YNU-HPCC at IJCNLP-2017 Task 5: Multi-choice Question Answering in Exams Using an Attention-based LSTM Model

Hang Yuan, You Zhang, Jin Wang and **Xuejie Zhang**
School of Information Science and Engineering
Yunnan University
Kunming, P.R. China
Contact:xjzhang@ynu.edu.cn

Abstract

A shared task is a typical question answering task that aims to test how accurately the participants can answer the questions in exams. Typically, for each question, there are four candidate answers, and only one of the answers is correct. The existing methods for such a task usually implement a recurrent neural network (RNN) or long short-term memory (LSTM). However, both RNN and LSTM are biased models in which the words in the tail of a sentence are more dominant than the words in the header. In this paper, we propose the use of an attention-based LSTM (AT-LSTM) model for these tasks. By adding an attention mechanism to the standard LSTM, this model can more easily capture long contextual information. Our submission ranked first among 35 teams in terms of the accuracy at the IJCNLP-2017 multi-choice question answering in Exams for all datasets.

1 Introduction

Designing an intelligent question answering system that can answer general scientific questions has always been an important research direction in natural language processing. In this field, various scholars have made very important contributions before, for example, IBM insuranceQA and The Allen AI Science Challenge on the Kaggle (Schoenick et al., 2017). Multi-choice question answering in exams is a typical natural language processing task. For this task, it is required to design a question and answer system that can solve the examination of a general subject, such as biology and chemistry. The task can be considered as a binary classification that requires a system for determining whether the answer of the candidate is correct or not.

In the recent research field of question answering, various methods have proved to be highly useful. The difference between the existing methods is mainly reflected in the access to the knowledge and reasoning framework. Clark et al. (2013) proposed a method based on text statistical rules. Clark (2015) described how to obtain more information from the background knowledge base, i.e., they introduced the use of background knowledge to build the best scene. Sachan et al. (2016) presented a unified max-margin framework that learns to detect the hidden structures that explain the correctness of an answer when provided with the question and instructional materials. A system that extracts information from the corpus for automatic generation of test questions was designed by Khot et al. (2015), whereas a structured inference system based on integer linear programming was proposed by Khashabi et al. (2016). A more complex method is presented in (Clark et al., 2016). This model operates at three levels of representation and reasoning: information retrieval, corpus statistics, and simple inference over a semi-automatically constructed knowledge base.

In this paper, we mainly focus on an attention-based long short-term memory (AT-LSTM) model. Two different word embeddings are used to learn the word vectors in both the Chinese and English corpora. Subsequently, the word vectors are fed into the long short-term memory (LSTM) layer, and the attention mechanism is combined. The prediction results are output via softmax activation. There are two probabilities for each candidate: positive probability (probability of a correct answer) and negative probability (probability of a wrong answer). The sum of the positive and negative probabilities is one. The candidate answer with the highest probability of positive probabili-

Proceedings of the 8th International Joint Conference on Natural Language Processing, Shared Tasks, pages 208–212,
Taipei, Taiwan, November 27 – December 1, 2017.

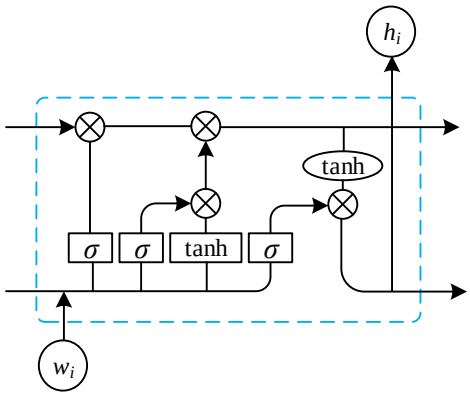

Figure 1: Architecture of a standard LSTM cell.

ty will be considered as our predictive answer. To obtain better experimental results, several models such as convolution neural network (CNN), LSTM, and AT-LSTM are employed for comparison. We also attempt to use different types of word embedding in the process. The experimental results show that the AT-LSTM model yields the best results when using GoogleNews for word embedding. The results of this model are presented in this paper.

The rest of our paper is structured as follows: Section 2 introduces the LSTM and AT-LSTM models. Section 3 provides a detailed description of our experiments and evaluation. The conclusions are drawn in Section 4.

2 Model

Three models are implemented in this competition for comparison: CNN, LSTM, and AT-LSTM models. For the two different subsets of English and Chinese, two different word embeddings are used to process the input data. The experimental results also reflect the theoretical analysis, and the AT-LSTM model achieves better results. Compared with a standard LSTM model, this model adds an attention mechanism after the LSTM layer. The input questions and answers are converted into word vectors after the embedding layer. The function of the LSTM layer is to train the input word vectors into hidden vectors. The key to this model is that the attention mechanism generates a weight for each hidden vector. The hidden vectors and attention weights are combined and passed to the following layer for calculation.

