Noise Reduction Methods for Distantly Supervised Biomedical Relation Extraction

Gang Li[1] **Cathy H. Wu**[1,2] **K. Vijay-Shanker**[1]
[1]Department of Computer Information & Sciences
[2]Center for Bioinformatics and Computational Biology
University of Delaware
Newark, DE 19716, USA
{ligang,wuc,vijay}@udel.edu

Abstract

Distant supervision has been applied to automatically generate labeled data for biomedical relation extraction. Noise exists in both positively and negatively-labeled data and affects the performance of supervised machine learning methods. In this paper, we propose three novel heuristics based on the notion of proximity, trigger word and confidence of patterns to leverage lexical and syntactic information to reduce the level of noise in the distantly labeled data. Experiments on three different tasks, extraction of protein-protein-interaction, miRNA-gene regulation relation and protein-localization event, show that the proposed methods can improve the F-score over the baseline by 6, 10 and 14 points for the three tasks, respectively. We also show that when the models are configured to output high-confidence results, high precisions can be obtained using the proposed methods, making them promising for facilitating manual curation for databases.

1 Introduction

Biomedical relation extraction is a widely studied field that is concerned with the detection of different kinds of relations between bio-entities mentioned in text. With the rapid growth of biomedical literature, it has attracted much research interest as it makes possible to automatically extract structured information from large amounts of text. Biomedical relation extraction has helped facilitate manual curation of many biomedical databases as well as biological hypothesis generation.

Various tasks have been studied for biomedical relation extraction, e.g., extraction of protein-protein interaction (Airola et al., 2008), drug-drug interaction (Segura-Bedmar et al., 2013) and mutation-disease association (Singhal et al., 2016). In recent years, community-organized events, such as BioNLP (Kim et al., 2012, 2013) and BioCreative (Arighi et al., 2014; Wei et al., 2015b), provide comprehensive evaluation for extraction systems of a wide range of biomedical relations and events. In these tasks, supervised learning methods are commonly used and achieve state-of-the-art results.

When applying supervised learning methods, a training corpus is required to train the extraction model. The creation of a training corpus usually requires curators with domain knowledge, and is a time-consuming and labor-intensive process. Thus, it is one of the main obstacles in the use of supervised learning methods for relation extraction. To address this issue, recently researchers have been using distant supervision to construct training data automatically.

In distant supervision, a heuristic labeling process is used to label a text corpus using known related entity pairs from a database. Text containing these entity mentions or their different name variations are labeled as positive instances. To illustrate the labeling process, we show two example sentences labeled using interacting protein pairs from the database IntAct (Orchard et al., 2014).

- ⟨NgR, p75⟩: **NgR** interacted with **p75** in lipid rafts
- ⟨Mdm2, p53⟩: As a consequence, N-terminally truncated **Mdm2** binds **p53** and promotes its stability.

The above sentences are labeled as positive instances and express protein-protein interaction re-

Proceedings of the BioNLP 2017 workshop, pages 184–193,
Vancouver, Canada, August 4, 2017.

lation between the protein mention pair. When a protein pair mentioned in a sentence is not recorded by IntAct, the sentence is then labeled as a negative instance. The positively and negatively-labeled data generated by this process can potentially be used by supervised learning algorithms to train a model. Various existing biological databases and the large amount of Medline abstracts and PMC full-length articles can support applying distant supervision for many biomedical relation extraction tasks. However, the main drawback of distant supervision is that the created data can be very noisy, due to the guideless heuristic labeling process. Wrongly labeled instances exist in both positively and negatively-labeled data. For example, consider the two labeled sentences below for protein-protein interaction.

- ⟨Mdm2, p53⟩: Ribosomal protein S3: A multi-functional protein that interacts with both **p53** and **MDM2** through its KH domain.
- ⟨LRAP35a, MYO18A⟩: **LRAP35a** binds independently to **MYO18A** and MRCK.

