The Thirty-First International Florida
Artificial Intelligence Research Society Conference (FLAIRS-31)

Chinese Relation Classification via Convolutional Neural Networks

Linrui Zhang, Dan Moldovan
The University of Texas at Dallas
800 West Campbell Road ; MS EC31, Richardson, TX 75080 U.S.A
linrui.zhang@utdallas.edu, moldovan@hlt.utdallas.edu

Abstract

Relation classification is an important task in natural language processing. Traditional relation classification techniques suffer from extensive use of linguistic features and external toolkits. In recent years, deep learning models that can automatically learn features from text are playing a more essential role in this area. In this paper we present a novel convolutional neural network (CNN) approach along shortest dependency paths (SDP) for Chinese relation classification. We first propose a baseline end-to-end model that only takes sentence-level features, and then improve its performance by joint use of pre-extracted linguistic features. The performance of the system is evaluated on the ACE 2005 Multilingual Training Corpus Chinese dataset. The baseline model achieved a 74.93% F-score on six general type relations and a 66.29% F-score on eighteen subtype relations, and the performance improved 10.71% and 13.60% respectively by incorporating linguistic features into the baseline system.

Introduction

The goal of relation classification is to predict a semantic relation r between a head entity e_h and a tail entity e_t from a given sentence s. For instance, in the phrase "U.N. Arms Inspectors", the relation between the PER entity "Inspectors" and the ORG entity "U.N." is ORG-AFF, which indicates that the "Inspectors" are affiliated to the organization "U.N.". This relation could be represented as ORG-AFF (U.N., Inspectors). To be more specic, "Inspectors" are the employees of "U.N.", so the relation between them could be sub-categorized into EMPLOYMENT (U.N., Inspectors). The ACE 2005 Multilingual Training Corpus defines six general relations between entity pairs, such as "PART-WHOLE", "ORG-AFF", and "PER-SOC", and eighteen subtype relations such as "GEOGRAPHICAL", "MEMBERSHIP", and "BUSINESS".

In the literature, two popular methods have dominated this area: the feature-based approach and the kernel-based approach. The feature-based methods treat the task as a classification problem. Linguistic features are first extracted from sentences and then fed into a machine learning classifier such as a SVM for classification. Such work includes (Kambhatla 2004), (Boschee, Weischedel, and Zamanian 2005), (GuoDong et al. 2005), and (Jiang and Zhai 2007). However, if the features are not well selected, the error will propagate through the entire system, leading to errors in the results. More importantly, most of the effort in feature-based approaches is spent on feature engineering work, and they also greatly depend on the availability of external toolkits and resources. The kernel-based approaches compute kernel functions to measure the similarity between two data objects. Such work includes (Zelenko, Aone, and Richardella 2003), (Culotta and Sorensen 2004), (Bunescu and Mooney 2005), (Zhang et al. 2006), (Zhou et al. 2007), (Wang 2008) and (Plank and Moschitti 2013). The key issue of the kernel-based approaches is the slow training and prediction time, making them inefficient for processing big data.

In addition to feature/kernel-based approaches, several neural network-based approaches have been proposed in recent years, including recursive neural network-based (Socher et al. 2012), (Ebrahimi and Dou 2015), recurrent neural network-based (Xu et al. 2015b), and convolutional neural network-based approaches (Zeng et al. 2014), (Xu et al. 2015a), (Nguyen and Grishman 2015). These deep learning models have all focused on English datasets, but since deep learning models have the advantage of being able to do end-to-end training with little or no domain knowledge needed, the work on English should be easily extended to other languages. Intrigued by this idea, we developed a novel convolutional neural network model together with shortest dependency path for Chinese relation classification.

Convolutional neural networks (CNN) may be viewed as computing vectors for all the N-grams of a sentence. However, some elements of sentences, such as auxiliary words, are less informative than others, and may even introduce noise into the entire system. For instance, in the sentence "Then you could take the train or bus back to Boston", if the relation between "you" and "bus" was to be extracted, the words that really matter are "you", "take", "bus" and "train"; and compared with these words, "Then", "could", "to", and "Boston" are less important. Shortest dependency paths (SDP) can completely remove redundant and irrelevant words from sentences, and this has proven to be very useful for relation extraction (Bunescu and Mooney 2005), (Xu et al. 2015b) and (Ebrahimi and Dou 2015). In this paper, we use CNN jointly with SDP, so the model may be viewed as computing vectors for the N-grams of the SDP

