YUN-HPCC at SemEval-2018 Task 12: The Argument Reasoning Comprehension Task Using a Bi-directional LSTM with Attention Model

Quanlei Liao, Xutao Yang, Jin Wang and **Xuejie Zhang**
School of Information Science and Engineering
Yunnan University
Kunming, P.R. China
Contact:xjzhang@ynu.edu.cn

Abstract

An argument is divided into two parts, the claim and the reason. To obtain a clearer conclusion, some additional explanation is required. In this task, the explanations are called warrants. This paper introduces a bi-directional long short term memory (Bi-LSTM) with an attention model to select a correct warrant from two to explain an argument. We address this question as a question-answering system. For each warrant, the model produces a probability that it is correct. Finally, the system chooses the highest correct probability as the answer. Ensemble learning is used to enhance the performance of the model. Among all of the participants, we ranked 15th on the test results.

1 Introduction

Reasoning is an important part of human logical thinking. It gives us the ability to draw fresh conclusions from some of the known points (Judea, 1988). Argument is the basis for reasoning. Except for the argument's claim and reason, usually, it needs some additional information. Therefore, what we know is the additional information and arguments reason. The claim also needs warrants for an explanation. An example is shown in Table 1.

Obviously, A is a reasonable explanation. The task is to get the reader to find a reasonable explanation for the known messages and claims in the two warrants. Due to the small number of alternative warrants, this problem can be considered to be a binary classification problem. This idea can be used as the baseline model. However, for system scalability and effectiveness, we treat this problem as the regression problem of probability prediction. The idea calculates the probability for each warrant that it is correct. Because of the diversity of natural language expression, there are many ways in which the same meaning can be expressed. Thus, this approach can be better to address this situation (Collobert et al., 2011).

Topic	Should It Be Illegal to Declaw Your Cat?
Additional Info	With legislation pending, New York could become the first state to make removing the claws of a cat a crime.
Argument	Declawing is a crime; instead, people should be educated on proper care and training. And since ...,
Claim	It should be illegal to declaw your cat .
Warrant0	**A) owners should not have the right to be in charge of their animals.**
Warrant1	B) owners should have the right to be in charge of their animals.

Table 1: An Example of the Task.

Another benefit of addressing the problem in this way is to make the problem similar in form to the multi-choice question-answering system. The question-answering system is a classic problem of natural language processing. Many methods and models can be used for reference.

The traditional question-answering system is based on semantic and statistical methods (Alfonseca et al., 2002). This method requires an enormous background knowledge base. In addition, it is not very effective for nonstandard language expression. The state-of-the-art methods are usually based on neural networks. The trained word

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 1109–1113
New Orleans, Louisiana, June 5–6, 2018. ©2018 Association for Computational Linguistics

embedding can fully express the semantics and knowledge. Therefore, the new method is usually better than the traditional statistical-based method.

In this paper, we proposed a bi-directional LSTM with an attention model. The model uses a bi-LSTM network to encode the original word embedding. Then, the semantic outputs are fed into the dense decoder with an attention mechanism. Due to the uncertainty of a single model, ensemble learning is used to enhance the performance of the model.

The remainder of the paper consists of 3 parts. The second part introduces the proposed model in detail, and the implementation is presented in the third part, while the last part presents our conclusions.

2 Model

The model contains several elements, word embeddings, the bi-directional recurrent neural network (Bi-RNN), a semantic encoder (Chen et al., 2016), the attention mechanism and dense layer decoder. Word embedding is a layer before Bi-RNN. This layer contains a map from a word index to the word embedding. This map is a pre-trained word vector look-up table. This task is dependent on semantics, and the question-answering system relies more on knowledge.Therefore, the choice of the word embedding training corpus must pay more attention to the correct grammar. Since a sentence is a whole, a single word in a semantic expression is context dependent. On the other hand, due to the variable length of the input, the Bi-RNN is the choice to complete this encoding task.

Attention mechanisms are used to remind the model of the claim's information. The final result is the probability of a fixed length. In a simple consideration, we use the full connection layer to decode under a softmax function.

2.1 Bi-directional LSTM

The RNN has a powerful ability to extract full-text features (Schuster and Paliwal, 1997), and thus, it is a good tool to obtain the word semantic information. Based on past experience, an LSTM cell is selected to avoid vanishing and exploding gradients. The LSTM is an improved RNN cell. It has two distinct improvements over traditional RNN cells. The first is the gradient problem mentioned above, and the second is the ability to carry long-term information. This arrangement is chosen because LSTM uses a gate structure to make the useful information available for long-term transmission, and the useless information can be filtered out over time. The Bi-directional network allows the forward and backward information to both be expressed.

