Revisiting Distant Supervision for Relation Extraction

Tingsong Jiang[†], **Jing Liu**[‡], **Chin-Yew Lin**[§], **Zhifang Sui**[†]
†Key Laboratory of Computational Linguistics, Ministry of Education
School of Electronics Engineering and Computer Science, Peking University
‡Baidu Inc., Beijing, China
§Microsoft Research
tingsong@pku.edu.cn, liujing46@baidu.com, cyl@microsoft.com,szf@pku.edu.cn

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 two issues: (i) The compared models were trained on different training datasets. (ii) The existing testing data contains noise and bias issues. 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. We draw new conclusions based on the new testing dataset. The new testing data can be obtained from *http://aka.ms/relationie*.
Keywords: relation extraction, distant supervision

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. 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 attention. For example, here is a sentence: `Paris is the capital city of France`.
where `Paris` and `France` 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 `capitalOf` relation.

Most existing approaches formulate RE as a classification task and 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 mainly classified into two categories:

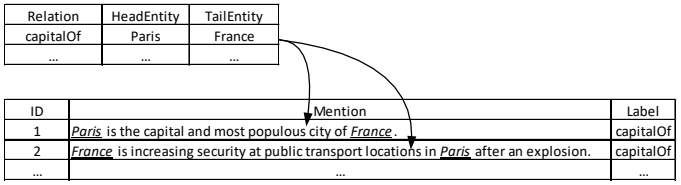


Figure 1: Training instances generated via distant supervision. The first sentence has a correct label, but the second sentence has a wrong label.

(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.

To deal with the two major issues, multi-instance multi-label learning (MIML) was proposed for RE 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 (Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012). The previous work of MIML-based approaches can be mainly categorized into two folds: (i) **feature-based approaches** (Hoffmann et al., 2011; Surdeanu et al., 2012) and (ii) **neural network-based approaches** (Zeng et al., 2015; Lin et al., 2016).

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: (i) 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, the neural network-based models (Lin et al., 2016) actually used a large training dataset containing $522,611$ sentences, while feature-based models (Hoffmann et al., 2011)

This work was done when Tingsong Jiang and Jing Liu were working at Microsoft Research.

were trained on a small dataset containing $126,184$ sentences. We will give the details of the two training datasets in Section 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 3.4.. (ii) 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). To address these issues, similar to the slot filling task of TAC KBP, we 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 3.4..

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.
- 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. We draw new conclusions based on the new testing dataset.

2. Datasets

As we discussed in Section 1., we find two issues of previous MIML-based approaches to relation extraction: (i) The compared models were trained on different training datasets. (ii) The existing testing data contains noise and bias issues. In this section, we will give the details of the datasets used in previous work, and create a new testing dataset.

Data Source. Most previous work uses New York Times (NYT) dataset [2] developed by (Riedel et al., 2010)[3]. The NYT corpus contains about 1.8 million news articles.

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
OurTest	2,040	1,666	547

Table 1: Statistics about the datasets.

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. The Freebase relations were divided 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.

The Existing 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 to-

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

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

[3] 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 dim.	Position dim.	Batch size	Learning rate	Dropout prob.	Sentence dim.
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.

wards MultiR.

A New Testing Data. Since the existing testing datasets have noise and bias issues, 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 that will be described in Section 3.1.). 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 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 $3/5$ or more 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. Experiments

In this section, we will examine that how the two data issues will affect the conclusions in previous work, and we will give our new observations based on the new testing dataset.

3.1. Systems

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

- **MultiR** (Hoffmann et al., 2011) which is a feature-based MIML approach.
- **CNNONE** (Zeng et al., 2015) which is a convolutional neural networks (CNN) based MIML 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) based MIML 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 et al., 2016)[4].

3.2. 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.

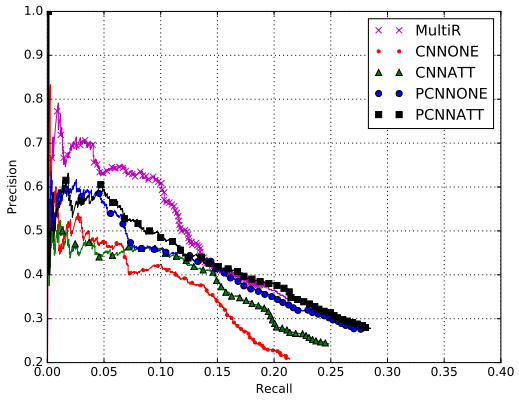
3.3. Evaluation Metrics

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.

