Bi-directional Capsule Network Model for Chinese Biomedical Community Question Answering

Tongxuan Zhang[1], Yuqi Ren[1], Michael Mesfin Tadessem[1], Bo Xu[1,2], Xikai Liu[1], Liang Yang[1], Zhihao Yang[1], Jian Wang[1], and Hongfei Lin[1](✉)

[1] Dalian University of Technology, Dalian 116024, People's Republic of China
hflin@dlut.edu.cn

[2] State Key Laboratory of Cognitive Intelligence, iFLYTEK, Dalian, People's Republic of China

Abstract. With the rapid development of the Internet, community question answering (CQA) platforms have attracted increasing attention over recent years, particularly in the biomedical field. On biomedical CQA platforms, patients share information about diseases, drugs and symptoms by communicating with each other. Therefore, the biomedical CQA platforms become particularly valuable resources for information and knowledge acquisition of patients. To accurately acquire relevant information, question answering techniques have been introduced in biomedical CQA. However, existing approaches cannot achieve the ideal performance due to the domain-specific characteristics. For example, biomedical CQA involves more complex interactive information between askers and answerers, while CQA techniques designed for the general field can only deal with single interactions between questions and candidate answers within a similar topic. To address the problem, we propose a novel neural network model for biomedical CQA. Our model adopts the bidirectional capsule network to focus on different aspects of biomedical questions and candidate answers, and merges high-level vector representations of questions and answers to capture abundant semantic information. Furthermore, to capture the meaning of Chinese characters, we incorporate the radical of Chinese characters embedding as auxiliary information to improve the performance of Chinese biomedical CQA. We conduct extensive experiments, and demonstrate that our model achieves significant improvement on the performance of answer selection in the Chinese biomedical CQA task.

Keywords: Community question answering (CQA) · Biomedical question answering · Answer selection · Capsule network

1 Introduction

Question answering (QA) system [1, 2], as an advanced form of information retrieval system, has attracted intense research interest in the field of information retrieval (IR) and natural language processing (NLP) in recent years. Different from search engines, QA system aims to obtain more concise answers instead of relevant documents

J. Tang et al. (Eds.): NLPCC 2019, LNAI 11838, pp. 105–116, 2019.
https://doi.org/10.1007/978-3-030-32233-5_9

for submitted questions by askers. With the rapid development of the Internet, the community-based question answering (CQA) platforms, such as Yahoo Answers, Wiki Answers, and Baidu Zhidao, have become popular and practical Internet-based web services for satisfying user information needs [3]. With the increasing scale of the CQA archive, the large amount of questions and corresponding answers pose a great challenge for exactly matching the candidate answers with the submitted questions. Therefore, it is necessary to design effective methods for selecting the optimal answer to the given question and meeting the information needs.

Answer selection is an important research problem in the open domain for many years [4, 5]. Related studies have focused on improving CQA in general fields from different respects [6, 7]. With the increasing popularity of online health-related platforms, biomedical CQA has greatly facilitated people' life and attract much attention of medical practitioners and interdisciplinary researchers. Related research has attempted to develop effective approaches for accurate CQA matching in the biomedical field. To help better understand biomedical CQA, we illustrate an example of biomedical question and its candidate answers in Table 1. In the example, Answer 1 can better match the question and Answer 2 is an irrelevant answer. Previous work on biomedical answer selection has mostly relied on feature engineering [8]. Recent advances in deep learning have provided a new direction for enhancing biomedical CQA [9]. Compared with feature engineering, deep learning does not need handcrafted feature, which can reduce much manual labor on feature extraction.

