Medical Answer Selection Based on Two Attention Mechanisms with BiRNN

Jiajia Ma, Chao Che and Qiang Zhang*

Key Laboratory of Advanced Design and Intelligent Computing Ministry of Education, Dalian University, Dalian, China

Abstract. The contradiction between the large population of China and the limited medical resources lead to the difficulty of getting medical services. The emergence of question answering (QA) system in the medical field allows people to receive timely treatment at home and alleviates the burden on hospitals and doctors. To this end, this paper proposes a new model called Att-BiRNN-Att which combines the Bidirectional RNN (Recurrent Neural Network) with two attention mechanisms. The model employs BiRNN to capture more information in the context instead of the traditional directional RNN. Also, two attention mechanisms are used in the model to produce better feature representation of the answer. One attention is used before the input of BiRNN, and the other is used after the output of BiRNN. The combination of two attentions makes full use of the relevant information between the answer and question. The experiment on the HealthTap medical QA dataset shows that our model outperforms four state-of-the-art deep learning models, which confirm the effectiveness of Att-BiRNN-Att model.

1 INTRODUCTION

The difficulty of getting medical services is a hot and difficult issue in China. It is also a major livelihood problem that the masses of people want to solve urgently. China's large population and the limited medical resources is the main reason for the problem of "the difficulty of getting medical services". However, "many people always want to go to the hospital regardless of their illness state," this also increases the pressure of hospitals and doctors. Moreover, answering the same question every day to different patients is sure to consume the doctors' energy of treating patients. Automatic QA system allows patients describe their situations on the Internet directly and get doctor's professional advice. Patients can be treated promptly at home which decreases the amount of outpatients and alleviates the burden on hospitals and doctors. An important module in the medical automatic QA system is the answer selection. Once a patient asks a question about his illness, lots of answers related to the question are found. Finding the correct answer what the patient really needs exactly is a very difficult task.

Traditional answer selection methods usually rely on the language processing tools or other external methods(Wang and Manning, 2010). Because many steps need to be carried out before the text data are input into the models, semantic information may be easily lost and mixed. But the imprisonment of the traditional approaches have been broke by deep learning methods, which do not rely on language processing tools and input the data into the models directly. The deep learning methods convert a sentence into a vector space and extract its features to represent the sentence, which is then implemented to accomplish a task. For example, Feng, et.al(Feng et al., 2015) extracted the features of the questions and answers by difficult CNN models and adopted the cosine function to measure their similarity. In recent years, attention mechanism(Bahdanau et al., 2014; Luong et al., 2015; Yin et al., 2015) has good performance in many areas of Natural Language Processing, and more and more studies have applied the attention mechanism to the answer selection task. Candidate answers in medical selection task usually are very long and contain a lot of words that are not related to the question. If we calculate the question and answer features separately, the answer will be affected by the similar phrases or noise words thus the relevant semantic information to the question and answer will be ignored, which is undoubtedly very serious for patients with consulting questions. Attention mechanism can solve this problem by assigning different weights to different part of the answers when calculating the answer features. Attention mechanism gives a bigger weight to the parts of answer associated with the question, while assigns a smaller weight to the parts of answer not related to question so the representation of the answer features will have better correlation with the question.

In order to generate better answer feature representation take advantage of the relationship between

Corresponding author: jiabao6861@163.com, chechao101@163.com, zhangq26@126.com

answer and question, this paper proposes a new deep learning model Att-BiRNN-Att based on two attention mechanisms for the medical answer selection problem. Different from the previous studies, we perform two attention mechanisms to calculate the answer features. One joins the representation before the input of BiRNN. Another is after the output of BiRNN. The combination of the two attention mechanisms can better generate the answer feature representation based on the question. The Att-BiRNN-Att model is tested with other four baseline deep learning methods on the HealthTap database.

The rest of the paper is organized as follows: Section2 describes the related work for deep learning in the answer selection; Section3 provides the details of the proposed models; Experimental setting and results of HealthTap database are discussed in section4; finally, we draw conclusions in section5 and acknowledgments in section6.

