Revisiting Distant Supervision for Relation Extraction

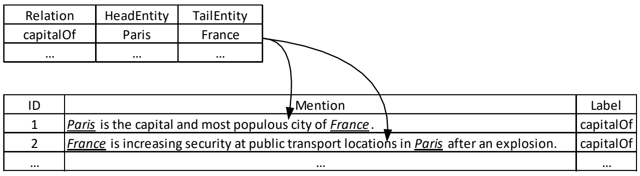
Anonymous EMNLP submission

Abstract

Distant supervision has been widely used in the task of relation extraction (RE). However, when we carefully examine the experimental settings of previous work, we find several issues: (i) The compared models were trained on different training datasets. (ii) The existing testing data is constructed automatically and inevitably contains noise. These issues may affect the conclusions in previous work. In this paper, our primary aim is to re-examine the distant supervision-based approaches under the experimental settings without the above issues. We approach this by training models on the same dataset and creating a new testing dataset annotated by the workers on Amzaon Mechanical Turk. Our major new observations include: (i) Neural network-based approaches benefit more when the size of training data increases. (ii) The performance gap between feature-based approaches and neural network-based approaches is much smaller as compared to the observations in previous work. (iii) Sentence-level attention brings significant improvement for convolutional neural networks but not for piecewise convolutional neural networks. We will share the new testing data with the research community.

1 Introduction

In recent years, knowledge bases (KBs) like Freebase (Bollacker et al., 2008), DBPedia (Lehmann et al., 2015) and NELL (Carlson et al., 2010) have become extremely useful resources for many natural language processing (NLP) tasks, including named entity recognition (Luo et al., 2015), document similarity measurement (Peng et al., 2016), question answering (Bordes et al., 2014), etc. These KBs are mostly composed of relational facts between entities, which are typically represented as triples with the format (head entity, relation, tail entity), e.g., (`Paris`, `capitalOf`, `France`). Although existing KBs may contain billions of relational facts, they are still far from complete and missing many crucial facts. To enrich KBs, relation extraction (RE), *i.e.*, the task of extracting relations between entities from plain texts, has thus attracted increasing attentions.



Relation	HeadEntity	TailEntity
capitalOf	Paris	France
...	...	...

ID	Mention	Label
1	*Paris* is the capital and most populous city of *France* .	capitalOf
2	*France* is increasing security at public transport locations in *Paris* after an explosion.	capitalOf
...	...	...

Figure 1: Training instances generated via distant supervision. The first sentence has a correct label, but the second sentence has a wrong label.

Most existing approaches to RE use supervised learning on relation-specific training data, which is very expensive to acquire. To address this issue, distant supervision is proposed to automatically generate training data via aligning facts in KBs and texts (Wu and Weld, 2007; Mintz et al., 2009). The *distant supervision assumption* is that if two entities preserve a relation in a KB, then *all sentences* that mention the two entities express the relation. Figure 1 shows an example of the automatic labeling of data via distant supervision. In this example, `Paris` and `France` are two entities with a relation type `capitalOf` in a KB. All sentences contain these two entities are labeled with `capitalOf`. Although distant supervision provides a cheap way to automatically label training data, it leads to a noise problem with the data.

The noisy data can be classified into three categories: (i) False positive instances. Not necessarily all sentences that mention an entity pair express the target relation. As shown in Figure 1, the second sentence is a false positive instance. (ii) Multiple labels instances. An entity pair may preserve multiple relation types in a KB. For example, (`Bill Gates`, `founderOf`, `Microsoft`) and (`Bill Gates`, `ceoOf`, `Microsoft`) are clearly true. Hence, distant supervision should assign the sentence that mentions the entity pair with multiple labels. (iii) False negative instances. Because KBs are not complete, the sentences that mention an entity pair which is not in a KB may still express the target relation.

