# BLCU_NLP at SemEval-2018 Task 12: An Ensemble Model for Argument Reasoning Based on Hierarchical Attention

**Meiqian Zhao, Chunhua Liu, Lu Liu, Yan Zhao, Dong Yu***
Beijing Language and Culture University
{zhaomq195, chunhualiu596, luliu.nlp, zhaoyan.nlp}@gmail.com,
yudong@blcu.edu.cn

## Abstract

The argument comprehension reasoning task aims to reconstruct and analyze the argument reasoning. To comprehend an argument and fill the gap between claims and reasons, it is vital to find the implicit supporting warrants behind. In this paper, we propose a hierarchical attention model to identify the right warrant which explains why the reason stands for the claim. Our model focuses not only on the similar part between warrants and other information but also on the contradictory part between two opposing warrants. In addition, we use the ensemble method for different models. Our model achieves an accuracy of 61%, ranking second in this task. Experimental results demonstrate that our model is effective to make correct choices.

## 1 Introduction

The argument reasoning comprehension is a crucial part for natural language understanding and inference, since argument comprehension requires reconstruction and analysis for its reasoning. Lack of common sense makes it difficult to infer claims from corresponding reasons directly. To fill the gap between claims and reasons, several explorations have been performed on argumentative structure of a debate (Hastings, 1963; Walton et al., 2008; Walton, 1990). Argument reasoning comprehension, a new task in SemEval 2018, sheds some light on the core of reasoning in natural language argumentation: implicit warrants, which are seen as a bridge between claims and reasons. Given a reason R, a claim C and two alternative warrants W0 and W1, the goal of this task is to identify the right warrant which can justify the use of R as support for C. The difficulty of the task is the warrants are plausible and lexically close but lead to contradicting claims (Habernal et al., 2018).

To be more specific, the reason R and the claim C are propositions extracted from a natural language argument. And warrant W is the relation between R and C which is characterized by a rule of inference (Newman and Marshall, 1992). Walton proposed that an argument refers to a claim based on reasons given in the premises (Walton, 1990). The most central part of this task is how to find the warrant for the given R and C. In the argument reasoning comprehension task, the organizer extracts the instances from Room for Debate section of the New York Times. After a complex crowd-sourcing process, 1970 valid instances are provided for the task. Two alternative warrants are provided as candidates, where one can justify the use of R as support for C and the other one can justify R as support for the opposite side of C. We need to reconstruct the reasoning and select the right warrant which stands for claim in this task. The average score for human is 79.8%, while for those with extensive formal training is 90.9% (Habernal et al., 2018).

In this paper, we not only pay attention to the similar part between each warrant and other information but also pay close attention to the contradictory part between two warrants. We propose a model for this task which consists of four components: sentence representation layer, attended warrant layer, enhanced attention layer and prediction layer. All the sentences are represented with word embeddings in the sentence representation layer. And to better understand the meaning of warrant, we incorporate the additional information to re-represent the two warrants. Then we apply an enhanced attention layer to emphasize the similar and contradictory part between the two alternative warrants W0 and W1. At last, we make prediction through a feedforward neural network layer (Fine,

*The corresponding author

*Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, pages 1104–1108
New Orleans, Louisiana, June 5–6, 2018. ©2018 Association for Computational Linguistics

2001). In addition to the primary model, we propose an ensemble method to achieve a stable and credible accuracy. This method is well established for obtaining highly accurate classifiers by combining less accurate ones (Dietterich, 2000). Our model improves the accuracy by 4% compared to the baseline model.

## 2 Model Description

Our hierarchical attention model is composed of the following major components: sentence representation, attened warrant layer, enhanced attention layer and prediction layer. Figure 1 shows a high-level view of sentence representation and attended warrant layer. And Figure 2 shows a high-level view of the enhanced attention layer and prediction layer. Our code is implemented with TensorFlow and is available on github [1].

In the attended warrant layer, we pay attention to the relevant part between each warrant and other information, and get two attended representation for the two warrants. While in the enhanced attention layer, we focus on the similarities and differences between two attended warrants. The output of our model is the predicted label for each instance, where label 0 means warrant0 stands for the claim and label 1 means warrant1 does.

