Relation Extraction with Multi-instance Multi-label Convolutional Neural Networks

Xiaotian Jiang[†‡], Quan Wang[†‡*], Peng Li[†‡], Bin Wang[†‡]
[†]Institute of Information Engineering, Chinese Academy of Sciences
No.89A Minzhuang Road, Beijing 100093, China
[‡]University of Chinese Academy of Sciences
No.19A Yuquan Road, Beijing 100049, China
{jiangxiaotian,wangquan,lipeng,wangbin}@iie.ac.cn

Abstract

Distant supervision is an efficient approach that automatically generates labeled data for relation extraction (RE). Traditional distantly supervised RE systems rely heavily on handcrafted features, and hence suffer from error propagation. Recently, a neural network architecture has been proposed to automatically extract features for relation classification. However, this approach follows the traditional expressed-at-least-once assumption, and fails to make full use of information across different sentences. Moreover, it ignores the fact that there can be multiple relations holding between the same entity pair. In this paper, we propose a multi-instance multi-label convolutional neural network for distantly supervised RE. It first relaxes the expressed-at-least-once assumption, and employs cross-sentence max-pooling so as to enable information sharing across different sentences. Then it handles overlapping relations by multi-label learning with a neural network classifier. Experimental results show that our approach performs significantly and consistently better than state-of-the-art methods.

1 Introduction

Relation extraction (RE), defined as the task of extracting binary relations from plain text, has long been a crucial task in natural language processing. Supervised methods are widely used for this task due to their relatively high performance (Zhou et al., 2005; Surdeanu and Ciaramita, 2007). Such methods, however, usually require intensive human annotation and can be time-consuming. To address this issue, distant supervision is proposed to generate labeled data automatically, by aligning facts in a knowledge base (KB) with sentences mentioning these facts (Mintz et al., 2009; Riedel et al., 2010; Riedel et al., 2013).

Traditional (distantly) supervised RE methods use as input numerous lexical and syntactic features, e.g., POS tags, dependency paths, and named entity tags (Mintz et al., 2009; Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012). These features are extracted from sentences using various NLP algorithms, thus inevitably have errors. The induced errors become more serious for long sentences (McDonald and Nivre, 2007), which is unfortunately very common in real-world relation extraction corpus (Zeng et al., 2015). Building distant supervision methods on faulty features inevitably leads to error propagation, the main culprit responsible for performance degradation. Recent studies have shown promising results on using deep neural networks for automatic feature extraction (Zeng et al., 2014; Liu et al., 2015; Xu et al., 2015). Particularly, Zeng et al. (2015) proposed a piecewise convolutional neural network (PCNN) architecture, which can build an extractor based on distant supervision. PCNN automatically extracts features with convolutional neural networks, and introduces piecewise max-pooling to better fit the RE scenario. Although PCNN achieves substantial improvements in distantly supervised relation extraction, it still has the following deficiencies.

First, PCNN uses the expressed-at-least-once assumption (Riedel et al., 2010) for labeled data generation, which states that "if two entities participate in a relation, at least one sentence that mentions

*Corresponding author: Quan Wang (wangquan@iie.ac.cn).

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1471–1480, Osaka, Japan, December 11-17 2016.

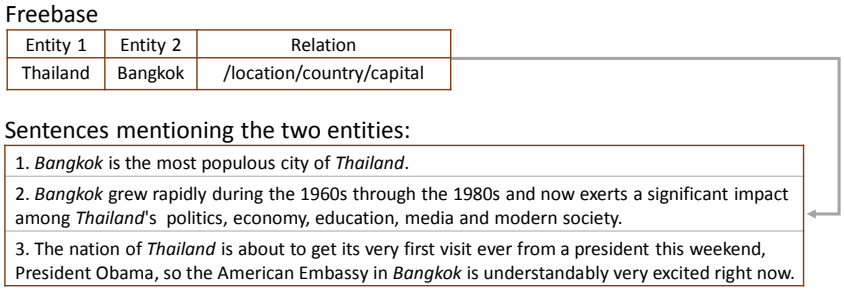


Figure 1: The new assumption states that a relation holding between two entities can be either expressed explicitly or inferred implicitly from all sentences that mention these two entities.

these two entities will express that relation". According to this assumption, PCNN selects only the most likely sentence for each entity pair in training and prediction. We argue, however, that the expressed-at-least-once assumption might be too strong, and selecting a single sentence will definitely lose rich information contained in other sentences. Actually, given two entities participating in a KB relation, it might be difficult to find from the training text the exact single sentence that expresses the relation. Aggregating information available in multiple sentences would probably make the alignment an easier task. Take Figure 1 for example. Given the KB fact (`Thailand`, `/location/country/capital`, `Bangkok`), none of the three sentences mentioning `Thailand` and `Bangkok` expresses the relation of `/location/country/capital`. But if we consider these sentences collectively, we will get more evidence supporting the fact, profiting from the relevant information available in different sentences.

