Copyright © 2019 Tech Science Press JIHPP, vol.1, no.2, pp.69-75, 2019

Joint Self-Attention Based Neural Networks for Semantic Relation Extraction

Jun Sun[1], Yan Li[1], Yatian Shen[1, *], Wenke Ding[1], Xianjin Shi[1], Lei Zhang[1], Xiajiong Shen[1] and Jing He[2]

Abstract: Relation extraction is an important task in NLP community. However, some models often fail in capturing Long-distance dependence on semantics, and the interaction between semantics of two entities is ignored. In this paper, we propose a novel neural network model for semantic relation classification called joint self-attention bi-LSTM (SA-Bi-LSTM) to model the internal structure of the sentence to obtain the importance of each word of the sentence without relying on additional information, and capture Long-distance dependence on semantics. We conduct experiments using the SemEval-2010 Task 8 dataset. Extensive experiments and the results demonstrated that the proposed method is effective against relation classification, which can obtain state-of-the-art classification accuracy just with minimal feature engineering.

Keywords: Self-attention, relation extraction, neural networks.

1 Introduction

Relation extraction is a fundamental task in information extraction, which has important applications in question answering, information retrieval, big data analysis etc. Traditional approaches to relation extraction take entity recognition as a predecessor step in the pipeline predicting relations between given entities. In recent years, there has been a surge of interest in relation extraction task.

The traditional methods are mainly based on supervised relation extraction [Suchanek, Ifrim and Weikum (2006); Qian, Zhou, Kong et al. (2008)], which usually suffer from the issue that lacks sufficient labelled relation-specific training data. If a large number of data sets are tagged, it is a time-consuming and laboring work. Meanwhile, artificial feature extraction methods need some tools of natural language processing, which lead to the propagation of the errors in the existing tools and hinders the performance of some systems [Bach and Badaskar (2007)].

Recently, neural network models have been verified to be effective against classifying relations between plain text [Collobert, Weston, Bottou et al. (2011)]. Moreover, some

[1] School of Computer and Information Engineering, Henan University, Kaifeng, 475000, China.

[2] The Corporate and Investment, Bank Technology, J. P. Morgan Chase N. A. 25 Bank St, Canary Wharf, London, E145JP, United Kingdom.

[*] Corresponding Author: Yatian Shen. Email: sy602@126.com.

JIHPP. doi:10.32604/jihpp.2019.06357 www.techscience.com/jihpp

researchers also explore different deep learning methods [Meng, Rice, Wang et al. (2018); Xiong, Shen, Wang et al. (2018)]. in the field of relation extraction, such as recursive neural network [Hashimoto, Miwa, Tsuruoka et al. (2013)], convolutional deep neural network [Huang and Shen (2016)], long short-term memory neural network (LSTM) [Xu, Mou, Li et al. (2015)] and adversarial learning [Zeng, Dai, Li et al. (2018)]. Still, these models often fail in capturing long distance dependencies on semantics, and the interaction between semantics of two entities is ignored. How to effectively model the interaction between semantics of two entities in a sentence and capture Long-distance dependence on semantics are important task.

Inspired by the idea mentioned above, we encode the text segment of every entity to its feature representation to bi-LSTM [Kiperwasser and Goldberg (2016)]. Then, we use self-attention mechanism to get semantic representation of text segment that is related to every entity, which can attention the sentence itself to extract relevant information and capture Long-distance dependence on semantics to capture the interaction between semantics of two entities in a sentence we propose joint self-attention bi-LSTM(SA-Bi-LSTM) to model the internal structure of the sentence to obtain the importance of each word in the sentence without relying on additional information. Empirical results from the SemEval-2010 Task 8 dataset show that the proposed approach just with minimal feature engineering obtains state-of-the-art classification accuracy about 85.3% F1 value.

The main contribution of this paper can be summarized as follows:

(1) In order to preserve the contextual information about the entity, we encode the text segment of every entity to its semantic representation through a bi-LSTM.

(2) To capture long distance dependencies of semantics, and the interaction between semantics of two entities we propose joint self-attention bi-LSTM(SA-Bi-LSTM) to model the internal structure of the sentence to obtain the importance of each word with the sentence without relying on additional information.

(3) We conduct experiments using the SemEval-2010 Task 8 dataset. Extensive experiments and the results demonstrate that the proposed joint self-attention bi-LSTM(SA-Bi-LSTM) is effective for relation classification, which can obtain classification result from 85.3% F1 value.