### 2.1 Sentence Representation

We implement five BiLSTM networks with shared weights to learn each of the sentence representation. It is reasonable to use shared weights among BiLSTM networks because all the sentences are in the same vector space. The input of BiLSTM network is the sentence represented with word embeddings. For later use, we use *w0* for the hidden states of W0 generated by the BiLSTM at time *i* over the input sequence. And *w1*, *c*, *r*, *d* for the hidden states of W1, C, R and D information at each time step.

$$w0_i = BiLSTM(W0, i), \forall i \in [1, ..., l_{W0}] \quad (1)$$

$$w1_j = BiLSTM(W1, j), \forall j \in [1, ..., l_{W1}] \quad (2)$$

$$c_k = BiLSTM(C, k), \forall k \in [1, ..., l_C] \quad (3)$$

$$r_m = BiLSTM(R, m), \forall m \in [1, ..., l_R] \quad (4)$$

$$d_n = BiLSTM(D, n), \forall n \in [1, ..., l_D] \quad (5)$$

Where $l_{W0}, l_{W1}, l_C, l_R$ and $l_D$ are the length of W0, W1, C, R and D respectively.

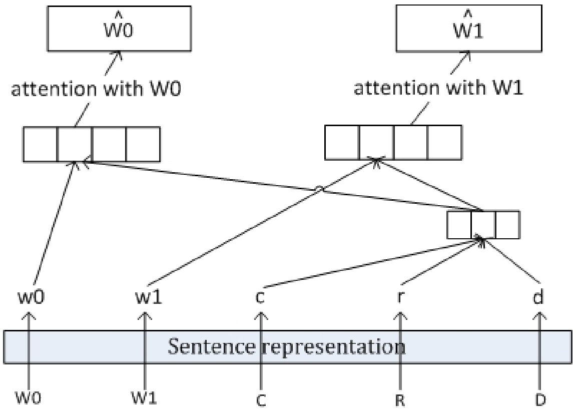

Figure 1: Sentence representation and attended warrant layer

### 2.2 Attended Warrant Layer

Considering the importance of the hint of reason sentence and claim sentence, we try to pay attention to the most relevant part between each warrant and these additional information. So we implement standard attention mechanism over each warrant with additional information including C, R and D, which is represented with A.

$$A = [c; r; d] \quad (6)$$

For better comprehension, we apply intra-attention over each warrant. For convenience, we combine the two processes above by concatenating the two vectors for attention. For further use, we denote attention vectors for W0 and W1 as $\overline{W0}$ and $\overline{W1}$.

$$\overline{W0} = [w0; A] \quad (7)$$

$$\overline{W1} = [w1; A] \quad (8)$$

Specifically, we apply attention over W0 and the concatenate vector for W0 as the following formulas (Tan et al., 2015). We can get $\hat{W0}_i$ to represent the vector for W0 after attention.

$$M0_i = \tanh(Uw0_i + V\overline{W0}) \quad (9)$$

$$S0_i \propto \exp(wM0_i(t)) \quad (10)$$

$$\hat{W0}_i = w0_i S0_i \quad (11)$$

Where U, V and w are attention weight parameters. In the same way, we can compute $\hat{W1}_j$ to represent the vector for W1 after attention.

[1] https://github.com/blcunlp/SemEval2018–argument_reasoning_comprehension

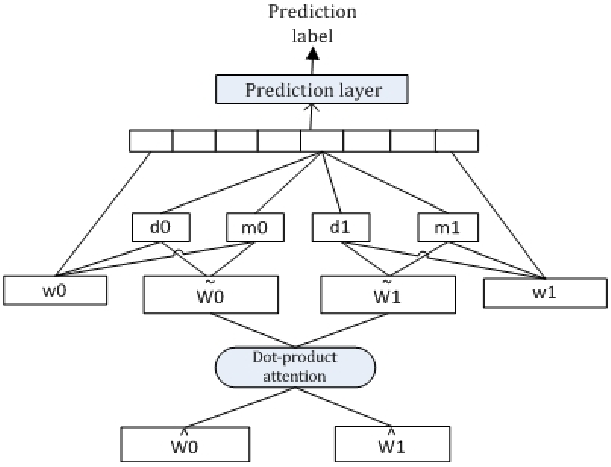

Figure 2: Enhanced attention layer and prediction layer

### 2.3 Enhanced Attention Layer

The similar and contradictory parts of the two warrants are the most important information for argument reasoning comprehension. In the enhanced attention layer, our model align the two warrants to focus on their similarities and differences.