Table 1. An example question with candidate answers

Question	脑血管硬化吃什么食物?不该吃哪些? Which kind of food can be eaten by people who has cerebral arteriosclerosis? What food can't?
Answer 1	'1': 脑血管硬化患者饮食上应多吃绿色蔬菜和新鲜水果, 减少动物脂肪的摄入, 烹调时最好用植物油, …… (✓) **'1': Cerebral arteriosclerosis patients should eat more green vegetables and fresh fruits, reduce animal fat intake, it is best to use vegetable oil when cooking …… (✓)**
Answer 2	'2': 问题分析: 您好; 这种情况一般考虑偏头疼, 一般是功能性因素引起的头痛, 无器质性病变, …… (✗) '2': Problem analysis: Hello, this situation is generally considered to be migraine, a kind of headache caused by functional cases, it's no organic lesion …… (✗) ……

To improve the performance of CQA, Tan et al. [10] employed a bidirectional long short-term memory (Bi-LSTM) network [11] to represent the input questions and answers, respectively, aiming to match the questions with candidate answers by accommodating their semantic relations. There are also studies on Chinese question answering. Yuan et al. [12] proposed a deep feature selection method for Chinese questions classification. Yu et al. [13] developed a model based on a CNN, and applied it to the task of answer sentence selection. However, these studies have partly ignored the specific characteristics of the Chinese language. Meanwhile, the existing neural models have mostly encoded sentence information into one vector representation by optimization strategies, such as the max-pooling, which may partly overlook the complex semantic relationship between the question and answers.

To deal with these problems, Sabour et al. [14] proposed the framework capsule network, which can be used enrich the feature representations. Subsequent studies have further combined deep neural networks and capsule networks to build effective deep learning architectures in NLP tasks, such as CNN+Capsule [15] and RNN+Capsule [16]. However, the original capsule network only captures the useful information from the insider of one sentence. To improve the performance of biomedical CQA using capsule networks, we need to consider the information from both the questions and answers. Meanwhile, some of the methods only explored the answer selection in a single direction but neglected the reverse direction.

In this work, we focus on the task of Chinese biomedical CQA. Chinese language processing has its unique difficulties. For example, Chinese sentence is written continuously and biomedical entities are more implicitly presented than those in English. To overcome these difficulties, we design a novel neural network model based on capsule network for obtaining high-quality answers. Our model adopts the bi-directional capsule network to focus on different aspects in two directions, and merges high-level vector representations of questions and answers to capture abundant semantic interactive information. Furthermore, we incorporate the radical of Chinese characters embedding bring with additional information to improve the performance of Chinese biomedical CQA. We summarize the contributions of this work as follows:

- We propose a bi-directional capsule network-based framework for Chinese biomedical CQA. In each interactive direction, our model captures contextual information and focuses on different aspects from the question and candidate answer, respectively.
- We extract Chinese component-level features to capture additional useful information by radical-CNN in the assumption that similar radical sequences can encode similar semantic information.
- We conducted experiments on a Chinese biomedical CQA dataset, and demonstrated the effectiveness of our model. Experimental results show that our model can significantly outperform the state-of-the-art methods.

2 Methods

In this section, we provide more details on the proposed model for biomedical answer selection. We first illustrate the entire architecture in Fig. 1. The architecture of our model consists of three components: (1) the Bi-LSTM layer, which is used to encode contextual information of the question and candidate answers and formulate the sentence representation with abundant semantic information; (2) the self-attention layer, which extracts the features in the question-answer pair to highlight different aspects of the matching between each pair of question and answer; (3) the bi-directional capsule layer, which learns the final representations. The number of capsule networks equals the number of classification categories. The length of the output vector represents the probability for each category.

