Lyb3b at SemEval-2018 Task 12: Ensemble-based Deep Learning Models for Argument Reasoning Comprehension Task

Yongbin Li[1,2], **Xiaobing Zhou**[1,*]
[1]Yunnan University, Kunming, Yunnan, P.R. China
[2]Zunyi Medical University, Zunyi, Guizhou, P.R. China
[*] Corresponding author, zhouxb.cn@gmail.com

Abstract

Reasoning is a crucial part of natural language argumentation. In order to comprehend an argument, we have to reconstruct and analyze its reasoning. In this task, given a natural language argument with a reason and a claim, the goal is to choose the correct implicit reasoning from two options, in order to form a reasonable structure of (Reason, Warrant, Claim). Our approach is to build distributed word embedding of reason, warrant and claim respectively, meanwhile, we use a series of frameworks such as CNN model, LSTM model, GRU with attention model and biLSTM with attention model for processing word vector. Finally, ensemble mechanism is used to integrate the results of each framework to improve the final accuracy. Experiments demonstrate superior performance of ensemble mechanism compared to each separate framework. We are the 11th in official results, the final model can reach a 0.568 accuracy rate on the test dataset.

1 Introduction

Argument reasoning comprehension is a crucial part of natural language argumentation, and the realization of argument reasoning requires the understanding of the deep meaning of the text by the computer. At the same time, argument reasoning is also an important evaluation criterion for the understanding of natural language by computer. This paper is based on the argument reasoning comprehension task proposed by (Habernal et al., 2018), which proposed a complex, yet scalable crowdsourcing process, and created a new freely licensed dataset based on authentic arguments from news comments. The dataset consists of three parts: train dataset, validation dataset and test dataset, with the quantity being 1210, 316 and 444 respectively.

The task is formally defined as follow: given an argument consisting of a reason *R* and a claim *C* along with the title and a short description of the debate they occur in, identify the correct warrant *W* from two candidates, the goal is to select the correct warrant *W* that explains reasoning of this particular argument. There are only two options given and only one answer is correct. The key point of the task is that it is difficult to find the answer through the shallow semantics, and the answer is usually implicit.

Being a binary classification task, through preliminary experiment, our approach is combining debate title and description into reason, and splitting a sample {**R (with debate title and description) ; C; W_0; W_1; correct_label**} into two quadruples, which are {**R; C; W_0; label**} and {**R; C; W_1; label**}. On the validation, we employ the same processing mode, determining the matching degree of fit between a quadruples, the highest will be chosen. The four main deep learning(DL) frameworks we employed are based on Convolutional Neural Network(CNN) model, Long Short-Term Memory(LSTM) model, GRU with attention model and Bidirectional Long Short-Term Memory (biLSTM) with attention model, which are based on utilizing word distributed representation on *R* and *C* and *W* respectively, on top of models, a dense layer is used to determine the matching degree.

In our paper, an ensemble mechanism is introduced into neural network (NN) models, we integrate the results of each model to improve the final accuracy. Experiments demonstrate superior performance of ensemble mechanism compared to each separate model. For confirming the effect of ensemble method, we use each separate model as a reference.

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 1137–1141
New Orleans, Louisiana, June 5–6, 2018.

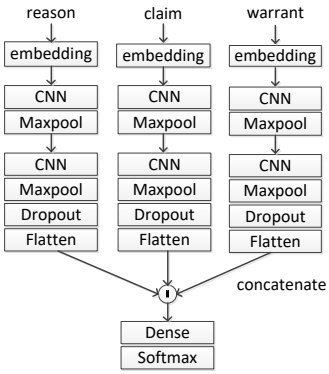

Figure 1: The architecture of CNN framework

2 Related work

An argument consists of a claim and multiple premises was pointed out in (Damer, 2009). Toulmin elaborated on a model of argument in which the reason supports the claim on behalf of a warrant. The abstract structure of an argument then is reason → (since) warrant → (therefore) claim. But, making comprehending and analyzing arguments is hard, for claims and warrants are usually implicit (Freeman, 2011). The phenomenon is referred to as common knowledge (Macagno and Walton, 2014).

Previous, feature extraction and semantic analysis are usually used in natural language argumentation. With the development of neural networks, we adopt the method of word distributed representation from (Hinton, 1986), CNN model refers to (Kim, 2014), the LSTM model from (Hochreiter and Schmidhuber, 1997) and be improved by (Graves et al., 2013), the attention mechanism from (Hermann et al., 2015), eventually converted the task into a classification problem (Wang and Nyberg, 2015). Meanwhile, inspired by the ensemble method in statistical learning domain, we develop a very simple but efficient integrated method.

