Effective Piecewise CNN with Attention Mechanism for Distant Supervision on Relation Extraction Task

Yuming Li[1], Pin Ni[1], Gangmin Li[2] and Victor Chang[3]
[1]*University of Liverpool, U.K.*
[2]*Xi'an Jiaotong-Liverpool University, China*
[3]*Teesside University, U.K.*

Keywords: Relation Extraction, Distant Supervision, Piecewise Convolutional Neural Networks, Attention, Convolutional Neural Networks.

Abstract: Relation Extraction is an important sub-task in the field of information extraction. Its goal is to identify entities from text and extract semantic relationships between entities. However, the current Relationship Extraction task based on deep learning methods generally have practical problems such as insufficient amount of manually labeled data, so training under weak supervision has become a big challenge. Distant Supervision is a novel idea that can automatically annotate a large number of unlabeled data based on a small amount of labeled data. Based on this idea, this paper proposes a method combining the Piecewise Convolutional Neural Networks and Attention mechanism for automatically annotating the data of Relation Extraction task. The experiments proved that the proposed method achieved the highest precision is 76.24% on NYT-FB (New York Times - Freebase) dataset (top 100 relation categories). The results show that the proposed method performed better than CNN-based models in most cases.

1 INTRODUCTION

1.1 Background

Relation extraction aims to extract the relation between entity pairs based on text contents, such as the employment relation between people and organizations, affiliation between organizations and institutions, the geographical relation between a building and its location, etc.

Neural networks-based relation extraction requires a large amount of training data, but manually annotating these training data is too time-consuming and expensive. Most of the existing methods use public data sets, such as nyt10, semeval-2010 task-8, ACE2005, etc. Although the training data are annotated by professionals, they are still based on a fixed number of established data, which is not flexible enough, easily making the model oversaturated and unable to guarantee that the model can achieve the same good effect in real cases. Therefore, aiming at the above problems, distant supervision method is introduced.

In our proposed model, we use Piecewise Convolutional Neural Networks (PCNN) with Attention mechanism and get the precision of 76.24% on NYT-FB (New York Times - Freebase) dataset (top 100 relation categories), which is higher than conventional Convolutional Neural Networks with 3.96% improvement. And other results also show that the proposed method performed better than CNN-based models in most cases.

1.2 Contribution

Our specific contribution is: we propose a hybrid Distant Supervision model based on the PCNN and Attention mechanism, which only needs a small number of labeled samples as the basis to realize automatic annotation of the remaining large amounts of unlabeled data. The experiments proved that compared with the conventional methods, our model has a certain improvement in precision. This provides more trainable resources for deep learning-based relation extraction.

Li, Y., Ni, P., Li, G. and Chang, V.
Effective Piecewise CNN with Attention Mechanism for Distant Supervision on Relation Extraction Task.
DOI: 10.5220/0009582700530060
In *Proceedings of the 5th International Conference on Complexity, Future Information Systems and Risk (COMPLEXIS 2020)*, pages 53-60
ISBN: 978-989-758-427-5

2 RELATED WORK

Faced with a large number of data without labels, supervised relation extraction consumes a large amount of manpower and appears to be inadequate. Therefore, distant supervision entity relation extraction emerges. Mintz et al. (Mintz et al., 2009) first proposed to apply distant supervision to relation extraction tasks in 2009, which solved the problem of automatic annotation of a large number of unlabeled data in the open domain by automatically aligning knowledge base with data. There are two main problems in distant supervision and marking of data: noise and feature extraction error propagation. The noise problem is due to the strong hypothesis condition of distant supervision, which causes the relation of a large number of data to be wrongly marked so that there is a large amount of noise in the training data. The problem of error propagation in feature extraction is that the conventional feature extraction mainly uses the general NLP tool to extract the features of the data set, so a large number of propagation errors will be introduced. As for the problem of false label, the Attention mechanism proposed by Surdeanu et al. (Surdeanu et al., 2012) in 2010 and Lin et al. (Lin et al., 2016) has effectively weakened the influence of distant supervision of false label on extraction performance.

Since the rise of deep learning and its irreplaceable results, it has become a trend to replace conventional feature engineering with the idea of feature extraction through deep learning. For example, the extended CNN model, PCNN+MIL (Zeng et al., 2015) and LSTM (He et al., 2017) gets the orientation information of the entity; COTYPE (Ren et al., 2017) joint extraction of entity and relation information; The deep residual network (Huang and Wang, 2017) prevents layer-by-layer accumulation of mislabeled noise. At present, the research points of distant supervised entity relation extraction based on deep learning mainly focus on the noise problem of distant supervision and error propagation of feature extraction.