Second, PCNN treats distantly supervised RE as a single-label learning problem and selects for each entity pair a single relation label, ignoring the fact that there might be multiple relations holding between the same entity pair. In fact, as pointed out by Hoffmann et al. (2011), about 18.3% of the distant supervision facts in Freebase that match sentences in the New York Times 2007 corpus have overlapping relations.

In this paper, we propose a multi-instance multi-label convolutional neural network (MIMLCNN) architecture to address the two problems described above. For the first problem, we relax the expressed-at-least-once assumption, and instead assume that "*a relation holding between two entities can be either expressed explicitly or inferred implicitly from all sentences that mention these two entities*" (see Figure 1 for a simple illustration). Therefore, after automatically extracting features within each sentence using a convolutional architecture, we employ cross-sentence max-pooling to select features across different sentences, and then aggregate the most significant features into a vector representation for each entity pair. Since the resultant representation consists of features from different sentences, we successfully make full use of all available information contained in these sentences. For the second problem, we handle overlapping relations by designing various multi-label loss functions in the neural network classifier. The overall architecture is sketched in Figure 2.

The main contributions of this paper can be summarized as follows: (1) We relax the expressed-at-least-once assumption, and propose a more realistic one that naturally enables information sharing from multiple sentences for relation extraction. (2) We propose a multi-instance multi-label convolutional neural network architecture, which handles the multi-label nature of relation extraction. (3) We evaluate our approach on a real-world dataset, and show significant and consistent improvements over state-of-the-art methods.

2 Related Work

Relation extraction is one of the most important tasks in NLP, and has been applied in many practical scenarios (Kordjamshidi et al., 2011; Madaan et al., 2016). Supervised methods has relatively high performance and better practicability, but require massive human annotation, which is both expensive and time consuming. Distant supervision solves this problem by using heuristic assumptions to align triples in a knowledge base with sentences in real-world text corpus, and has been employed in building

large-scale knowledge bases like Knowledge Vault (Dong et al., 2014). A well-known approach in distant supervision is Mintz et al. (2009), which aligns Freebase with Wikipedia articles and extracts relations with logistic regression. Follow-up studies use the feature set developed in this approach, but with deeper understanding on the nature of distant supervision. For example, Riedel et al. (2010) relaxes the assumption used in Mintz et al. (2009) and formulates distant supervision as a multi-instance learning issue; Hoffmann et al. (2011) and Surdeanu et al. (2012) consider overlapping relations between an entity pair. Further effects are also made to model missing data (Ritter et al., 2013), reduce noise (Roth et al., 2013), inject logical background knowledge (Rocktäschel et al., 2015), etc.

In recent years, deep neural network has proven its ability to learn task-specific representation automatically, so that avoiding error propagation suffered by traditional feature-based models. In particular, many neural network approaches have been proposed and shown better performance in relation classification (Zeng et al., 2014; Liu et al., 2015; Xu et al., 2015) and relation extraction (Nguyen and Grishman, 2015). However, these two tasks differ from ours in that relations are extracted at sentence-level, while annotation data is readily available. In distant supervision paradigm, Zeng et al. (2015) is a known neural network model that uses expressed-at-least-once assumption for multi-instance single-label learning. Nevertheless, it selects only one sentence as the representation of an entity pair in training phrase, which wastes the information in the neglected sentences. Besides, it also fails to consider other relations that might hold between this entity pair. The proposed method, on the other hand, leverages evidences collected from all the aligned sentences, and models overlapping relations with multi-label learning.