To deal with the issue of false positive instances, Riedel et al. (2010) introduces multi-instance learning (MIL) by relaxing the distant supervision assumption and making the *at-least-one assumption*: if two entities preserve a relation in a KB, *at least one sentence* that mentions the entity pair expresses the relation. However, the model cannot handle multiple labels for one instance. Hoffmann et al. (2011); Surdeanu et al. (2012) follow MIL and introduce multi-instance multi-label learning (MIML). These approaches (Hoffmann et al., 2011; Surdeanu et al., 2012) depend on handcrafted features derived from NLP toolkits, such as part-of-speech tagging, dependency parsing, etc. We call these methods as **feature-based approaches**. Zeng et al. (2015) follows MIML and propose piecewise convolutional neural networks (PCNN) for RE without handcrafted features. Lin et al. (2016) further introduces a sentence-level attention schema for both convolutional neural network (CNN) and PCNN. We call these methods as **neural network-based approaches**. To deal with the issue of false negative instances, Min et al. (2013) follows MIML and proposes a semi-supervised approach by assuming that each entity pair has a fixed proportion of true positive instances. Xu et al. (2013) also follows MIML and adds missing triples to enlarge the KB. In summary, most of previous work follow MIML to deal with the noisy data problem.

The major conclusions of previous MIML-based work include: (i) Neural network-based approaches significantly outperform the feature-based approaches. (ii) For both CNN and PCNN, sentence-level attention brings performance gains. However, we carefully examine the experimental settings of the previous MIML-based work (Hoffmann et al., 2011; Zeng et al., 2015; Lin et al., 2016) and we find two main issues with the experimental settings:

First, when Zeng et al. (2015); Lin et al. (2016) conducted the comparision experiments, the compared models were trained on datasets with different size. Specifically, when comparing neural network-based approaches with feature-based approaches, the neural network-based models (Zeng et al., 2015; Lin et al., 2016) actually used a large training dataset containing $522,611$ sentences, while feature-based models (Hoffmann et al., 2011) were trained on a small dataset containing $126,184$ sentences.

Second, most MIML-based approaches (Hoffmann et al., 2011; Zeng et al., 2015; Lin et al., 2016) were evaluated on the testing data generated by distant supervision. However, as we discussed previously, such data has a noise problem. Although Hoffmann et al. (2011) released a testing dataset which was sampled from NYT corpus and manually annotated, the sampled data is biased towards the model proposed by Hoffmann et al. (2011).

The above issues may affect the conclusions in previous work. In this paper, we revisit the distant supervision for relation extraction. Specifically, our contributions include:

- We carefully re-examine the experimental settings of previous MIML-based work for RE. We find the issues with the training data size and the testing data used in the experiments.
- We create a new testing dataset by pooling the extraction results from all compared models (Hoffmann et al., 2011; Zeng et al., 2015; Lin et al., 2016) and the data generated by distant supervision on NYT corpus, and using Amazon Mechanical Turk[1] (MTurk) to annotate the data in a crowdsourcing way. The dataset will be shared with the research community.
- We conduct extensive experiments to examine the MIML-based approaches (Hoffmann et al., 2011; Zeng et al., 2015; Lin et al., 2016). All models are trained on the same training dataset and evaluated on our new testing dataset. Our new conclusions include: (i) Neural network-based approaches benefit more when the size of training data increases. (ii) Additionally, neural network-based ap-

[1] `https://https://www.mturk.com/`

proaches outperforms the feature-based approaches, but the gap is much smaller as compared to the observations in previous work. (iii) Sentence-level attention brings significant improvement for CNN but not for PCNN.

The remaining of this paper is organized as follows. Section 2 re-examines the experimental settings of previous MIML-based approaches and finds the issues of the experimental settings in previous work. Section 3 describes the experimental settings and our new testing dataset, and reports our new observations from the experimental results. Section 4 summarizes other related work on distant supervision for relation extraction. We conclude this paper in Section 5 and discuss future work.

2 Revisit Distant Supervision for RE

Relation Extraction is considered as a task of predicting the relations expressed in text. For example, here is a sentence:

`Bill Gates founded Microsoft in 1975.`

where `Bill Gates` and `Microsoft` are two entity mentions. A relation extractor or classifier takes the sentence and the two entity mentions as inputs, and determines the semantic relation that it expresses, if any. In the above example, a correct prediction may be `founderOf` relation.

In previous work, relation extraction is mostly formulated as a classification task. However, it is expensive to acquire relation-specific training data. **Distant supervision** was introduced to deal with this challenge. Distant supervision was first introduced by Craven et al. (1999). They aligned a knowledge base of yeast protein with the abstracts of biomedical papers to obtain the training data. Then, they train a naive Bayes classifier to extract biological relations from texts. Wu and Weld (2007) heuristically annotated the text of Wikipedia articles with the attribute values in Infobox to build the training data for learning attribute value extractors. Mintz et al. (2009) aligned a set of most frequent relations in Freebase with the text of Wikipedia articles and trained a classifier for relation extraction.

