of a graphical model for an entity pair (hippocampus, BDNF) along with its three occurrences is shown in Fig. 4 Those two entities have *geneExpression*, *regulation*, and *noRelation* relationships observed from the UMLS Semantic Network. As our goal is to extract the *geneExpression* relationship, we assume sentences do not have such type of relation fall into the category of *otherRelation*. According to the assumption, at least one from the three occurrences is expected to have the relationship observed from the knowledge base. In this example, the first sentence is a positive sentence for the *geneExpression* relationship for the (hippocampus, BDNF) pair; while the last two are positive sentences for the *otherRelation* relationship.

Formally, we define the graphical model as follows. Given an entity dictionary E and a set of relation types R, i.e. $R = \{geneExpression, otherRelation\}$, for each

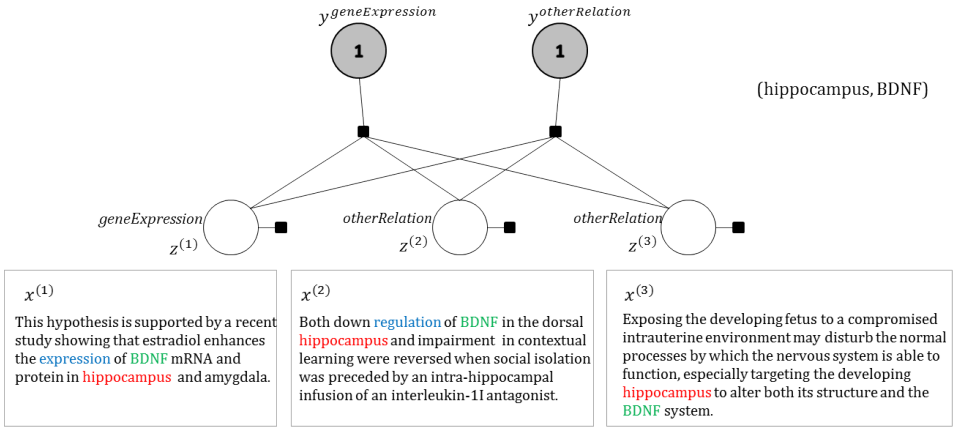

Fig. 4: An example of graphical model for an entity pair (hippocampus, BDNF)

entity pair, we define the model consisting of three random variables:

$\boldsymbol{x}$: ranging over all the natural language sentences. Each sentence is characterized by a feature vector. We use a set of feature vectors $\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, ..., \boldsymbol{x}^{(m)}\}$ representing sentences containing the particular pair of entities $(e_1, e_2) \in E \times E$.

$\boldsymbol{y}$: a set of two Boolean variable $\{y^{geneExpression}, y^{otherRelation}\}$. The value of y^r where $r \in R$ indicates whether $r(e_1, e_2)$ is true.

$\boldsymbol{z}$: a set of variables $\{z^{(1)}, z^{(2)}, ..., z^{(m)}\}$, representing the relation types for those corresponding m sentences. The value of $z^{(i)}(i = 1...m)$ indicates the relation type expressed in sentence $x^{(i)}$. Currently, we have two relation types, so $z^{(i)} \in \{geneExpression, otherRelation\}$.

The model defines a conditional probability for extraction as written in Eq. (1):

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) &= p(\boldsymbol{y}, \boldsymbol{z} | \boldsymbol{x}; \theta) \\ &= \frac{1}{Z_x} \Pi_{r \in R} \Phi^{corpus}(y^r, \boldsymbol{z}) \Pi_{i=1}^{m} \Phi^{sentence}(z^{(i)}, \boldsymbol{x}^{(i)}) \end{aligned} \tag{1}$$

where m is the number of sentences containing the particular entity pair; and Z_x is the normalization term which ensures that the sum over all possible values of random variables should be one.