In traditional supervised learning, an example is usually represented by one instance and one class label. However, there are real-world issues that an example contains multiple instances and has a set of labels. This multi-instance multi-label (MIML) learning scenario was formulated in Zhou et al. (2012), and get widely employed in various tasks (Zha et al., 2008; Zhou and Zhang, 2006; Li et al., 2012). Distant supervised relation extraction is by nature a MIML learning issue, where example is entity pair, instance is sentence aligned with the pair, and label denotes relations. Among previous distant supervision methods, (Surdeanu et al., 2012) formally proposed a multi-instance multi-label framework in a Bayesian framework. In contrast, our method is constructed under a neural network architecture, with the merit of no dependency on lexical and syntactic features.

3 Our Approach

The proposed model takes as input an entity pair (e_1, e_2) as well as all the sentences aligned to this pair, and outputs a set of KB relations that hold between the two entities. As illustrated in Figure 2, our approach consists of three key steps: (1) sentence-level feature extraction, (2) cross-sentence max-pooling, and (3) multi-label relation modeling, detailed as follows.

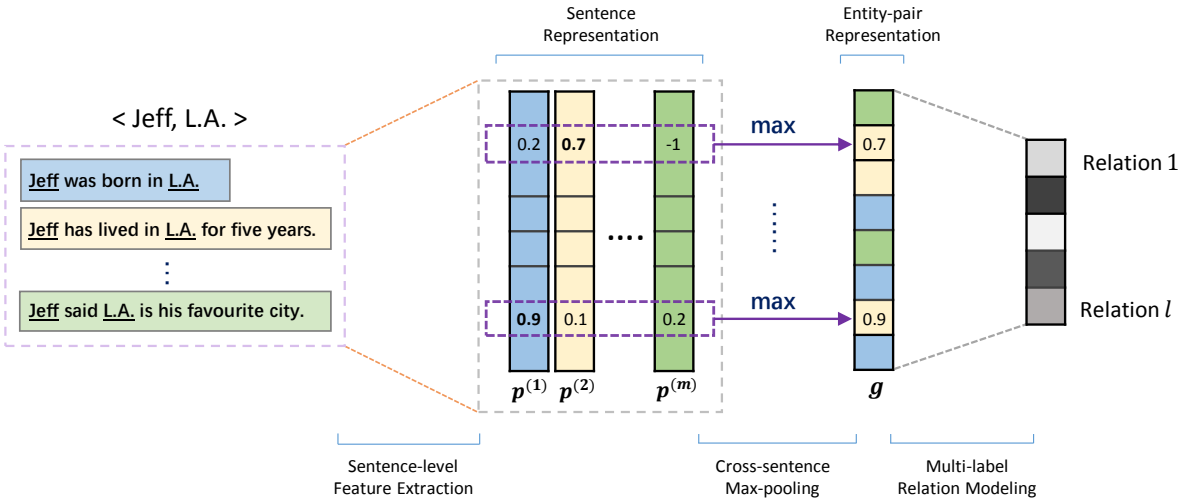


Figure 2: Overall architecture of MIMLCNN.

3.1 Sentence-level Feature Extraction

Sentence-level feature extraction aims to produce a vector feature for each of the aligned sentences. We first pad sentence length to h with zero and transform it to a matrix representation, where each row represents a word token. Convolution, piecewise max-pooling operations are then applied on the matrix to get the vector representation, as illustrated in Figure 3.

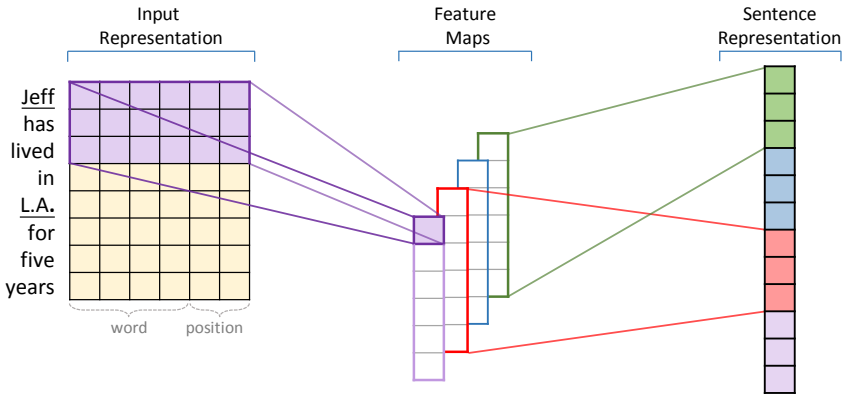


Figure 3: Sentence-level feature extraction using a convolutional architecture.