In the first sentence, although the protein pair ⟨Mdm2, p53⟩ are interacting with each other according to IntAct, no explicit description in the sentence expresses such an interaction relation. It is labeled as a positive instance by the heuristic labeling process, which is a wrong annotation. On the other hand, if a related entity pair has not been recorded in the database, all the sentences containing their mentions will be labeled as negative instances, which may also contain wrong annotations. As an example, the protein pair ⟨LRAP35a, MYO18A⟩ in the second sentence is not recorded by IntAct. The sentence is labeled as negative, while it expresses an interaction relation between the two proteins. Thus, it is a wrong annotation in the negatively-labeled data.

In this paper, we propose three novel heuristics that attempt to reduce the noise in the positively-labeled data set P as well as the negatively-labeled data set N. First, noise can be removed from P using lexical and syntactic information of the entity mention pairs. Next, high-confidence patterns can be discovered using the purified P, which can then be used to remove noise from N. Experiments on three tasks, extraction of protein-protein interaction, miRNA-gene regulation relation and protein-localization event, show that our methods can improve the F-score by 6, 10 and 14 points over the baseline for the three tasks, respectively. Furthermore, we show that our methods obtain 0.71, 0.95 and 0.77 precision at recall level 0.30 for the three tasks, respectively, making them promising for facilitating database curation.

In the rest of the paper, we first discuss the related work in Section 2. Section 3 describes the three tasks for experiments, as well as the databases and text corpora used in our experiments for applying distant supervision. In Section 4, we describe the details of the proposed methods. Experiments results will be reported in Section 5. We conclude with future work in Section 6.

2 Related Work

Distant supervision for relation extraction was first proposed by Craven and Kumlien (1999) to extract protein-localization relation. Mintz et al. (2009) used Freebase relations to annotate articles in Wikipedia and trained a logistic regression model to extract 102 different types of relations. Riedel et al. (2010) proposed to use multi-instance learning to tolerate noise in the positively-labeled data. They relaxed the original assumption in distant supervision that all the positively-labeled sentences of an entity pair express the relation of interest and instead, they assume that at least one of the sentences does. Hoffmann et al. (2011) and Surdeanu et al. (2012) 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 these approaches, researchers focus on developing models that can tolerate noise and improve extraction performance on entity pair level. However, it is important to note that the noise is not explicitly removed from the labeled data, and extraction on sentence level is not optimized directly.

Focusing on explicitly reducing noise from the distantly-labeled training data, Intxaurrondo et al. (2013) proposed three simple heuristics to remove noise from the positively-labeled data. They tried to filter out positively-labeled instances that appear too frequently or have a large distance from their cluster centroid, or positive entity pairs that have a low partial mutual information. Takamatsu et al. (2012) proposed a statistical model to estimate $P(relation|pattern)$, and removed positively-labeled instances that match a low-probability pattern. Xu et al. (2013)

used pseudo-relevance feedback to discover high-confidence related entity pairs which do not exist in the database, and removed negatively-labeled instances of these entity pairs. Roller et al. (2015) tried to reduce noise in the negatively-labeled data by inferring new relations of a knowledge graph using a random-walk algorithm. Roth et al. (2013) gave a nice review of some of the above methods.

Distant supervision has also been applied to extract biomedical relation. Zheng and Blake (2015) 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. (2011) used a list of words which are frequently employed to indicate protein interaction to filter out noise for protein-protein interaction extraction. Roller and Stevenson (2015) tried to combine existing hand-labeled data with distantly labeled data to improve the performance for drug-condition relations. Multi-instance learning was used by Roller et al. (2015) to extract two subsets of relations in UMLS database with reduced noise by a path ranking algorithm, and by Lamurias et al. (2017) to extract miRNA-gene relations.