In these work, the KBs and text corpus used for distant supervision are highly related. By highly related, we mean that both the attribute values in Infobox and the facts in Freebase are derived from Wikipedia articles. However, Riedel et al. (2010) pointed out that the training data generated via distant supervision becomes more noisy, particularly when the KB is not directly related to the text, e.g. Freebase and news articles. Riedel et al. (2010) developed and released a challenging corpus that was generated by aligning the text of New York Times articles and the facts of Freebase. As we introduced in the previous section, the noisy data can be classified into three categories: (i) False positive instances; (ii) Multiple labels instances; (iii) False negative instances. To deal with these challenges, a serious of multi-instance multi-learning based approaches have been proposed.

2.1 MIML-based Approaches

To deal with the issue of false positive instances, Riedel et al. (2010) introduce multi-instance learning (MIL) (Amores, 2013). MIL is a variant technique of supervised learning. Instead of receiving a set of instances which are individually labeled, the learner receives a set of labeled bags, each containing many instances. A bag is labeled positive if there is at least one instance in it which is positive, while a bag is labeled negative if all the instances in it are negative. In the scenario of relation extraction, an instance is a sentence that mention an entity pair. A bag is composed of all sentences that mention the same entity pair. Given a relation, a bag will be labeled as positive if the corresponding entity pair preserves the target relation in a KB. On the other hand, a bag will be labeled as negative if the corresponding entity pair has no target relation in the KB. Riedel et al. (2010) introduce a novel graphical model which is learned from a set of labeled bags.

To further handle the issue of multiple labels instances, Hoffmann et al. (2011); Surdeanu et al. (2012) extend the models of Riedel et al. (2010) and introduce multi-instance multi-label learning (MIML). These approaches are based on the features derived by NLP toolkits, e.g. chunking, dependency parsing, named entity recognition, etc. We call these as **feature-based approaches**. To overcome the problem of error propagations and reduce the efforts of feature engineering, neural network models are applied to automatically learn features for relation extraction (Socher et al., 2012; Zeng et al., 2014; Santos et al., 2015). Zeng et al. (2015) first combines MIML with convolutional neural networks (CNN) and picewise convolutional neural networks (PCNN) for relation extraction. The experimental results show that **neural network-based approaches** significantly outperform the feature-based approaches. However, Zeng et al. (2015) assume that only one sentence is active in each bag (for one entity pair), and it will lose the information of other neglected sentences in each bag. Lin et al. (2016) try to extend the models of CNN and PCNN, and propose sentence-level attention over multiple instances in one bag, which can utilize all informative sentences. The experimental results show that sentence-level attention brings performance gains for both CNN and PCNN.

To address the issue of false negative instances, Min et al. (2013) follow MIML and propose a semi-supervised approach by assuming that each entity pair has a fixed proportion of true positive instances. Xu et al. (2013) also follow MIML and allow feedback from a coarse relation extractor to add missing triples and enlarge the KB. In summary, most of previous work follow MIML to deal with the noisy data problem.

In a summary, MIML-based approaches are the main stream ones to solve the noisy data problem under the distant supervision assumption. The major conclusions of previous MIML-based work include: (i) Neural network-based approaches significantly outperform the feature-based approaches. (ii) For both CNN and PCNN, sentence-level attention brings performance gains.

2.2 The Issues with the Experimental Settings

However, when we carefully examine the experimental settings of the previous MIML-based work (Hoffmann et al., 2011; Zeng et al., 2015; Lin et al., 2016), we find the following issues which may affect the existing conclusions.

When the model comparison experiments were conducted by Zeng et al. (2015); Lin et al. (2016), the compared models were trained on the datasets with different size. Particularly, when comparing neural network-based approaches with feature-based approaches, the neural network-based models (Zeng et al., 2015; Lin et al., 2016) actually used a large training dataset containing $522,611$ sentences, while feature-based models (Hoffmann et al., 2011) were trained on a small dataset containing $126,184$ sentences. We will give the details of the two training datasets in Section 3.2. It is important to re-examine the performances of the models that are trained on the same training dataset. The new experimental results will be shown in Section 3.5.