2 RELATED WORK

Traditional answer selection methods usually rely on language processing tools or other external methods. In this paper(Wang and Manning, 2010), the answer selection problem was transformed to a syntactical matching between the question/answer parse trees. Some work tried to fulfill the matching using minimal edit sequences between dependency parse trees(Heilman and Smith, 2010). Recently, discriminative tree-edit features extraction and engineering over parsing trees were automated in(Severyn and Moschitti, 2013).

Handling the answer selection task by using deep learning approaches has been very mature in recent years. Deep learning models convert text into vectors and model them in the vector space. There are several common deep learning models: Convolutional Neural Network (CNN)(Yin et al., 2015), Recurrent Neural Network (RNN)(Socher et al., 2013) and so on. When RNN processes the long-time dependence it will bring the gradient disappearance problem while the deformation of RNN---LSTM(Graves, 1997; Wang and Jiang, 2016) and GRU(Gated Recurrent Unit)(Cho et al., 2014) can solve this problem effectively, so they have good performance in the text.

There are several ways to resolve the answer selection tasks as follows: constructed a joint feature vector based on question and answer and then transformed the task into a classification or sorting problem(Wang and Nyberg, 2015). What's more, Bahdanau, et.al(Bahdanau et al., 2014) applies the idea of text generation model to answer selection task. Papers (Yu et al., 2014; Feng et al., 2015; Santos et al., 2016; Tan et al., 2016; Wang et al., 2016)learned the features of the question and answer, and then match them used similarity measures. The papers (Feng et al., 2015; Tan et al., 2016; Wang et al., 2016) used the deformable models of LSTM to deal with the problems. Single direction LSTM cannot use context information in the future while BiLSTM is able to use the past and future context information and process the text in two directions. Most papers (Mnih et al., 2014; Wang and Nyberg, 2015; Mueller and Thyagarajan, 2016; Neculoiu et al., 2016; Tan et al., 2016) are modeled by BiLSTM.

Even though BiLSTM can take the relation of word order into account, the effect of distance words is reduced after long distance calculations. The attention mechanism is introduced to enhance the weight of the associated words and reduce the weight of the less associated words.

Recent studies show that models based on the attention mechanism improve the performance of the deep learning models(Mnih et al., 2014; Zheng et al., 2015). However, these models separated the question from the answer, which ignored the question's relevant information corresponded with answer. Such as, Tan et al(Tan et al., 2016) tried to add the attention mechanism to the RNN structure. It dynamically aligned the parts of answers relevant to the questions. The paper applied attention mechanism after the feature extraction. Nevertheless, when RNN processes the sequence the state of the t moment contains all the information from the start to the t moment. When we add the question attention information to find the most useful information, the features closer to the end of the sequence are easier to pick out because it contains all of the previous information. Therefore, the attention mechanism is more biased towards the latter state features. In order to deal with the problem caused by the above computation, paper(Wang et al., 2016)proposed to add attention mechanism before the feature extraction.

In view of the fact that the representation of answer has not fully utilized the relevant information about the question and answer in the previous researches, we propose a model Att-BiRNN-Att to deal with this problem in this paper. The model uses two attention mechanisms with BiRNN, which avoids the output features tending to the back state features and gives full consideration to the relationship between questions and answers.

3 APPROACH

The answer selection problem in medical field solved in this paper can be expressed as follows: Given a question and an answer candidate pool for a question, the goal is to find the best answer candidate. If the selected answer is inside the ground truth set (one question could have more than one correct answers) of question, the question is considered to be answered correctly, otherwise it is answered incorrectly.

Our model combines two attention mechanisms with BiRNN to better produce the feature representation of the answer. The algorithm framework is shown in figure 1. After obtaining the question features by BiRNN and the max pooling, attention1 are calculated by combining the answer input and the question features. Then, we utilize the output of BiRNN in answer and the question features to calculate attention2 and obtain the answer features by max pooling. Finally, we employ the cosine function to match the question and answer. The combination of two attentions can get the answer features expressed with the relevant the question information.