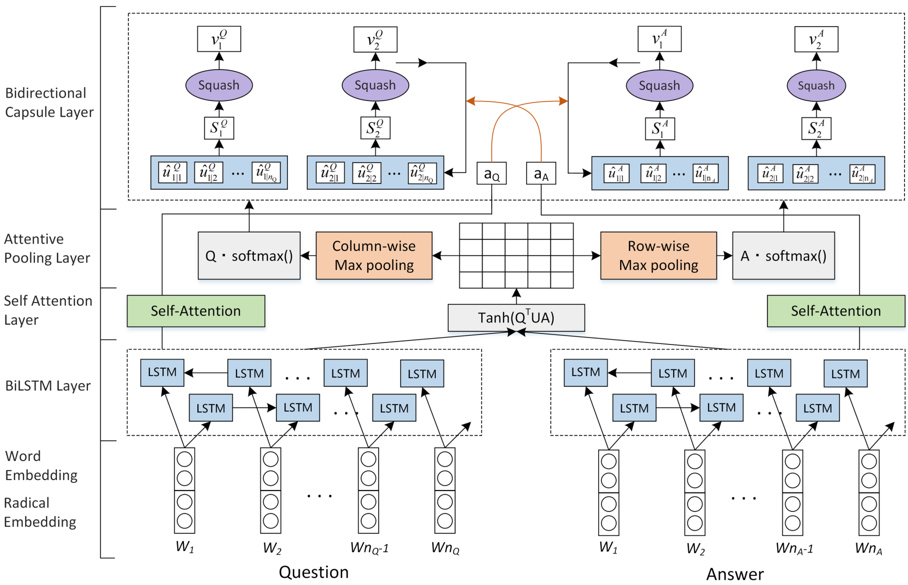

Fig. 1. Overview of our answer selection model

2.1 Input Representations

We perform Chinese word segmentation using an open source Chinese word segmentation tool and remove the stopwords in advance. For the questions and answers, each word is encoded into a real-valued vector representation by looking up the pre-trained word embedding. Given a question Q and a set of candidate answers $\{A_i\}$, we can represent the inputs as embedding matrix $W^Q \in R^{d \times n_Q}$ and $W^A \in R^{d \times n_A}$, where d denotes the dimension of the word embedding and n_Q is the number of words in question Q and n_A is the number of words in the answer A_i. In the input word embedding, each word w_i is embedded into a vector $w_i = \left[w_i^{word} + w_i^{rad}\right]$, which is composed of two sub-vectors: the word embedding W^{word}, the Chinese radical embedding W^{rad}.

For word embedding, we adopt word2vec based embedding [17], which learns low-dimensional continuous vector representations of words. In this paper, we used the CBOW model to per-train Chinese word embedding, and trained the Chinese word embedding using texts from Chinese medical literature.

For radical embedding, we use a convolutional approach to extracting local features around each radical of the character, and combine them using max-pooling to generate a fixed-sized radical embedding for each Chinese character. In Chinese, characters are composed of specific radicals, which serve as the basic unit for building character meanings [18, 19]. The radicals are particularly useful in medical text processing, because the radicals can indicate different descriptions of diseases and symptoms, and similar radical sequences usually convey similar semantic information. For example, “呕吐” and “喉咙痛” are Chinese medical entities, we obtain the radical of Chinese characters from online Xinhua Dictionary[1]. The radical of them is "□ □" and "□ □疒". They all have the meaning related to mouth because they share the same radical "□ □". Therefore, we believe the radical embedding could contribute the representations of questions and answers.

2.2 Interactive Information Extraction

To capture the interactive information between questions and answers, we adopt Bi-LSTM to obtain the sentence representations. Long Short-Term Memory (LSTM) is designed to deal with the long-distance sequences and tackle the gradient vanishing problems of RNN [20]. The bi-directional LSTM (Bi-LSTM) [11] seeks to obtain two directions of information from word sequences. Specifically, we employ Bi-LSTM to encode the question Q and the candidate answer A as $H^Q \in R^{2d \times n_Q}$ and $H^A \in R^{2d \times n_A}$.