Most MIML-based approaches (Hoffmann et al., 2011; Zeng et al., 2015; Lin et al., 2016) were evaluated on the testing data generated by distant supervision. However, the automatically generated labels in the testing data could be wrong due to the limitation of distant supervision assumption. The quality of the testing data may affect the experimental results. Although Hoffmann et al. (2011) released a testing dataset which was manually annotated, the dataset was sampled from the union of the extraction results by the model of (Hoffmann et al., 2011) and the data generated by distant supervision. If the experiments are conducted on this testing data, the results may be biased towards to the model of (Hoffmann et al., 2011). In this paper, similar to the slot filling task of TAC KBP, we should develop a new testing dataset which is sampled from a set by pooling the extraction results from all compared models and the data generated by distant supervision. We will show our new observations based on the new testing data in Section 3.5.

3 Experiments

3.1 Systems

In our experiments, we compare the following feature-based and neural network-based systems:

- **MultiR** (Hoffmann et al., 2011) which is a feature-based approach.
- **CNNONE** (Zeng et al., 2015) which is a convolutional neural networks (CNN) model. ONE means that only one sentence is active in each bag (for one entity pair).
- **PCNNONE** (Zeng et al., 2015) which is a picewise convolutional neural networks (PCNN) model.
- **CNNATT** (Lin et al., 2016) which extends the model of CNNONE by introducing sentence-level attention over multiple instances.
- **PCNNATT** (Lin et al., 2016) which extends the model of PCNNONE by introducing sentence-level attention over multiple instances.

We use the implementations of these systems shared by the authors (Hoffmann et al., 2011; Lin

Dataset	#sentences	#pairs	#facts
DSTrainSmall	126,184	67,946	4,700
DSTrainLarge	522,611	279,786	18,252
DSTest	172,448	96,678	1,950
HoffmannTest	881	565	259
Ours	2,040	1,666	547

Table 1: Statistics about the datasets.

et al., 2016)[2].

3.2 Dataset

In this paper, we use the widely-adopted New York Times (NYT) dataset [3] developed by (Riedel et al., 2010)[4]. The NYT corpus contains about 1.8 million news articles. When constructing the dataset, named entity mentions were first extracted from the text of NYT articles by using Stanford Named Entity Tagger (Finkel et al., 2005). Then, the named entity mentions were linked to the entities in Freebase by using exact string matching. If a sentence mentions two entities that have a relation in Freebase, then a corresponding instance will be generated and labeled as the relation type. Otherwise, an instance with a label *NA* which indicates that there is no relation between the entity pair, will be generated. Riedel et al. (2010) mainly focus on the relations related to "people", "business", "person" and "location". There are 53 relation labels including the special label *NA* in the corpus. Following (Riedel et al., 2010), we divide the Freebase relations into two parts, one for training and one for testing. The former is aligned to the $2005-2006$ articles of NYT corpus, and the latter to the 2007 articles.

Training Data. In previous work, there are two training datasets sampled from the aligned sentences of $2005-2006$ NYT articles. (i) Riedel et al. (2010) sampled a small training dataset containing $126,184$ sentences, $67,946$ entity pairs and $4,700$ facts. We denote this dataset as *DSTrainSmall*. (ii) Zeng et al. (2015); Lin et al. (2016) sampled a large training dataset containing $522,611$ sentences, $279,786$ entity pairs and $18,252$ facts which covers all sentences in *DSTrainSmall*. We denote this dataset as *DSTrainLarge*. The details of these two training datasets have been given in Table 1. In the experiments of Zeng et al. (2015); Lin et al. (2016), the neural network-based models were trained on *DSTrainLarge*, while the feature-based models were trained on *DSTrainSmall*. The comparison might be not fair. We will train and compare these models on the two training datasets respectively.

Testing Data. In previous work, there are two popular testing datasets. (i) One is the dataset generated by distant supervision. We denote this dataset as *DSTest*. However, as we discussed previously, the automatically generated labels in the testing data could be wrong due to the limitation of distant supervision assumption. The quality of the testing data may affect the experimental results. (ii) Although Hoffmann et al. (2011) released a testing dataset which was manually annotated, the dataset was sampled from the union of the extraction results from MultiR (Hoffmann et al., 2011) and the data generated from distant supervision. We denote this dataset as *HoffmannTest*. The evaluation conducted on this dataset may be biased towards MultiR.

Besides the existing two testing datasets, we develop a new testing dataset. Our aim is to guarantee the quality of the data and make it not biased towards any of the compared models. Because the instances with non-"NA" labels are quite sparse in the data, it is difficult for us to directly sample enough non-"NA" instances from the corpus to annotate. Similar to the slot filling task of TAC KBP, the key idea of creating the dataset is pooling the top predicted results from all compared models and the data generated from distant supervision.