2 Methodology

In this part, we first introduce the basic self-attention model, and then introduce the semantic relation extraction model of joint self-attention in detail.

2.1 Self-attention

Self-attention is also called intra-attention, it is a special attention mechanism. The main application of our model is multi-head attention. Multi-head attention is a variant of scaled dot-product attention. Scaled dot-product attention is that adds a scale dot product function of the basis of dot-product attention. Give the query matrix and key matrix with dimension dk and the value matrix with dimension dv as input, the calculation formula for scaled dot-product attention is as follows:

$$\mathrm{Attention}(Q,K,V) = soft\max(\frac{QK^T}{\sqrt{d^k}})V \tag{1}$$

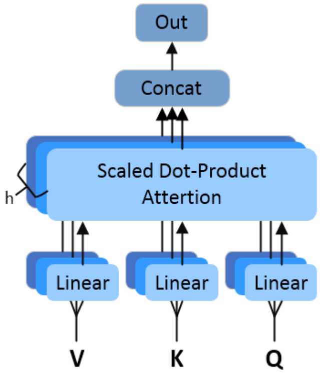

Figure 1: The structure of the multi-head attention

The structure of the multi-head attention is shown in Fig. 1. The specific calculation formula is as follows:

$$\mathrm{head}_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2}$$

$$\mathrm{Multi}Head(Q,K,V) = Conect(head_1,\cdots,head_h)W^O \tag{3}$$

where $\mathrm{W}_\mathrm{i}^Q \in R^{\frac{n\times d}{h}}$, $\mathrm{W}_\mathrm{i}^K \in R^{\frac{n\times d}{h}}$, $\mathrm{W}_\mathrm{i}^\mathrm{V} \in R^{\frac{n\times d}{h}}$ and $\mathrm{W}^\mathrm{O} \in R^{d\times d}$.

2.2 Joint self-attention Bi-LSTM(SA-Bi-LSTM)

The joint self-attention semantic relationship extraction model can explore the internal structure of the sentence to obtain the importance of each word of the sentence without relying on additional information. The model can directly calculate the dependence on words with considering the distance between words, so as to get the influence of each word with the semantics of sentences. We believe that capturing the contribution to different words to sentence semantics in a sentence is effective against improving the accuracy of relation classification.

Specifically, we use two Bi-LSTM neural networks to model two entities respectively. Suppose we have a sentence with n tokens represented in a word embedding sequence:

$$\mathrm{S} = (w_1, w_2, w_3, \cdots, e_1, \cdots, e_2, \cdots, w_n) \tag{4}$$

where e_1 , e_2 represent the entities in the sentence. For the context of entities, we will use Bi-LSTM modeling into sequential word embedding:

$$\mathrm{context}_{pre} = (w_1, w_2, w_3, \cdots, e_1, \cdots) \tag{5}$$

$$\mathrm{context}_{fol} = (\cdots, e_2, \cdots, w_{n-2}, w_{n-1}, w_n) \tag{6}$$

Here we use two different Bi-LSTM neural networks to encode the context, and then we keep all of these hidden layers' information:

$$\mathrm{H}_{pre} = (h_1^{pre}, h_2^{pre}, h_3^{pre}, \cdots, h_n^{pre}) \tag{7}$$

$$\mathrm{H}_{\mathrm{fol}} = (h_1^{fol}, h_2^{fol}, h_3^{fol}, \cdots, h_n^{fol}) \tag{8}$$

Then we input H^{pre} and H^{fol} into the multi-head attention module to get the multi-layered attention representation of the sentence. We connect the attention matrix together and input it into the dense net to get the overall feature representation of the context and provide them to the softmax layer to classify the semantic relation. An illustration of the model is shown in Fig. 2.