3 METHODOLOGY

Distant Supervision is a common practice in relation extraction at present, which was first proposed by Mintz et al. (Mintz et al., 2009) in ACL2009. It is neither a simple supervision method in the conventional sense nor an unsupervised method, but a tagging method that aligns plain text with Knowledge Base to reduce the dependence of the model on manual marking data and enhance the model's adaptability across fields. The proposed distant supervision is based on the following assumptions: If there is a relation between two entities in the Knowledge Base, the unstructured sentence containing the two entities can represent such a relation, and the unstructured sentence can be used as a training positive instance to train the model.

The specific implementation steps of remote supervision method can be summarized as the following two points:

1. Extract the relation entity pairs from the knowledge base.
2. Extract sentences containing entity pairs from the unstructured text as training examples.

However, such methods still have some shortcomings, firstly, some sentences containing specific entities cannot reflect any relation and are invalid data for the training set, which inevitably introduces a lot of noise.

For example, as shown in Figure 1, 'Charlie Chaplin' and 'Vevey' are two related entities with the relation 'place_of_death' in Freebase (Bollacker et al., 2008). The sentences listed below are training corpus generated through distant supervision, but only the second sentence describes the expected relation 'place_of_death', the upper sentence is invalid training corpus, which can be defined as 'wrong label problem'.

Secondly, the process of data construction depends on NLP tasks like Named Entity Recognition (NER), and errors in the intermediate process will cause error propagation problems. McDonald and Nivre (McDonald and Nivre, 2007) proposed that the accuracy of syntactic analysis decreased with the increase of sentence length, while long sentences accounted for a large proportion in the corpus, so the error accumulation and transmission would greatly reduce the task accuracy.

To solve these problems, there are four main methods :

1. Introduce prior knowledge as the limit in the process of constructing data sets;
2. Use the relation between reference and reference to score data samples with graph model to filter out sentences with low confidence;
3. Label the test package with multi-example learning method;
4. The attention mechanism is used to assign different weights to sentences with different degrees of confidence.

In this work, we use Piecewise Convolutional Neural Networks with the Attention mechanism as the

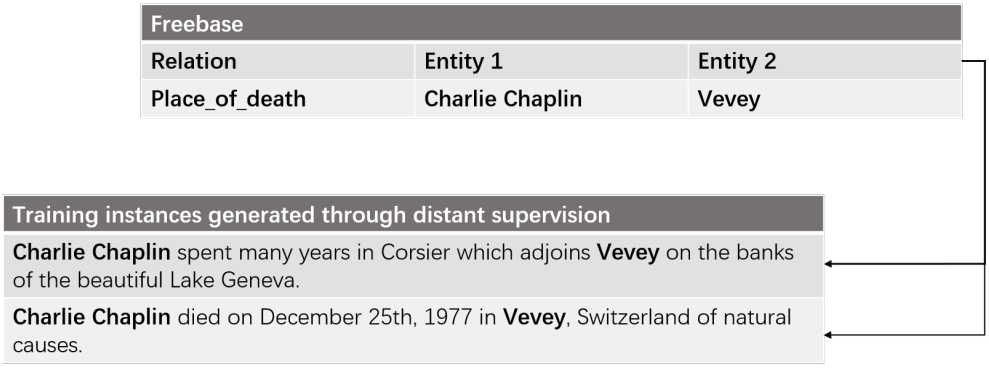

Figure 1: A Sample of Distant Supervision. Upper sentence: incorrect instance; lower sentence: correct instance.

model of Distant Supervision for Relation Extraction. The use of the NLP tool in extracting features in the classic entity relation extraction task will lead to error propagation layer by layer and affect the final effect of relation extraction. PCNN method in deep learning effectively minimizes the influence of feature extraction error propagation. The method regards the relation extraction problem as a multi-instance problem. The training set is composed of multiple packages, and each package contains multiple instances. The sample with the highest confidence in each bag is selected as the positive sample for training, so as to effectively remove useless samples and reduce data noise. The details of the PCNN model can be discussed as follow.

