TRANSRW at SemEval-2018 Task 12: Transforming Semantic Representations for Argument Reasoning Comprehension

Zhimin Chen, Wei Song*, Lizhen Liu
Information Engineering College
Capital Normal University
Beijing 100048, China
{zmchen, wsong, liz_liu7480}@cnu.edu.cn

Abstract

This paper describes our system in SemEval-2018 task 12: Argument Reasoning Comprehension. The task is to select the correct warrant that explains reasoning of a particular argument consisting of a claim and a reason. The main idea of our methods is based on the assumption that the semantic composition of the reason and the warrant should be close to the semantic representation of the corresponding claim. We propose two neural network models. The first one considers two warrant candidates simultaneously, while the second one processes each candidate separately and then chooses the best one. We also incorporate sentiment polarity by assuming that there are kinds of sentiment associations between the reason, the warrant and the claim. The experiments show that the first framework is more effective and sentiment polarity is useful.

1 Introduction

Argument reasoning is a key step in the process of argumentation mining and is a very challenging task in natural language processing and artificial intelligence. Maccartney and Manning (2008) suggested that the key factor in the study of natural language understanding is the mastery of natural language reasoning. When we argue for an argument, it is necessary to reconstruct the implicit reasoning (Newman and Marshall, 1992; Habernal et al., 2017) under the relevant assumption and premise to get a simple and concise explanation of the whole reasoning process.

The Argument Reasoning Comprehension task is defined as following:

Given an argument consisting of a claim C and a reason R, the goal is to select the correct warrant that explains reasoning of this particular argument. There are only two options W_0 and W_1 given and only one answer is correct.

Our solution is based on the assumption that the semantic composition of the reason and the true warrant should be close to the semantic representation of the claim. We propose two frameworks. First one is dependent on the task settings that two warrant candidates are considered simultaneously to make a decision. The second one is more general that the task is simplified as determine whether a warrant candidate can explain argument reasoning. We found that the first one performed better.

In addition, we attempt to incorporate sentiment polarity to capture the sentiment association between the reason, the warrant and the claim. The experimental results demonstrate that adding sentiment polarity can improve the performance.

The final result is produced by an ensemble approach that combines the outputs of multiple single models. Our system achieves an accuracy of 0.67 on development dataset and 0.57 on test dataset.

2 System Description

2.1 Model1: Competitive Model

The first proposed model is designed depending on the specific task setting. The architecture of Modell is shown in Figure 1. We first get the representations of the claim C, the reason R and the two warrant candidates W_0 and W_1. Then we use a transformer to get the representation of a pseudo claim, by compositing the reason R and a warrant candidate W. Finally, the model predicts which warrant candidate is the correct one by considering the claim C and two pseudo claims.

2.1.1 Sentence Representation

Figure 2 illustrates the architecture for getting the sentence representation.

*corresponding author

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 1142–1145
New Orleans, Louisiana, June 5–6, 2018. ©2018 Association for Computational Linguistics

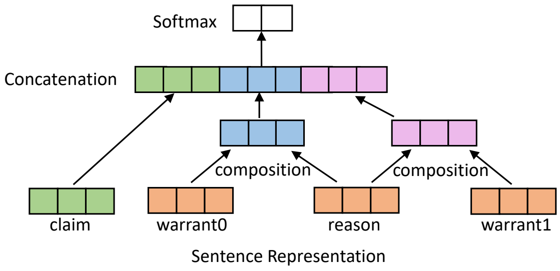

Figure 1: The architecture of Model1: Competitive Model.

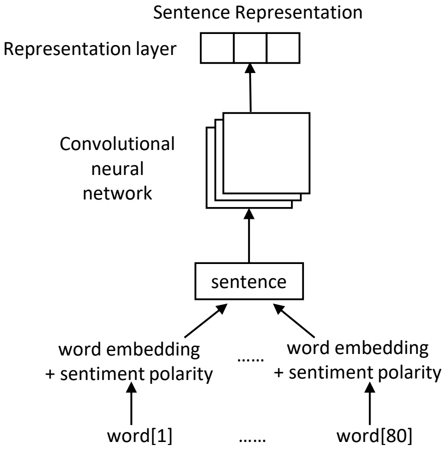
Sentence Representation
Representation layer
Convolutional neural network
sentence
word embedding + sentiment polarity
......
word embedding + sentiment polarity
word[1]
......
word[80]

Figure 2: The representation of the sentence obtained by the convolutional neural network.