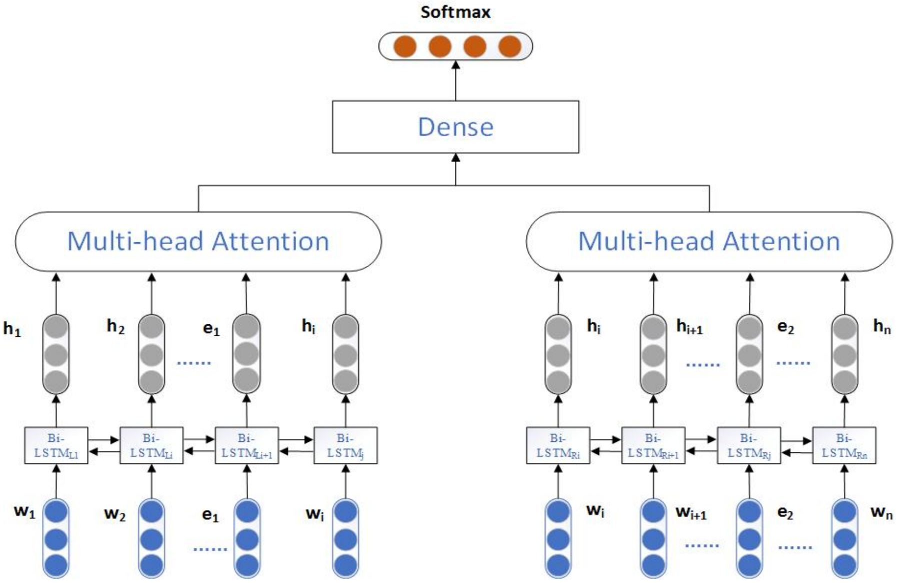

Figure 2: Bi-LSTM model frame diagram of joint self-attention

3 Experiments

3.1 Dataset and evaluation metrics

We use the SemEval-2010 Task 8 dataset Hendrickx et al. [Hendrickx, Kim, Kozareva et al. (2009)] as the required for our experiments. This dataset is public and contains a total of 10,717 annotation examples, including 8,000 training instances and 2,717 test instances. The data has nine directional relationship classes and one other class with no orientation. The data onto SemEval-2010 Task 8 focuses on the semantic relationship between named pairs. For example, thief and screwdriver are in an INSTRUMENT-AGENCY relation in 'A thief who tried to steal the truck broke the ignition with screwdriver'. In the experiment, we do not distinguish the direction of the relationship, using 10 kinds of tags. In order to compare with the previous research results, we used the macro-averaged F1-score value as the evaluation criterion in our experiment.

$$F_1 = \frac{2 * precision * recall}{precision + recall} \tag{9}$$

3.2 Results of comparison experiments

We select some approaches as competitors to be compared with our method in Tab. 1. Kambhatla [Kambhatla (2004)] use traditional features and employ SVM as the classifier. Gormley et al. [Gormley, Yu and Dredze (2015)] proposes feature combination of handcrafted features and word embeddings. Socher et al. [Socher, Huval, Manning et al. (2012)] assigns a matrix to every word in the recursive procedure. Zeng et al. [Zeng, Dai, Li et al. (2018)] used a convolutional neural network to extract features, and Xu et al. [Xu, Feng, Huang et al. (2015)] considered more robuster relation representations from shortest dependency paths. The model of Xu et al. [Xu, Mou, Li et al. (2015)] also takes into account shortest dependency paths. However, Xu et al. [Xu, Mou, Li et al. (2015)] used another neural network structure, which is long short-term memory (LSTM) model.

The model proposed by us is called joint self-attention bi-LSTM(SA-Bi-LSTM) which can obtain the importance of each word in the sentence without relying on additional information, and capture long distance dependencies of semantics. The experiment demonstrate that it is very important for semantic classification, our proposed SA-Bi-LSTM model yields an F1-score of 85.3%, whereas the previous best model achieved only F1-score of 84.1% [Xu, Feng, Huang et al. (2015)].

Table 1: Comparison of the proposed method with existing methods in the SemEval-2010 Task 8 dataset

Model	Feature Sets	F_1
SVM	POS, stemming, syntactic patterns, WordNet	78.8
[Socher, Huval, Manning et al. (2012)]	POS, NER, WordNet	82.4
[Zeng, Liu, Lai et al. (2014)]	position embeddings, WordNet	82.7
[Gormley, Yu and Dredze (2015)]	dependency parsing, NE	83.0
[Xu, Feng, Huang et al. (2015)]	Word embeddings, position embeddings	84.1
[Xu, Mou, Li et al. (2015)]	POS embeddings, WordNet	83.7
Ours	Word2Vec	85.3

Tab. 1 illustrates the macro-averaged F1 measure results for these competing methods along with the resources, features and classifier used by each method. Based on these results, we make the following observations:

It is relatively difficult to manually choose the best feature sets, which depends on human ingenuity and prior NLP knowledge. Socher et al. [Socher, Huval, Manning et al. (2012)] depend on the syntactic tree used in the recursive procedures. Errors in syntactic parsing inhibit the ability of these methods to learn high quality features. The position encoding is also another way of feature extraction, which encode position information from each entity to all the tokens in a sentence. So Zeng et al. [Zeng, Liu, Lai et al. (2014)] gain a lot of improvement about 82.7%. The model Gormley et al. [Gormley, Yu, Dredze et al. (2015)] connects word embedding with arbitrary linguistic structure, as expressed by hand crafted features, which get the advancement of classification result. Xu et al. [Xu, Feng, Huang et al. (2015)] and Xu et al. [Xu, Mou, Li et al. (2015)] learn more robust relation representations to the shortest dependency paths through a convolution neural network and long short-term memory network (LSTM), and two models mentioned above demonstrate effectiveness and practicability of the dependency paths in semantic relation classification task. Our method achieves the best result about 85.3%, and this is the best performance among all of the compared methods. The performance demonstrates the effectiveness of the self-attention mechanism, which can model the internal structure of the sentence to obtain the importance of each word in the sentence without relying on additional information, and capture long distance dependencies of semantics.

4 Conclusion

In this paper, we proposed a novel neural network model for semantic relation classification called joint self-attention Bi-LSTM(SA-Bi-LSTM) to model the internal structure of the sentence, which can obtain the importance of each word of the sentence without relying on additional information and capture Long-distance dependence on semantics. We conduct experiments using the SemEval-2010 Task 8 dataset. Extensive experiments and the results demonstrate that the proposed methods are effective against relation classification, which can obtain state-of-the-art classification accuracy just with minimal feature engineering.

In the future work, we will focus on exploring better neural network structure of eature extraction in relation extraction task. Meanwhile, knowledge base is an important tool for improving relation extraction performance, we will seek better methods of mutual enhancement of knowledge base complement and relation extraction.

References

Bach, N.; Badaskar, S. (2007): A review of relation extraction. *Literature Review for Language and Statistics II.*

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K. et al. (2011): Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, vol. 12, pp. 2493-2537.

Gormley, M. R.; Yu, M.; Dredze, M. (2015): Improved relation extraction with feature-rich compositional embedding models. *Computer Science*.

Hashimoto, K.; Miwa, M.; Tsuruoka, Y.; Chikayama, T. (2013): Simple customization of recursive neural networks for semantic relation classification. *Empirical Methods on Natural Language Processing*, pp. 1372-1376.

Hendrickx, I.; Kim, S. N.; Kozareva, Z.; Nakov, P.; Ó Séaghdha, D. et al. (2009): Semeval-2010 task 8: multi-way classification of semantic relations between pairs of nominals. *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pp. 94-99.

Huang, X.; Shen, Y. (2016): Attention-based convolutional neural network for semantic relation extraction. *26th International Conference on Computational Linguistics: Technical Papers*, pp. 2526-2536.

Kambhatla, N. (2004): Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*, pp. 22.

Kiperwasser, E.; Goldberg, Y. (2016): Simple and accurate dependency parsing using bidirectional lstm feature representations. *Transactions of the Association for Computational Linguistics*, pp. 313-327.

Meng, R.; Rice, S. G.; Wang J.; Sun X. (2018): A fusion steganographic algorithm based on faster R-CNN. *Computers, Materials & Continua*, vol. 55, no. 1, pp. 1.

Qian, L.; Zhou, G.; Kong, F.; Zhu, Q.; Qian, P. (2008): Exploiting constituent dependencies for tree kernel-based semantic relation extraction. *Proceedings of the 22nd International Conference on Computational Linguistics*, vol. 1, pp. 697-704.

Socher, R.; Huval, B.; Manning, C. D.; Ng, A. Y. (2012): Semantic compositionality through recursive matrix-vector spaces. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1201-1211.

Suchanek, F. M.; Ifrim, G.; Weikum, G. (2006): Combining linguistic and statistical analysis to extract relations from web documents. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 712-717

Xiong, Z.; Shen, Q.; Wang, Y.; Zhu, C. (2018): Paragraph vector representation based on word to vector and CNN learning. *Computers, Materials & Continua*, vol. 55, no. 2, pp. 213-227.

Xu, K.; Feng, Y. S.; Huang, S. F.; Zhao, D. Y. (2015): Semantic relation classification via convolutional neural networks with simple negative sampling. *Computer Science*, vol. 71, no. 7, pp. 941-949.

Xu, Y.; Mou, L.; Li, G.; Chen, Y.; Peng, H. et al. (2015): Classifying relations via long short term memory networks along shortest dependency paths. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1785-1794.

Zeng, D.; Liu, K.; Lai, S.; Zhou, G.; Zhao, J. (2014): Relation classification via convolutional deep neural network. *Proceedings of COLING*, pp. 2335-2344.