Research on Question Answering Technology Based on Bi-LSTM

Xiaoya Suna, Xinmeng Lib

College of Systems Engineering, National University of Defense Technology, Changsha, China

asunxiaoya_nudt@126.com, bxml.nudt@hotmail.com

Abstract. Question Answering (QA) is always the key issue in the Natural Language Process (NLP). This paper mainly researches the question answering model based on Bi-directional Long Short Term Memory networks (Bi-LSTM), with the use of the WebQA dataset to train the model. Experiment results show that ACC@1 in Bi-LSTM model is 55.00%, ACC@3 is 73.24%, and ACC@10 reaches to 86.64%. The Bi-LSTM model outperforms Best Match 25 (BM25) algorithm in all of the three metrics, which proves the superiority of Bi-LSTM model.

1. Introduction

From early expert systems to today's search engines, people have been pursuing accuracy and fast access to information resources. The question answering (QA) technology based on deep semantic understanding and the research in related fields are of great significance. With the development of big data technology, the retrieval system based on shallow semantic analysis cannot meet the needs gradually. In recent years, the boom of deep learning technology has brought an opportunity for the development of Natural Language Process (NLP) technology [1]. The application of neural network technology and large-scale data accumulated on forums to solve NLP related tasks has become a hot research topic.

Question answering needs to obtain more information from the question, while the model needs to have deeper memory and reasoning ability, which is different from the information retrieval system. According to the implementation principle, the question answering technology can be divided into three categories: rule-based methods, keyword-based retrieval methods and neural network-based methods. The neural network-based method avoids cumbersome feature engineering and can be used for end-to-end training, which is favored by many researchers. At this point, the development of various neural network models, such as Recurrent Neural Networks (RNN), Long Short Term Memory Networks (LSTM), Gated Recurrent Unit (GRU), Attention Mechanism, Memory Network have helped to make great progress in many natural language related tasks [2-7].

This paper builds a question answering model based on Bi-LSTM, and uses Baidu's WebQA dataset for training. The model is able to retrieve the best matching answer from the answer set based on the question. The BM25 algorithm is used as a benchmark method to verify the superiority of the Bi-LSTM model.

The structure of this paper is as follows: Section II discusses related work. Section III shows the structure of LSTM, Bi-LSTM and BM25. Section IV details our experiments and discussions. Section V concludes this work.
2. Related Work
In this section, we will elaborate on traditional question answering technology and some question answering techniques based on neural networks.

2.1 Traditional Question Answering Technology
The traditional question answering technology generally consists of three parts: problem analysis, information retrieval and answer extraction. As shown in Fig. 1. Problem analysis is to classify the problem, extract keywords from the problem and expand the keywords in time. If it is a Chinese question answering, you also need to perform word segmentation. Information retrieval is to search the relevant documents in the data source through the keyword set obtained by the problem analysis, and finally form a candidate set. The answer extraction is to find candidate answers in the candidate set, sort the candidate answers according to some rules and return the best answer.

Cui et al. proposed a method based on the results returned by an external search engine to obtain keywords, and the point mutual information between the finally returned words can be used to explore the topic [8]. Stoyanchev et al. proposed the use of named entities, noun phrases, verb phrases and prepositional phrases for semantic match, which can greatly improve the accuracy and recall rate of information retrieval [9]. Ravichandran and Hovy manually constructed some features to extract answers based on the surface information of the text [10].

Traditional question answering techniques are mostly based on analyzing sentence structure and performing keyword matching. Although these methods are effective, there are some disadvantages. For example, there are a large number of manual pre-processing operations and excessive reliance on syntactic parsing of questions, resulting in such methods being inefficient and poorly scalable.

2.2 Neural Network Based Question Answering Technology
Kalchbrenner et al. firstly applied Convolutional Neural Networks (CNN) to the field of NLP [11]. Hu et al. used convolutional neural networks to extract semantic information between sentence contexts and constructed a model to solve sentence similarity matching [12]. M. Feng et al. proposed a deep learning framework for sentence matching using single-layer CNN and multi-layer CNN [13]. W. Yin et al. used CNN with attention mechanism to process the semantic information expression of generated sentences, taking into account the characteristics of different granularities at the same time [14].