$$e_{i,j} = \hat{W0}_i^T \hat{W1}_j \tag{12}$$

$$\widetilde{W0} = \sum_{j=1}^{l_{W1}} \frac{\exp(e_{i,j})}{\sum_{k=1}^{l_{W1}} \exp(e_{i,k})} \hat{W1}_j \tag{13}$$

$$\widetilde{W1} = \sum_{i=1}^{l_{W0}} \frac{\exp(e_{i,j})}{\sum_{k=1}^{l_{W0}} \exp(e_{k,j})} \hat{W0}_i \tag{14}$$

Where $\widetilde{W0}$ is a weighted summation of each element in W1, the relevant element in W1 is selected and represented into $\widetilde{W0}$. We can get $\widetilde{W1}$ in the same way.

To extract the different part in W0, we let $\hat{W0}_i$ subtract $\widetilde{W0}$. To get the similar part in warrant0, we let $\hat{W0}_i$ multiply $\widetilde{W0}$ (Chen et al., 2017). In the formulas, $\odot$ is element-wise multiply of two vectors. The same is performed for W1.

$$d0 = \hat{W0} - \widetilde{W0} \tag{15}$$

$$d1 = \hat{W1} - \widetilde{W1} \tag{16}$$

$$m0 = \hat{W0} \odot \widetilde{W0} \tag{17}$$

$$m1 = \hat{W1} \odot \widetilde{W1} \tag{18}$$

### 2.4 Prediction Layer

To make best use of all provided information, we concatenate all the representations.

$$W0\_inf = [\hat{W0}; d0; m0] \tag{19}$$

$$W1\_inf = [\hat{W1}; d1; m1] \tag{20}$$

$$all\_info = [W0\_inf; W1\_inf] \tag{21}$$

We implement a simple feedforward neural network to make the final prediction.

### 2.5 Model Ensemble

Ensemble learning helps improve machine learning performance by combining several models (Zhou, 2012). This approach allows the production of a better predictive performance compared to a single model. To improve and stabilize the performance of our model, we ensemble different models by majority voting strategy. We choose five models which have best performance on development data. The final result is better than every single model.

## 3 Experiments

### 3.1 Dataset

The argument reasoning comprehension task chooses the Room for Debate section of the New York Times as source data. The given reason sentences come from stance-taking comments, and the false warrant is generated by annotators. It is validated manually that the created false warrant can prove the reason which will lead to the opposite claim. And the right warrant is written by minimal modifications to the false one, which can ensure that this warrant can be inferred from the reason and stands for the claim. And also, to evaluate the performance of the models, each instance in the dataset is represented as a tuple (R;C;W0;W1) along with a label (0 or 1). If the label is 0, W0 is the correct warrant, otherwise W1. More details about dataset can be found in the work of Habernal (2018)

All the data of the argument reasoning comprehension task is provided in the GitHub by task organizers[2]. For the availability and validity, after complex manual processing, only 1955 instances are selected from 11k comments.

### 3.2 Experimental Setup

We use the development set to select models for testing. Training details are as follows. We use ADAM optimizer (Kingma and Ba, 2014) for training, setting the first hyperparameter to be

[2] the data can be found in github https://github.com/habernal/semeval2018-task12/tree/master/data

0.9 and the second 0.999. The initial learning rate is 0.001 and the batch size is 32. We use word2vec((Mikolov et al., 2013)) to pre-train the word embedding of 300 dimention and keep them from updating while training. The numbers of hidden units and layers of biLSTM networks are 64 and 1 respectively. And the dropout rate is set to be 0.1, and is applied to all biLSTM networks. For the prediction layer, we choose standard FNN with 1 layer and set the hidden cells number to 64.

### 3.3 Evaluation Method

We use accuracy to evaluate the performance of the models, that is, computing the ratio of the right predicted labels of all instances. A scorer and detail information are described in the task introduction website[3].The scorer can give us the expected accuracy of the model.