3.1 Conventional Convolutional Neural Networks

PCNN is a variant of Convolutional Neural Networks (CNN), so we describe the core of CNN first. CNN is a type of feed-forward artificial neural networks which was first developed in the 1980s. Layers of the networks are formed by a convolution operation followed by a pooling operation (LeCun et al., 1998; Kalchbrenner et al., 2014). Since there is a surge of interest in deep learning recently, CNNs are applied effectively in various NLP tasks, including relation extraction.

Due to an input sentence that could not predict labels for each word in relation extraction, it is essential to utilize all local features and perform this prediction. The convolution operation is a natural method of merging all these features when using a neural network (Collobert et al., 2011). In this step, the matrix x representing the input relation is fed into the convolutional layer to extract higher-level features. A filter whose window size is w can be represented as a weight matrix $f = [f_1, f_2, ..., f_w]$. The purpose of this layer is to produce a score sequence $s = [s_1, s_2, ..., s_{n-w+1}]$ by obtaining from two matrices x and f:

$$s_i = g(\sum_{j=0}^{w-1} f_{j+1}^T x_{j+i}^T + b) \quad (1)$$

where g is some non-linear function and b is a bias term. It could as well replicate this process to increase the n-gram coverage of the model for various filters with different window sizes.

After the convolution operation, pooling operation is applied to combine convolution layers being independent of the sentence length such that these layers can be used in subsequent layers (Collobert et al., 2011; Zeng et al., 2014). Max pooling operation is widely applied as it can identify the most important or relevant features from the score sequence. More formally, for each filter f, its score sequence s is passed through the max function to obtain a single number: $p_f = max s = max s_1, s_2, ..., s_{n-w+1}$ which can be seen as estimating the possible some augmented n-gram of the hidden class of f appears in the context.

PCNN added segment operation on the basis of CNN, which is more suitable for Natural Language Processing tasks. The details of PCNN are shown as follow sections.

3.2 Piecewise Convolutional Neural Networks

3.2.1 Data Pre-processing and Sentence Segmentation

Firstly, the data is encoded by position, according to the distance from the entity of each word in the sentence. For example, in sentence:

'In just a decade and a half **Jack Ma**, a man from modest beginnings founded and built **Alibaba** into one of the world's largest companies.'

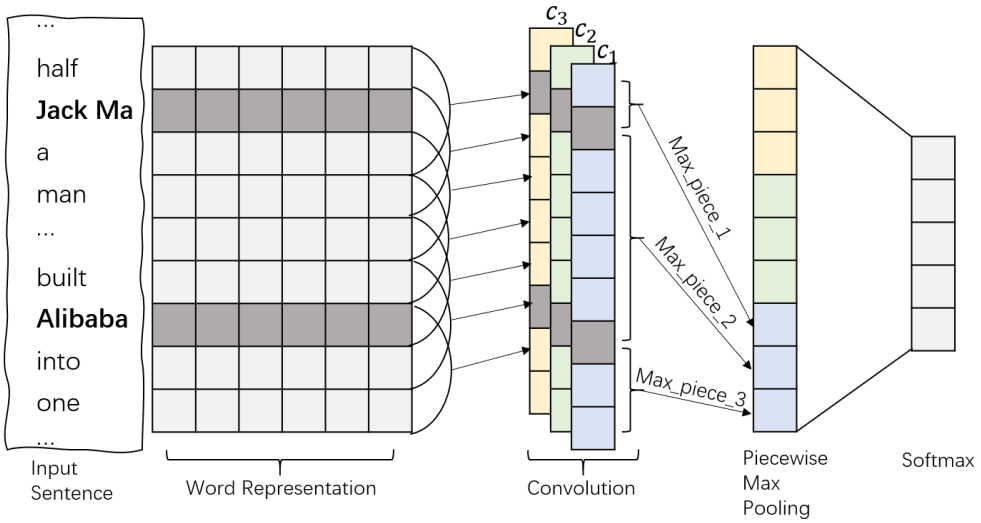

Figure 2: Architecture of Piecewise Convolutional Neural Networks with input sentence 'In just a decade and half Jack Ma, a man from modest beginnings founded and built Alibaba into one of the world's largest companies'.