Tan et al. took the advantage of recurrent neural networks for sentence sequence instead of convolutional neural networks, and used bidirectional recurrent neural networks to extract semantic information of sentences [15] [16]. LSTM is one of the earliest effective variants of RNN. Hermann et al. used multiple machine reading mechanisms on LSTM to effectively extract the relationship between questions and sentences [17]. Wang et al. proposed using LSTM to learn the semantics of question and answer to select the correct answer [18]. However, LSTM can only use the previous context. Bi-LSTM can perform better by using both previous and posterior context.

In the more fine-grained classification, for the classification tasks of positive, neutral, and negative, it is necessary to highline the importance of the interaction between emotional words, degree words, and negative words. Bi-LSTM can capture the Bi-directional semantic dependence by using both previous and posterior context. Accordingly, this paper uses Bi-LSTM to make better use of the semantic features of the sentence level.
3 Methodology

RNN are widely used in NLP tasks, but it is difficult to train deeper RNN model because of gradient disappearance and gradient explosion problems [19]. The gated recurrent unit network such as LSTM solves this problem perfectly [20].

Fig. 2 shows the basic structure of the LSTM model. The outer layer of the LSTM network is a sequence model. There are multiple control gates and memory units inside the node. The introduction of control gates and memory units allow the LSTM to selectively forget or enhance the information of the input sequence, which can solve the gradient disappearance and gradient explosion problems effectively.

Figure 2 LSTM model.

LSTM network include three control gates: forget gate $f_t$, the input gate $i_t$ and the output gate $o_t$, which are calculated as follows respectively:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$  (1)

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$  (2)

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$  (3)

The input of each gate represents as the input $x_i$ of the current node and the output $h_{t-1}$ of the previous node. The output value is mapped to the range of (0, 1) by the sigmoid function. $W$ is the weight parameter, while $b$ is the offset parameter.

The information of memory unit $C_t$ is determined by $x_i$, $h_{t-1}$ as well as historical memory $C_{t-1}$, and is adjusted by the forget gate $f_t$ and the input gate $i_t$:

$$\bar{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$  (4)

$$C_t = f_t \times C_{t-1} + i_t \times \bar{C}_t$$  (5)

The final output of the node is determined by memory unit $C_t$ and output gate $o_t$:

$$h_t = o_t \times \tanh(C_t)$$  (6)

The memory unit $C_t$ and output $h_t$ of the current node will have an impact on the next node as historical information iteratively.

The Bi-LSTM can process the context of a sentence from the forward directions and backward directions to find accurate contextual associations of words. The output of Bi-LSTM is $[\overrightarrow{h}_t, \overleftarrow{h}_t]$, which is obtained by concatenating the output of the forward LSTM $\overrightarrow{h}_t$ and the output of the backward LSTM $\overleftarrow{h}_t$.

The result of Bi-LSTM is maximized and then mapped to the range of (-1, 1) by the activation function tanh to obtain the model’s output:
\[
Bi = \tanh \left( \text{Pool}_{\text{max}} \left( \vec{h}_i, \vec{h}_i \right) \right) \quad (7)
\]

The match score between the question and the answer is measured by the cosine similarity. Let \( \vec{q} = Bi(q) \) and \( \vec{a} = Bi(a) \) represent the Bi-LSTM encoded vector of the question \( q \) and the answer \( a \) respectively, then the match score \( M \) can be expressed as:

\[
M = \cos(q,a) = \frac{\vec{q}^T \vec{a}}{||\vec{q}|| ||\vec{a}||} \quad (8)
\]

\( ||\vec{a}|| \) represents the norm of the vector.

This paper uses the BM25 algorithm as a benchmark method to compare with the Bi-LSTM model results. BM25 is a sorting algorithm used in information retrieval, while it is also widely used in the semantic matching task of question answering. The definition of the BM25 algorithm is:

\[
\text{BM25}(q,a) = \sum_{w_j \in q} \text{IDF}(w_j) \frac{f(w_j,a) \cdot (k+1)}{f(w_j,a) + k \cdot (1 - b + b \cdot \frac{|q|}{a_{avg}})} \quad (9)
\]

where \( w_j \) represents the \( j \)th word in the question \( q \). \( \text{IDF}(w_j) \) is the inverse document frequency (IDF) of \( w_j \). \( |a| \) is the length of the answer \( a \). \( a_{avg} \) is the average length of all the answers. \( f(w_j,a) \) is the frequency of the words \( w_j \) in the answer \( a \). \( k \) and \( b \) are the model parameters.