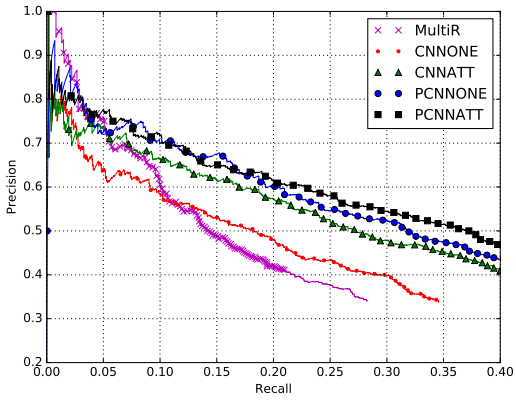
Additionally, in previous work, the evaluation were usually conducted in two levels: **entity pair level** and **sentence level**. By entity pair level, we mean that the system should determine the relation of one bag (i.e., a set of sentences that mention the same entity pair). When using the testing data generated by distant supervision (*DSTest*), we use entity pair level evaluation. Because *DSTest* has less noise at

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

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

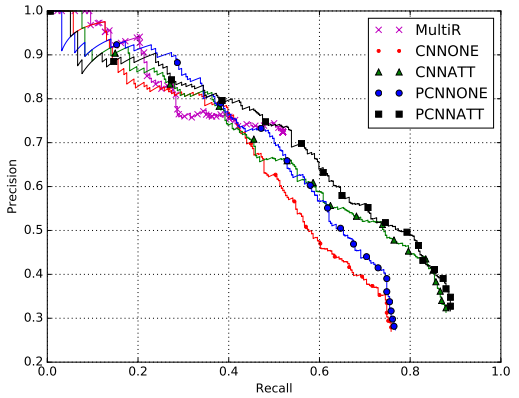


(a) The results of models trained on *DSTrainSmall*.

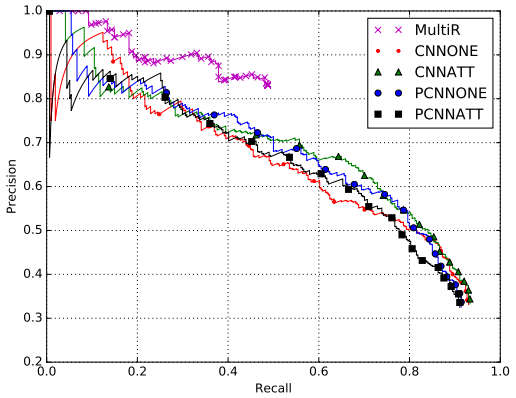


(b) The results of models trained on *DSTrainLarge*.

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

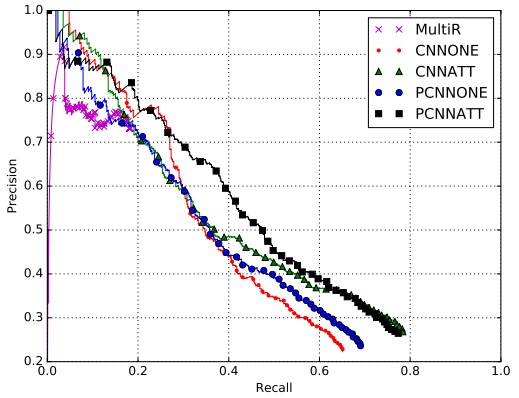


(a) The results of models trained on *DSTrainSmall*.

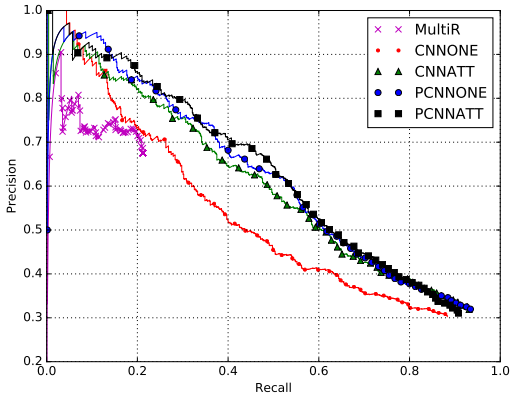


(b) The results of models trained on *DSTrainLarge*.

Figure 3: The experimental results on the manual testing data created by Hoffmann (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*).

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. Experimental Results

As we discussed in the Section 1., 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 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. 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 also share the new testing data with the research community.

5. Acknowledgments

We thank the anonymous reviewers for their valuable comments. The research work is supported by the National Key Research and Development Program of China under Grant No. 2017YFB1002101 and the National Science Foundation of China under Grant No. 61772040. The contact author is Zhifang Sui.

6. Bibliographical References

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of SIGMOD*, pages 1247–1250. AcM.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr, E. R., and Mitchell, T. M. (2010). Toward an architecture for never-ending language learning. In *Proceedings of AAAI*, page 3.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics.

Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., and Weld, D. S. (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, pages 541–550. Association for Computational Linguistics.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al. (2015). Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Lin, Y., Shen, S., Liu, Z., Luan, H., and Sun, M. (2016). Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2124–2133.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (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*, pages 1003–1011. Association for Computational Linguistics.

Riedel, S., Yao, L., and McCallum, A. (2010). Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.

Surdeanu, M., Tibshirani, J., Nallapati, R., and Manning, C. D. (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*, pages 455–465. Association for Computational Linguistics.

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

Zeng, D., Liu, K., Chen, Y., and Zhao, J. (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.