Since there are two entities in the sentence, the sentence produces two encodings of the same length as the sentence:

pos_1:$[-6,-5,-4,-3,-2,-1,0,1,2,3...]$, where 0 is the location of the first entity '**Jack Ma**';

pos_2:$[-15,-14,-13,...,0,1,2,3...]$, where 0 is the location of the second entity '**Alibaba**';

The model then cuts the text data into three segments at each entity location (position 0), the above sentences will be divided into:

1. In just a decade and a half **Jack Ma**.
2. **Jack Ma**, a man from modest beginnings founded and built **Alibaba**.
3. **Alibaba** into one of the world's largest companies.

3.2.2 Convolution

In order to clearly describe the convolution operation, we first define two matrix convolution operations of the same dimension. For $A, B \in \mathbf{R}^{m_1 \times m_2}$, the convolution operation between A and B is

$$A \otimes B = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} a_{ij} b_{ij} \tag{2}$$

We assume the length of the filter for the convolution is w, and the width of the filter is d, which is equal to the dimension of the word vector, due to the smallest unit of the convolution operation is the word. Then the filter of the model is a two-dimensional matrix $W \in \mathbf{R}^{w \times d}$. We define the input sentence S as $S = \{s_1, s_2, \cdots, s_{|s|}\}$, in which $|s|$ represent the number of words in S, s_i is the vector representation of the i^{th} word. Then define $S_{i,j} = [s_i : s_{i+1} : ...s_j$ as a matrix spliced horizontally from s_i to s_j, thus the convolution operation between the sentence S and the filter produces a vector $c \in \mathbf{R}^{|s|-w+1}$:

$$c_j = W \otimes S_{j:j+w-1} \tag{3}$$

In which $1 \leqslant j \leqslant |S| - w + 1$.

However, in order to capture more abundant text features, $n(n > 1)$ filters will be used in the real world experiment, so the convolution parameter is an n-dimensional tensor composed by n matrices, the whole convolution operation can be expressed as:

$$c_{i,j} = W_i \otimes S_{j:j+w-1} \tag{4}$$

In which $1 \leqslant i \leqslant n, 1 \leqslant j \leqslant |S| - w + 1$. The final convolution vector is

$$c_i = \{c_{i,1}, c_{i,2}, \cdots, c_i, |s| - w + 1\} (1 \leqslant i \leqslant n) \tag{5}$$

Figure 2 shows an example of using three different filters.

3.2.3 Piecewise Max Pooling

Conventional max-pooling divides the text to be processed into several small parts of the same size without overlapping. In each small part, only the largest number of the region is taken, and then the original structure is retained to output after discarding other nodes. At first, CNN was widely used in the field of Computer Vision, however, different from images, text has different main features under different situations and tasks. Therefore, the disadvantage of conventional max pooling is that it is difficult to identify the important information needed in the text, and dropout might cause the loss of key information.

In the piecewise max-pooling, we segment the sentence from the position of entities that needs to extract the relation, then take the maximum value *max_piece_n* (n is the number of segments) in each segment, and perform the same operation for all convolution vectors. Then the extracted maxima are spliced into a new vector, and a nonlinear operation is performed on the vector. Finally, the vector obtained is represented as the feature of the current text sentence.

3.2.4 Dropout and Softmax Classifier

The conventional neural networks connection mode is full connection, the connection method of Dropout algorithm is to randomly set the original input data (in this paper, the result vector of segmented pooling) to 0 according to a certain proportion, and only other elements without 0 are involved in the calculation and connection.

We assume that only one sample is taken for each parameter update. The process can be described as follows: Firstly, set some elements of the input vector to 0 according to the proportion ρ, and the elements without 0 are involved in the operation and optimization of the classifier; Then the input vector for the second sample is accepted, and the training elements are selected in the same way as random 0 setting until all samples have learned once. Since only one sample is entered at a time, the 0 setting is random, so the network weight parameters are different for each update. In the process of final prediction, the parameters of the whole network are multiplied by $1-\rho$, and the final classifier network parameters can be obtained. The final network parameters of the Dropout algorithm are composed of the parameters of multiple modules, which is a process to extract the important information and remove the useless information, so it has better generalization ability.