LSTM. Recurrent neural networks (RNNs) are associated with the gradient vanishing or exploding problems. To overcome these possible problems, the LSTM method was developed and it exhibited a better performance. The most significant difference between LSTM and RNN is that the former combined a processor to determine whether the information is useful or not (Sainath et al., 2015). Such processor is a memory cell. Each cell has three gates to control the transmission of information. They are called the input gate, forget gate, and output gate.

After the input data is fed into an LSTM system, the system will determine the usefulness of the information according to established rules. Only the information that is identified as useful will be retained, and the rest will be abandoned. Figure 1 illustrates the architecture of a standard LSTM cell. In the figure, w_i and h_i represent the cell unit input and hidden layer vector, respectively, and σ denotes a sigmoid function. The output of the hidden layer can be considered as the representation of a sentence. The hidden vectors will eventually be passed to the softmax layer for the classification prediction. For each candidate answer, the predicted result will consist of two parts: probability of a correct answer and probability of a wrong answer. The sum of these two probabilities is one. The answer with the highest correct probability among the four answers will be accepted.

AT-LSTM. Although LSTM addresses the problem of gradient vanishing and explosion, it is not very suitable for solving QA problems because there are very long distances involved in the QA context (Wang et al., 2016). One of the solutions is to add a mechanism of attention.

The original questions and answers are converted into vector representations by the embedding layer, and these word vectors are fed into the LSTM layer. Subsequently, the word vectors are expressed as hidden vectors. Then, the attention mechanism assigns a weight to each hidden vector. The attention mechanism produces attention weight vector α and weighted hidden representation r. Both the attention weight vector and hidden vectors are fed into the softmax layer. Figure 2 illustrates the architecture of the proposed AT-LSTM.

The attention mechanism allows the model to retain some important hidden information when the sentences are quite long. In our task, the questions and answers are relatively long sentences. The use of a standard LSTM will result in the loss

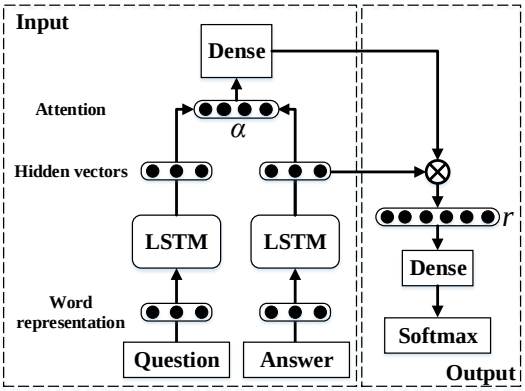

Figure 2: Architecture of the proposed AT-LSTM.

English Subset	Acc
CNN (GoogleNews)	0.275
LSTM (GoogleNews)	0.289
AT-LSTM (GoogleNews)	0.353
CNN (GloVe)	0.261
LSTM (GloVe)	0.271
AT-LSTM (GloVe)	0.307
Chinese Subset	**Acc**
CNN (character vector)	0.297
LSTM (character vector)	0.313
AT-LSTM (character vector)	0.332
CNN (word vector)	0.308
LSTM (word vector)	0.347
AT-LSTM (word vector)	0.465

Table 1: Comparative experiment results

of hidden information. To solve this possible problem, AT-LSTM is used to design the question and answer system.

3 Experiments and Evaluation

Data pre-processing. The competition is divided into two contests: English subset and Chinese subset. Our team participates in both the contests. The datasets of the organizer include training datasets, validation datasets, and test datasets. The English training data mainly contains five subject corpora: biology, chemistry, earth-science, life-science, and physical-science. The Chinese training data mainly contains biology and history. All these corpora contain the question ID, question content, four candidate answers, and correct answers. The validation data is used to initially assess the quality of the trained model and assist in the selection of the model parameters. As with the test data, the validation data is not provided the correct answer. For the English subset, we used a tokenizer to process the questions and answers into an array of tokens. The English word embedding is GoogleNews. Here, all the punctuations are ignored and all non-English letters are treated as unknown words. In the word vectors, unknown word vectors are randomly generated from a uniform distribution $U(-0.25, 0.25)$. For the Chinese subset, first, we use the Jieba toolkit to implement word segmentation on the original corpus. Then the sentences are changed into the word vectors through our own training word embedding. To obtain better training results, we increase the training data. We crwaled numerous junior high school and high school corresponding subject examination questions from the Internet. These are processed into the original training data format to train the model (Yi et al., 2015).

In this experiment, for the English subset, the original corpus is transformed into a word vector by two different word embeddings: GoogleNews and GloVe (Pennington et al., 2014). The results show that GoogleNews can be used to obtain better results. It is used to initialize the weight of the embedding layer in build 300-dimension word vectors for all the questions and answers. For the Chinese subset, we also use a character vector and word vector with two different word embeddings. The character word embedding is trained from the Chinese version of Wikipedia, whereas the word vector embedding is trained from the news (12G), Baidu Encyclopedia (20G), and a novel (90G). The dimensions of the character vector and word vector are 200 and 64, respectively. In the experiment, we notice that after the Jieba toolkit word segmentation, the accuracy of the Chinese subset is significantly improved. We combine the best results of English subsets and Chinese subsets to form our final submissions.