3.1 Recurrent Neural Network (RNN)

RNN have been widely exploited to deal with variable-length sequence input. The long-distance history is stored in a recurrent hidden vector which is dependent on the immediate previous hidden vector. LSTM and GRU are the popular variations of RNN to mitigate the gradient vanishing problem of RNN.

In this paper, BiLSTM and BiGRU are used to calculate the question and answer features. Because BiLSTM/BiGRU is the improved model based on the LSTM/GRU, we first introduce LSTM/GRU, and then introduce BiLSTM/BiGRU.

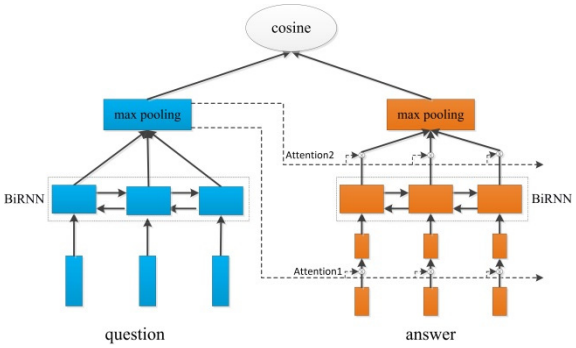

Figure 1: Algorithm framework in this paper.

Given a sentence $X=\{x_1,x_2.....x_n\}$, x_t is k dimension vector which is 100 dimensions in this paper. LSTM hidden vector h_t at the time step *t* is updated as follows:

$$i_t = sigmoid(W_i x_t + U_i h_{t-1} + b_i) \quad (1)$$

$$f_t = sigmoid(W_f x_t + U_f h_{t-1} + b_f) \quad (2)$$

$$o_t = sigmoid(W_o x_t + U_o h_{t-1} + b_o) \quad (3)$$

$$\tilde{C}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (4)$$

$$C_t = i_c * \tilde{C}_t + f_t * C_{t-1} \quad (5)$$

$$h_t = o_c * \tanh(C_t) \quad (6)$$

There are three gates, i.e. input gate *i*, forget gate *f*, output gate *o* and a cell memory C in the LSTM architecture. The input gate can determine how incoming vectors x_t alter the state of the memory cell. The output gate can allow the memory cell to have an effect on the outputs. Finally, the forget gate allows the cell to remember or forget its previous state. The formula (4) is an input transformation, and formula (5) is an update of the cell state.

GRU is formulated as follows:

$$z_t = sigmoid(W_z x_t + U_z h_{t-1}) \quad (7)$$

$$f_t = sigmoid(W_f x_t + U_f h_{t-1}) \quad (8)$$

$$h_t = \tanh(W_h \tilde{x}_t + U_h(f_t \odot h_{t-1})) \quad (9)$$

$$h_t = (1-z_t) \odot h_{t-1} + z_t \tilde{h}_t) \quad (10)$$

Where W_z, W_f, W_h, U_z, U_h are weight matrices and $\odot$ stands for element-wise multiplication.

LSTM/GRU only uses the information about the past, while Bidirectional LSTM/GRU can take the past and future sequence information into account. Bi LSTM/BiGRU utilizes both the previous

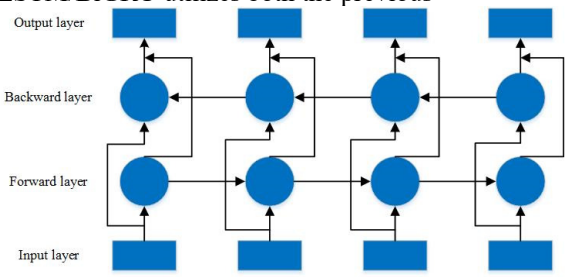

Figure 2: Bidirectional LSTM/GRU architectures.

and future context by processing the sequence on two directions, and generates two independent sequences of LSTM/GRU output vectors. One processes the input sequence in the forward direction, while the other processes the input in the reverse direction. The output at each time step is the concatenation of the two output vectors from both directions, i.e., $h_t = \overrightarrow{h_t} \| \overleftarrow{h_t}$ 。