3.1.1 Input Representation

Two kinds of information are used to construct the input representation for each sentence:

- Raw tokens: We first split the sentence into a sequence of word tokens, then map each token to a d_w dimensional vector called word embedding. The embedding vectors are learned by model training.
- Position features: we use position features (Zeng et al., 2014) to point out the relative positions of a token to e_1 and e_2 in the sentence. Each token has two relative positions, and they are mapped to two different d_p dimensional vectors, separately.

We concatenate the result of these two parts and get matrix $\mathbf{X} \in \mathbb{R}^{h \times d_s}$ as input representation, where $d_s = d_w + 2 * d_p$.

3.1.2 Convolution

The convolution operation aims to extract features from input matrix $\mathbf{X}$, and can be formulated as:

$$c_i = f(\sum_{j=1}^{w_c} \sum_{k=1}^{d_s} \mathbf{W}_{j,k} \mathbf{X}_{j+i-1,k} + b) \tag{1}$$

Here $\mathbf{W} = \mathbb{R}^{w_c \times d_s}$ is a convolutional matrix, where w_c is the width of convolution window; $b \in \mathbb{R}$ is a bias; $f(\cdot)$ is a non-linear function such as Tanh, ReLU. A feature map $\mathbf{c} = [c_1, c_2, \cdots, c_{(h-w_c+1)}]$ is produced by sliding convolution window down the sentence and applying this function at each valid position. To extract n features from the sentence, we repeat the above process with different $\mathbf{W}, b$ for n times. The resultant feature maps are then stacked to construct matrix $\mathbf{C} \in \mathbb{R}^{n \times (h-w_c+1)}$.

3.1.3 Piecewise Max-pooling

To capture the most important feature, max-over-time pooling is often used to select the maximum activation value in each feature map. Piecewise max-pooling (Zeng et al., 2015) improves this idea by first dividing each feature map $\mathbf{C}_i$ into three components $\{\mathbf{c}_{i1}, \mathbf{c}_{i2}, \mathbf{c}_{i3}\}$ based on the positions of the two entities, and then applying max-over-time pooling on each component. This process is formulated as:

$$p_{ij} = max(\mathbf{c}_{ij}) \qquad 1 \le i \le n, 1 \le j \le 3 \tag{2}$$

When piecewise max-pooling is finished, the results of each feature map are concatenated to form vector $\mathbf{p} \in \mathbb{R}^{3n}$, as the feature representation for this sentence.

3.2 Cross-sentence Max-pooling

In the last subsection, we obtain a feature vector $\mathbf{p}$ for each single sentence, but how to take fully usage of the information across sentences is still worth attention. In this paper, we solve this problem by relaxing expressed-at-least-once assumption as:

Assumption: *A relation holding between two entities can be either expressed explicitly or inferred implicitly from all sentences that mention these two entities.*

That is, we relax the expressed-at-least-once assumption by not only allowing making predictions from evidences in each single sentence, but also allowing making predictions by inferring from evidences in all sentences collectively. By nature of this assumption, we skip sentence-level relation extraction and directly make prediction at entity-pair-level, which is more concerned for downstream application and beneficial for evidence aggregation, as described in Riedel et al. (2010).

We propose cross-sentence max-pooling to take the advantage of this assumption. Suppose there are m sentences aligned with the entity pair, and $p_i^{(j)}$ denotes the i^{th} component of the vector representation of the j^{th} sentence, cross-sentence max-pooling aggregates all sentence representations into an entity-pair-level representation $\mathbf{g} = [g_1, g_2, ..., g_{3n}]$, where:

$$g_i = \max(p_i^{(1)}, p_i^{(2)}, \cdots, p_i^{(m)}) \tag{3}$$

This operation brings the following benefits: First, it aggregates features from each sentence, thus supporting entity-pair-level relation extraction directly. Second, it can collect evidence from different sentences, which enables classifiers to make prediction with evidences from different sentences. Besides, compared with Zeng et al. (2015) who only selects one sentence for training at one time, we take advantage of information from all available sentences in each training iteration.

Other approaches, such as mean-pooling, can also be applied in this phrase, but we use cross-sentence max-pooling for the following reason: We consider that multiple occurrences of a feature do not supply much extra information in entity-pair-level relation extraction. That is, a discriminative signal that appears only once can also be sufficient for extracting a relation. This thinking is embodied in the cross-sentence max-pooling operation, where the maximum activation level of each feature is collected across sentences. In contrast, mean pooling averages activation signals by the number of sentences, so that predictive features may be diluted in the representation of entity-pairs that have multiple mentions. This claim is supported by the experimental results.