3 Resources

3.1 Task Definition

In this paper, we use three tasks, extraction of protein-protein interaction (PPI), miRNA-gene regulation relation (MIRGENE) and protein-localization event (PLOC), to evaluate our methods. Extraction of PPIs is a well-studied task (Miwa et al., 2009; Peng et al., 2016). We aim to extract interacting protein pairs from text using distant supervision, and evaluate it on one of the public corpora used by previous work. Extraction of miRNA-gene regulation relations have attracted much interest recently because of the rapid growth of miRNA-related literature (Bagewadi et al., 2014; Li et al., 2015). In a MIRGENE relation, a miRNA regulates gene expression via direct binding to the gene's 3' UTR or indirect pathway effect. Extraction of protein-localization event has been a subtask in BioNLP shared task from 2009 to 2013 in the Genia track (Kim et al., 2013). It describes the event that a protein is localized to a subcellular location. We only consider extraction of such events when the sentence mentions the protein and the location, same with Zheng and Blake (2015). We list an example sentence for each task below.

- PPI: Interaction of **Shc** with **Grb2** regulates association of Grb2 with mSOS.
- MIRGENE: **MicroRNA-223** regulates **FOXO1** expression and cell proliferation.
- PLOC: The **cyclin G1** protein was localized in **nucleus**.

3.2 Training Data Construction

To construct the training set, we need a database containing related entity pairs and a large amount of text for the heuristic labeling. Table 1 lists the databases, text corpora and numbers of positively/negatively-labeled instances produced by the heuristic labeling process for the three tasks.

Task	Database	Abstracts	Positive / Negative
PPI	IntAct	14,769	67,099 / 108,016
MIRGENE	Tarbase, miRTarBase	30,000	75,632 / 97,118
PLOC	UniProt	30,000	28,985 / 82,132

Table 1: Databases, text corpora and distantly labeled data for the three tasks.

From all the Medline abstracts, we randomly sampled 30,000 abstracts with sentences mentioning a pair of miRNA and gene for miRNA-gene regulation relation, and 30,000 abstracts with sentences mentioning a pair of protein and subcellular location for protein-localization event. We tried sampling more abstracts but the experiment results were not significantly different. For protein-protein interaction, using Medline abstracts leads 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 are still able to improve performance over the baseline in the experiments.

In the heuristic labeling process, we need to recognize entity mentions in text and map them to their database entry. For gene/protein, we use the output from GenNorm++ (Wei et al., 2015a). We use simple regular expressions to recognize miRNA mentions, and map them to a miRNA entry in TarBase (Vlachos et al., 2014) or miRTarBase (Hsu et al., 2014) using the number in the miRNA name. For subcellular location, similar to Zheng and Blake (2015), we use a dictionary from

UniProt (UniProt Consortium, 2014) and perform string matching to find subcellular location mentions. The entry "secreted" is removed as it is not a specific subcellular location. The dictionary contains name variants for each location, and we normalize a matched variant in text to its standard name.

3.3 Test Data

We evaluate the baselines and proposed methods on a test set directly for the three tasks. 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 the purpose of parameter tuning. Thus, when a method has multiple possible choices for a parameter, we will report the results using different parameter values.

For the test set, we use the AIMed corpus (Bunescu et al., 2005) for PPI extraction, same with Bobic et al. (2012). We extend the corpus in our work (Li et al., 2015) to include relation mention annotations, and use the development set to evaluate MIRGENE extraction. For PLOC extraction we use BioNLP 2011 Genia training and development set, same with Zheng and Blake (2015). Gold entity annotations in these corpora are used except for subcellular location, we use the dictionary from UniProt to recognize them, as BioNLP Genia corpus only annotates subcellular locations that participate in an event. The characteristics of the three test corpora are listed in Table 2. We ensure that the test sets do not overlap with the training sets. Specifically, all the abstracts used by the test sets are removed from the document pools from where the training sets are sampled.

Task	Documents	Annotations (P/N)
PPI	225	1,000 / 4,611
MIRGENE	200	464 / 775
PLOC	1,167	125 / 1,783

Table 2: Test sets for the three tasks.

