End-to-end Answer Selection via Attention-Based Bi-LSTM Network

Yuqi Ren
College of Computer Science and Technology
Dalian University of Technology
Dalian, China
ryq13@mail.dlut.edu.cn

Tongxuan Zhang & Xikai Liu
College of Computer Science and Technology
Dalian University of Technology
Dalian, China
zhang_tongxuan@mail.dlut.edu.cn

Hongfei Lin
College of Computer Science and Technology
Dalian University of Technology
Dalian, China
hflin@dlut.edu.cn

***Abstract*— Many people ask medical questions online, finding the most suitable answer from candidate answers is an important research area in health care. The IEEE HotICN Knowledge Graph Academic Competition given a question and several candidate answers, then sort the candidate answers to get the best answer. We treated this subtask as a binary classification task, sorted the answers by calculating similarity between the question and each answer. In this work, we proposed a neural selection model trained on the training dataset. Our network architecture is based on the combination of Bi-LSTM and Attention mechanism, extended with biomedical word embeddings. Based on this fact, our model achieve state-of-the-art results on answer selection of medical community.**

Keywords—biomedical question answer, answer selection, Bi-LSTM, Attention mechanism

I. Introduction

Question Answering (QA) system, as an extension of Information Retrieval, accurately responds to questions in natural language. Finding useful information in candidate passages is a challenge. the existing QA researches are mainly focused on general field, without fully considering the difference across disciplines. Among all topics available on the internet, medicine is one of the most frequently searched area[1].

According to different dataset, the QA system is divided into three categories: based on structured data, based on free text and based on question answers pairs. Many studies showed successful cases of exploring general search engines. Collins-Thompson et al. [2] discussed that re-ranking results based on general search engine for difference user. Bevan Koopman et al. [3] used the inference so that improved the performance on hard queries. Gia-Hung Nguyen et al. [4] optimized the document representation by leveraging neural-based approaches to document-to-document matching.

The IEEE HotICN Knowledge Graph Academic Competition is co-organized by IEEE HotICN2018 and Shenzhen Medical Information Center, which provided question answers pairs from medical community.

In this paper, the problem we are trying to solve is re-ranking the answers from several candidate answers, and the most relevant answer is the first. Figure 1 gives an overview of our answer selection model which is mainly composed of six parts including word embed layer, input layer, drop out layer, Bi-LSTM layer, attention layer and output layer.

The remainder of this paper is structured as follows: Sect. 2 introduces the problem domain as well as our experimental setup. Sect. 3 empirically evaluates the merit of our method on the medical community data. finally, we discuss the implications of this work in Sect. 4.

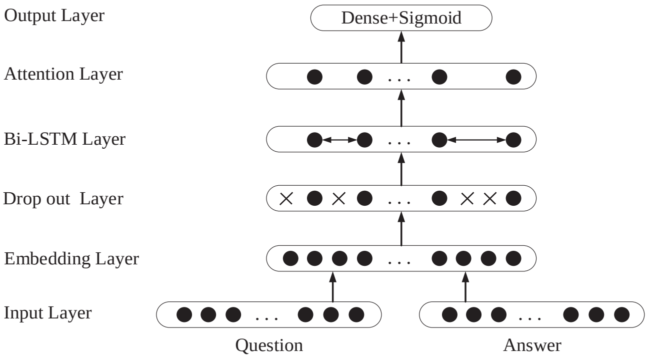

Figure 1: Overview of our answer selection model

II. Methods

We regard this task as a binary classification task, the question-correct answer pair is labeled positive, the question-wrong answer pair is negative. For sorting the candidate answers, we took the probability value of the classification as the score of the candidate answer .We propose a deep learning framework to predict the candidate answers scores by classify question-answer (QA) pair. The architecture of our model consists of three components: (1)the word embedding layer to encode sentence representation; (2)Bi-LSTM layer and Attention layer to obtain sentence features ; (3) Output layer to predict the classification of QA pair.

A. Embedding Layer

Distributed word representations have been widely used in deep learning base text mining, which is convenient for us to capture semantic and key information. For the questions and answers, each word is encoded into a real-values vector by looking up the pre-trained embeddings. Then, the concatenation of question vector and each corresponding answer vector as sample embedding matrix $I^{emb} \in V \times N$, where V denotes the dimension of the vector and N is the number of words. In this work, we adopted a common Chinese word embeddings pre-trained by Shen Li et al. [5]

B. Bi-LSTM Layer

Recurrent neural networks (RNNs) performs well in processing sequential data benefits from the RNNs have the advantage of limited short-term memory. However, with the long-distance sequences, RNNs will cause the gradient vanishing/exploding problems. [6] LSTMs [7] are designed to

978-1-5386-4870-4/18/$31.00 ©2018 IEEE

deal with the long-distance sequences, solve the gradient vanishing problems. The architecture of a LSTM unit is incorporating three gates: input gate, forget gate, output gate.