Inspired by attentive pooling networks [21], we also use a two-way attention mechanism to represent the questions and answers. Through the attention mechanism, the information from the question Q can influence the computation of the answer representations, and vice versa. After we obtain the question and answer hidden features H^Q and H^A by Bi-LSTM, we compute the interactive matrix G as follows:

$$G = tanh\left({H^Q}^T U H^A\right) \tag{1}$$

Where U is the parameter matrix. Then we apply the Softmax function to the vector $g^Q \in R^{n_Q}$ and $g^A \in R^{n_A}$ by the column-wise and row-wise max-pooling over G. Finally, the new representations $H^{Q'}$ and $H^{A'}$ are computed as $H^{Q'} = H^Q softmax(g^Q)$ and $H^{A'} = H^A softmax(g^A)$, where $H^{Q'} \in R^{2d*n_Q}$ and $H^{A'} \in R^{2d*n_A}$. We compute the question representation $H^{Q'}$ with answer information and the answer representation $H^{A'}$ with question information, so that the interactive information can be fully encoded in the final representations.

2.3 Bi-directional Capsule Network

To further capture the relevance of candidate answers, we propose a bi-directional capsule network model for the answer selection. Capsule network was originally proposed for digit recognition from images by Hinton et al. [22] and exhibited powerful capability in related tasks. Specifically, a capsule involves a group of neurons, and the number of the capsule equals the number of classification categories in specific tasks. We adopt capsule networks to generate the capsules using the dynamic routing algorithm. The algorithm converts the low-level sentence information to the high-level vector representation by eliminating trivial features of sentences. The process replaces the max-pooling in the original model with feature clustering, which greatly contribute to the improvement of classification accuracy.

[1] http://tool.httpcn.com/Zi/.

In the modified model for answer selection, the low-level capsule is denoted as $h_i^{Q'}$. Between each two layers l and l + 1, the prediction vectors $\hat{u}_{j|i}^{Q}$ is produced by multiplying the output $h_i^{Q'}$ of the capsule layer l and a weight matrix W_{ij}^{Q} as follows.

$$\hat{u}_{j|i}^{Q} = W_{ij}^{Q} h_i^{Q'} \tag{2}$$

Then, the total input s_j^{Q} to the layer $l + 1$ question capsule is generated by the weighted sum over all $\hat{u}_{j|i}^{Q}$. The obtained s_j^{Q} can focus on both local features and contextual information from the question word sequence, which is formalized as follows.

$$s_j^{Q} = \sum\nolimits_i c_{ij}^{Q} \hat{u}_{j|i}^{Q} \tag{3}$$

Where the c_{ij}^{Q} is a coupling coefficient that is determined by the dynamic routing algorithm with the number of iterations as r. Furthermore, the capsules $l + 1$ are generated by a non-linear squashing function as follow:

$$v_j^{Q} = \frac{\left\| s_j^{Q} \right\|^2}{1 + \left\| s_j^{Q} \right\|^2} \frac{s_j^{Q}}{\left\| s_j^{Q} \right\|} \tag{4}$$

Where v_j^{Q} is the outputted question vector in the $l + 1$ capsule layer. The value of the outputted capsule vector v_j^{Q} indicates the final classification probability. The non-linear squashing function is used to limit the value of v_j^{Q} in the range [0, 1]. We also obtain the answer vector v_j^{A} in a similar manner. Then, we average the sum of v_j^{Q} and v_j^{A} as the result v_j. In our work, the number of the capsule is set as 2, namely $j = 2$.

$$v_j = \frac{1}{2}\left(v_j^{Q} + v_j^{A}\right) \tag{5}$$

Dynamic Routing Algorithm. Inspired by the attention-based routing algorithm [16], we propose a bi-directional capsule network with dynamic routing for answer selection, which is a variant of the original capsule network. The question dynamic routing algorithm focuses more on the matching of each pair of question and answer. Moreover, we use the self-attention mechanism to combine local information at a higher level through local perception, which helps to reduce useless information in matching.