4 Methods

4.1 Model and Feature Set

Logistic regression (LR) model is used for all our proposed methods in the experiments. An example sentence with relevant dependency relations and its extracted features are shown in Fig. 1 and Table 3. E-walk and v-walk features are ⟨edge, stem, edge⟩ and ⟨stem, edge, stem⟩ triples including the direction extracted from the shortest dependency path. They preserve partial structure information and are more generalizable than the full dependency path.

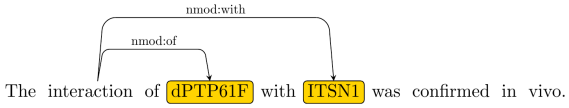


Figure 1: Example sentence for feature extraction.

No.	Feature
1	P1←nmod:of←→nmod:with→P2
2	nmod:of←interact→nmod:with
3	P1←nmod:of→interact interact→nmod:with→P2
4	P1_with_P2 of_P2_with_P2_be interact_of_P1_with_P2_be_confirm
5	2
6	1

Table 3: Features extracted from the example sentence. P1 and P2 represent the two protein mentions. 1: unlexicalized shortest dependency path; 2: e-walk features; 3: v-walks features; 4: three stem sequences, 5: number of edges on the shortest dependency path; 6: number of stems on the first stem sequence.

For all the lexical terms, we use their stems produced by Porter's stemmer (Porter, 1980). Charniak parser (Charniak, 2000; Charniak and Johnson, 2005) with the biomedical model (Mcclosky, 2010) is used to produce constituency parse for each sentence, which is converted to collapsed dependency parse using Stanford CoreNLP converter (Manning et al., 2014) with *CCprocessed* setting. We remove features that only appear once in the whole training set.

4.2 Baselines

The baseline is a LR model trained on the distantly labeled set without any filtering of noise. We also implement two previous methods for comparison. First, we train a LR model on the distantly labeled set filtered by a heuristic (DPFreq) proposed by Zheng and Blake (2015), 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 and Blake (2015). In addition, we implement a widely-used multi-instance model described in Surdeanu et al. (2012) and train it on unfiltered distantly labeled data.

4.3 Proposed Heuristics

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 $\langle e_1, e_2 \rangle$, if their shortest dependency path has the smallest length among all the positively-labeled instances that involve either e_1 or e_2, the pair $\langle e_1, e_2 \rangle$ is considered as a closest pair. When computing the dependency path length, we skip the *appos* 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. We propose to automatically mine such trigger words from the large distantly-labeled corpus, 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 4.

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

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**.

Table 4: The top 10 verb stems and top pattern and example sentence for the three tasks.

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 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 Results and Discussions

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. Specificity is also presented. We will discuss how different numbers of trigger stems and patterns may affect the results later.

Table 5 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 Bobic et al. (2012), and 11 points for PLOC extraction compared to Zheng and Blake (2015).

Different numbers of trigger stems: as different numbers of trigger stems can be used in heuristic TW and HP, we investigated how they affect

	PPI				MIRGENE				PLOC			
Method	P	R	F	S	P	R	F	S	P	R	F	S
Bobic et al. (2012)	0.26	**0.78**	0.39	-	-	-	-	-	-	-	-	-
Zheng and Blake (2015)	-	-	-	-	-	-	-	-	**0.43**	0.25	0.31	-
Baseline	0.37	0.52	0.43	0.86	0.56	0.58	0.57	0.74	0.18	**0.57**	0.28	0.94
Multi-instance (MI)	0.57	0.35	0.43	0.91	0.64	0.56	0.59	0.78	0.22	0.38	0.29	0.94
DPFreq	0.42	0.41	0.41	0.87	0.63	0.50	0.56	0.78	0.21	0.39	0.29	0.94
CP	0.55	0.34	0.42	0.95	0.68	0.50	0.57	0.81	0.26	0.51	0.35	0.95
CP+TW	**0.69**	0.28	0.40	0.93	0.72	0.44	0.55	0.83	0.34	0.42	0.37	0.95
CP+TW+HP	0.65	0.39	**0.49**	0.93	**0.73**	**0.61**	**0.67**	0.84	0.35	0.53	**0.42**	0.95

Table 5: Precision, recall, F-score and specificity of all the methods for three extraction tasks.

the performance for the three tasks. In Fig. 2 (a)-(c), precisions, recalls and F-scores are shown for 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 precision decreased with increased recall when more trigger stems were used, indicating that the quality of the trigger stems can be improved. 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. 2 (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 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.