The model introduces two factors, $\Phi^{sentence}$ and Φ^{corpus}. $\Phi^{sentence}$ represents a log-linear model that predicts the highest probability of the relation $z^{(i)}$ of the entity pair in a given sentence $x^{(i)}$. The Φ^{corpus} factor is defined in Eq. (2):

$$\begin{aligned} \Phi^{sentence}(z^{(i)}, \boldsymbol{x}^{(i)}) &= p(z^{(i)} | \boldsymbol{x}^{(i)}; \theta) \\ &= exp(\theta^T \cdot \phi(z^{(i)}, \boldsymbol{x}^{(i)})) \end{aligned} \tag{2}$$

where $\phi(z^{(i)}, \boldsymbol{x}^{(i)})$ defines the feature function (see section III.C) for each of the sentences. The parameter $\boldsymbol{\theta}$ determines the relation type for each sentence.

Φ^{corpus} presents the aggregated type of relation based on the "at-least-one" assumption. A relation type $r(e_1, e2)$ fires if and only if one of the sentences containing the entity pair expresses that relation. The Φ^{corpus} factor is defined in Eq.(3):

$$\Phi^{corpus}(y^r, z) = \begin{cases} 1 \ if \ y^r = 1 \wedge \exists i \ z_i = r \\ 0 \ otherwise \end{cases}, \quad for \ i = 1...m \tag{3}$$

where m is the number of occurrences in the dataset for a particular entity pair.

When training the model, we obtain training examples by matching the $\boldsymbol{y}$ variables against the results from SemRep which extracts existing relations from the UMLS knowledge base. We treat the $\boldsymbol{z}$ variables as hidden variables that can take any value, as long as they produce the correct aggregate predictions against $\boldsymbol{y}$. Subsequently, the parameter $\boldsymbol{\theta}$ will be estimated using the training examples and knowledgebase. We adopt the same inference technique and perceptron-style parameter estimation method in Hoffmann et al's work [8].

E. Relation Extraction

The resultant distant supervised model can be used to predict the relation types at both sentential-level and corpus-level given an entity pair along with a set of sentences in which it appears. The model first predicts the relation types for the entity pair corresponding to each of its evidences. Then, the predictions of relation types at sentential-level are aggregated to approximate the relation types for the entity pair at corpus-level. As long as one of the sentences is classified into the *geneExpression* category, the *geneExpression* relation is returned at corpus-level; otherwise, only *otherRelation* is returned.

IV. Experiments

A. Data Set

Our experimental dataset includes 10,000 randomly selected full-text articles from Elsevier Neuroscience corpus. Those articles are published in journals such as *Pharmacology*, *Biochemistry and Behavior*, *Applied Research in Mental Retardation*, *Progress in Neuro-Psychopharmacology*, etc., in the year between 1973 and 2013. We first use the sentence tokenizer in NLTK to split the text, and then use the BioTagger [11] and brain dictionaries mentioned in Section III.A to

annotate sentences with genes and brains. The F_1-score of entity annotation is 0.8 for 300 manually examined examples. We retain only sentences containing both genes and brain regions, which yield approximately 30,000 sentences. Those sentences containing 7,700 unique entity pairs are used to train the distant supervised model.

As there are no gold standard examples for testing, we ask two students under domain expert guidance to manually label 259 randomly chosen sentences in which 215 unique pairs of entities co-occur. 95 out of the 259 sentences contain entities with coordinating structure. 114 sentences are labeled as examples of *geneExpression*; and 143 sentences are labeled as examples of *otherRelation*. The initial percent agreement between the two coders is greater than 0.8. After examining the inconsistent labels together, the two coders resolve the conflicts. The total amount of time for annotation and examination is up to 5 hours.

B. Experimental Setup

We conduct the following experiments to compare the performance of our proposed approach and state-of-the-art approaches. First of all, we compare the performance of the distant supervised approach with that of semi-supervised approach, as both rely on minimal human intervention. For the baseline, we use SVM-light[6] to implement the idea in Erkan et al.'s work [5], which trained transductive Support Vector Machines (TSVMs) with only a small set of labeled examples for protein-protein extraction. We manually label 60 examples that are randomly chosen from the 30,000 sentences for training the semi-supervised models. As the performance of a semi-supervised model is dependent on the choice of seed examples [12], which may contribute to biased models, we train two TSVMs with different sets of label of size 10 and 20, respectively, each for three times. We use the 7,700 unique entity pairs along with corresponding 30,000 labeled and unlabeled evidences to train those models, and use the labeled 215 entity pairs along with their 259 occurrences as test examples. 8,310 dimensional of features are extracted to represent the data.