In the testing corpus *DSTest*, distant supervision labels $172,448$ sentences and only $6,444$ sentences are labeled as non-"NA". We first pool the $6,444$ non-"NA" sentences given by distant supervision and the top $10,000$ non-"NA" sentences predicted by each of the compared systems (including MultiR, CNNONE, PCNNONE, CNNATT and PCNNATT). The pooling results contain $17,147$ sentences. Then we randomly sample $2,040$ sentences from the pooling results, and utilize Amazon Mechanical Turk to annotate the dataset in a crowdsourcing way. We divide the $2,040$ sentences into 120 tasks, and each task contains 17 sentences to be labeled and 3 controls. Each control is a sentence that we already know

[2] http://www.cs.washington.edu/ai/raphaelh/mr and https://github.com/thunlp/NRE

[3] http://iesl.cs.umass.edu/riedel/ecml/

[4] Apart from the NYT dataset released by (Riedel et al., 2010), Surdeanu et al. (2012) has released a KBP dataset with a manually labeled testing data. Unfortunately, KBP dataset only contains feature information for each sentence while lack of original plain texts. Neural network-based approaches cannot be compared on the KBP dataset due to lack of plain texts.

Window size	Word dimension	Position dimension	Batch size	Learning rate	Dropout probability	Sentence dimension
l=3	d_w=50	d_p=5	160	0.001	0.5	d_c=230

Table 2: The parameters of neural networks-based approaches used in our experiments.

its label. All the controlled sentences are sampled from the set of *HoffmannTest* (Hoffmann et al., 2011). Since it is important to control the annotation quality, we use the controls to detect the unqualified workers. If a worker fails on more than one control in a task, we will discard all his annotations for the task. Then the task will be automatically re-assigned to a new worker to complete. Besides, we request 5 workers to annotate each sentence, and use the majority votes to get the ground truth label. The agreements between workers are high. There are 99.7% sentences to which more than 2 workers give the same label. If there is a tie, we will ask another annotator to break it. There are only 6 tie sentences.

The details of the three testing datasets have been shown in Table 1.

3.3 Evaluation Metrics

In the experiments, we compare the precision and recall curve of each system. The curve of each system is drawn by (i) ranking all predicted instances according to their confidence scores given by the system, and (ii) traversing the ranking list from the high score to low score to measure the precision and recall at each position.

Additionally, in previous work, the evaluation were usually conducted in two levels: **entity pair level** and **sentence level**.

By entity pair level, we mean that the system should determine the relation of one bag (i.e., a set of sentences that mention the same entity pair). When using the testing data generated by distant supervision (*DSTest*), we use entity pair level evaluation. Because *DSTest* has less noise at entity pair level while more noise at sentence level, it is better to use the bag level label under the at-least-one assumption.

By sentence level, we mean that the system should determine the relation of one instance (i.e., a sentence that mentions an entity pair). When using the testing data that contains the manually labeled sentences (*HoffmannTest* and *Ours*), we use sentence level evaluation.

3.4 Parameters Settings

In this section, we describe the parameters settings of the neural network-based approaches.

Word Embeddings. In this paper, we follow Lin et al. (2016) and use the word2vec tool[5] to train the word embeddings to on NYT corpus. We keep the words which appear more than 100 times in the corpus as vocabulary. Besides, when training the word embeddings, an entity mention will be considered as one token if it has multiple words.