3.2 Attention mechanism

When quickly read long text, our attention is focused on key events or entities; when looking for an answer, we make use of the information of the question to find the most relevant answer in the answer pool. The attention mechanism is designed to take advantage of the information about questions and answers. In this paper, the two attention calculations we use are as follows:

Attention1： It is performed before the input of BiLSTM/BiGRU. Attention1 can adjust the effective information parts of question corresponding to the answer and find the link between question and answer, which was then applied to feature representation of the answers.

$$\alpha_t = sigmoid(r_q^T M_{qi} x_t) \quad (11)$$

$$\tilde{x}_t = \alpha_t * x_t \quad (12)$$

M_{qi} is a attention matrix that converts the question into word embedding space.

Attention2： BiLSTM/BiGRU can take the word orders into consider, but when the sentence is very long, BiLSTM/BiGRU may be influenced by the less important words in the candidate answers. So we introduce attention2 to increase the weights of strong associated words, and reduce the weight of weak associated words in the output of BiLSTM/BiGRU. In BiLSTM/BiGRU, a word at each time step is added and the hidden state is updated recurrently, those hidden states near the end of the sentence are expected to capture more information. Consequently, when the attention2 information is added to the time sequence hidden representations, the near-the-end hidden variables will be more easily selected because of the abundant semantic

accumulation, which may lead to a biased attentive weight towards the later coming words in RNN.

Attention1 calculation also has carried on the correction. The combined use of the attention mechanism produces a better answer representation. Attention function F_a calculation methods and answer features are calculated as follows:

$$m(t) = \tanh(W_{hm} h_a(t) + W_{qm} r_q) \quad (13)$$

$$F_a(r_q, h_a(t)) = \exp(w_{ms}^T m(t)) \quad (14)$$

$$S_t \propto F_a(r_q, h_a(t)) \quad (15)$$

$$\tilde{h}_a(t) = h_a(t) * S_t) \quad (16)$$

$$r_a = \sum_{t=1}^{m} \tilde{h}_a(t) \quad (17)$$

Among them, $h_a(t)$ is the hidden state of the answer at the *t* moment, W_{hm}, W_{qm} is the attention weight matrix, and w_{ms} is the attention weight vector. It can be argued that when the hidden state of answer is not related to the question, it will play a very small role in the final feature representation, and conversely, it contributes greatly to the representation of the answer.

3.3 Objective function

The model is constructed in the form of (q, a_+, a_-). Q stands for questions, a_+ stands for positive answers, and a_- stands for negative answers. The negative answers are chosen randomly in training, and are combined into the form of (q, a_+, a_-). We use the objective function L as follows:

$$L = \max\{0, M - \cos(q, a_+) + \cos(q, a_-)\} \quad (18)$$

The cosine function calculates the similarity of the question and answer and compares it with the range M., $M - \cos(q, a_+) + \cos(q, a_-) > 0$ the similarity of positive answer is less than the negative answer. If $M - \cos(q, a_+) + \cos(q, a_-) \le 0$ there is no need to update the parameters and we will train a new negative example until the margin is less than M.

4 EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Setup

We evaluate the proposed model on the HealthTap database. HealthTap is the first world's global health care platform. It allows users to access via the video, text, or voice all day, with more than 108000 top doctors online. HealthTap's answers all come from real doctors, and basic questions can be answered free. We use a part of this data(Nie et al., 2017). Among it, there are 11915 questions in the training data, and 1601 problems in the test data. The creation format of test data are same with paper(Feng et al., 2015). The answer pool for each question are composed of positive answers, and then randomly select the answers from all the answer pool until the total number of answers to 500. The length of the question and answer sentences are 100.

Our method is implemented by the Tensorflow framework. Word2vec is used for training the word vector and the vector dimension is 100. The optimization method is Stochastic Gradient Descent (SGD), batch size is 256, and learning rate is 0.01. The value of margin is set as 0.05. Hidden size is 220. We use the Top1 accuracy (P@1)to evaluate the effectiveness of our method.