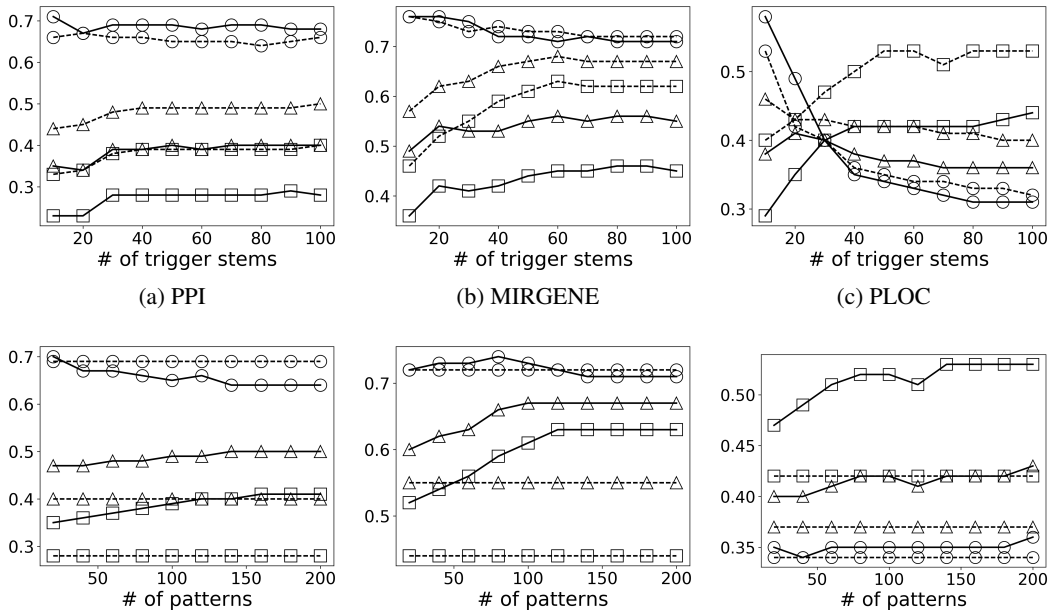
To investigate the performance of the proposed methods for the high-confidence extractions, we draw precision-recall curves using the probability produced by the logistic regression model. By definition, logistic regression model predicts an instance as positive if the probability is greater than 0.5. 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 task, six curves are drawn in Fig. 3. 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.

6 Conclusion

In this paper, we proposed three novel heuristics that use lexical and syntactic information to remove noise from labeled data generated by distant supervision. Experiments showed that the proposed methods achieved significantly higher F-scores than the baseline and previous works for the three tasks, and high precision can be obtained for high-confidence results. For future work, we plan to improve the trigger stem list by asking curators to remove non-informative stems. Aggregating evidences from all the sentences for entity pair level extraction or incorporating direct supervision (Wallace et al., 2016) are two interesting directions.

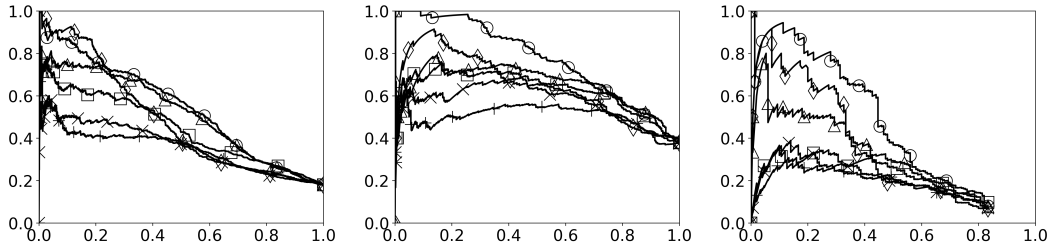
The code and data used in the experiments



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

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

Figure 2: Results of using different numbers of trigger stems (a)-(c) and patterns (d)-(f). 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).