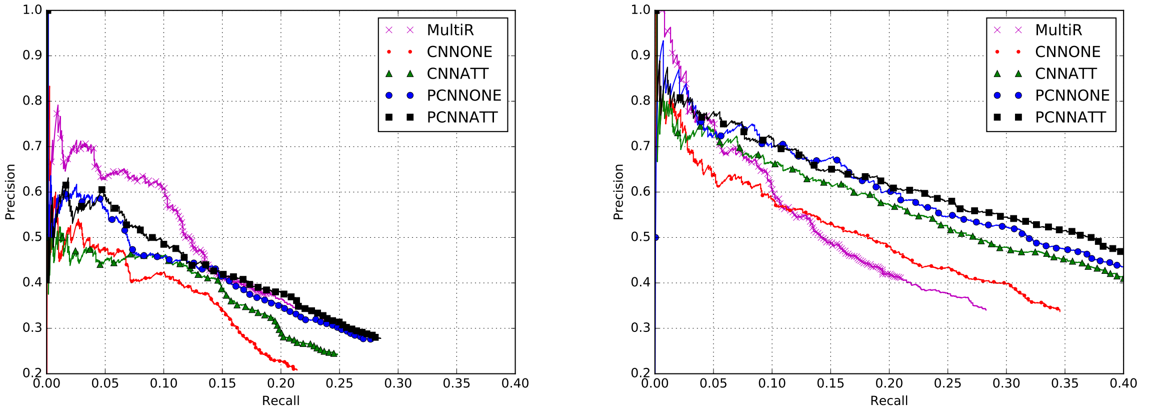
Model Parameters. Following (Surdeanu et al., 2012), we use three-fold validation on the training set to tune the parameters of all models. We use grid search to determine the optimal parameters and manually specify spaces of the following parameters: the sliding window size $l \in \{1,2,...,8\}$, the size of sentence embedding $n \in \{50,60,..., 300\}$ and the batch size B among $\{40, 160, 640, 1280\}$. For other parameters, we follow the settings used in (Zeng et al., 2015; Lin et al., 2016). Table 2 summarizes the parameters of neural networks-based approaches used in our experiments.

3.5 Experimental Results

As we discussed in the Section 2.2, there are two issues with the experimental settings of previous work (Zeng et al., 2015; Lin et al., 2016): (i) the compared models were trained on the data with different size. (ii) the quality of the existing testing data is not good. In this paper, we conduct three experiments. In the first two experiments, we try to examine that how the two issues may affect the conclusions in previous work (Zeng et al., 2015; Lin et al., 2016). The third experiment is conducted on our new testing dataset, and we will give our new observations based on the results.

Experiment 1. In the experiments of (Zeng et al., 2015; Lin et al., 2016), the main evaluations were conducted on the testing data of *DSTest*. When Zeng et al. (2015); Lin et al. (2016) compare different models, the feature-based model (MultiR) was trained on the small training dataset *DSTrainSmall*, while the neural network-based approaches (including CNNONE, PCNNONE, CNNATT and PCNNATT) were trained on the large training dataset *DSTrainLarge*. Their major conclusion is that neural network-based approaches significantly outperform the feature-based approaches. However, it might be not fair to compare

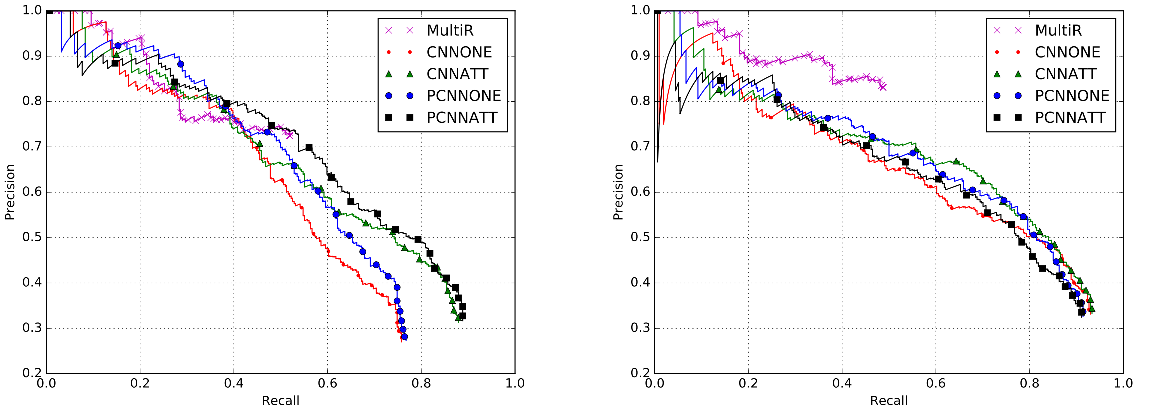
[5] https://code.google.com/p/word2vec/



(a) The results of models trained on *DSTrainSmall*.

(b) The results of models trained on *DSTrainLarge*.