Word Embeddings. We first map each word to a word embedding, which is a dense distributed vector. Since the dataset of this task is relatively small, we hope the word embeddings can improve generalization. The word embeddings we used were pre-trained and released by Huang et al. (2012).

Sentiment Polarity of Words. We expect that there are relations between the claim, the reason and the warrant. For example, they may have the same sentiment polarity, while there may be polarity conflicts when involving a false warrant.

Therefore, we use the sentiment polarity of words as a kind of common sense knowledge. The negative words and positive words come from the dictionary provided by Hu and Liu (2004). The polarity representation of each word is a two dimensional vector. A positive word is represented as $[1, 0]$ and a negative word is represented as $[0, 1]$. The representation of out-of dictionary words is $[0, 0]$.

We concatenate the word embedding and sentiment polarity representation together as the final representation of a word.

Convolutional Neural Networks. The word embeddings are feed into a convolutional neural network (CNN) to get the representation of a sentence. We mainly follow the architecture of Kim (2014), which reports excellent performance for several sentence classification tasks. The dimension of word embeddings is k and the sentence length is fixed. A sentence s consisting of n words can be represented as the concatenation of the embeddings of the n words:

$$s = \vec{e}_1 \oplus \vec{e}_2 \oplus \cdots \oplus \vec{e}_n, \tag{1}$$

where $\vec{e}_i$ is the k-dimensional word vector of the i_{th} word. A convolution operation involves a filter $\mathbf{w} \in R^{h,k}$, which is applied to a window of h words to produce a new feature a_i:

$$a_i = f(\mathbf{w} \cdot s_{i:i+h-1} + b), \tag{2}$$

where f is a non-linear activation function, which is set to Relu (Nair and Hinton, 2010) and b is a bias term. The feature map $\mathbf{a}$ can be represented as

$$\mathbf{a} = [a_1, a_2, \cdots, a_{n-h+1}]. \tag{3}$$

Then a max-pooling operation is applied to the resulted feature map to get the sentence representation.

In experiments, $k = 52$, $h = 3$, and we used 64 filters. A dropout layer is added after the word embedding layer with a probability 0.25. The representations of the claim, reason and warrant candidates are all learned in this way.

2.1.2 Pseudo Claim Representation

We assume that the claim is a semantic composition of a reason and a warrant. Therefore, we use a composition operator to combine the representations of the reason and a warrant candidate to get the representation of a *pseudo* claim, noted as C_0 and C_1 respectively.

We have tried four composition operators: ADD, INNERPRODUCT, CONCATENATION and FULLYCONNECTEDNETWORK. In experiments, ADD performs best.

2.1.3 Prediction

Finally, we connect the representations of C, C_0 and C_1 to fully connect layers and concatenate them into one representation. And we connect the representation to the output layer through a non-linear transformation layer (Relu). The output is expected to be 1 if W_1 is right and expected to be 0 if W_0 is right. In experiments, we permuted the order of W_0 and W_1 to enlarge the training dataset.

2.2 Model2: Isolation Model

We consider a more general setting: Given the claim and reason of argument, determine whether a warrant candidate can support the argumentation. As shown in Figure 3, Model2 is just a simplification of Model1. It processes one warrant candidate individually. The output indicates whether the given warrant candidate is right.

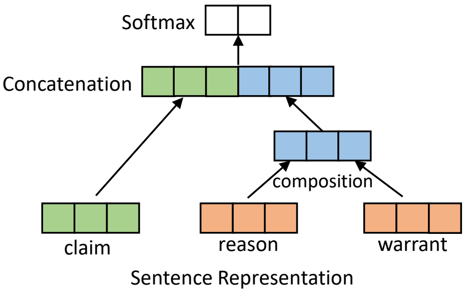

Figure 3: The architecture of Model2: Isolation Model.

For W_0 and W_1, we can get their corresponding probability $p(1|W_0)$ and $p(1|W_1)$. We choose the one with a higher probability as the predicted right warrant for the task.

2.3 Ensemble Model

We have already described the two proposed models. For each model, we trained several times with different initialization parameters. Finally, we chose 3 models from Model1 and 3 models from Model2, which performed well on the development dataset. We used the prediction probabilities of the 6 models as features and trained a random forest classifier for ensemble (Pal, 2005).

3 Evaluation

We conducted experiments on the official datasets of SemEval-2018 Task 12. The model parameters are trained using the training dataset and tuned based on the performance of development dataset. We will report the results on both development dataset and test dataset. Accuracy is the official evaluation metric. We also report precision, recall and F_1 score.