Second, we test the effectiveness of the distant supervised method with the grouping strategy. We first compare the performance of sentence-level relation extraction using distant supervised method with and without the grouping strategy. We use the same 30,000 training examples and 259 test examples in the first experiment. In addition to sentence-level comparison, we also compare the performance of those models at corpus-level. The 7,700 entity pairs are randomly split into training and test sets with equal size. Then, the same feature engineering procedure is applied to extract 5,300 dimensional of features. After representing the data, we train the distant supervised model and use it to predict the relations for each of the entity pairs in the test set.

C. Evaluation

We use precision, recall, and F_1-score to measure the performance of using the models to extract the geneExpression relationship, given the confusion matrix shown in Table II. The precision, recall and F_1-score are defined in Eq. (4)-(6):

TABLE II: Confusion Matrix for Classification

Prediction \ Ground Truth	**geneExpression**	**otherRelation**
geneExpression	True Positive (TP)	False Positive (FP)
otherRelation	False Negative (FN)	True Negative (TN)

$$precision = \frac{TP}{TP + FP} \tag{4}$$

$$recall = \frac{TP}{TP + FN} \tag{5}$$

$$F_1\text{-}score = \frac{2 \times precision \times recall}{precision + recall} \tag{6}$$

V. Results and Discussions

A. Distant Supervised Method vs. Semi-supervised Method

The results of the relation extraction using distant supervised model and semi-supervised models are shown in Table III. The best performance of semi-supervised models is achieved by a TSVM trained with 10 labeled examples with an F_1-score of 0.512. However, due to the bias of seed examples, the results achieved by those TSVMs are not highly consistent. Because of this, it is more important to compare the average metrics as opposed to that achieved by individual experiments. Our distant supervised model yields a precision score of 0.477 and a recall score of 0.459 (boldface). Overall, those scores are superior to the average ones by both of the TSVMs; although they are slightly lower than that by some of the semi-supervised settings (Italic).

In summary, our distant supervised model surpasses the semi-supervised baseline. However, the performance is not as good as that achieved in general domain. In Hoffmann et al.'s [8] study, the distant supervised model obtained an F_1-score of 0.6 while ours achieve an F_1-score of 0.468 for sentential-level relation extraction. This may be hindered by the performance of named entity recognition in biomedical domain. As mentioned by Jiang [13], the performance of relation extraction is dependent on that of named entities recognition. In general domain, the NER tool can achieve around 0.9 of F_1-scores [13]; while in biomedical domain, our NER taggers achieve only an F_1-score of 0.8. Another possible explanation lies in the limited amount of training examples. Zhang et al. [14] found that an increasing number of training examples for distant supervision would yield a statistically significant improvement over F_1-score. As we only utilize a 30,000-size corpus in this study, it is hopeful that the performance will increase after a larger corpus is applied.

B. Distant Supervised Methods with and without Grouping Strategy

Table IV shows the results of sentence-level relation extraction achieved by the distant supervised model before and after taking into consideration the coordinating structure among entities. Without entity grouping, the model yields only a

[6]http://svmlight.joachims.org/

TABLE III: Results using Semi-supervised Method and Distant Supervised Method

	TSVM(10 labels)				TSVM(20 labels)				Distant Supervision Method
	Exp1	Exp2	Exp3	Avg.	Exp4	Exp5	Exp6	Avg.	
Precision	*0.529*	0.468	0.4	0.466	0.452	0.449	0.409	0.437	**0.477**
Recall	*0.496*	0.324	0.324	0.381	0.342	0.36	*0.568*	0.423	**0.459**
F_1-score	*0.512*	0.383	0.358	0.418	0.389	0.400	*0.476*	0.422	**0.468**

TABLE IV: Results using Distant Supervision Approach

	Before grouping entities	After grouping entities
Precision	0.357	**0.477**
Recall	0.36	**0.459**
F_1-score	0.358	**0.468**