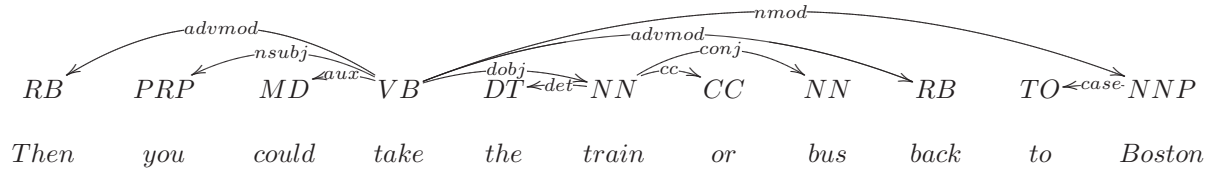

Figure 1: The dependency graph of the sentence "Then you could take the train or bus back to Boston."

instead of the whole sentence. We believe this will improve performance on measures of both accuracy and speed.

Related Work

Research on Chinese relation extraction and classification is quite limited compared to the progress with English. Traditional kernel-based approaches include (Che et al. 2005), (Kebin et al. 2007), (Huang, Sun, and Feng 2008), (Yu et al. 2010) and (Dandan, Yanan, and Longhua 2012), and feature-based approaches include (Li et al. 2008), (Zhang et al. 2008), (Zhang et al. 2011) and (Chen, Zheng, and Zhang 2014).

A convolutional neural network (LeCun et al. 1998) is a deep feed-forward artificial neural network that has been applied successfully to many NLP tasks. In relation classification, (Zeng et al. 2014) made the first attempt to apply CNNs to relation classification problems, and (Xu et al. 2015a) proposed a CNN-based approach with negative sampling. (Santos, Xiang, and Zhou 2015) tackled the relation classification task using a convolutional neural network that performs classification by ranking.

Approach

Shortest Dependency Paths

Dependency parsing captures the dependence relations between words. In dependency parsing, the dependency relations and words will form a dependency graph. The edges are the dependency relations and the vertices are the words. Finding the shortest dependency path between words may be mapped into finding the shortest path between two vertices in the dependency graph. Figures 1 shows the dependency graph of the sentence "Then you could take the train or bus back to Boston".

Feature Representation

The baseline end-to-end CNN model only takes sentence-level features which are the word embeddings in the shortest dependency path. Since CNNs can only work with fixed-length input, but the lengths of SDPs are variable, we unified the input sequences by padding shorter sequences and trimming longer sequences. Three extra features are also extracted from the text: the Part-of-Speech tags, entity type and entity subtype. We randomly initialized their embeddings and concatenated them together into the corresponding word embeddings. The mathematical representation of a single word vector is W_i = [w_i, p_i, t_i, s_i], where w_i, p_i, t_i and s_i are the word, Part-of-Speech tag, entity type, and entity subtype embeddings of the i-th word in the SDP. The single input matrix is the concatenation of all the word vectors in the SDP, and its dimensionality is (m_w + m_p + m_t + m_s) × n, where m is the length of the corresponding feature, (m_w + m_p + m_t + m_s) is the length of a single word vector and n is the number of words of the fixed-length SDP.

The Structure of the Model

We implemented a system based on (Kim 2014). The structure of the system has three layers: a convolutional layer, a max pooling layer, and a softmax layer. In the convolutional layer we implemented three filters with window sizes of 1, 2 and 3. Figure 2 shows the basic structure of the system for the sentence "Then you could take the train or bus back to Boston".

Convolutional Layer

The input of the convolutional layer is in three dimensions: (m_w + m_p + m_t + m_s) × n × b, where b is the batch size. A convolutional filter w may be viewed as a weighted matrix that passes over an input matrix. During each time step it generates a score from the application of the convolutional operator of the filter matrix w and a portion of the input matrix x. After the filter passes over the entire input matrix, it will end up with a score sequence C = [c_1, c_2, ... , c_{n-h+1}]. The mathematical representation of the convolutional layer is as follows:

$$C_i = f(W^T X_{i:i+h-1} + b)$$

f is the activation function, W is the weighted matrix of the convolutional filter, X is the input matrix, h is the window size of the filter, b is a bias term, and i is the i-th time step.

Pooling Layer

A pooling layer is used to further extract features from the output of the convolutional layer. A popular strategy is to use max pooling, which aims to identify the most significant feature from the score sequence C. In mathematics, the score sequence C of each filter in the convolutional layer will be passed through a Max function to pick up the max score in this sequence. The pooling score is calculated as $Max(x) = max([c_1, c_2, ..., c_{h-w+1}])$.