Specifically, let $h^{Q} = \left\{ h_1^{Q}, h_2^{Q}, h_3^{Q}, \ldots, h_{an_Q}^{Q} \right\}$ denote the hidden vectors of the question after passing through the Bi-LSTM layer. We compute the weighted representation of each sentence as:

$$\mathrm{T} = tanh\left(\mathrm{W}_Q h_i^Q + \mathrm{b}\right) \tag{6}$$

$$a_Q = \sum\nolimits_t \mathrm{softmax}\left(w^T \mathrm{T}\right)^T h_i^Q \tag{7}$$

Based on the self-attention mechanism, a_Q is determined on h_i^Q. W_Q is the attentive weight matrices and w is the attentive weight vector. With the help of the self-attention mechanism, our model learns to represent question vector representation by focusing on different aspects of questions, and represent answer vector a_Q representation by self-attention weights. The detailed algorithm is summarized in Algorithm 1.

Algorithm 1: Question Capsule Dynamic Routing Algorithm

procedure ROUTING($\hat{u}_{j|i}^Q$, r, l)
for the question capsule i in layer l and the question capsule j in layer $l+1$: initialize the logits of coupling coefficients $b_{ij}^Q = 0$
for r iterations **do**
  $c_i^Q = softmax(b_{ij}^Q)$
  $\alpha_{j|i} = \alpha({a_Q}^T \hat{u}_{j|i}^Q)$
  $\hat{u}_{j|i}^{Q'} = \alpha_{j|i} \hat{u}_{j|i}^Q + (1 - \alpha_{j|i})\, a_Q$
  $s_j^Q = \sum_i c_{ij}^Q\, a_Q \hat{u}_{j|i}^{Q'}$
  $v_j^Q = g(s_j^Q)$ non-linear squashing function
  $b_{ij}^Q = b_{ij}^Q + \hat{u}_{j|i}^Q \cdot v_j^Q$
end for
return v_j^Q

In the algorithm, the coupling coefficient c between all the question capsule i in l layer and all the question capsule j in $l + 1$ layer is determined by Softmax function with initial logits b_i^Q. The answer capsule network is in the same form as the question capsule. The variant of the capsule network is more suitable for CQA, which can make the most use of interactive information. To train the proposed model, we use a margin loss for the answer selection as follows.

$$L_j = Y_j max\left(0,\, m^+ - \left\|v_j\right\|\right)^2 + \lambda\left(1 - Y_j\right) max\left(0,\, \left\|v_j\right\| - m^-\right)^2 \tag{8}$$

We minimize the margin loss L_j. Namely, if the candidate answer is the correct answer, we set $Y_j = 1$. Otherwise, we set $Y_j = 0$. λ is the weight on the absent classes, and m^+ is the top margin and m^- is the bottom margin. We set $\lambda = 0.5$, $m^+ = 0.9$ and $m^- = 0.1$ in our implementation.

3 Experiment

3.1 Experimental Dataset and Setting

We use the data from the “2018 IEEE HotICN Knowledge Graph Academic Competition” evaluation task to evaluate the proposed model. The training data contain

1000 questions, each question involves 10 candidate answers. There is only one correct answer for each question. The data are designed by IEEE HotICN2018 and Shenzhen Medical Information Center. In our experiments, we use random over-sampling to increase the number of positive samples. We use 20% of the training data as the development set. The test set consists of 200 questions each with 10 candidate answers. The dataset is publicly available at https://hoticn.com/competition.html.

In our experiments, we use Keras to implement our proposed model. We examine the effect of parameters on the report results, and the parameter settings of the final model have shown in Table 2. In this paper, we set the length of a question as 25, and the length of an answer to 150. We use the official evaluation measure for the competition, including mean reciprocal rank (MRR) and P@1. Because there is only one correct answer, the results of the MAP and MRR are the same.