We calculate the matching scores of question with 500 answers in the answer pool. If the tag of the answer with the highest matching score is 1, answer selection is right. Otherwise, it is wrong.

4.2 Comparison algorithms and experimental results analysis

We employed four state-of-the-art algorithms to compare with our method in this paper:

1.BiLSTM(Tan et al., 2016): The model does not use the attention.

2.QA-LSTM(Tan et al., 2016): The model employs bidirectional LSTM model and uses the attention after the feature extraction.

3.IARNN-WORD(Wang et al., 2016) and IALSTM-WORD: Add attention information before RNN hidden representation. IARNN-WORD and IALSTM-WORD employ bidirectional GRU and LSTM model to perform feature representation, respectively.

Our proposed model is also tested using BiLSTM and BiGRU, which is called Att-BiGRU-Att and Att-BiLSTM-Att for short, respectively. The experimental results and analysis are in Table 1:

Table 1: Experimental results

	Model	**P@1**
1	BiLSTM	60.39
2	QA-LSTM	62.46
3	IARNN-WORD	61.64
4	IALSTM-WORD	61.71
5	Att-BiGRU-Att	63.58
6	Att-BiLSTM-Att	**63.83**

As we can see from table 1, LSTM model performs slightly better than the GRU model, and the performance of the model based on two attention mechanisms is superior to other models. QA-LSTM and IARNN-WORD outperform the model BiLSTM, which demonstrates the importance of the attention mechanism in the answer representation. For the non-attention models, the features of question and answer are represented individually, and the similarity scores of questions and answers represent their correlation. In the

model with attention, the answer feature is calculated with the question feature. Answer representation is related to the question information. So the similarity scores contain relevant information.

Two attention mechanisms are applied in our model. The model using two attentions can generate better answer representation according to the question. Compared with models with one attention, Att-BiLSTM-Att model can extract more question

Table 2: Examples (questions, correct and wrong answers) selected in the HealthTap dataset.

Q: I am having extreme pain and burning in my index finger and morning stiffness in other fingers. It gets better after the morning. It is only in one hand and iblood work. Does it indicate rheumatoid arthritis or not?
Answer 1: Carpal tunnel is Associated with altered sensation in the thumb, index, middle and part of the ring. It can affect the muscles of the thumb. Symtoms can be intermittent, occasional or severe and long standing. Testing such as a nerve study/emg is often used to aid diagnosis and gauge severity. Treatment can range from splinting to injections and even surgery. A very common disorder; many functions well despite it.
Positive Answer: Many possibilities. This could be carpal tunnel syndrome, or another form of arthritis associated with psoriasis or other diseases, or could even be early rheumatoid arthritis. Ask your doctor for help with diagnosis and consider physical therapy if prescribed.

to represent answer, and makes the answer representation more targeted when selecting answer. As shown in Table 2, with regard to the questions Q, we should answer "Does it indicate rheumatoid arthritis or not". QA-LSTM and IALSTM-WORD with one attention pay more attention to the symptoms, and there is no clear answer to the question. Although answer1 has a relationship between the question and answer, the question needed to be answered is only mentioned as "A very common disorder". While the model with two attention focus more attention on the correlation between question and answer, such as "Many possibilities. This could be" and "or could even be" are all interrelated with "whether rheumatoid arthritis" in the correct answer. Therefore it has a better prediction effect.

5 CONCLUSIONS

In this paper, we propose a bidirectional RNN model with two attentions to perform answer selection for the healthcare QA. The two attentions is embedded before the input and after the output of BiRNN respectively based, which gives full consideration to the relation between the questions and answers. The experiment results on the HealthTap database show that the model performs better than several models. In the future work, we would like to study the answer ranking in health QA.

ACKNOWLEDGEMENTS

We acknowledge the support from the National Natural Science Foundation of China (No.91546123), Program for Changjiang Scholars and Innovative Research Team in University (No. IRT15R07).