4. Experiments

4.1 Datasets

The model is trained using the WebQA dataset, which is constructed by Baidu Company using Baidu Knows [21]. Table 1 is the statistics of the WebQA dataset. The dataset has 36,181 questions. The number of each question’s answers is from 2 to 12 and there are 140,897 answers totally. The data set is divided into three parts: training set, development set and test set. Table 2 is an example of a question and answer in the data set. \( q \) is the problem and \( a \) is the answer that matches the question.

| Data Set     | Number of Questions | Number of Answers |
|--------------|---------------------|-------------------|
| training set | 30000               | 116476            |
| development set | 3090             | 12298             |
| test set     | 3091                | 12123             |
| total        | 36181               | 140897            |

Table 2 A Sample of WebQA Data Set

| \( q \) | “March Street” is a traditional festival in which minority of China? |
|---------|---------------------------------------------------------------------|
| \( a1 \) | The minority of Bai \(<e>\) The Bai Nationality have many traditional festivals. The “March Street” with thousands of years of history is the biggest annual festival of the Bai people. It is now called “March Street National Festival”. |
| \( a2 \) | It mainly lives in the Dali Bai Autonomous Prefecture of Yunnan Province, and the rest are distributed in various parts of Yunnan, Bijie Prefecture of Guizhou Province and Liangshan Prefecture of Sichuan. The Bai nationality claim to be “Bai Huo”, “Bai Ni”, “Bai Zi”, etc which means “the Bai People” in Chinese. In 1956, “Bai” was officially named according to the willingness of its people. There are many traditional festivals of the Bai people. The “March Street”, |

4
which has a history of more than a thousand years, is the biggest annual festival of the Bai people. It is now called the “March Street National Festival”.

In addition, all ethnic minorities also keep their own traditional festivals, such as the Songkran Festival of the Dai people, the Nadam meeting of the Mongolian people, the Torch Festival of the Yi people, the Danu Festival of the Yao people, the March Street of the Bai people, the songs of the Zhuang people, and the Tibetans. Tibetan calendar year and hope fruit festival, Miao nationality flower jumping festival and so on.

Bai Festival March Street Dali March Street, held every year from March 15th to 22nd, is a traditional event of Dali national material culture exchange, called Guanyin City or Guanyin, with more than 1,000 years of history. In 1991, the Standing Committee of the Dali State People's Congress designated this most traditional festival of the minorities in Dali as “Dali Bai Autonomous Prefecture National Festival.”

| Evaluation Metric |
|--------------------|
| For each question in the test set, the match score to each answer in the answer library is calculated by cosine similarity. The result is measured by ACC@k. The match is considered successful when the correct answers are in the top k candidate answers with the highest similarity. |

| Setup |
|-------|
| The experiment applies Tensorflow Toolkit to build the neural network framework. All questions and answers are segmented by jieba, a widely used word segmentation tool. The word vectors are initialized by zhwiki, which is pre-trained by Wikipedia Simplified Chinese corpus. The dimension of word vector is 50 dimensions, and the word vector remains unchanged during training. |
| The experiment uses the hinge loss function for training. For each question, a correct answer and a randomly selected answer are paired as a training sample. To increase the training effect, each correct answer expands 5 training samples. There are totally 582,380 training samples. The experiment uses stochastic gradient descent (SGD) algorithm to update the parameters, and the batch size is set to 20. The model is trained for 80 epochs and the initial training learning rate is set to 0.4. The learning rate is halved after every 20 epochs. |

| Results |
|---------|
| Table 3 compares the results of the Bi-LSTM model and BM25 model, with k values of 1, 3, and 10. |

| Table 3 Experimental Results of Bi-LSTM and BM25 |
|-----------------------------------------------|
|                  | ACC@1 | ACC@3 | ACC@10 |
| Bi-LSTM          | 55.00% | 73.24% | 86.64%  |
| BM25             | 50.98% | 71.60% | 82.49%  |

The results show that the accuracy of this two algorithms increases with the loosening of the matching restriction. If the restriction is limited to match the correct answer at one click, the correct rate of the Bi-LSTM model is 55.00%, and the correct rate of the BM25 algorithm is 50.98%. Under the same conditions, Bi-LSTM has a higher accuracy rate than BM25.