Assume the vector generated by piecewise max-pooling is $c^{'}$, the way the drop-out algorithm sets its element to 0 can be represented by Bernoulli distribution, which first used to generate a binary vector (elements only have 0 or 1) r with dimensions equal to $c^{'}$:

$$r \sim Bernoulli(\rho) \tag{6}$$

The vector entered into the softmax classifier is denoted as:

$$c'_d = c' \cdot r \tag{7}$$

The network parameters of the softmax classifier are defined as W_c and the bias vector is b_c, then the output of the network is:

$$o = f\left(W_c c'_d + b_c\right) \tag{8}$$

Where f is sigmoid function or *tanh* function, then the probability that the current sentence belongs to the i^{th} category is:

$$p(i|S) = e^{o_i} / \sum_{j=1}^{N} e^{o_j} \tag{9}$$

Where o_i represents the i^{th} element of vector o, and N represents the number of categories.

3.2.5 Objective Function

The parameters to be optimized include two parts: word vector and network parameters.The word vector is defined as E, the parameter of convolution operation is $\hat{W}$, and the parameter of classifier is W_c, set $\theta = \{E, \hat{W}, W_c\}$. For samples of training set $\Omega = \{(S_1, y_1), (S_2, y_2), \cdots, (S_{|\Omega|}, y_{|\Omega|})\}$, in which S_i represents the i^th sentence and y_i represents its category label. $|\Omega|$ represents the number of training set samples, $p(y_i|S_i, \theta)$ represents the probability of classifying a sentence S_i into the relation category y_i when the parameter theta is given. Then the objective function of optimization is:

$$L = \sum_{i=1}^{|\Omega|} \log p(y_i|S_i, \theta) + \lambda\theta_2^2 \tag{10}$$

Where λ is the parameter to the regular term.

Although PCNN has made great progress in distant supervision relation extraction, it still has the following deficiencies:

First, if there is a relation between two entities, at least one sentence will express it. According to this assumption, PCNN only selects the most likely sentence of each entity pair in the training and prediction, ignoring other meanings of the entity pair in other sentences, which is difficult to apply in the realistic context of polysemy.

Secondly, PCNN regards remote supervision as a single label task, only selecting the most possible relation for each entity pair, but in reality, there are many cases of multiple relations between an entity pair, e.g.: 'As **Alice**'s best friend and husband, **Bob** and she lived happily after marriage for more than 50 years', in the above sentence, entity **Alice** and entity **Bob** have two relations: 'spouse' and 'friend'. Therefore, Multi-instance and Multi-labeled Relation Extraction can be seen as our future work.

4 EXPERIMENT

We use several models based on Piecewise Convolutional Neural Networks for distant supervision. Since

PCNN is the improved method of conventional Convolutional Neural Networks (CNN), we compared this 2 methods together to verify the advantages of PCNN, and use 3 different selectors: Attention, Average, Maximum respectively to evaluate the precision value in top 100 categories, top 200 categories and top 300 categories (P@100, P@200, and P@300).

In which the essence of the Attention mechanism is the process of assigning attention weight to features. The function of the Attention mechanism is that it can learn the hidden states that need to be used and how the weight of hidden states be allocated. Attention selectors are suitable for tasks that require grasping key points or contain keywords. Average refers to taking the mean as the final value, which is suitable for cases where all sentences of information need to be considered comprehensively. Maximum means that the maximum value is taken as the final value, which is applicable to the case with large noise.

For the relation extraction task through this study, we use the NYT-FB dataset, which is an open dataset obtained by the Freebase (Bollacker et al., 2008) repository aligned with the text of the New York Times. The training data were gotten from the New York Times in 2005 and 2006, and the test library data were the Times text in 2007. NYT-FB data set with a total of 53 kinds of relations, a total of 695,059 data (the training set contains 522,611 statement, nearly 80% of the sentences in the training data labels for N/A, test statements set contains 172,448), by combining FreeBase with NYT corpus for entity linking, relation alignment and other operations for annotation, a widely used relation extraction data set is finally obtained.

5 RESULT

Table 1, Figure 3 and Figure 4 show the result and its Precision-Recall curve of Distant Supervision for Relation Extraction task. From Table 1, we can find that Piecewise Convolutional Neural Networks with Attention Mechanism shows the best result with the precision of 76.24% in the top 100 relation categories in the NYT dataset and achieves the 71.39% precision on average. Compare with conventional CNN, PCNN gets a 3.2% improvement with the Attention mechanism, a 2.87% improvement with the Average selector, but a 1.15% decrease with maximum selector. The slightly decrease mainly due to the Piece-wise operation cut the whole text into 3 pieces, and Maximum selector chooses the max value in each piece, and get the average in the final calculation. The key information is weakened according to the piece-wise process, and finally get the worse result than conventional CNN. In addition, as the range of test categories increased from 100 to 300, the accuracy of all models decreased. This reason is obvious, as the range increased, the uncertainty increased, and eventually, the accuracy decreased according to the indeterminacy.