### 3.4 Results

| models | dev acc | test acc |
|---|---|---|
| baseline model | 0.632 | 0.548 |
| attention model | 0.653 | 0.585 |
| hierarchical model | 0.672 | 0.599 |
| ensemble model-3 | 0.670 | 0.602 |
| ensemble model-5 | 0.686 | 0.606 |
| ensemble model-7 | 0.681 | 0.610 |

Table 1: results of different models

Table 1 shows the accuracy of different models on development dataset and test dataset. The first row is a baseline model which uses intra-warrant attention between two warrants (Habernal et al., 2018). In this model, all the representations of input sentences except warrant0 are concatenated as an attention vector for warrant0 , and all sentences except warrant1 are concatenated as an attention vector for warrant1. And the attentive representations is used for prediction .

To evaluate the performance of each part of our model, we implement experiments on different parts of our model. The attention model in Table1 includes three parts: sentence representation, attended warrant layer and prediction layer. In the prediction layer, the two attended warrant vectors are used to predict the right label. And the third model is a single hierarchical attention model, which adds the enhanced attention layer based on the second model. The accuracy improved by 1.4% on the test data.

We use an ensemble method for different models with majority voting strategy. As we can see in Table 1, the ensemble model for 5 single models achieves the highest accuracy on development dataset.

### 3.5 Results Analysis

For the attended warrant layer, we use shared weights of biLSTM networks to get the representations of sentences. Also we implement intra-attention over each warrant and implement attention over each warrant and other sentences. These changes help to better understand and represent the warrant meaning itself. We introduce intra-attention to emphasize the important meaning in the warrant sentence. The alignment over warrants and the additional information provide the relativity of words in warrant and other sentences. With these attention information, we can see there is an improvement of 2%.

In the enhanced attention layer based on attended warrant layer, we emphasize the similar part and the contrary part between two warrants. The difference of attended warrant0 and warrant1 focuses on the contrary information and the multiplication of attended warrant0 and warrant1 emphasizes the similar part. So the model assigns different weights to different words of warrants according to their relativity. We can see there is about 2% points improvement for this part.

All the results mentioned above are based on the single model. And to get more stable performance, we use an ensemble method for different single models. In our experiment, the top 5 models voting result shows the best performance on development dataset, and the top 7 models voting result shows the best performance on test dataset.

## 4 Conclusion

In this paper, we propose a hierarchical attention model to select the supporting warrant for the argument. The model performs well in SemEval-2018 Task12: The Argument Reasoning Comprehension Task. We find that the information from both the similar part and contrary part between two alternative warrants is crucial to reconstruct the argument reasoning. Moreover, the ensemble method is of great help for the good performance and stability of our model.

[3] scorer can be found in https://github.com/habernal/semeval2018-task12

## Acknowledgments

We would like to thank the anonymous reviewers for their helpful suggestions and comments.The research work is funded by the Natural Science Foundation of China (No.61300081), and the Fundamental Research Funds for the Central Universities in BLCU (No. 17PT05).

## References

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. pages 1657–1668.

Thomas G Dietterich. 2000. Ensemble methods in machine learning. *Proc International Workshgp on Multiple Classifier Systems*, 1857(1):1–15.

Terrence L. Fine. 2001. Feedforward neural network methodology. *Technometrics*, 42(4):432–433.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page (to appear), New Orleans, LA, USA. Association for Computational Linguistics.

Arthur Hastings. 1963. A reformulation of the modes of reasoning in argumentation.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *Computer Science*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Computer Science*.

Susan E. Newman and Catherine C. Marshall. 1992. Pushing toulmin too far: Learning from an argument representation scheme. *Xerox Parc Tech Rpt Ssl*.

Ming Tan, Bing Xiang, and Bowen Zhou. 2015. Lstm-based deep learning models for non-factoid answer selection. *CoRR*, abs/1511.04108.

Douglas Walton, Chris Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*.

Douglas N. Walton. 1990. What is reasoning? what is an argument? *Journal of Philosophy*, 87(8):399–419.

Zhi Hua Zhou. 2012. *Ensemble Methods: Foundations and Algorithms*. Taylor Francis.