

ECNU at SemEval-2018 Task 12: An End-to-End Attention-based Neural Network for the Argument Reasoning Comprehension Task

Junfeng Tian¹, Man Lan^{1,2*} and Yuanbin Wu^{1,2}

¹ School of Computer Science and Software Engineering,
East China Normal University, Shanghai, P.R.China

² Shanghai Key Laboratory of Multidimensional Information Processing
51151201048@stu.ecnu.edu.cn, {mlan, ybwu}@cs.ecnu.edu.cn

Abstract

This paper presents our submissions to SemEval 2018 Task 12: the Argument Reasoning Comprehension Task. We investigate an end-to-end attention-based neural network to represent the two lexically close candidate warrants. On the one hand, we extract their different parts as attention vectors to obtain distinguishable representations. On the other hand, we use their surrounds (i.e., claim, reason, debate context) as another attention vectors to get contextual representations, which work as final clues to select the correct warrant. Our model achieves 60.4% accuracy and ranks 3rd among 22 participating systems.

1 Introduction

Reasoning is a crucial part of natural language argumentation. In order to comprehend an argument, one must analyze its *warrant*, which explains why its *claim* follows from its premises (aka *reasons*) (Habernal et al., 2018a).

SemEval-2018 Task 12 provides the argument reasoning comprehension task (Habernal et al., 2018b). Given a reason and a *claim* along with the *title* and a short *description* of the debate they occur in, the goal is to identify the correct *warrant* from two candidates. Figure 1 gives an example. The abstract structure of an argument is: *reason* → (since) *warrant* → (therefore) *claim*.

The challenging factor is that both candidate warrants are plausible and lexically very close while leading to contradicting claims (Habernal et al., 2018a). Here we give three examples of the two candidate warrants:

Ex1: A huge pandemic would (not) be a great news story.

Ex2: The role of a citizen and a supreme court justice are inseparable /separable.

Ex3: The rest of the comments can be skipped easily /make the section unreadable.

Title: Have Comment Section Failed?

Description: In recent years, many media companies have disabled them because of widespread abuse and obscenity.

Reason: Many comments are thoughtful and insightful. And since {Warrant0 | Warrant1},

Claim: Comment sections have not failed.

✓ **Warrant0:** The rest of the comments can be skipped easily.

✗ **Warrant1:** The rest of the comments make the section unreadable.

Figure 1: An example of a debate in the argument reasoning comprehension task.

The differences are either negative words, or antonyms, or opposite phrases. Therefore, it is important to emphasize the different parts to obtain distinguishable representations of the two warrants, which express the opposite meanings.

To address this factor, we proposed an end-to-end attention-based neural network. On the one hand, we extract the different parts of the two warrants and use them as attention vectors to obtain warrants' distinguishable representations. On the other hand, we represent their surrounds (i.e., reason, claim, debate context) as another attention vector to get the contextual representations.

2 System Architecture

Formally, given an instance containing two candidate warrants (W_0, W_1) and the context around the warrants (i.e., R, C, T, I), the goal is to choose the correct warrant $y \in \{0, 1\}$, where $y = 0$ means W_0 is the correct answer, and $y = 1$ otherwise.

2.1 Overview

The network is inspired by Siamese network (Mueller and Thyagarajan, 2016). The two candidate warrants are modeled in the same structure. Figure 2 illustrates our system architecture.