Table 1: Results of Distant Supervision for Relation Extraction task.

Model	Range	Attention	Average	Maximum
CNN	P@100	0.7228	0.6931	0.7723
	P@200	0.6716	0.6766	0.6866
	P@300	0.6512	0.6412	0.6312
	Mean	0.6819	0.6703	0.6967
PCNN	P@100	***0.7624***	0.7426	0.7327
	P@200	0.7015	0.6965	0.6816
	P@300	0.6777	0.6578	0.6412
	Mean	0.7139	0.6990	0.6852

Figure 3 compares each selector in CNN and PCNN, which is intended to compare the differences between the conventional CNN and the improved PCNN in the same model, the first figure (CNN and PCNN with attention mechanism) has the most obvious effect. PCNN represented by the red line is significantly higher than CNN represented by the blue line at first, however, due to the increasing recall rate, the gap gradually becomes smaller. The comparison of the middle figure is not as obvious as the first figure, mainly due to the average selector has the function of weakening the emphasis, which can narrow the effect gap caused by the essence of the model to some extent. Although the third figure shows a contradictive result, the reason has been discussed in the last paragraph. However, from a global perspective, the Piecewise Convolutional Neural Networks with Attention mechanism works best in Distant Supervision for Relation Extraction task.

Figure 4 collects the Precision-Recall curves of all models on one figure, which intends to make the comparison between each model more intuitive and obvious. We can learn from the line chart, the red line represents the PCNN-Attention model accuracy is better than other models, at the same time CNN-Maximum model is showed relatively poor results, in which the precision of the PCNN-Maximum model improved significantly with the increase of recall.

In our proposed model, we use PCNN with the Attention mechanism and get the precision of 76.24%, which is higher than Convolutional Neural Networks with 3.96% precision improvement.

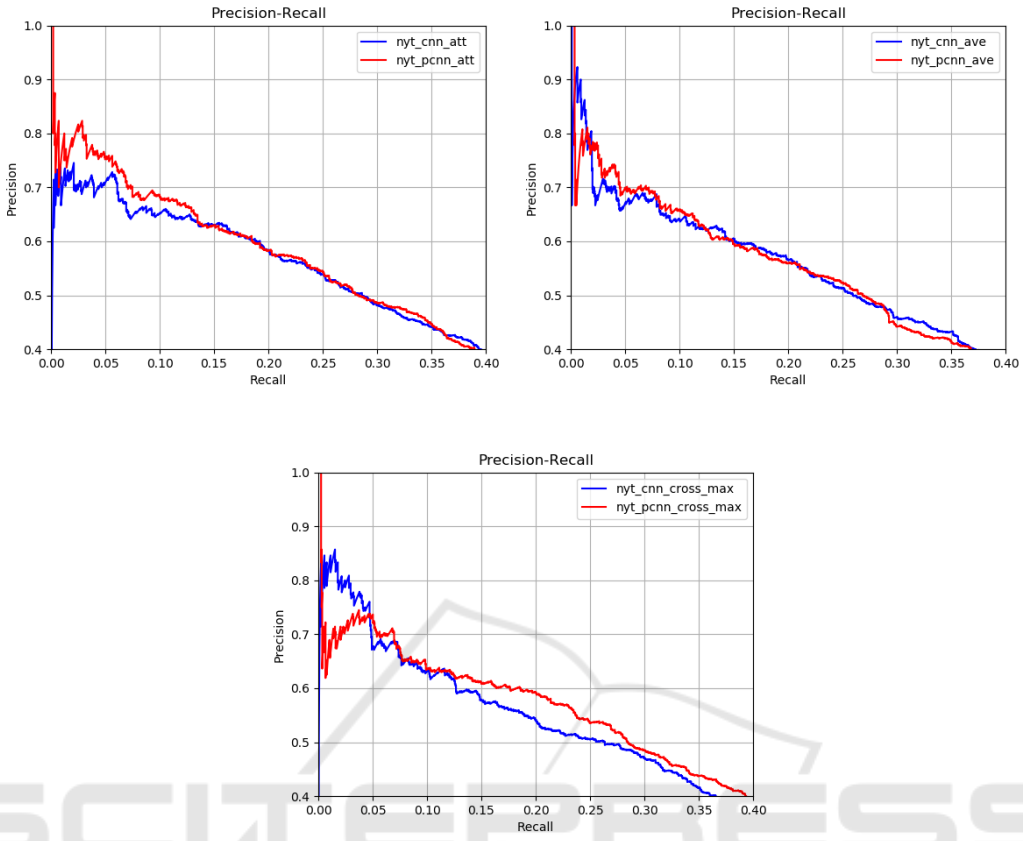

Figure 3: Comparison of CNN and PCNN model in different selectors.