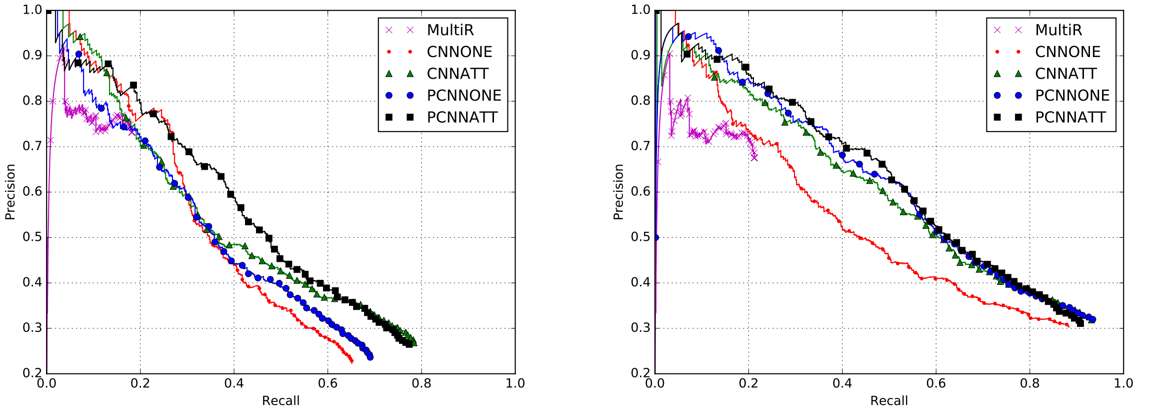
Figure 2: The experimental results on the testing data generated by distant supervision (i.e. *DSTest*).



(a) The results of models trained on *DSTrainSmall*.

(b) The results of models trained on *DSTrainLarge*.

Figure 3: The experimental results on the manual testing data created by Hoffmann et al. (2011) (i.e. *HoffmannTest*).



(a) The results of models trained on *DSTrainSmall*.

(b) The results of models trained on *DSTrainLarge*.

Figure 4: The experimental results on the manual testing data created in this paper (i.e. *Ours*).

these models that were trained on the datasets with different size.

In Experiment 1, we train all models on two training datasets (*DSTrainSmall, DSTrainLarge*) respectively and compare their performance on testing dataset *DSTest*. Figure 2 shows the experimental results. We can observe that (i) In Figure 2a, the feature-based approach MultiR outperforms the neural network-based approaches when all models are trained on the small training data *DSTrainSmall*. (ii) In Figure 2b, when all models are trained on *DSTrainLarge*, MultiR outperforms the neural network-based approaches at the low recall positions. While MultiR performs worse at the high recall positions. (iii) All models benefit from enlarging the training data.

Experiment 2. Since there is a noise problem with the testing dataset *DSTest*, we further conduct the evaluation based on the testing data *HoffmannTest* which was manually annotated by Hoffmann et al. (2011). Figure 3 shows the experimental results. From Figure 3, we can observe that (i) MultiR is comparable to the the neural network-based approaches, when all models are trained on *DSTrainSmall*. (ii) MultiR significantly outperforms the neural network-based approaches when all models are trained on *DSTrainLarge*. (iii) Only MultiR benefits from enlarging the training data. The reason might be that the sampling strategy of the testing data *HoffmanTest* is biased towards MultiR.

Experiment 3. In this experiment, we conduct the evaluation on our new manual testing dataset, which tries to avoid the bias issue. Figure 4 shows the experimental results. From Figure 4, we have the following observations:

- Comparing to Figure 4a and Figure 4b, neural network-based approaches benefit more when the size of training data increases. The gap between MultiR and neural network-based approaches becomes larger when increasing the training data.
- According to Figure 4b, neural network-based approaches outperforms the feature-based approaches, but the gap is much smaller as compared to the observations in previous work. In the experimental results of (Lin et al., 2016), the precision gap between MultiR and PCNNATT is more than 0.3 given the recall 0.2. In contract, the precision gap is around 0.1 at the same recall position in our experiments according to Figure 4b.
- Lin et al. (2016) concludes that sentence-level attention brings performance gains for both CNN and PCNN. However, according to Figure 4b, sentence-level attention brings significant gains for CNN only. We cannot observe significantly improvements on PCNN.

4 Related Work

We have already given an introduction of MIML-based RE approaches in previous sections. In this section, we describe other related work on distant supervision for relation extraction.

Aside from MIML approaches, two branch approaches tackle the noise problem without at-least-one assumption. Alfonseca et al. (2012) propose hierarchical topic models based on the assumption that a context pattern matching a KB fact is either typical for the entity pair, the relation, or neither. This assumption is used to infer the distributions of entity pairs, relations and background text. Takamatsu et al. (2012) propose a generate model to estimate the probability of each pattern showing each relation, based on the automatically labeled dataset. It can be viewed as a pre-processing process to remove false positive matches by filtering mentions with low-probability patterns.