Table 2. The parameter settings of model

Parameter name	Description	Value
EMBEDDING_WORD_DIM	Word embedding dimension	200
EMBEDDING_EAD_DIM	Radical embedding dimension	20
LSTM	Number of units	50
r	Number of iterations	2
Epochs	Maximum of epochs	10
Batch size	Batch size	8

3.2 Performance Evaluation

In order to evaluate the proposed model, we compare our results with state-of-the-art baseline models from Tan et al. [10], Royal et al. [5], Alexis et al. [23] and Ren et al. [24]. These methods are implemented for biomedical CQA tasks in our experiments, which obtain the best performances in the competition. Tan et al. [10] developed an attention mechanism for the purpose of constructing better answer representations according to the input question. Royal et al. [5] proposed a model which CNN and idf-weighted were joint learning for answer selection. Alexis et al. [23] proposed BiLSTM with max-pooling for answer selection. Ren et al. [24] proposed a neural selection model based on the combination of Bi-LSTM and Attention mechanism which is the top result in the academic competition. The performance of each model is shown in Table 3.

Table 3. Performance comparison with existing methods

Method	MRR	P@1
Tan et al.# [10]	52.08%	37%
Royal et al.# [5]	53.84%	37.5%
Alexis et al.# [23]	50.59%	37%
Ren et al. [24]	50.06%	35%
Our model	**54.05%**	**38.5%**

Models with # are our implementations.

From Table 3, we observe that MRR and P@1 reach 54.05% and 38.5% by our model, which is better than other existing methods. This is because these baseline methods process the questions and answers independently at the encoding and matching steps, which may ignore prominent implicit relationship between questions and answers. Our method outperforms other baselines for certain reasons. First, we extracted the interactive information from the question and answer. Second, bi-directional capsule network interacts with the question and answer vector representations, which jointly learns the representations of questions and the candidate answers. Therefore, our method obtained state-of-the-art performance in the test set. Our model has exceeded the performance of the top result in the Academic Competition (improvements of 3.99% and 3.5% in MRR and P@1, respectively).

3.3 Effect of Different Layers in the Proposed Model

In order to verify the contribution of bi-directional capsule networks and radical-level feature in our model, we implement extra baselines on the Chinese CQA dataset to analyze the improvement contributed by each part of our model. We train the neural networks model separately from each other. The result is shown in Table 4.

Table 4. Performance on different layers in our model

Method	MRR	Δ	P@1	Δ
Our Model	**54.05%**	–	**38.5%**	–
w/o question capsule network	53.67%	−0.38	38%	−0.5
w/o answer capsule network	52.54%	−1.51	38%	−0.5
w/o radical embedding	51.63%	−2.42	37.5%	−1

From Table 4, we observe that our models with all the layers achieve the best performance, which shows that the bi-directional capsule networks can capture abundant information from different aspects. Specifically, we have the following observations, (1) our bi-directional capsule network obtains better results on MRR and P@1 scores than the other methods. (2) the model with answer capsule networks obtains better results than the model with question capsule networks. Because the length of the answer sentence is longer than the length of the question sentence, the sentence vector contains more information. (3) the radical-level features can incorporate additional information that benefits semantic representation of the Chinese sentence, and improve the results.

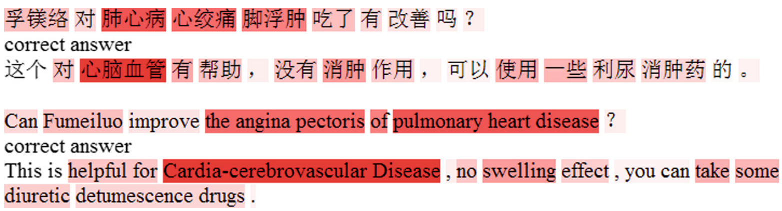

Fig. 2. Self-attention heat map of dataset

From Fig. 2, we can see that the heat map of a question and the correct answer by the output of Self-Attention from dataset. The stronger red color of a word in the sentence, the larger weight of that word. The Self-Attention can focus on the important part of a sentence for semantic representation.

3.4 The Effect of Random Over-Sampling

In our model, we use random over-sampling to increase the number of positive data in the training set to relieve the problem caused by the imbalance of positive and negative data. The imbalance between positive and negative proportions of the sample may seriously affect the accuracy of the experimental results so that the model could not learn the dataset information well. A plot comparing the random over-sampling parameter by the increase is shown in Fig. 3.