Softmax Layer

In deep learning, a softmax layer is usually used as the output layer in classification problem. The pooling scores for each filter will be flattened and concatenated into a single vector V = [S_1, S_2, ... , S_n], where S_i is the flattened pooling score of each filter and n is the number of filters. The

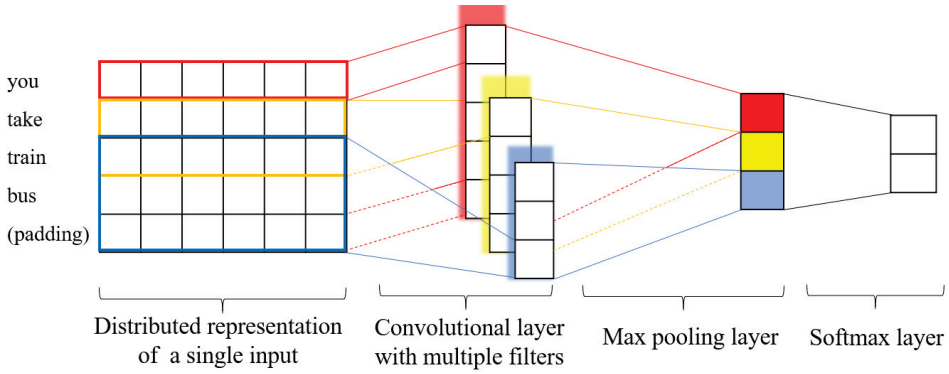

Figure 2: The basic structure of the CNN model with the example "Then you could take the train or bus back to Boston".

softmax layer takes V as input and outputs the probability distribution of the candidate classes, which in our case are the six general type relations and the eighteen subtype relations between the entities.

The training goal is to minimize the cross-entropy error between the probabilistic distribution of the predicated relations and the one-hot representation of the gold annotated relations. The mathematic representation of the training process is as follows:

$$E(\theta) = -\sum_{n}\sum_{k} t_n^k \log y_n^k + \frac{\lambda}{2}||\theta||^2$$

t is the gold annotated relations. y is the predicated relations from the softmax layer. λ is the regularization rate. θ is the model parameters we are trying to learn.

Experiment and Results

The system was tested with the ACE 2005 Multilingual Training Corpus Chinese dataset. We extracted 7985 positive instances and divided them into 80% for training and 20% for testing. We used pretrained Chinese word embeddings on 10G Chinese Wikipedia dataset, and set the word embedding length to 60. We randomly initialized the POS, relation type and relation subtype embeddings and set the length to 20 for each. We set the output space of each filter to 16 and used the Adam optimizer, setting the learning rate to 0.01. The maximum number of iterations was set to 2000.

Experiment Results

We designed two experiments. The first experiment aims to illustrate the influence of linguistic features on the CNN model. Table 1 shows the system performance with only sentence-level features and the performance increasing rate by incorporating various linguistic features. From the results we can conclude that the performance of the baseline system may be greatly improved by joining linguistic features, among which entity type is the most important.

The second experiment aims to illustrate the influence of various filter window sizes on the performance of the model. We used window sizes of one (unigram filter), two (bigram filter), three (trigram filter) and a combination of them. Table 2 shows the performance of system with various filter window sizes. From the results, we could observe that, among single filters, the unigram filter achieved the best results. However, the system achieved the best performance using two mixed filters with window sizes of one and two, which outperforms the system's performance using any single filter.

Features	Type F1	SubType F1
WordEmbedding	74.93%	66.29%
+POS	-0.37%	+2.63%
+Entity Type	+8.04%	+9.33%
+Entity Subtype	+5.43%	+8.64%
Overall	85.64%	79.89%

Table 1: System performance with various linguistic features.

No.	Window	Type F1	SubType F1
1	1	85.24%	80.19%
2	2	85.33%	79.62%
3	3	84.84%	77.20%
4	1-2	**87.23%**	**81.46%**
5	1-3	86.84%	80.04%
6	2-3	85.02%	79.76%
7	1-2-3	85.64%	79.89%

Table 2: System performance with various filter window size.

Conclusion

In this paper, we presented a novel CNN model for Chinese relation classification. We systematically analyzed the influence of linguistic features and filter window size on the CNN model of the task. The experimental results confirmed the effectiveness and advantages of the CNN approach when applied to Chinese relation classification.