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

Figure 3: Precision-recall curves for the 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).

of this paper are available at `http://biotm.cis.udel.edu/biotm/projects/ds`.

Acknowledgments

Research reported in this manuscript was supported by the National Institutes of Health under the Grant No. U01GM120953. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Antti Airola, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski. 2008. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinformatics* 9 Suppl 11:S2. https://doi.org/10.1186/1471-2105-9-S11-S2.

Cecilia N Arighi, Cathy H Wu, Kevin B Cohen, Lynette Hirschman, Martin Krallinger, Alfonso Valencia, Zhiyong Lu, John W Wilbur, and Thomas C Wiegers. 2014. BioCreative-IV virtual issue. *Database* 2014. https://doi.org/10.1093/database/bau039.

Shweta Bagewadi, Tamara Bobić, Martin Hofmann-Apitius, Juliane Fluck, and Roman Klinger. 2014. Detecting miRNA mentions and relations in biomedical literature. *F1000Res.* 3. https://doi.org/10.12688/f1000research.4591.3.

Tamara Bobic, Roman Klinger, Philippe Thomas, and Martin Hofmann-Apitius. 2012. *Proceedings of the Joint Workshop on Unsupervised*

and Semi-Supervised Learning in NLP, Association for Computational Linguistics, chapter Improving Distantly Supervised Extraction of Drug-Drug and Protein-Protein Interactions, pages 35–43. http://aclweb.org/anthology/W12-0705.

Razvan Bunescu, Ruifang Ge, Rohit J Kate, Edward M Marcotte, Raymond J Mooney, Arun K Ramani, and Yuk Wah Wong. 2005. Comparative experiments on learning information extractors for proteins and their interactions. *Artif. Intell. Med.* 33(2):139–155. https://doi.org/10.1016/j.artmed.2004.07.016.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*. http://aclweb.org/anthology/A00-2018.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, volume 1 of *ACL '05*, pages 173–180. http://aclweb.org/anthology/P05-1022.

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

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and S. Daniel Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 541–550. http://aclweb.org/anthology/P11-1055.

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. 2014. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res.* 42(Database issue):D78–85. https://doi.org/10.1093/nar/gkt1266.

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

Jin-Dong Kim, Ngan Nguyen, Yue Wang, Jun'ichi Tsujii, Toshihisa Takagi, and Akinori Yonezawa. 2012. The genia event and protein coreference tasks of the BioNLP shared task 2011. *BMC Bioinformatics* 13 Suppl 1(Suppl 11):S1. https://doi.org/10.1186/1471-2105-13-S11-S1.

Jin-Dong Kim, Yue Wang, and Yamamoto Yasunori. 2013. *Proceedings of the BioNLP Shared Task 2013 Workshop*, Association for Computational Linguistics, chapter The Genia Event Extraction Shared Task, 2013 Edition - Overview, pages 8–15. http://aclweb.org/anthology/W13-2002.

Andre Lamurias, Luka A Clarke, and Francisco M Couto. 2017. Extracting microRNA-gene relations from biomedical literature using distant supervision. *PLoS One* 12(3):e0171929. https://doi.org/10.1371/journal.pone.0171929.

Gang Li, Karen E Ross, Cecilia N Arighi, Yifan Peng, Cathy H Wu, and K Vijay-Shanker. 2015. miRTex: A text mining system for miRNA-Gene relation extraction. *PLoS Comput. Biol.* 11(9):e1004391. https://doi.org/10.1371/journal.pcbi.1004391.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, pages 55–60. https://doi.org/10.3115/v1/P14-5010.

