YNU_Deep at SemEval-2018 Task 12: A BiLSTM Model with Neural Attention for Argument Reasoning Comprehension

Peng Ding, Xiaobing Zhou*
School of Information Science and Engineering
Yunnan University, Yunnan, P.R. China
*Corresponding author, zhouxb.cn@gmail.com

Abstract

This paper describes the system submitted to SemEval-2018 Task 12 (The Argument Reasoning Comprehension Task). Enabling a computer to understand a text so that it can answer comprehension questions is still a challenging goal of NLP. We propose a Bidirectional LSTM (BiLSTM) model that reads two sentences separated by a delimiter to determine which warrant is correct. We extend this model with a neural attention mechanism that encourages the model to make reasoning over the given claims and reasons. Officially released results show that our system ranks 6th among 22 submissions to this task.

1 Introduction

Machine comprehension of text is an important problem in natural language processing. Traditional approaches to machine comprehension are based on either hand engineered grammars (Riloff and Thelen, 2000), or information extraction methods (Poon et al., 2010).

Recently, recurrent neural networks (RNNs) with long short-term memory (LSTM) cells (Hochreiter and Schmidhuber, 1997) have been successfully applied to a wide range of NLP tasks, such as machine translation (Sutskever et al., 2014), constituency parsing (Vinyals et al., 2015), language modeling (Zaremba et al., 2014) and machine comprehension (Hermann et al., 2015). A potential issue with the LSTM models is that a neural network needs to be able to compress all the necessary information of a source sentence into a fixed-length vector (Bahdanau et al., 2014). This may make it difficult for the neural network to cope with long sentences. In order to address this issue, attention mechanisms have been successfully extended to the LSTMs. Attentive Reader (Hermann et al., 2015) used a $tanh$ layer to compute the attention between document and question embeddings. This allows a model to focus on the aspects of a document that it believes helpful to answer a question. The attention-based LSTM models have achieved state-of-the-art results in machine comprehension tasks (Kadlec et al., 2016; Chen et al., 2016; Tseng et al., 2016).

The argument reasoning comprehension task has been presented by (Habernal et al., 2018). The problem can be described as follows: Given an argument consisting of a claim and a reason, the goal is to select the correct warrant that explains reasoning of this particular argument. Compared to traditional machine comprehension task, argument reasoning comprehension requires models to possess extra reasoning abilities. Some models increase the depth of the network, continuously updating the representations of the documents and questions to realize the reasoning process (Sukhbaatar et al., 2015; Tseng et al., 2016; Dhingra et al., 2017; Sordoni et al., 2016).

In this paper, we use a BiLSTM model to encode the reason and claim pairs (reason-claim) and warrants. Then a word-to-sentence neural attention mechanism is implemented to improve the model performance.

The rest of the paper is organized as follows: Section 2 provides the details of the proposed model; Experimental settings and results are discussed in section 3. Finally, we draw conclusions in section 4.

2 System Description

Firstly, we concatenate the reason-claim and warrants with a delimiter, then we encode the reason-claim via a BiLSTM. A second BiLSTM with different parameters is used to encode the delimiter and the warrants, but its memory state is initialized with the last cell state of the previous BiLSTM. The attention mechanism is implemented by the

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 1120–1123
New Orleans, Louisiana, June 5–6, 2018. ©2018 Association for Computational Linguistics

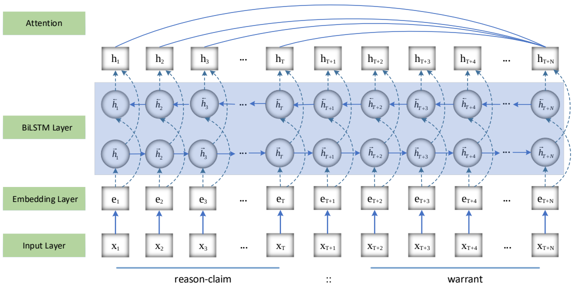

Figure 1: Our BiLSTM model with neural attention for argument reasoning comprehension, basically follows the attention model described in (Rocktäschel et al., 2015).

last output vector of the second BiLSTM and the output vector at each time step produced by the first BiLSTM. Then we use a $tanh$ activation to obtain the final representation. Finally, we predict the correct label via a fully connected layer and a $softmax$ activation.

2.1 LSTM & BiLSTM

Recurrent Neural Networks (RNNs) have been widely exploited to deal with variable-length sequence input. RNNs are networks with loops in them, allowing information to persist. A potential issue of RNNs is that they become unable to learn to connect the previous information when the length of the document grows. LSTM (Hochreiter and Schmidhuber, 1997) is one of the popular variations of RNN to mitigate the gradient vanish problem. LSTMs have three gates: input gate, forget gate and output gate. Gates are a way to optionally let information through. With these gates, LSTMs can remember information for long periods of time and avoid the long-term dependency problem. Given an input vector x_t at time step t, the previous output h_{t-1} and cell state c_{t-1}, an LSTM with hidden state size k computes the next output h_t and new cell state c_t as:

$$H = \begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix} \tag{1}$$

$$i_t = \sigma(W_i H + b_i) \tag{2}$$

$$f_t = \sigma(W_f H + b_f) \tag{3}$$

$$o_t = \sigma(W_o H + b_o) \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c H + b_c) \tag{5}$$

$$h_t = o_t \odot \tanh(c_t) \tag{6}$$

where W_i, W_f, W_o, W_c are trained matrices, b_i, b_f, b_o, b_c are trained biases, σ and $\odot$ denote the sigmoid function and the element-wise multiplication of two vectors, respectively.