First, we first extract the different parts of war-

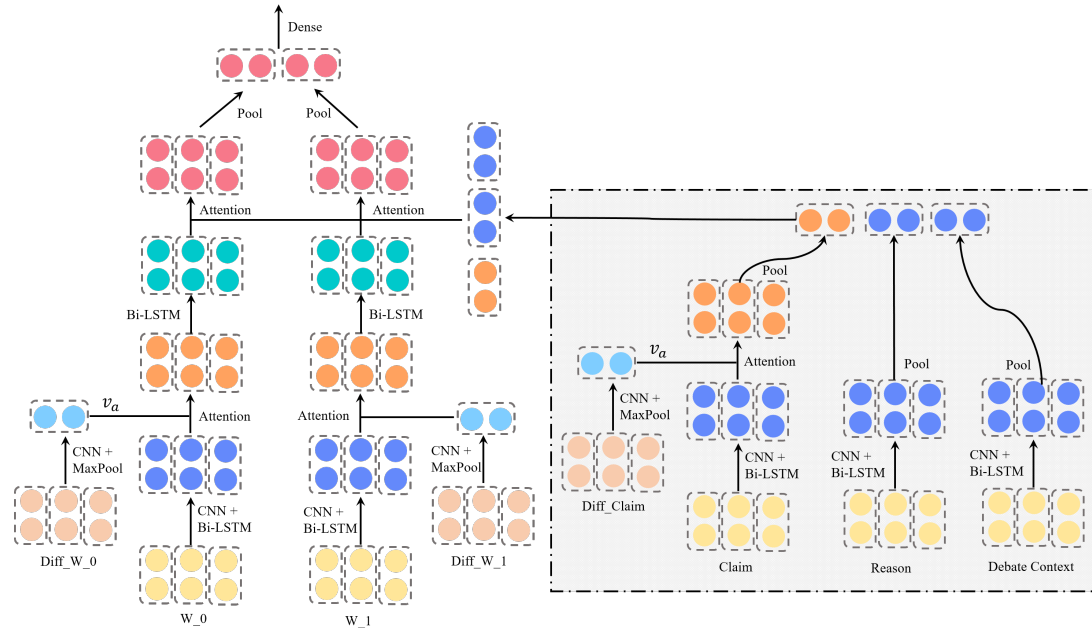


Figure 2: The system architecture

warrant0, warrant1 and claim to get $\text{Diff_}W_0$, $\text{Diff_}W_1$ and $\text{Diff_}Claim$ (see in Sec. 2.2).

Then, we stack a CNN and a RNN to represent each component to obtain representation of each component, because the combination of CNN and RNN can make use of the spatial and temporal context information (in Sec. 2.3).

Next, the intra-temporal attention is adopted to obtain distinguishable representations of warrants. Similarly, we apply the same strategy to the claim. The intra-temporal attention is introduced in Sec. 2.4).

After that, we concatenate the representation of surrounds (i.e., reason, claim, debate context and warrant) as another attention vector to obtain the contextual representations of the warrants.

Last, we adopt a dense layer to obtain the probability of the two candidate warrants (in Sec. 2.5). The contextual representations work as hidden clues to select the correct warrant.

2.2 Extract the Different Part

The two candidate warrants are lexically close (since they often mean the opposite), thus we extract the different part between them to serve as attention vector to guide the neural network to generate distinguishable representation for the warrants.

To do this, we remove the longest common prefix and suffix, and let the remain part as the different part, denoted as $\text{Diff_}W_0$, $\text{Diff_}W_1$. Take cases

mentioned in Sec. 1 as examples, it would extract “not be ” as $\text{Diff_}W_0$ and “be” as $\text{Diff_}W_1$ in Ex1; “inseparable” as $\text{Diff_}W_0$ and “separable” as $\text{Diff_}W_1$ in Ex2; “can be skipped easily” as $\text{Diff_}W_0$ and “make the section unreadable” as $\text{Diff_}W_1$ in Ex3. Note that if the remain part is empty, we use the word after the prefix as the different part.

Similarly, we also get the different part between the claim and its opposite, denoted as $\text{Diff_}Claim$. There always exists the opposite claims in debates, since the reason chains $R \rightarrow W \rightarrow C$ and $R \rightarrow \neg W \rightarrow \neg C$ both exists. We collected the claims and warrants under the same debate. If the warrants express the opposite meaning, then the two claims are opposite. Besides, the organizers also provide similar dataset in “data/train-w-swap-doubled.tsv”.

2.3 Context Representation

To incorporate contextual information of each components in a debate, we combine Convolutional Neural Network (CNN) and Recurrent neural network (RNN) to encode the input word vectors. CNN is good at dealing with spatially related data, such as “sometimes warranted” and “rarely warranted”, while RNN is good at temporal signals. Instead of using a typical vanilla RNN, we use Long Short-Term Memory Network (Hochreiter and Schmidhuber, 1997) for eliminating the issue of long term dependencies.