3 Model description

In this section, we describe the four main proposed deep learning frameworks, the framework architectures are shown in figure 1 to 4. The main idea of these different systems is the same: learn a distributed vector representation of given reason and claim and warrant candidates, and use a dense layer to measure the matching degree.

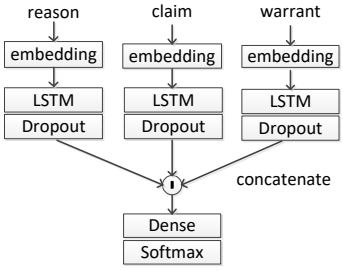

Figure 2: The architecture of LSTM framework

3.1 CNN framework

The first framework is based on CNN model. Step one is to obtain word embedding by pre-trained word2vec (Mikolov et al., 2013), the word embedding provides the distributed representation for each token in reason, claim and warrant candidates respectively. The vectors have dimensionality of 300 and were trained by 100 billion words of Google News, and was initialized from an unsupervised neural language model.

Reason, claim and warrant will be transformed to a word vector matrix and be entered into CNN layer respectively. In order to get more composite representation of semantic features, we adopted double layer CNN model. The numbers of filters are 64 and 32, and the filter size is set as 3, after each CNN layer, we resort to a MaxPooling layer of size 2.

Above the CNN layer, the output of reason, claim and warrant are merged to one and performed flatten operations, through a dense layer, the final output is passed through a two-dimensional softmax layer.

3.2 LSTM framework

LSTM is a special type of RNN that can learn to rely on long-distance history and the immediate previous hidden vector, its a remarkable variation of RNN to alleviate the gradient vanish problem.

In the same way of producing word distribution vector representation in embedding layer, the difference is that, as the LSTM model can process variable length sequences, so we employ masking method to skip (filter out) time steps whose tokens are equal to zero. Above the embedding layer, we introduced the LSTM layer with unit number of 64. Through the LSTM layer, reason, claim and warrant will be transformed to a vector respectively.

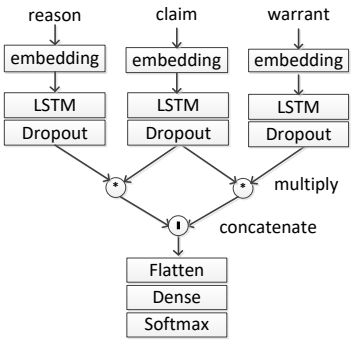

Figure 3: The architecture of GRU framework

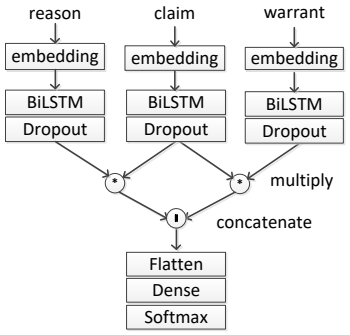

Figure 4: The architecture of biLSTM framework

3.3 GRU with attention

Like the LSTM framework, the masking method is used in the embedding layer. Reason, claim and warrant will be transformed to a word vector matrix and be entered into GRU layer respectively. Gated recurrent unit (GRU), was proposed by (Cho et al., 2014) to make each recurrent unit to adaptively capture dependencies of different time scales. Similarly to the LSTM unit, the GRU has gating units that modulates the flow of information inside the unit, without having a separate memory cells. Unlike the LSTM framework, GRU with attention return full output sequences, instead of the final output of the model.

Through the GRU layer, reason, claim and warrant are converted into three vector matrices. Now, we investigate a state-of-the-art attention model for the warrant vector generation based on claim which is called fact matrix, and the claim vector generation based on warrant which is called attention matrix. An attention mechanism are used to alleviate weakness by dynamically aligning the more informative parts. Specifically, attention model gives more weights on certain words, just like tf-idf for each word, however, the weight is calculated by another vector. Final, we merge the two sequences to one and perform flatten operations.

3.4 biLSTM with attention

The framework is similar to the above one, just changing the GRU model into a biLSTM model, Single direction LSTMs suffers a weakness of not utilizing the contextual information from the future tokens. biLSTM utilizes both the previous and future context by processing the sequence on two directions, and generates two independent sequences of LSTM output vectors.

4 Ensemble Methods

Ensemble methods is an important method in statistical learning, which combines a number of weak models into a strong model through a certain combination. The most famous of them are the Bagging algorithm and the Boosting algorithm.

Our method is inspired by the bagging method, the differences between our method and the traditional bagging method are as follows: we use several strong classifiers for combination, and use all the data to train a single model.