References

Boschee, E.; Weischedel, R.; and Zamanian, A. 2005. Automatic information extraction. In *Proceedings of the International Conference on Intelligence Analysis*, volume 71. Citeseer.

Bunescu, R. C., and Mooney, R. J. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, 724–731. Association for Computational Linguistics.

Che, W.; Jiang, J.; Su, Z.; Pan, Y.; and Liu, T. 2005. Improved-edit-distance kernel for chinese relation extraction. In *Proceedings of IJCNLP*, 132–137.

Chen, Y.; Zheng, Q.; and Zhang, W. 2014. Omni-word feature and soft constraint for chinese relation extraction. In *ACL (1)*, 572–581.

Culotta, A., and Sorensen, J. 2004. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd annual meeting on association for computational linguistics*, 423. Association for Computational Linguistics.

Dandan, L.; Yanan, H.; and Longhua, Q. 2012. Exploiting lexical semantic resource for tree kernel-based chinese relation extraction. *Natural Language Processing and Chinese Computing* 213–224.

Ebrahimi, J., and Dou, D. 2015. Chain based rnn for relation classification. In *HLT-NAACL*, 1244–1249.

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

Huang, R.; Sun, L.; and Feng, Y. 2008. Study of kernel-based methods for chinese relation extraction. *Information Retrieval Technology* 598–604.

Jiang, J., and Zhai, C. 2007. A systematic exploration of the feature space for relation extraction. In *HLT-NAACL*, 113–120.

Kambhatla, N. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, 22. Association for Computational Linguistics.

Kebin, L.; Fang, L.; Lei, L.; and Ying, H. 2007. Implementation of a kernel-based chinese relation extraction system [j]. *Journal of Computer Research and Development* 44(8):1406–1411.

Kim, Y. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

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

Li, W.; Zhang, P.; Wei, F.; Hou, Y.; and Lu, Q. 2008. A novel feature-based approach to chinese entity relation extraction. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, 89–92. Association for Computational Linguistics.

Nguyen, T. H., and Grishman, R. 2015. Relation extraction: Perspective from convolutional neural networks. In *VS@ HLT-NAACL*, 39–48.

Plank, B., and Moschitti, A. 2013. Embedding semantic similarity in tree kernels for domain adaptation of relation extraction. In *ACL (1)*, 1498–1507.

Santos, C. N. d.; Xiang, B.; and Zhou, B. 2015. Classifying relations by ranking with convolutional neural networks. *arXiv preprint arXiv:1504.06580*.

Socher, R.; Huval, B.; Manning, C. D.; and Ng, A. Y. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, 1201–1211. Association for Computational Linguistics.

Wang, M. 2008. A re-examination of dependency path kernels for relation extraction. In *IJCNLP*, 841–846.

Xu, K.; Feng, Y.; Huang, S.; and Zhao, D. 2015a. Semantic relation classification via convolutional neural networks with simple negative sampling. *arXiv preprint arXiv:1506.07650*.

Xu, Y.; Mou, L.; Li, G.; Chen, Y.; Peng, H.; and Jin, Z. 2015b. Classifying relations via long short term memory networks along shortest dependency paths. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1785–1794.

Yu, H.; Qian, L.; Zhou, G.; and Zhu, Q. 2010. Chinese semantic relation extraction based on unified syntactic and entity semantic tree. *Journal of Chinese Information Processing* 24(5):17–23.

Zelenko, D.; Aone, C.; and Richardella, A. 2003. Kernel methods for relation extraction. *Journal of machine learning research* 3(Feb):1083–1106.

Zeng, D.; Liu, K.; Lai, S.; Zhou, G.; Zhao, J.; et al. 2014. Relation classification via convolutional deep neural network. In *COLING*, 2335–2344.

Zhang, M.; Zhang, J.; Su, J.; and Zhou, G. 2006. A composite kernel to extract relations between entities with both flat and structured features. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 825–832. Association for Computational Linguistics.

Zhang, P.; Li, W.; Wei, F.; Lu, Q.; and Hou, Y. 2008. Exploiting the role of position feature in chinese relation extraction. In *LREC*.

Zhang, P.; Li, W.; Hou, Y.; and Song, D. 2011. Developing position structure-based framework for chinese entity relation extraction. *ACM Transactions on Asian Language Information Processing (TALIP)* 10(3):14.

Zhou, G.; Zhang, M.; Ji, D.; and Zhu, Q. 2007. Tree kernel-based relation extraction with context-sensitive structured parse tree information. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.