Given a sentence $S = \{w_i\}_1^n$, we first map each word w_i into its vector representation $x_i \in \mathbb{R}^d$ via a look-up table of word embeddings (d is the dimension of the word embeddings).

Then, we adopt CNN on the input sequence $\{x_i\}_1^n$ to obtain the spatial representation $\{x'_i\}_1^n$:

$$e_i^j = \text{ReLU}(w^j[x_i, \dots, x_{i+k-1}]) \quad (1)$$

$$x'_i = [e_i^1, \dots, e_i^m] (1 \leq j \leq m) \quad (2)$$

where k is the window size, w^j is the parameter of a filter, m is the number of the filters. We also adopt padding before the convolution operation. As a result, we obtain the spatial representations $x'_i \in \mathbb{R}^m$, which has the same length as the input sequence.

After that, we utilize a bi-directional LSTM (Bi-LSTM) to obtain the temporal information. For each time step t , the LSTM unit computation corresponds to :

$$i_t = \sigma(W_i x'_t + U_i h_{t-1} + b_i) \quad (3)$$

$$f_t = \sigma(W_f x'_t + U_f h_{t-1} + b_f) \quad (4)$$

$$o_t = \sigma(W_o x'_t + U_o h_{t-1} + b_o) \quad (5)$$

$$\tilde{c}_t = \tanh(W_c x'_t + U_c h_{t-1} + b_c) \quad (6)$$

$$c_t = i_t \odot \tilde{c}_t + f_t \odot c_{t-1} \quad (7)$$

$$h_t = o_t \odot \tanh(c_t) \quad (8)$$

where σ is the element-wise sigmoid function, \odot is the element-wise product and i_t, f_t, o_t, c_t denote the input gate, forget gate, output gate and memory cell respectively.

2.4 Intra-Temporal Attention

Inspired from Habernal et al. (2018a), we use an intra-temporal attention function to attend over specific parts of the input sequence. This kind of attention encourages the model to generate different representation according to the attention vector. Habernal et al. (2018a) have shown that such intra-temporal attention outperforms standard attention.

We define v_a as the attention vector, and h_t as the hidden states at time step t :

$$m_t = \tanh(U_m h_t \odot v_a + b_m) \quad (9)$$

$$a_t = \sigma(W_a m_t + b_a) \quad (10)$$

$$h_t = h_t \odot a_t \quad (11)$$

where a_t is the attention weights over the hidden states h_t , \odot is element-wise multiplication.

We first apply the intra-temporal attention over W_0 and W_1 , in order to obtain different warrant representations from $\text{Diff}_\cdot W_0$ and $\text{Diff}_\cdot W_1$. As a result, the model can easily distinguish the two candidate warrants. Similarly, we apply the attention over claim to make the claim representation distinguishable.

Moreover, we adopt another intra-temporal attention over W_0 and W_1 , with the concatenation of {claim, reason, debate context} representations as attention vector. The candidate warrants receive the information from the claim, reason and debate context, and the model would select the correct warrant which satisfies the reasoning chain $R \rightarrow W \rightarrow C$.

Finally, we obtain two attended warrant vectors att_{W_0}, att_{W_1} .

2.5 Output

To evaluate the probability distribution of the two candidate warrants, we employ a feed-forward neural network with one dense layer, and apply the *Softmax* function to predict the probability.

$$h_o = \text{ReLU}(W_o[att_{W_0}, att_{W_1}]) \quad (12)$$

$$\hat{p} = \text{Softmax}(W_p h_o) \quad (13)$$

As for the optimization, cross-entropy loss is used as the loss function since we are handling a binary classification problem:

$$\mathcal{L} = -\left(y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)\right) \quad (14)$$

where y_i is the gold label.

3 Experiments

3.1 Datasets

SemEval 2018 provided 1,970 instances for the argument reasoning comprehension task (Habernal et al., 2018b). The instances are divided into three sets based on the year when the debates are taken from. Table 1 lists the statistics of the datasets. We also include the number of debate topics of each set.

Dataset	# Pairs	# Topics	Source Year
Train	1,210	111	2011-2015
Dev	316	31	2016
Test	444	30	2017

Table 1: The statistics of the datasets

Being a binary task, accuracy (Acc) is adopted as the evaluation metric.