5. Conclusion
This paper mainly discusses the question answering technology based on Bi-LSTM model and uses WebQA data set for training. The final result shows that the Bi-LSTM model outperforms BM25 remarkably. The model will be optimized continuously and we look forward to achieving better performance on QA tasks.
Acknowledgment
The work described in this paper is sponsored by the National Natural Science Foundation of China under Grant No. 61473300 and the National Science Foundation of Hunan Province (NSFHP) project 2017JJ3371, China.

References
[1] Y. Lecun, Y. Bengio, G. Hinton, “Deep learning,” Nature, vol. 521, pp. 436-444, 2015.
[2] T. Mikolov, M. Karafiát, L. Burget, J. H. Cernocký, S. Khudanpur, “Recurrent neural network based language model,” Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, pp. 1045-1048, September 2010.
[3] T. Ming, B. Xiang, B. Zhou, “LSTM-based Deep Learning Models for non-factoid answer selection,” 2015.
[4] Y. Cui, T. Liu, Z. Chen, S. Wang, G. Hu, “Consensus Attention-based Neural Networks for Chinese Reading Comprehension,” 2016.
[5] Y. Cui, Z. Chen, S. Wei, S. Wang, T. Liu, G. Hu, “Attention-over-Attention Neural Networks for Reading Comprehension,” 2016.
[6] B. Hu, Z. Lu, H. Li, Q. Chen, “Convolutional Neural Network Architectures for Matching Natural Language Sentences,” Advances in Neural Information Processing Systems, pp. 2042-2050, 2015. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.
[7] S. Zhang, X. Zhang, H. Wang, J. Cheng, P. Li, Z. Ding, “Chinese Medical Question Answer Matching Using End-to-End Character-Level Multi-Scale CNNs,” Applied Sciences, vol. 7, pp. 767, 2017.
[8] H. Cui, M. Y. Kan, T. S. Chua, “Unsupervised Learning of Soft Patterns for Generating Definitions from Online News,” Proceedings of the 13th international conference on World Wide Web, New York, pp. 17-20, May 2004.
[9] S. Stoyanchev, Y. C. Song, W. Lahti, “Exact phrases in information retrieval for question answering,” Coling 2008: Proceedings of the 2nd workshop on Information Retrieval for Question Answering, Association for Computational Linguistics, 2008.
[10] D. Ravichandran, E. Hovy, “Learning surface text patterns for a question answering system,” Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, 2002.
[11] N. Kalchbrenner, E. Grefenstette, P. Blunsom, “A convolutional neural network for modelling sentences,” Eprint Arxiv, 2014.
[12] B. Hu, Z. Lu, H. Li, Q. Chen, “Convolutional Neural Network Architectures for Matching Natural Language Sentences,” Advances in Neural Information Processing Systems, vol. 3, pp. 2042-2050, 2015.
[13] M. Feng, B. Xiang, M. R. Glass, L. Wang, B. Zhou, “Applying deep learning to answer selection: A study and an open task,” Automatic Speech Recognition and Understanding, 2015.
[14] W. Yin, H. Schütze, B. Xiang, B. Zhou, “ABCNN: Attention-Based Convolutional Neural Network for Modeling Sentence Pairs,” Computer Science, 2015.
[15] M. Tan, C. d. Santos, B. Xiang, B. Zhou, “LSTM-based Deep Learning Models for Non-factoid Answer Selection,” Computer Science, 2016.
[16] M. Tan, C. d. Santos, B. Xiang, B. Zhou, “LSTM-based deep learning models for non-factoid answer selection,” 2015.
[17] K. M. Hermann, T. Kočiský, E.Grefenstette, L. Espeholt, W. Kay, M. Suleyman, et al., “Teaching Machines to Read and Comprehend,” Computer Science, 2015.
[18] D. Wang, E. Nyberg, “A long short-term memory model for answer sentence selection in question answering,” Meeting of the Association for Computational Linguistics and the International Joint Conference Natural Language Processing, pp.707-712, 2015.
[19] J. F. Kolen, S. C. Kremer, “Gradient flow in recurrent nets: the difficulty of learning long-term dependencies,” 2001.
[20] S. Hochreiter, J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, pp. 1735-1780, 1997.
[21] P. Li, W. Li, Z. He, X. Wang, Y. Cao, J. Zhou, et al., “Dataset and Neural Recurrent Sequence Labeling Model for Open-Domain Factoid Question Answering,” Computer Science, 2016.