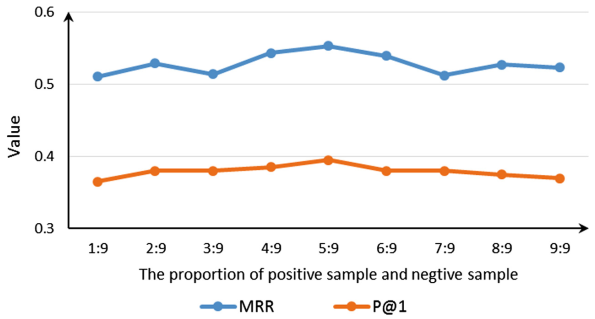

Fig. 3. The MRR and P@1 trend with random over-sampling parameter

In Fig. 3, when the random over-sampling parameter is set as 4, we achieve the optimal performance in terms of the MRR and P@1. Namely, the proportion of positive and negative data in the training dataset is 5:9. The performance increases with the increase of the proportion of positive samples, but deceases due to overfitting when setting larger than 5:9. In our method, we did not use under-sampling because of the small amount of data.

3.5 Effect of Routing Iteration

The coupling coefficient c is updated by the dynamic routing algorithm, which is the connections between capsule i in l layer and the capsule j in $l + 1$ layer. To analyze the effect of the number of iteration, we test the bi-directional capsule networks with a series of interactions on Chinese CQA corpus. We also plot a learning curve to show the training loss and the evaluation loss over epochs with different iterations of routing. As shown in Fig. 4(a), bi-directional capsule networks with 2 iterations of routing converge to a lower loss at the end than the capsule network with 1 or 3 iteration. From Fig. 4(b), we observe that the bi-directional capsule networks with 2 iterations of routing obtain the best performance and the result is more stable. So we utilize 2 iterations in our experiments.

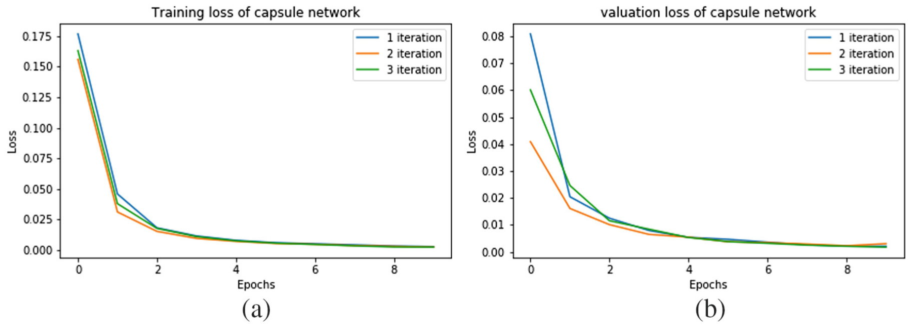

Fig. 4. (a) Training loss of capsule network. (b) Valuation loss of capsule network

4 Conclusion

In this paper, to better capture the interaction between the question-answer pair, we propose a novel neural network model for Chinese biomedical CQA. Our model adopts the bi-directional capsule networks to give more attention to the different aspects of the matching between the answers and questions so that the model captures detailed information and ignore useless information in the learned representations. Meanwhile, we incorporate Chinese radical-level information to Bi-LSTM and obtain additional semantic information. We conduct extensive experiments, and demonstrate that our model achieves the state-of-the-art performance on answer selection in the Chinese biomedical CQA task.

Acknowledgments. This work is partially supported by grant from the Natural Science Foundation of China (No. 61572102, 61702080, 61772103), the Postdoctoral Science Foundation of China (No. 2018M641691), the Foundation of State Key Laboratory of Cognitive Intelligence, iFLYTEK, P.R. China (COGOS-20190001).