Approach	Dev
Intra-warrant attention	0.638 (± 0.024)
Intra-warrant attention w/ context	0.637 (± 0.040)
Our basic model	0.666 (± 0.019)
· + Diff_{ $\{W_0, W_1\}$ }	0.678 (± 0.001)
· + Diff_{ $\{W_0, W_1\}$ } + CNN	0.675 (± 0.010)
· + Diff_{ $\{W_0, W_1, Claim\}$ } + CNN	0.676 (± 0.010)
Ensemble (Vote)	0.708

Table 2: Accuracy of each approach on the developing dataset.

3.2 Parameters Setting

The word embeddings are initialized with the 300d pre-trained word2vec (Mikolov et al., 2013), and do not fine-tune during training. The window sizes of CNN is (1,2,3) and the kernel size is 50. The dimensions of the hidden size in Bi-LSTM and Att-LSTM are set to 50. The dense layer in Output is 25. We train the model using Adam (Kingma and Ba, 2014) with gradient clipping (the max norm is set to 30, batch size is 32), The networks are regularized by dropout (the dropout ratio equals 0.8). We ran each model three times with random initializations. Our code is available at <https://github.com/rgtjff/SemEval2018-Task12>.

3.3 Results on Training Data

Table 2 shows the results of each components of our attention-based neural network. We have the following findings:

(1) Comparing with the Intra-warrant attention (w/ context) provided by the organizer, our basic model obtains 2.8% improvement through sharing parameters in Bi-LSTM. It indicates that the neural network need sufficient training data and parameters sharing could alleviate the demand.

(2) All of the three improvements achieves improved accuracy. It suggests that utilizing the different part as attention vector can obtain discriminative representation, which is beneficial for choosing the correct answer.

(3) The introduction of CNN does not seem to improve the performance of the model. The possible reason may be that RNN actually learn any computational function and capture the spatial information.

(4) The ensemble of the three networks can further improve the performance. Therefore, we configure the ensemble model as our final submission.

Approach	Test
Human average	0.798 (± 0.162)
Human w/ training in reasoning	0.909 (± 0.114)
Random baseline	0.508 (± 0.015)
Intra-warrant attention w/ context	0.584 (± 0.015)
Rank 1: GIST	0.712
Rank 2: blcu_nlp	0.606
Rank 3: ECNU	0.604

Table 3: Accuracy of each approach (humans and systems) on the test set.

3.4 Results on Test Data

Table 3 lists the results of three top systems and several baselines provided by the organizer. We find that: (1) Comparing with Intra-warrant attention w/ context model, our model outperforms it by 2% in terms of accuracy, which demonstrates the efficiency of the proposed attention-based model. (2) Comparing with GIST and blcu_nlp, our result is comparable to blcu_nlp but worse than GIST. Both of them use the pre-trained ESIM model (Chen et al., 2017) trained on SNLI (Bowman et al., 2015) and MultiNLI (Nangia et al., 2017) dataset. Our model only uses the training dataset and does not require any extra resources. However, this is also the limitation of our model because this small-size dataset is insufficient to learn parameters in our model.

4 Conclusion

In this work, we propose an end-to-end neural network for the reading comprehension task. We stack a CNN and a RNN to represent each component in a debate and extract the warrants' and claim's different part as attention vector to obtain their distinguish representation. Moreover, we use another attention network to incorporate the information of reason, claim, debate context into the contextual representation of the warrants for final decisions. Our model achieves 60.4% accuracy and ranks 3rd among 22 participating systems.

Acknowledgments

The authors would like to thank Changzhi Sun for his valuable suggestions and the anonymous reviewers for their helpful comments. This work is supported by grants from Science and Technology Commission of Shanghai Municipality (15ZR1410700), the Open Project of Shanghai Key Laboratory of Trustworthy Computing (No.07dz22304201604).

References

- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.
- Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. [Enhanced lstm for natural language inference](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018a. 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*.
- Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018b. Semeval-2018 task 12: The argument reasoning comprehension task. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA. Association for Computational Linguistics.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.
- 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.
- Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *Thirtieth AAAI Conference on Artificial Intelligence*.
- Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel Bowman. 2017. [The repeval 2017 shared task: Multi-genre natural language inference with sentence representations](#). In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 1–10.