References

1. Bahdanau, D., Cho, K. and Bengio, Y., 2014. *Neural Machine Translation by Jointly Learning to Align and Translate*. Computer Science.
2. Cho, K., Merrienboer, B. V., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y., 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *In EMNLP, 2014 Conference on Empirical Methods in Natural Language Processing* (pp.1724–1734). ACL.
3. Feng, M., Xiang, B., Glass, M. R., Wang, L. and Zhou, B., 2015. Applying deep learning to answer selection: A study and an open task. *In ASRU, Automatic Speech Recognition and Understanding* (pp.813-820).
4. Graves, A., 1997. *Long Short-Term Memory*. Neural Computation. 9(8), p.1735.
5. Heilman, M. and Smith, N. A., 2010. Tree edit models for recognizing textual entailments, paraphrases, and answers to questions. *In Human Language Technologies, The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp.1011-1019). ACL.
6. Luong, M. T., Pham, H. and Manning, C. D., 2015. Effective Approaches to Attention-based Neural Machine Translation. *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. ACL.
7. Mnih, V., Heess, N., Graves, A. and Kavukcuoglu, K., 2014. Recurrent models of visual attention. *In Proceedings of the 27th International Conference on Neural Information Processing Systems* (pp.2204-2212). MIT Press.
8. Mueller, J. and Thyagarajan, A., 2016. Siamese recurrent architectures for learning sentence similarity. *In AAAI-16, 3th AAAI Conference on Artificial Intelligence* (pp.2786-2792). AAAI Press.
9. Neculoiu, P., Versteegh, M. and Rotaru, M., 2016. Learning Text Similarity with Siamese Recurrent Networks. *In ACL2016, 1st Workshop on Representation Learning for NLP*.
10. Nie, L., Wei, X., Zhang, D., Wang, X., Gao, Z. and Yang, Y., 2017. *Data-Driven Answer Selection in Community QA Systems*. IEEE Transactions on Knowledge & Data Engineering. 29(6):1186-1198.
11. Santos, C. D., Tan, M., Xiang, B. and Zhou, B., 2016. *Attentive Pooling Networks*. Computer Science.
12. Severyn, A. and Moschitti, A., 2013. Automatic feature engineering for answer selection and extraction. *In EMNLP, 2013 Conference on Empirical Methods in Natural Language Processing* (pp.458-467). ACL.
13. Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y. and Potts, C., 2013. Recursive deep models for semantic compositionality over a sentiment treebank. *In EMNLP, Conference on Empirical Methods in Natural Language Processing* (pp.1631-1642). ACL.
14. Tan, M., Santos, C. D., Xiang, B. and Zhou, B., 2016. Improved Representation Learning for Question Answer Matching. *In ACL, 54th Annual Meeting of the*

Association for Computational Linguistics (pp.464-473). ACL.

15. Wang, B., Liu, K. and Zhao, J., 2016. Inner Attention based Recurrent Neural Networks for Answer Selection. *In ACL, 54th Annual Meeting of the Association for Computational Linguistics* (pp.1288-1297). ACL.
16. Wang, D. and Nyberg, E., 2015. A Long Short-Term Memory Model for Answer Sentence Selection in Question Answering. *53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (pp.707-712). ACL.
17. Wang, M. and Manning, C. D., 2010. Probabilistic tree-edit models with structured latent variables for textual entailment and question answering. *23rd International Conference on Computational Linguistics* (pp.1164-1172). Coling 2010 Organizing Committee.
18. Wang, S. and Jiang, J., 2016. Learning Natural Language Inference with LSTM. *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL.
19. Yin, W., Schütze, H., Xiang, B. and Zhou, B., 2015. *ABCNN: Attention-Based Convolutional Neural Network for Modeling Sentence Pairs*. Transactions of the Association of Computational Linguistics.
20. Yu, L., Hermann, K. M., Blunsom, P. and Pulman, S., 2014. *Deep Learning for Answer Sentence Selection*. Computer Science.
21. Zheng, Y., Zemel, R. S., Zhang, Y. J. and Larochelle, H., 2015. *A Neural Autoregressive Approach to Attention-based Recognition*. International Journal of Computer Vision. 113(1):67-79.