For LSTM unit, it only learned the past information, can't take information from future. Therefore, we used bi-directional LSTM [8] to obtain each directional information of sequences. Except a forward output, we also calculated the backward output. Then the two outputs are concatenated as the final output.

C. Attention Layer

Since attention mechanism is effective in machine translation [9], lots of researchers applied it to natural language processing. In our model, we used attention mechanism to extract sentence level features from the output of Bi-LSTM. Each sentence consists of several words which have same weight. Many words carry noise or useless information, we trained weights for each word by attention mechanism to focus on the key words. Then we form the sentence level features by element-wise weighted sum.
The formulas to compute output w_j are:

$$\mathrm{T} = \tanh(H) \quad (1)$$

$$\beta = \mathrm{softmax}(\alpha^T \mathrm{T}) \quad (2)$$

$$w_j = H\beta^T \quad (3)$$

where H is the input matrix, which consists of $[h_1, h_2, \ldots, h_n]$ produced by Bi-LSTM layer. The α is the trained parameter vector, the dimensions of α is the dimensions of the word vectors. The dimensions of β is the length of the sentences.

D. Output Layer

After we obtained the sentence representation, we predicted the classification of the sample by fully connected network. The sigmoid function is chosen as the activation function, which calculation result is between 0 and 1. If result is greater than 0.5, the sample is positive, otherwise the sample is negative. The formula of sigmoid function is:

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}} \quad (4)$$

where θ is a trained parameter. Prior to the prediction, we added a full connected layer to extract key features. The cost function of the model is the cross-entropy of the true class label y defined as:

$$\mathrm{Cost} = -\frac{1}{n}\sum(y_i lns_i + (1-y)\ln(1-lns_i)) + \frac{\lambda}{2n}\sum w^2 \quad (5)$$

For preventing overfitting, we added L2 regularization term to the cost function. w is the weight in the softmax function. The λ is regularization parameter. We trained the parameters by minimizing the loss function.

III. Experiments and results

Dataset: The IEEE HotICN Knowledge Graph Academic Competition provided 1000 questions with 10 candidate answers for each, which marks the correct answer. The test set consists of 200 questions with 10 candidate answers for each without marks. We combined the questions with each corresponding answer as sample datasets. Then we shuffled all data randomly, split out 80 percent of the data as training set, the rest of data as develop set. We used publicly available word embeddings trained on 1.3 million words. The dimension of word vector is 300.

Evaluation metrics: The official measure for the answer selection will be the mean average precision (MAP) and mean reciprocal rank (MRR), which are often used to evaluate re-ranking in question answering challenges. The average accuracy of a single topic is the average of the accuracy of each relevant document retrieved. The average accuracy (MAP) of the primary set is the average of the average accuracy of each topic.

Results: Our method obtained state-of-the-art performance in test set. The MRR score is 0.5006, the P@1 is 0.3538. We tried multiple neural networks in this work. We evaluated the performance of our model in dev set. The model word embedding+Bi-LSTM+Attention has the best effect. The performance of each model are shown in Table 1:

Table 1 Performance comparison of multiple methods

Method	F-score	MAP	MRR
Word+Bi-LSTM	51.2%	39%	48.7%
Word+CNN	50.6%	39.4%	49.9%
Word+CNN+Attention	51.5%	40.5%	51.1%
Word+Bi-LSTM+Attention	53.2%	41.1%	52.3%

IV. Conclusion

In this paper, we propose a framework to deal with answer selection task. Our results provide experimental evidence of the usefulness of answer selection of medical community. It re-ranks the answers from several candidate answers with considering the question information. At the same time, our method could get the best answer from the sort of new and solve the problem of several answers in medical community. The experimental results demonstrate that our model outperforms the based models on the real medical community data.

References

[1] Palotti, João, et al. "How users search and what they search for in the medical domain." Information Retrieval Journal 19.1-2 (2016): 189-224.

[2] Collins-Thompson, Kevyn, et al. "Personalizing web search results by reading level." Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011.

[3] Koopman, Bevan, et al. "Information retrieval as semantic inference: A graph inference model applied to medical search." Information Retrieval Journal 19.1-2 (2016): 6-37.

[4] Nguyen, Gia-Hung, et al. "Learning Concept-Driven Document Embeddings for Medical Information Search." Conference on Artificial Intelligence in Medicine in Europe. Springer, Cham, 2017: 160-170.

[5] Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, et al. "Analogical Reasoning on Chinese Morphological and Semantic Relations" ACL 2018.

[6] Bengio, Yoshua, Patrice Simard, and Paolo Frasconi. "Learning long-term dependencies with gradient descent is difficult." IEEE transactions on neural networks 5.2 (1994): 157-166.

[7] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

[8] Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. "Transition-based dependency parsing with stack long short-term memory" In Proceedings of ACL-2015 (Volume 1: Long Papers), pages 334–343.

[9] Luong, Minh Thang, H. Pham, and C. D. Manning. "Effective Approa ches to Attention-based Neural Machine Translation." Computer Science (2015).