Specifically, we use three ensemble methods. The first method is the soft voting based on Bagging method, that is, each framework outputs a class probability and takes the average to judge the category. The second method is hard voting, each framework predicts categories separately, and finally votes. Because four frameworks are used in the paper, the weight of the best framework is set to 2. The third method is finding the best weight which is the best accuracy on the validation dataset by the exhaustive method.

5 Experimental setup

Our approach in this task is realized by keras, we use the accuracy on validation dataset to locate the best parameters. All the results are taken three times, and the average value is taken.

In the experiment, we use the loss function of categorical cross entropy and the optimizer of adaptive moment estimation. The length of reason, claim and warrant tokens sequence all take the maximum length, if the length is not enough, then zero is added.

For comparison, we report the performance and analysis of four frameworks in Table 1. Rows (1) to (2), list accuracy of task originators models on the validation set and the test set respectively

	Framework	val	test
1	Intra-warrant attention	63.8	55.6
2	Intra-warrant attention w/context	63.7	56.0
3	CNN	63.92	52.93
4	LSTM	66.46	54.95
5	GRU with attention	67.72	56.19
6	biLSTM with attention	67.09	57.21

Table 1: Results of originator and the four main frameworks which our paper employed

	Weight	val	test
soft voting	1 1 1 1	68.35	56.85
hard voting	1 1 1 2	68.03	56.63
exhaustion weight	1 4 2 3	69.93	55.50
	0 2 1 5	69.62	56.41
	1 0 1 5	69.30	55.28

Table 2: Results of ensemble method

(Habernal et al., 2018), we take the results as the baseline to measure other frameworks. Rows (3) to (6) corresponds to our main four frameworks. Row (3) achieve acceptable results on validation set compared to the baseline. Row (4) has been greatly improved compared with CNN on the validation set and test set, especially on the validation set, reaching a 66.46% accuracy rate, proving that the LSTM model is more advantageous in processing sequence text. Row (5) has a better effect on the validation set and test set, the results have exceeded to the baseline. In these four major frameworks, the result of Row (6) is the most satisfying, improves over the baseline already, especially on the validation set, 4% higher than the baseline, on the test set, the accuracy rate of 57.21% is also reached. From the Table 1, we can perceive that the attention mechanism is beneficial to improve the capability of the model.

We report the performance of ensemble method in Table 2. As can be seen from the result, we can observe that Row (1) uses the soft voting method, which achieves a 68.35% accuracy rate on the validation set, which surpasses all the single frameworks. The performance on the test set is also good, reaching 56.85%, more than the baseline model 1%, though it is not as good as the best result of single framework, but it is also a good result. Row (2) is the hard voting method, which is slightly worse than the soft voting. The weight of Rows (3) to (5) which were found on validation set achieve the highest accuracy on the validation set, more than 69%, this is a huge improvement, but the performance on the test dataset is not the best, the reason may be the overfitting caused by this method. **It needs to be emphasized that soft voting method is adopted in the actual tasks, that is, the Row (1). The other methods listed here are only theoretical discussions on the ensemble method from the perspective of research**.

Through these experiments, we can conclude that remarkable results can be achieved through the ensemble method. At the same time, the soft voting method is better than other methods, although the results of the validation dataset are not the best in three method, but the effect on test dataset is the best.

6 Conclusion

In this paper, we solve the Argument Reasoning Comprehension Task by employing four main frameworks and ensemble mechanism. Through a series of attempts, our experimental results demonstrate that: comparing to a single framework, ensembling a series of models can effectively improve the accuracy of the model; the results of the single framework are unstable, and the stability of the model can be improved effectively through the ensemble method, the accuracy of the integrated system on the test set is guaranteed to be above 55%; in the experiment, soft voting is a good way to integrate and achieve better results.

Acknowledgments

This work was supported by the Natural Science Foundations of China No.61463050, No.61702443, No.61762091, the NSF of Yunnan Province No. 2015FB113, the Project of Innovative Research Team of Yunnan Province.

References

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

T Edward Damer. 2009. Attacking faulty reasoning: A practical guide to fallacy-free reasoning. *Wadsworth Cengage Learning,*.

James B Freeman. 2011. Argument structure: Representation and theory. *Argumentation Library*, 121(7):1194–1206.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and*

signal processing (icassp), 2013 IEEE international conference on, pages 6645–6649. IEEE.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The Argument Reasoning Comprehension Task: Identification and Reconstruction of Implicit Warrants. In *16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Geoffrey E Hinton. 1986. Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society*, volume 1, page 12. Amherst, MA.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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

Fabrizio Macagno and Douglas Walton. 2014. *Emotive language in argumentation*. Cambridge University Press.

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.

Di Wang and Eric Nyberg. 2015. A long short-term memory model for answer sentence selection in question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 707–712.