3.3 Multi-label Relation Modeling

In distant supervision, there are often multiple relations holding between an entity pair. Existing neural network method adopts multi-instance learning, but with single label. In this paper, we model distant supervision under neural network architecture as a multi-label learning problem.

We first calculate the confidence scores for each label by:

$$\mathbf{o} = \mathbf{W_1 g} + \mathbf{b_1} \tag{4}$$

where matrix $\mathbf{W}_1 \in \mathbb{R}^{3n \times l}$ is the collection of weight vectors for each label; $\mathbf{b}_1 \in \mathbb{R}^l$ is a bias. Afterwards, we apply sigmoid function on each element of the score vector $\mathbf{o}$ to calculate the probability of each relation:

$$p(i|M, \theta) = \frac{1}{1 + e^{-o_i}}, \quad i = \{1, 2, \cdots, l\} \tag{5}$$

where M denotes the set of the aligned sentences, and l is the number of relation labels.

A binary label vector $\mathbf{y}$ is set to indicate the set of true relations holding between the entity pair, where 1 means an relation in the set, and 0 otherwise. This way, NA (meaning there is no relation between the entity pair) is naturally represented as an all-zero vector, the complement of the combinations of positive relations.

It is worth noting that relations are often not independent. For example, if triple (A, *capital*, B) holds, another triple (A, *contains*, B) will hold as well. In our model, dependencies between relations are handled by using a shared entity-pair-level representation for all relation labels.

Following this setting, we design two loss functions for multi-label modeling:

$$Loss_{sigmoid} = -\sum_{i=1}^{l} y_i log(p_i) + (1 - y_i) log(1 - p_i) \tag{6}$$

$$Loss_{squared} = \sum_{i=1}^{l} (y_i - p_i)^2 \tag{7}$$

where $y_i \in \{0, 1\}$ is the true value on label i. In the rest sections of this paper, these two loss functions are denoted by sigmoid loss and squared loss, respectively.

The proposed method is trained in an end-to-end fashion. Loss functions are optimized with Adadelta (Zeiler, 2012), which is an robust variant of Stochastic Gradient Decent (SGD) method and features adaptive learning rate over time. Dropout (Srivastava et al., 2014) is also employed on formula (4) for regularization. Specifically, at training time, each element in $\mathbf{g}$ is randomly dropped out by multiplying a Bernoulli random variable with probability p of being 0. At test time, the learned matrix $\mathbf{W}_1$ is scaled by p (i.e. $\hat{\mathbf{W}}_1 = p\mathbf{W}_1$) before scoring. Given an entity pair, the proposed model selects relations whose probability exceeds 0.5 as predicted labels. If there is no such relation, NA is assigned to this entity pair.

4 Experiments

4.1 Dataset

We evaluate our approach on the basis of NYT10, a dataset developed by (Riedel et al., 2010) and then widely used in distantly supervised relation extraction (Hoffmann et al., 2011; Surdeanu et al., 2012; Zeng et al., 2015). NYT10 was generated by aligning Freebase relations with the New York Times (NYT) corpus, with sentences from the years of 2005 and 2006 used for training and sentences from 2007 used for testing.

We follow (Zeng et al., 2015) and use a filtered version of NYT10 released by them[1]. The filtered version prunes the original NYT10 data slightly by removing (1) duplicated sentences for each entity pair, (2) sentences which have more than 40 tokens between a pair of entities, and (3) sentences with entity names that are substrings of other entity names in Freebase. As a result, some relations with low frequency are removed. Statistics of this dataset is shown in Table 1.

	# EPs	# positive EPs	# negative EPs	# sentences	# relations
Training	65,726	4,266	61,460	112,941	26
Testing	93,574	1,732	91,842	152,416	26

Table 1: Statistics of the filtered NYT10 dataset, where EP denotes entity pair.

4.2 Evaluation Metrics

In the following experiments, we use held-out evaluation. At testing time, predicted triples are judged by comparing them with ground truth triples in the testset. We evaluate the performance of each model with Precision-Recall curve, a common used metric for the ranked retrieval results, and P@N metric.