We are interested in two research questions:

- RQ1: Which proposed model is more effective?
- RQ2: Whether incorporating sentiment polarity can benefit this task?

3.1 Results on Development Dataset

Table 1 shows the results on the development dataset. The accuracy of the random baseline is 0.503. The proposed models significantly outperform the baseline.

By adding sentiment polarity representations, Model1 and Model2 both improve a lot. The accuracy of Model1 increases 3.16%, while the accuracy of Model2 increases 2.43%. The precision, recall and F_1 score all have the same trend. With the sentiment polarity added, the Model1 performs better than Model2. Without the sentiment polarity, their performance is very close.

3.2 Results on Test Dataset

Table 2 shows the results on test dataset. The random baseline submitted by task organizer is 0.527. The accuracy of the ensemble model is 0.57, which outperforms the random baseline by 4.3%. After the task organizer released the gold test dataset, we predicted it again using the ensemble model and the accuracy is 0.5811.

Similar to the results on development dataset, with the sentiment polarity added, both Model1 and Model2 achieve a better performance. We can see that on test dataset, Model1 outperforms Model2 no matter using or removing sentiment polarity representations. When using sentiment polarity, the performance difference is larger.

3.3 Discussion

From the experimental results on the development dataset and the test dataset, we can see that sentiment polarity is always useful for distinguishing the correct warrant from the false one. Model1 performs slightly better than Model2. It is reasonable since Model1 considers richer information than Model2. But Model2 actually is a more general model. With sentiment polarity added, the advantage of Model1 is amplified. This also indicates the usefulness of sentiment polarity of words.

The model performance is better on the development dataset than on the test dataset. The proposed models may still suffer the overfitting problem, since the training dataset is not very large.

4 Conclusion

In this paper we presented our system that participated in the SemEval-2018 Task 12: Argument Reasoning Comprehension. Our assumption is

Model	Precision	Recall	F_1 score	Accuracy
Random Baseline	-	-	-	0.503
Model1	0.6380 ± 0.018	0.6276 ± 0.010	0.6216 ± 0.010	0.6276 ± 0.010
w/o polarity	0.5982 ± 0.007	0.5960 ± 0.006	0.5932 ± 0.005	0.5960 ± 0.006
Model2	0.6244 ± 0.013	0.6203 ± 0.016	0.6157 ± 0.019	0.6203 ± 0.016
w/o polarity	0.5968 ± 0.018	0.5960 ± 0.017	0.5946 ± 0.017	0.5960 ± 0.017

Table 1: Results on the development dataset and w/o polarity means sentiment polarity representations of words are removed.

Model	Precision	Recall	F_1 score	Accuracy
Random Baseline	-	-	-	0.527
Model1	0.5457 ± 0.013	0.5420 ± 0.013	0.5347 ± 0.007	0.5420 ± 0.013
w/o polarity	0.5381 ± 0.009	0.5338 ± 0.011	0.5277 ± 0.010	0.5338 ± 0.011
Model2	0.5392 ± 0.019	0.5338 ± 0.021	0.5263 ± 0.028	0.5338 ± 0.021
w/o polarity	0.5340 ± 0.016	0.5285 ± 0.018	0.5227 ± 0.022	0.5285 ± 0.018
Ensemble	**0.5806**	**0.5811**	**0.5805**	**0.5811**

Table 2: Results on the test dataset and w/o polarity means sentiment polarity representations of words are removed.

that the semantic composition of the reason and the warrant should be close to the semantic representation of the corresponding claim. We proposed two neural networks based models: a competitive model that knows two warrant candidates and an isolation model that only considers one candidate for classification. In particular, we incorporated sentiment polarity of words into the models. The experimental results demonstrate that incorporating sentiment polarity of words always improves the performance. The competitive model is slightly better than the isolation model. All proposed models outperform the random baseline by a large margin.

Acknowledgements

The research work is funded by the National Natural Science Foundation of China (No.61402304), Beijing Municipal Education Commission (K-M201610028015, Connotation Development) and Beijing Advanced Innovation Center for Imaging Technology.

References

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2017. The argument reasoning comprehension task.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, Usa, August*, pages 168–177.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Meeting of the Association for Computational Linguistics: Long Papers*, pages 873–882.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Bill Maccartney and Christopher D. Manning. 2008. Modeling semantic containment and exclusion in natural language inference. In *International Conference on Computational Linguistics*, pages 521–528.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *International Conference on International Conference on Machine Learning*, pages 807–814.

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

M. Pal. 2005. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1):217–222.