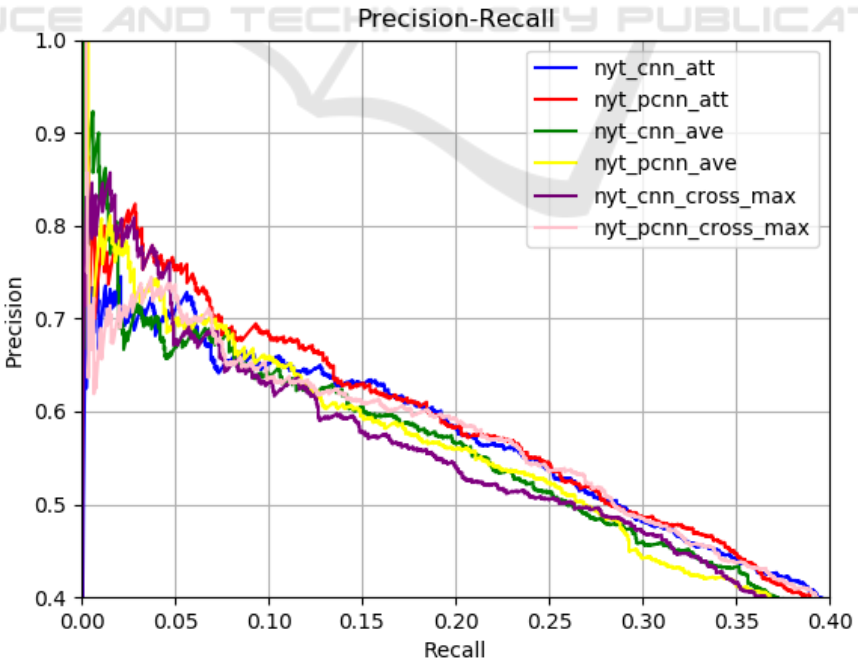

Figure 4: Precision-Recall of All Tested Models.

6 CONCLUSION

Relation extraction often faces the problem of lacking a sufficient amount of training data, so effective label learning under weak supervision becomes extremely challenging. The Distant Supervision as a novel idea that can solve the problem of training data annotation missing in the existing relation extraction task to a certain extent.

In this paper, we proposed a Distant Supervision method based on Piecewise Convolutional Neural Networks with Attentional mechanism for automatically annotating unlabeled data on Relation Extraction task, and achieved the highest precision is 76.24% on NYT-FB (New York Times - Freebase) dataset (top 100 relation categories). The results proved that our method performed better than CNN-based models in most cases. This helps with a more precise deep learning-based Relationship Extraction task.

ACKNOWLEDGMENT

We are grateful to VC Research (Funding No. VCR 0000040) to support this work.

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 the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. AcM.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537.

He, D., Zhang, H., Hao, W., Zhang, R., Chen, G., Jin, D., and Cheng, K. (2017). Distant supervised relation extraction via long short term memory networks with sentence embedding. *Intelligent Data Analysis*, 21(5):1213–1231.

Huang, Y. Y. and Wang, W. Y. (2017). Deep residual learning for weakly-supervised relation extraction. *arXiv preprint arXiv:1707.08866*.

Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

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)*, pages 2124–2133.

McDonald, R. and Nivre, J. (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 (EMNLP-CoNLL)*.

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.

Ren, X., Wu, Z., He, W., Qu, M., Voss, C. R., Ji, H., Abdelzaher, T. F., and Han, J. (2017). Cotype: Joint extraction of typed entities and relations with knowledge bases. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1015–1024. International World Wide Web Conferences Steering Committee.

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.

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*, pages 1753–1762.

Zeng, D., Liu, K., Lai, S., Zhou, G., and Zhao, J. (2014). Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344.