4.3 Baseline Methods

We select three popular feature-based traditional methods as well as the CNN-based method as baselines. We briefly introduce these baselines as follows:

- PCNN: employed a convolutional neural network based method for relation extraction. In contrast to traditional methods, this method allows for automatic feature extraction from raw text, hence avoiding error propagations. Besides, it also uses piecewise max-pooling for sentence-level relation

[1] http://www.nlpr.ia.ac.cn/cip/~liukang/liukangPageFile/code/ds_pcnns-master.zip

extraction. In the following experiments, we use the PCNN code[1] published on the authors' website, along with the dataset.

- Mintz++: (Mintz et al., 2009) proposed distant supervision paradigm that aligns knowledge base entity pairs with text corpus in relation extraction, thus voiding human annotation. The method uses as input lexical and syntactic features, and multi-class logistic regression for classification. [2]
- Multir: (Hoffmann et al., 2011) pointed out that many entity pairs have more than one relation. Their method models overlapping relations by combining sentence-level relation extraction results into entity-pair-level results, with a deterministic decision.
- MIMLRE: (Surdeanu et al., 2012) proposed a novel multi-instance multi-label approach for distant supervision using a graph model. For each entity pair, this method jointly models its multiple instances and multiple labels. Besides, it also models the correlation between labels.

4.4 Implementation Details

As a common practice in neural network models, word embeddings are initialized with pre-training. We run skip-gram model (Mikolov et al., 2013) on training dataset, and use the obtained word vector to initialize the word embedding part of model input. Position features are randomly initialized with uniform distribution between [-1, 1]. For convenience of comparing with baseline methods, our model uses the same parameter settings as (Zeng et al., 2015). Specifically, At model input layer, we use a mini-batch of 50 entity pairs, set the dimension of word embedding $d_w = 50$ and the dimension of position feature $d_p = 5$. At convolutional layers, windows size w_c is set to 3, and the number of feature maps to $n = 230$. Dropout rate p is set to 0.5. Two Adadelta parameters , $\rho = 0.95$ and $\epsilon = e^{-6}$, are set with default values. For baseline models, we use the codes released by (Surdeanu et al., 2012) [3] and Zeng et al. (2015). Since PCNN and MIMLCNN are influenced by random factors when running on GPU, we run both models with the above-mentioned settings for ten times and use the averaged results in the following comparisons.

4.5 Comparison with Baseline Methods

To evaluate the performance of the proposed method, we first compare it to baseline methods. In the following experiments, we use MIMLCNN to refer to the proposed model with cross-sentence max-pooling and sigmoid loss. Figure 4 shows the resulting precision-recall curve in the most concerned area.

From the curves, we observe that MIMCNN can consistently and significantly outperform all baseline methods in the entire range of recall. Comparing neural network methods with traditional feature-based methods, we can conclude that PCNN exceeds traditional methods for its alleviation of error propagation, while MIMCNN exceeds PCNN for its usage of cross-sentence max-pooling and multi-label modeling. The result indicates that the proposed method has the best sense of exploiting the characteristic of distant supervision in a neural network framework. It is worth emphasising that the best of baseline methods can keep a reasonable precision level (larger than 0.5) when recall is less than 0.17. In contrast, our model can keep the same precision level with recall at 0.28, amounting to a 64% increase. Also note that beyond the truncated recall level (0.35), the curve of our method can extend to recall at 0.66 without any loss of precision. This brings 103% increase at the maximum recall level in comparison with the best of baseline methods.

Table 2 further presents the results using P@N metric. In accordance with our observation in precision-recall curve, MIMLCNN is still the winner at most of the entire P@N levels. It is interesting that both of the neural network methods are all good at predicting top-ranked results compared with traditional feature-based methods, especially MIMLCNN. As N gets smaller, the superiority becomes more evident. At P@10, the precision of MIMLCNN can even reach to 0.90, while neither of the baseline methods can exceed 0.84. Also, MIMLCNN is the only method whose mean value of P@N exceeds 0.7.

[2]Note that in the following experiment we use the Surdeanu et al. (2012) implemented version, which has been reported significantly better performance than the original one.