References

1. Bhandwaldar, A., Zadrozny, W.: UNCC QA: biomedical question answering system. In: Proceedings of the 6th BioASQ Workshop A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, pp. 66–71 (2018)
2. Jin, Z.X., Zhang, B.W., Fang, F., et al.: Health assistant: answering your questions anytime from biomedical literature. Bioinformatics (2019)
3. Chen, L., Jose, J.M., Yu, H., et al.: A semantic graph based topic model for question retrieval in community question answering. In: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, pp. 287–296. ACM (2016)
4. Lai, T.M., Bui, T., Li, S.: A review on deep learning techniques applied to answer selection. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 2132–2144 (2018)
5. Sequiera, R., Baruah, G., Tu, Z., et al.: Exploring the effectiveness of convolutional neural networks for answer selection in end-to-end question answering. arXiv preprint arXiv:1707.07804 (2017)

6. Fang, H., Wu, F., Zhao, Z., et al.: Community-based question answering via heterogeneous social network learning. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
7. Zhao, Z., Lu, H., Zheng, V.W., et al.: Community-based question answering via asymmetric multi-faceted ranking network learning. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
8. Maleewong, K.: Predicting quality-assured consensual answers in community-based question answering systems. In: Meesad, P., Boonkrong, S., Unger, H. (eds.) Recent Advances in Information and Communication Technology. Advances in Intelligent Systems and Computing, pp. 117–127. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-40415-8_12
9. Tran, Q.H., Lai, T., Haffari, G., et al.: The context-dependent additive recurrent neural net. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 1274–1283 (2018)
10. Tan, M., Dos Santos, C., Xiang, B., et al.: Improved representation learning for question answer matching. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 464–473 (2016)
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
12. Liwei, Y., Lei, S., Peng, S.: Attribute reduction for Chinese question classification. In: 2016 Chinese Control and Decision Conference (CCDC), pp. 5488–5492. IEEE (2016
13. Yu, L., Hermann, K.M., Blunsom, P., et al.: Deep learning for answer sentence selection. arXiv preprint arXiv:1412.1632 (2014)
14. Sabour, S., Frosst, N., Hinton, G.E.: Dynamic routing between capsules. In: Advances in Neural Information Processing Systems, pp. 3856–3866 (2017)
15. Zhao, W., Ye, J., Yang, M., et al.: Investigating capsule networks with dynamic routing for text classification. arXiv preprint arXiv:1804.00538 (2018)
16. Zhang, X., Li, P., Jia, W., et al.: Multi-labeled relation extraction with attentive capsule network. arXiv preprint arXiv:1811.04354 (2018)
17. Mikolov, T., Chen, K., Corrado, G., et al.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
18. Li, Y., Li, W., Sun, F., et al.: Component-enhanced chinese character embeddings. arXiv preprint arXiv:1508.06669 (2015)
19. Dong, C., Zhang, J., Zong, C., Hattori, M., Di, H.: Character-based LSTM-CRF with radical-level features for Chinese named entity recognition. In: Lin, C.-Y., Xue, N., Zhao, D., Huang, X., Feng, Y. (eds.) ICCPOL/NLPCC-2016. LNCS (LNAI), vol. 10102, pp. 239–250. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-50496-4_20
20. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Networks **5**(2), 157–166 (1994)
21. Santos, C., Tan, M., Xiang, B., et al.: Attentive pooling networks. arXiv preprint arXiv:1602.03609 (2016)
22. Hinton, G.E., Krizhevsky, A., Wang, S.D.: Transforming auto-encoders. In: Honkela, T., Duch, W., Girolami, M., Kaski, S. (eds.) ICANN 2011. LNCS, vol. 6791, pp. 44–51. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21735-7_6
23. Conneau, A., Kiela, D., Schwenk, H., et al.: Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364 (2017)
24. Ren, Y., Zhang, T., Liu, X., et al.: End-to-end answer selection via attention-based Bi-LSTM network. In: 2018 1st IEEE International Conference on Hot Information-Centric Networking (HotICN), pp. 264–265. IEEE (2018)