Single direction LSTM has one drawback of not using the contextual information from the future tokens. BiLSTM exploits both the previous and future context by processing the sequence on two directions and generates two independent sequences of LSTM output vectors. One processes the input sequence in the forward direction, while the other processes the input in the backward direction. The output at each time step is the concatenation of the two output vectors from both directions, i.e. $h_t = \overrightarrow{h_t} \parallel \overleftarrow{h_t}$.

2.2 Attention

The LSTM model can alleviate the problem of gradient vanishing, but this problem persists in long range contexts. The attention mechanism is introduced to address this issue. Attention is the idea of freeing the encoder-decoder architecture from the fixed-length internal representation. This is achieved by keeping the intermediate outputs from the encoder LSTM and training the model to learn to pay selective attention to these inputs and relate them to items in the output sequence. These attention-based models have achieved state-of-the-art performance on many natural language processing tasks.

Let $C \in \mathbb{R}^{k \times T}$ be a matrix consisting of output vectors $[h_1, h_2, \ldots, h_T]$ produced by the first BiLSTM when reading the T words of the reason-claim, where k is a hyperparameter denoting the hidden units of LSTM. Moreover, let h_{T+N} be the last output vector after the reason-claim and warrant are processed by the two BiLSTMs, respectively. The attention mechanism will produce a vector of attention weights and a weighted representation r of the reason-claim via:

$$M = \tanh(W_c + W_h h_{T+N} \otimes e_T) \tag{7}$$

$$\alpha = softmax(W_m M) \tag{8}$$

$$r = C\alpha \tag{9}$$

where e_T is a vector of ones, $W_c, W_h \in \mathbb{R}^{k \times k}$ are trained projection matrices. $W_m \in \mathbb{R}^k$ is a trained parameter vector. The final sentence-pair representation is obtained from a non-linear combination of the attention-weighted representation r of the reason-claim and the last output vector

h_{T+N} using

$$h^* = \tanh(W_i r + W_j h_{T+N}) \tag{10}$$

where $W_i, W_j \in \mathbb{R}^{k \times k}$ are trained projection matrices.

3 Experiments

The organizers provided training, development, and test sets, containing 1210, 316, 444 instances, respectively. We combine the reason and claim to one sentence which can determine if the warrant is correct or not. The word tokenizer we adopted is $TweetTokenizer$ in Natural Language Toolkit (NLTK[1]).

We compare two word embedding tools, Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). Out-of-vocabulary words in the data sets are randomly initialized by sampling values uniformly from (-0.25, 0.25) and optimized during training. We set *epoch* = 10, *batchsize* = 256 and $LSTMUnits$ = 64. Optimization is carried out using Adaptive Moment Estimation (Adam). All models are attention-based LSTM or BiLSTM architecture.

Model	Tool	Dev Acc	Test Acc
LSTM	Word2Vec	0.626	**0.577**
LSTM	GloVe	**0.646**	0.567

Table 1: Comparison between Word2Vec and GloVe. GloVe performs better on dev data set, but Word2Vec outperforms GloVe on test data set.

We additionally try bidirectional LSTMs through experiments. Given the small scale of the data sets, we run each model 10 times, taking their average as the final result. We also use data augmentation such as shuffle the sentence order to expand the data set. Specifically, we randomize the word order of the reason-claims and the warrants to double the data set. A $randomseed$ is set to ensure our results are reproducible.

Model	Dev Acc	Test Acc
BiLSTM	**0.690**	**0.583**
BiLSTM+Shuffle	0.642	0.570

Table 2: Performance on models with or without shuffle. Both models are based on attention-based BiLSTM + GloVe architecture.

The results show that data augmentation like shuffling the sentence order does not have much effect on the performance of our models. So, we use the attention-based BiLSTM model as our final system to the task. Our final result on the test set is 0.583, which ranks 6th according to the official ranking.

4 Conclusion and Future Work

In this paper, we present a BiLSTM model for argument reasoning comprehension. We adopt a word-to-sentence attention mechanism to make model perform better. In the future, we will utilize external knowledge to enhance the reasoning ability of our models. We will also pay more attention to the generalization of models on small data sets.

Acknowledgments

This work was supported by the Natural Science Foundations of China No.61463050, No.617-02443, No.61762091, the NSF of Yunnan Province No. 2015FB113, the Project of Innovative Research Team of Yunnan Province.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Danqi Chen, Jason Bolton, and Christopher D Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2358–2367.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2017. Gated-attention readers for text comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1832–1846.

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.

[1] http://www.nltk.org/

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

Rudolf Kadlec, Martin Schmid, Ondřej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 908–918.

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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Hoifung Poon, Janara Christensen, Pedro Domingos, Oren Etzioni, Raphael Hoffmann, Chloe Kiddon, Thomas Lin, Xiao Ling, Alan Ritter, Stefan Schoenmackers, et al. 2010. Machine reading at the university of washington. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, pages 87–95. Association for Computational Linguistics.

Ellen Riloff and Michael Thelen. 2000. A rule-based question answering system for reading comprehension tests. In *Proceedings of the 2000 ANLP/NAACL Workshop on Reading comprehension tests as evaluation for computer-based language understanding sytems-Volume 6*, pages 13–19. Association for Computational Linguistics.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.

Alessandro Sordoni, Philip Bachman, Adam Trischler, and Yoshua Bengio. 2016. Iterative alternating neural attention for machine reading. *arXiv preprint arXiv:1606.02245*.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Bo-Hsiang Tseng, Sheng-syun Shen, Hung-Yi Lee, and Lin-Shan Lee. 2016. Towards machine comprehension of spoken content: Initial toefl listening comprehension test by machine. *Interspeech 2016*, pages 2731–2735.

Oriol Vinyals, Łukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. Grammar as a foreign language. In *Advances in Neural Information Processing Systems*, pages 2773–2781.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.