[3]http://nlp.stanford.edu/software/mimlre.shtml

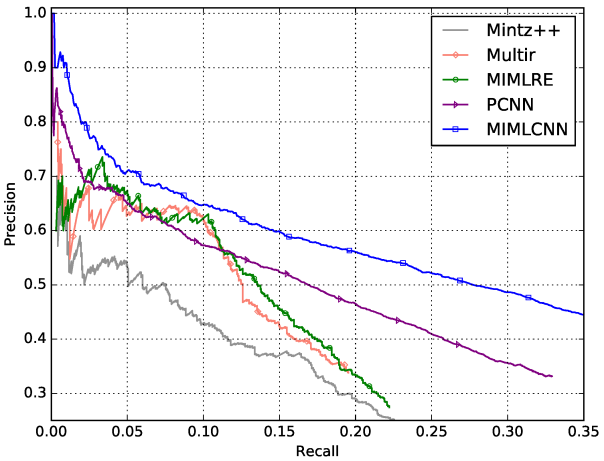


Figure 4: Precision-recall curves of the proposed method and four baselines.

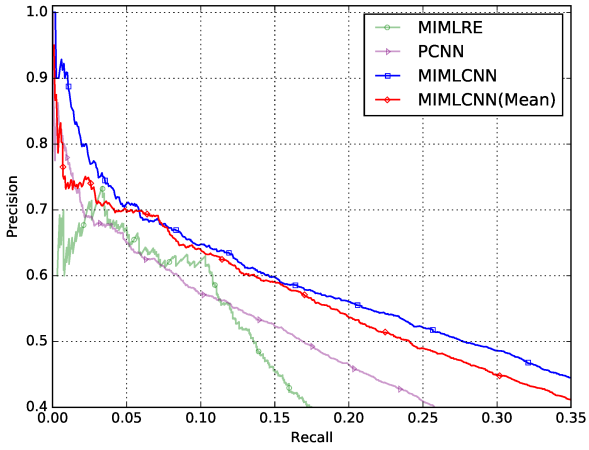
	Mintz++	MultiR	MIMLRE	PCNN	MIMLCNN
P@10	0.70	0.80	0.60	0.84	**0.90**
P@20	0.65	0.65	0.70	0.80	**0.83**
P@30	0.60	0.63	0.63	0.76	**0.80**
P@50	0.54	0.62	0.68	0.72	**0.75**
P@100	0.53	0.62	0.68	0.68	**0.69**
P@200	0.51	0.63	**0.64**	0.62	**0.64**
P@300	0.49	**0.63**	0.62	0.58	0.59
P@500	0.42	0.48	0.51	**0.53**	**0.53**
Mean	0.56	0.63	0.63	0.69	**0.72**

Table 2: P@N results.

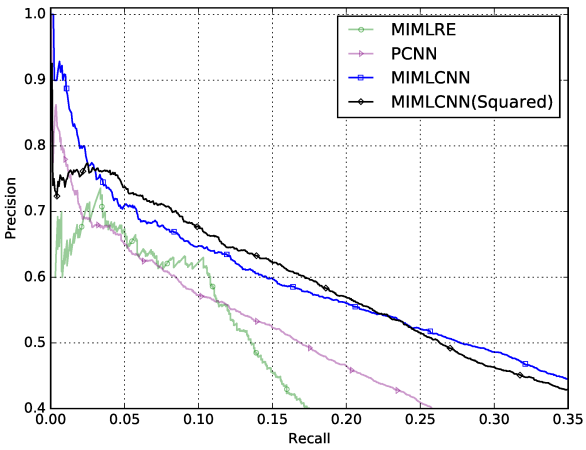
4.6 Effects of Cross-sentence Max-pooling and Multi-label Learning

In this subsection, we empirically prove the effects of cross-sentences max-pooling and multi-label learning, respectively.

In order to prove the effectiveness of cross-sentence max-pooling, we create a baseline method called MIMLCNN(Mean). Comparing with MIMLCNN, this method merely replaces cross-sentence max-pooling with the average of feature representations of all the aligned sentences. Experimental result is presented in Figure 5(b). In almost the entire curves of these two models, MIMLCNN shows better performance. The superiority is especially significant in the front and rearward part of recall levels. This observation supports our claim that cross-sentence max-pooling helps improving performance. It is also interesting that MIMLCNN(Mean) still shows improvements over the baseline methods, though not comparable with MIMLCNN.



(a) Cross-sentence max-pooling.



(b) Multi-label learning.

Figure 5: Effects of cross-sentence max-pooling and multi-label learning. PCNN and MIMLRE are used for reference.