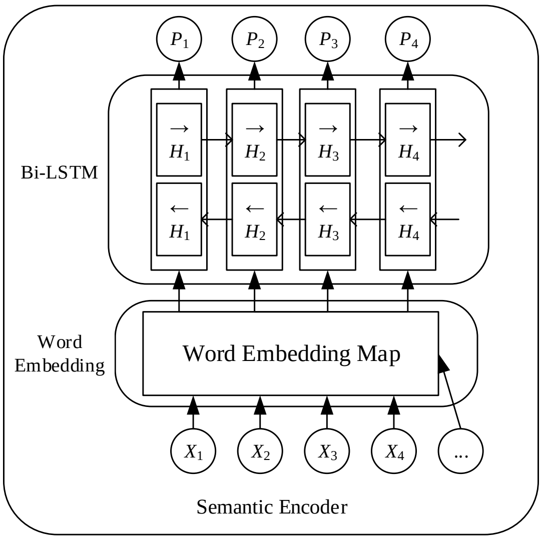

Figure 1: Semantic Encoder

The Bi-LSTM network is chosen to obtain semantic information over all locations in a sentence.

2.2 Semantic Encoder

Putting word embedding and Bi-LSTM together is a semantic encoder. The original text is encoded by it and has global information for each location. The structure of the semantic encoder is shown in Figure 1. The original text index sequence is fed into the encoder. Then, the word embedding layer turns it into a word embedding sequence by a pre-trained word embedding map. Bi-LSTM has forward and backward, 2 directions, to capture the global features. It outputs the combination of two directions results. The final outputs of encoder are the semantic information sequences for original text.

2.3 Attention Mechanism

Attentional mechanisms in natural language processing are usually used to provide the decoder with the source text information (Bahdanau et al., 2014). Semantic information in the original text can be fully obtained when decoding, instead of relying only on the semantic vector.

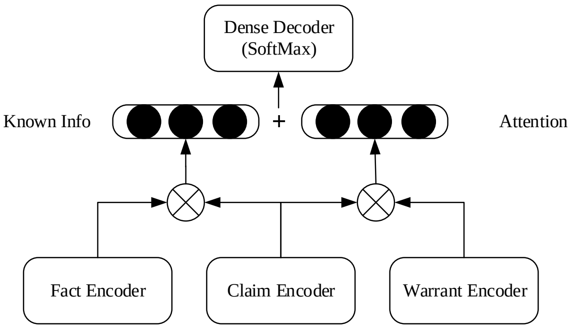

Figure 2: Model Summary

In this task, the attention mechanism is used to provide the model with the semantic information of the claim as the model is decoded.

Overall, the structure of the model is shown in Figure 2. As shown in Figure 2, the model contains 3 semantic encoders. The Fact, Claim, and Warrant are the 3 parts of the training data. Their clear definitions are presented in Part 3.1. The dot product of the two outputs of the fact encoder and claim encoder are combined with the dot product of the claim and warrants outputs. The fact and claim is referenced as known information. The result of the combination is the dense layer that is used to decode. In the decoding phase of the model, a softmax function arrives at the final prediction probability that we need. The output of the claim encoder is used twice during decoding. Its output is the attention mechanism to remind the model to focus on the valuable part of the claim.

3 Experiment

The experiment contains three parts. The first part is the selection and preprocess of experimental data. The second part is the implementation details. The third part is to show and analyze the results.

3.1 Dataset

The training corpus of the word vector, the training set of the model and the test set must be selected and processed.

As mentioned above, reading comprehension focuses more on semantic understanding (Tang et al., 2014), so GoogleNews is a good choice. Because news reports use more cautious words and more rigorous grammar. The mainstream word vector training tools are Word2Vec or GloVe. According to previous experimental results of related tasks, Word2Vec trained vector of words significantly better than GloVe (Yang et al., 2016). Thus, in this experiment, Word2vec was chosen to train the word vector.

The form of the task data is complicated. It is more difficult to obtain the data by artificial generation or online acquisition. Thus, the training data and test data are given by the official data set. Each row of the test data set is divided into several sections, including the id, topic, additional information, reason, claim, warrant0, warrant1 and label. For each row of data, it is processed into two test data of the model. Each training data contains four parts. They are fact, warrant, claim and label. Here, fact is the original topic, with additional information of reason. Warrant, claim and label are not changed. Because there are two warrants in one line, it generates two training data. This approach is similar to a question-answer system, where claim is the question, and warrant is the answer. Additionally, model is used to predict whether this answer is correct for the question.

Because English words have some special forms, such as past tense, past participle, abbreviation and so on, the lemmatisation is needed(Karr, 2006).

3.2 Implementation Detail

The model is implemented using the Keras framework with TensorFlow backend. The program based on python 3.6. The LSTM network and Bi-LSTM network are used as the baseline model.

The proposed model contains 3 semantic encoders, *i.e.*, 3 Bi-LSTM layers and 3 word embedding layers. Using the dense layer as the final decoder outputs the result. Thus, the model contains 3 hyper-parameters, including the number of units of Bi-LSTM layers (Bi-LSTM Unit Number), the dimension of the word embedding (Word Embedding Dimension) and the epochs of the training (Training Epoch).