David Mcclosky. 2010. *Any Domain Parsing: Automatic Domain Adaptation for Natural Language Parsing*. Ph.D. thesis, Brown University, Providence, RI, USA.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics, pages 1003–1011. http://aclweb.org/anthology/P09-1113.

Makoto Miwa, Rune Saetre, Yusuke Miyao, and Jun'ichi Tsujii. 2009. Protein-protein interaction extraction by leveraging multiple kernels and parsers. *Int. J. Med. Inform.* 78(12):e39–46. https://doi.org/10.1016/j.ijmedinf.2009.04.010.

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. 2014. The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* 42(Database issue):D358–63. https://doi.org/10.1093/nar/gkt1115.

Yifan Peng, Cecilia Arighi, Cathy H Wu, and K Vijay-Shanker. 2016. BioC-compatible full-text passage detection for protein-protein interactions using extended dependency graph. *Database* 2016. https://doi.org/10.1093/database/baw072.

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

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. *Modeling Relations and Their Mentions without Labeled Text*, Springer Berlin Heidelberg, volume 6323 of *Lecture Notes in Computer Science*, page 148–163. https://doi.org/10.1007/978-3-642-15939-8_10.

Roland Roller, Eneko Agirre, Aitor Soroa, and Mark Stevenson. 2015. Improving distant supervision using inference learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, pages 273–278. https://doi.org/10.3115/v1/P15-2045.

Roland Roller and Mark Stevenson. 2015. *Proceedings of BioNLP 15*, Association for Computational Linguistics, chapter Making the most of limited training data using distant supervision, pages 12–20. https://doi.org/10.18653/v1/W15-3802.

Benjamin Roth, Tassilo Barth, Michael Wiegand, and Dietrich Klakow. 2013. A survey of noise reduction methods for distant supervision. In *Proceedings of the 2013 workshop on Automated knowledge base construction*. ACM, pages 73–78. https://doi.org/10.1145/2509558.2509571.

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

Ayush Singhal, Michael Simmons, and Zhiyong Lu. 2016. Text mining Genotype-Phenotype relationships from biomedical literature for database curation and precision medicine. *PLoS Comput. Biol.* 12(11):e1005017. https://doi.org/10.1371/journal.pcbi.1005017.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and D. Christopher Manning. 2012. 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*. Association for Computational Linguistics, pages 455–465. http://aclweb.org/anthology/D12-1042.

Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. 2012. Reducing wrong labels in distant supervision for relation extraction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 721–729. http://aclweb.org/anthology/P12-1076.

Philippe Thomas, Illès Solt, Roman Klinger, and Ulf Leser. 2011. *Proceedings of Workshop on Robust Unsupervised and Semisupervised Methods in Natural Language Processing*, Association for Computational Linguistics, chapter Learning Protein Protein Interaction Extraction using Distant Supervision, pages 25–32. http://aclweb.org/anthology/W11-3904.

UniProt Consortium. 2014. Activities at the universal protein resource (UniProt). *Nucleic Acids Res.* 42(Database issue):D191–8. https://doi.org/10.1093/nar/gkt1140.

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. 2014. DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gku1215.

Byron C Wallace, Joël Kuiper, Aakash Sharma, Mingxi Brian Zhu, and Iain J Marshall. 2016. Extracting PICO sentences from clinical trial reports using supervised distant supervision. *J. Mach. Learn. Res.* 17.

Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2015a. GNormPlus: An integrative approach for tagging genes, gene families, and protein domains. *Biomed Res. Int.* 2015:918710. https://doi.org/10.1155/2015/918710.

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

Wei Xu, Raphael Hoffmann, Le Zhao, and Ralph Grishman. 2013. Filling knowledge base gaps for distant supervision of relation extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pages 665–670. http://aclweb.org/anthology/P13-2117.

Wu Zheng and Catherine Blake. 2015. Using distant supervised learning to identify protein subcellular localizations from full-text scientific articles. *J. Biomed. Inform.* https://doi.org/10.1016/j.jbi.2015.07.013.