We further compare the effect of using different loss functions in our model, as demonstrated in Figure 5(b). MIMLCNN(Squared) refers to the proposed model with cross-sentence max-pooling and squared loss. From the curves of these two models, we can see that different loss functions have diverse emphases. When we use sigmoid loss (MIMLCNN), most of the improvement resides in recall range [0.1, 0.3], but still remains competitive or slightly better in range [0, 0.1]. Compared with sigmoid loss, using

squared loss brings better performance at the middle area of the curve, but it is also less competitive with respect to the top-ranked results. In contrast with baseline methods, the superiority of MIMLCNN and MIMLCNN(Squared) indicates that multi-label modeling contributes to improving performance in distant supervision.

5 Conclusion

In this paper, we propose a novel neural network method for distant supervision with multi-instance multi-label learning. Given an entity pair, we relax the expressed-at-least-once assumption to take full usage of information from all the aligned sentences with cross-sentence max-pooling, and model multiple relations holding between the entity pair in a neural network architecture. We conduct experiments on a real-world dataset, and prove empirically (1) the proposed method has significantly and consistently better performance than state-of-the-art methods. (2) both cross-sentence max-pooling and multi-label learning take effects. In the future, we would like to further investigate how different loss functions influence performance, and enrich experiments by carrying out human evaluation as well as making detailed analysis on each relation.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments and suggestions. This research is supported by the National Natural Science Foundation of China (grant No. 61402465) and the Strategic Priority Research Program of the Chinese Academy of Sciences (grant No. XDA06030200).

References

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 601–610. ACM.

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*, pages 541–550. Association for Computational Linguistics.

Parisa Kordjamshidi, Paolo Frasconi, Martijn Van Otterlo, Marie-Francine Moens, and Luc De Raedt. 2011. Relational learning for spatial relation extraction from natural language. In *International Conference on Inductive Logic Programming*, pages 204–220. Springer.

Ying-Xin Li, Shuiwang Ji, Sudhir Kumar, Jieping Ye, and Zhi-Hua Zhou. 2012. Drosophila gene expression pattern annotation through multi-instance multi-label learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(1):98–112.

Yang Liu, Furu Wei, Sujian Li, Heng Ji, Ming Zhou, and Houfeng Wang. 2015. A dependency-based neural network for relation classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, page 285290.

Aman Madaan, Ashish Mittal, Ganesh Ramakrishnan, Sunita Sarawagi, et al. 2016. Numerical relation extraction with minimal supervision. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Ryan T McDonald and Joakim Nivre. 2007. Characterizing the errors of data-driven dependency parsing models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 122–131.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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*, pages 1003–1011. Association for Computational Linguistics.

Thien Huu Nguyen and Ralph Grishman. 2015. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of NAACL-HLT*, pages 39–48.

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

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. 2013. Relation extraction with matrix factorization and universal schemas. *NAACL HLT 2013*, pages 74–84.

Alan Ritter, Luke Zettlemoyer, Oren Etzioni, et al. 2013. Modeling missing data in distant supervision for information extraction. *Transactions of the Association for Computational Linguistics*, 1:367–378.

Tim Rocktäschel, Sameer Singh, and Sebastian Riedel. 2015. Injecting logical background knowledge into embeddings for relation extraction. In *Proceedings of the 2015 Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*.

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*, pages 73–78. ACM.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Mihai Surdeanu and Massimiliano Ciaramita. 2007. Robust information extraction with perceptrons. In *Proceedings of the NIST 2007 Automatic Content Extraction Workshop (ACE07)*.

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*, pages 455–465. Association for Computational Linguistics.

Kun Xu, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2015. Semantic relation classification via convolutional neural networks with simple negative sampling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 536–540.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

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.

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.

Zheng-Jun Zha, Xian-Sheng Hua, Tao Mei, Jingdong Wang, Guo-Jun Qi, and Zengfu Wang. 2008. Joint multi-label multi-instance learning for image classification. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.

Zhi-Hua Zhou and Min-Ling Zhang. 2006. Multi-instance multi-label learning with application to scene classification. In *Advances in neural information processing systems*, pages 1609–1616.

GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 427–434. Association for Computational Linguistics.

Zhi-Hua Zhou, Min-Ling Zhang, Sheng-Jun Huang, and Yu-Feng Li. 2012. Multi-instance multi-label learning. *Artificial Intelligence*, 176(1):2291–2320.