Due to the lack of training data, when there are more parameters of the model, it is easy to cause over-fitting. There are two improvements to avoid over-fitting. The first is dropout. Dropout is a classic way to avoid over-fitting. A dropout layer is added behind each Bi-LSTM. Thus, the model has one more hyper-parameter, which is the probability of dropout (Dropout Probability).

The second method is ensemble learning. Because of the implicit relationship between claim and reason, this task is very difficult (Habernal et al., 2018). To express all of the features of the

Parameter	Pre-set Values
Bi-LSTM Unit Number	64, 96, 128
Word Embedding Dimension	200, 300
Training Epoch	5, 8
Dropout Probability	0.3, 0.4, 0.5
Ensemble Model Number	5, 7 , 9 , 11

Table 2: Pre-set Parameter.

Model	Acc
LSTM	0.5126
Bi-LSTM w/o ATT	0.5253
Bi-LSTM w/ ATT	0.5696

Table 3: Results of Proposed Model and Baseline Model.

input, a sufficiently complex model is required. However, too little training data is not sufficient for the model to learn all of the features. This concern is a large limitation of a single model. Ensemble learning can effectively alleviate overfitting and greatly enhance the performance of the model.

The hard voting is chosen to implement the ensemble model (Dietterich, 2000). The hard voting means training multiple models at the same time. After training, it takes all the results of the model vote. The voting results are the result of the system.

Finally, the model has a total of 5 hyperparameters, including the ensemble model number. The grid search algorithm is used as the parameter tuning method. However, because the space of the parameters are too large, a few pre-set values are used to narrow the search. The pre-set values are shown in Table 2.

3.3 Result Analysis

Because of the lack of data, the following results are the result of dev data test under official train data training unless otherwise specified. Two baseline models are used to test the performance of the proposed model. In the case of no tuning parameters, finding the average number of test results in 3 times is shown in Table 3. As seen from the results in Table 3, the attention mechanism can effectively improve the accuracy of the prediction in this task. The result is also consistent with most experimental results.

In Table 4, **Epoch** is used for the Training Epoch, **Bi-LSTM** for Bi-LSTM Unit Number, **Emb Dim** for Word Embedding Dimension, **Dropout** for Dropout Probability, and **Acc** for Accuracy, the best 3 results are shown for the parameter tuning for the single model before the ensemble learning. The time spent to tune the parameters on multiple models is very large. Hence, during the implementation of hard voting, only the number of models will be tuned. The remaining parameters are the parameters that give the best result when there is only one model. (the first line in Table 4). The results are shown in Table 5.

Ensemble Model Number	Acc
6	0.6646
7	0.6741
9	**0.6803**
11	0.6772

Table 5: Results of Ensemble Learning.

It can be seen that the effect on the result tends to be stable when the model is over seven. However, the improvement from ensemble learning in the results is enormous. The accuracy increased by approximately 6 percentage points.

In the official test data of the competition, we chose the hard voting with nine models. The accuracy is 0.550. We rank 15th in all 22 teams.

4 Conclusions

Due to the complexity and abstraction of the logical system of human reasoning, it is not easy for a machine to learn its laws. Thus, this task is very challenging and difficult. The attention mechanism and ensemble learning are the key points to improve the performance of the model. In the experiment, both have a very large impact on the accuracy. The final result and rankings are not very good. After analysis, there could be two reasons. The first reason is that the data cleaning was not done well. The training data is mixed with a large amount of useless information. The second point is that the model parameter tuning was very limited.

This competition has benefited us greatly. We will continue to improve our model.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No.61702443 and No.61762091, and in part by Educational Commission of Yunnan Province of

Epoch	Bi-LSTM	Emb Dim	Dropout	Acc
8	64	300	0.3	0.6171
8	128	300	0.4	0.6107
5	64	300	0.4	0.6012

Table 4: Results of Parameter Tuning of Single Model.

China under Grant No.2017ZZX030.The authors would like to thank the anonymous reviewers and the area chairs for their constructive comments.

References

Enrique Alfonseca, Marco Boni, Jos Jara, and Suresh Manandhar. 2002. A prototype question answering system using syntactic and semantic information for answer retrieval. In *Nurs Stand*, pages 680–686.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *Computer Science*.

Tao Chen, Ruifeng Xu, Yulan He, and Xuan Wang. 2016. Improving sentiment analysis via sentence type classification using bilstm-crf and cnn. *Expert Systems with Applications*.

Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(1):2493–2537.

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

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.

Pearl Judea. 1988. *Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers,Inc.

Alan Karr. 2006. Exploratory data mining and data cleaning. *Proc International Workshop on Multiple Classifier Systems*, 101(473):399–399.

Mike Schuster and Kuldip Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Trans on Signal Processing*, 45(11):2673–2681.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Meeting of the Association for Computational Linguistics*, pages 155–161.

Jinnan Yang, Bo Peng, Jin Wang, Jixian Zhang, and Xuejie Zhang. 2016. Chinese grammatical error diagnosis using single word embedding. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA-16)*, pages 155–161.