Recent work has begun to explore additional information to augment the distantly supervised relation extraction. Koch et al. (2014); Zhang et al. (2013); Liu et al. (2014) incorporate fine-grained entity type information. Pershina et al. (2014) make use of labeled data to extract training guidelines, which are intended to generalize across many examples. Angeli et al. (2014) use active learning to sample and re-label a small number of difficult instances. Apart from using only facts in KB as distant supervision, additional knowledge such as human common sense (Han and Sun, 2016) or KB embedding knowedge (Weston et al., 2013; Riedel et al., 2013) are further incorporated to enhance distantly supervised RE.

5 Conclusions

In this paper, we carefully re-examine the experimental settings of previous work, and we find two issues: (i) the compared models were trained on the data with different size. (ii) the quality of the existing testing data is not good. We conduct experiments by training models on the same dataset and creating a new manual testing dataset annotated by the workers on Amzaon Mechanical Turk. Our major new observations include: (i) Neural network-based approaches benefit more when the size of training data increases. (ii) The performance gap between feature-based approaches and neural network-based approaches is much smaller as compared to the observations in previous work. (iii) Sentence-level attention brings significant improvement for CNN but not for PCNN. We will share the new testing data with the research community.

References

Enrique Alfonseca, Katja Filippova, Jean-Yves Delort, and Guillermo Garrido. 2012. Pattern learning for relation extraction with a hierarchical topic model. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, pages 54–59.

Jaume Amores. 2013. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence* 201:81–105.

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

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

Antoine Bordes, Jason Weston, and Nicolas Usunier. 2014. Open question answering with weakly supervised embedding models. In *Proceedings of ECML/PKDD*. pages 165–180.

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

Mark Craven, Johan Kumlien, et al. 1999. Constructing biological knowledge bases by extracting information from text sources. In *ISMB*. volume 1999, pages 77–86.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, pages 363–370.

Xianpei Han and Le Sun. 2016. Global distant supervision for relation extraction. In *AAAI*.

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

Mitchell Koch, John Gilmer, Stephen Soderland, and Daniel S Weld. 2014. Type-aware distantly supervised relation extraction with linked arguments. In *Proceedings of EMNLP*. Citeseer.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web* 6(2):167–195.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. *To be appeared in Proceedings of ACL* .

Yang Liu, Kang Liu, Liheng Xu, and Jun Zhao. 2014. Exploring fine-grained entity type constraints for distantly supervised relation extraction. In *COLING*.

Gang Luo, Xiaojiang Huang, Chin-Yew Lin, and Zaiqing Nie. 2015. Joint named entity recognition and disambiguation. In *Proceedings of EMNLP*.

Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. 2013. Distant supervision for relation extraction with an incomplete knowledge base. In *HLT-NAACL*. pages 777–782.

Mike Mintz, Steven Bills, Rion Snow, and Dan 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: Volume 2-Volume 2*. Association for Computational Linguistics, pages 1003–1011.

Hao Peng, Jing Liu, and Chin-Yew Lin. 2016. News citation recommendation with implicit and explicit semantics. In *Proceedings of ACL*.

Maria Pershina, Bonan Min, Wei Xu, and Ralph Grishman. 2014. Infusion of labeled data into distant supervision for relation extraction. In *ACL (2)*. pages 732–738.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, Springer, pages 148–163.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. 2013. Relation extraction with matrix factorization and universal schemas .

Cicero Nogueira dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying relations by ranking with convolutional neural networks. *arXiv preprint arXiv:1504.06580* .

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, pages 1201–1211.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D 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.

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: Long Papers-Volume 1*. Association for Computational Linguistics, pages 721–729.

Jason Weston, Antoine Bordes, Oksana Yakhnenko, and Nicolas Usunier. 2013. Connecting language and knowledge bases with embedding models for relation extraction. *arXiv preprint arXiv:1307.7973* .

Fei Wu and Daniel S Weld. 2007. Autonomously semantifying wikipedia. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM, pages 41–50.

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

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), Lisbon, Portugal*. pages 17–21.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao, et al. 2014. Relation classification via convolutional deep neural network. In *COLING*. pages 2335–2344.

Xingxing Zhang, Jianwen Zhang, Junyu Zeng, Jun Yan, Zheng Chen, and Zhifang Sui. 2013. Towards accurate distant supervision for relational facts extraction. In *ACL (2)*. pages 810–815.