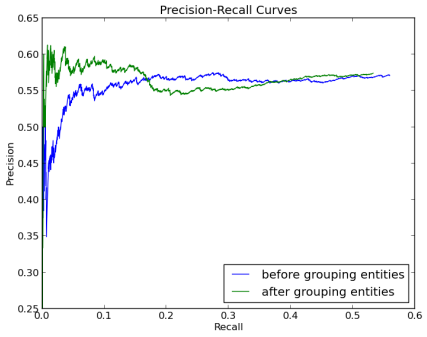

Fig. 5: Precision-Recall Curves of Distant Supervised Model

precision score of 0.357 and recall of 0.36. After incorporating the grouping strategy, the precision and recall scores can be boosted by 33.6% and 27.5%, respectively.

Fig. 5 shows the precision recall curves of relation extraction at corpus-level achieved by the model without/with the grouping method. The model incorporating the grouping strategy achieves better precision over the non-grouping baseline at the recall level between 0 and 0.15, except the very low recall range of approximately 0-1%. At the recall level between 0.15 and 0.4, the precision scores achieved by the model with the grouping strategy are slightly worse than those of the baseline. However, the precision score starts to increase from the recall level of 0.4, which is competitive with the baseline.

VI. Conclusion

In this paper, we present a novel relation extraction approach to extract the *gene expression* relationship between two types of biomedical entities from biomedical literature. This approach does not require any human labeled examples, and achieve achieves promising results compared with semi-supervised models. Moreover, our model takes into consideration of the coordinating structure between biomedical entities, which is superior to that without considering the characteristic of entities at both sentential-level and corpus-level. In the future, we would like to extend the model to target more types of relations involving more types of entities. In addition, we would also like to refine our model so it can be applied to larger datasets.

Acknowledgment

This work is funded in part by NSF grant IIP 1160960 for the Center for Visual and Decision Informatics (CVDI). We appreciate Prof. Zhihao Yang for insightful discussions and anonymous reviewers for valuable comments.

References

[1] G. Zhou, J. Su, J. Zhang, and M. Zhang, "Exploring various knowledge in relation extraction," in *Proceedings of the 43rd annual meeting on association for computational linguistics*, pp. 427–434.

[2] J. Jiang and C. Zhai, "A systematic exploration of the feature space for relation extraction." in *HLT-NAACL*, 2007, pp. 113–120.

[3] S. Zhao and R. Grishman, "Extracting relations with integrated information using kernel methods," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 419–426.

[4] M. Zhang, J. Zhang, J. Su, and G. Zhou, "A composite kernel to extract relations between entities with both flat and structured features," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 825–832.

[5] G. Erkan, A. zgr, and D. R. Radev, "Semi-supervised classification for extracting protein interaction sentences using dependency parsing." in *EMNLP-CoNLL*, vol. 7, 2007, pp. 228–237.

[6] M. Mintz, S. Bills, R. Snow, and D. 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*, pp. 1003–1011.

[7] S. Riedel, D. McClosky, M. Surdeanu, A. McCallum, and C. D. Manning, "Model combination for event extraction in BioNLP 2011," in *Proceedings of the BioNLP Shared Task 2011 Workshop*, pp. 51–55.

[8] R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. 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*, pp. 541–550.

[9] P. Thomas, I. Solt, R. Klinger, and U. Leser, "Learning protein protein interaction extraction using distant supervision," *Robust Unsupervised and Semi-Supervised Methods in Natural Language Processing*, pp. 34–41, 2011.

[10] T. C. Rindflesch and M. Fiszman, "The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text," *Journal of biomedical informatics*, vol. 36, no. 6, pp. 462–477, 2003.

[11] L. French, S. Lane, L. Xu, and P. Pavlidis, "Automated recognition of brain region mentions in neuroscience literature," *Frontiers in Neuroinformatics*, vol. 3, Sep. 2009.

[12] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no. 1, pp. 1–130, 2009.

[13] J. Jiang, "Information extraction from text," in *Mining text data*. Springer, 2012, pp. 11–41.

[14] C. Zhang, F. Niu, C. R, and J. Shavlik, "Big data versus the crowd: Looking for relationships in all the right places," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pp. 825–834.