Implementation. The source code for this experiment is written in Python, and the main framework of the program is Keras. The backend used in this experiment is TensorFlow. We use the same AT-LSTM to obtain the results for both the English and Chinese corpora. Both results outperform the baseline. We first use the CNN model to implement this system, but the result is not good. The reason is that some texts are extremely long, whereas a few are extremely short, making the CNN model inefficient. Next, we use the LSTM model to complete this task. The results of the LSTM model are better than the CNN model, but are still unable to reach the base-line. To better solve the problem of the longer distance dependent relationship, we added an attention mechanis-

Parameters	English	Chinese
Filter number	64	64
Filter length	3	3
Dropout rate	0.3	0.1
Epoch	20	20
Batch size	32	64
Word embedding dim	300	64
Score	0.353	0.465

Table 2: Optimal parameters

Corpura	English	Chinese	All
Our score	0.353	0.465	0.423
Rank 1 team	0.456	0.581	0.423
Baseline score	0.2945	0.4463	0.39
Our rank	4	2	1

Table 3: Final testing results and ranking

m to the LSTM model. The results show that, under the same experimental equipment conditions, the AT-LSTM model can yield better results. Table 1 presents the results of a comparative experiment for an English Subset and a Chinese Subset.

The Sklearn grid search function (Liu et al., 2015) is used to determine the best combination of the parameters. Although the same model is used for both the datasets, as the two datasets in the Chinese and English pretreatment are not the same, the parameters that achieve the best results may be different. Table 2 lists the parameters of the model when the best results are obtained.

For the English subset, the best-tuned parameters are as follows: number of the filters in CNN is 64, length of a filter is 3, dropout rate is 0.3, dimensionality of the hidden layer in AT-LSTM and LSTM is 300, batch size is 32, and number of epochs is 20. Simultaneously, the optimizer is Adam, loss function is the categorical cross-entropy, and activation function is softmax.

For the Chinese subset, the best-tuned parameters are as follows: The number of filters in CNN is 64, length of the filter is 3, dropout rate is 0.1, dimensionality of the hidden layer in AT-LSTM and LSTM is 64, batch size is 64, and number of epochs is 20. The rest of the model parameters are the same as for the English Subset.

Evaluation Metrics. For this experiment, the goal is to choose the correct answer from the four candidate answers. The results of the experiment are only the two categories of right and wrong. The baseline score of the organizer is also evaluated by the accuracy. Therefore, the system is evaluated by calculating the accuracy.

Results. According to the results provided by the organizers, a total of 35 teams enrolled in the competition. As can be seen from the comparison experiment in Table 2, the accuracy of using the AT-LSTM model is the highest for both the Chinese and English subsets. The difference between the length of the input sentence varies significantly; therefore, the AT-LSTM model is used to complete the task. Furthermore, the use of GoogleNews embedding for the English subset is better than the GloVe embedding. The main difference between the two embeddings is in the training sets. The training sets of GoogleNews are practically from the news, while the training sets of GloVe are from Twitter. Obviously, the GoogleNews data source is closer to this task. For the Chinese data sets, the use of word vectors is significantly better than the character vector. The meaning of Chinese words is not equivalent to the combination of the meaning of single characters. Compared with the character vector, the word vector can more accurately represent the original input information. Therefore, the results of the AT-LSTM model with GoogleNews embedding is chosen as the final uploaded English subset result. The Chinese subset selects the result of the AT-LSTM with our own training embedding as the final submission. Table 3 shows our final scores and ranking.

4 Conclusion

In this paper, we introduce the task of multi-choice question answering in exams. The AT-LSTM model is used to solve this problem. This model allows the extraction of the long distance dependencies. For more complex scientific questions, this model is proven to be superior to the standard LSTM. In our experiments also, the model exhibits a good performance (better than the standard CNN and standard LSTM models). In the future, we will attempt to improve the model or increase the knowledge of other corpus to enhance the accuracy of the system. Better preprocessing and more detailed word embedding are also helpful for improving the results.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants No.61702443 and No.61762091, and in part by Educational Commission of Yunnan Province of China under Grant No.2017ZZX030.

The authors would like to thank the anonymous reviewers and the area chairs for their constructive comments.

References

Peter Clark. 2015. Elementary school science and math tests as a driver for ai: Take the aristo challenge! In *Proceedings of the TwentyNinth AAAI Conference on Artificial Intelligence (AAAI-15)*, pages 4019–4021.

Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Turney, and Daniel Khashabi. 2016. Combining retrieval, statistics, and inference to answer elementary science questions. In *Proceedings of Thirtieth AAAI Conference on Artificial Intelligence*, pages 2580–2586.

Peter Clark, Philip Harrison, and Niranjan Balasubramanian. 2013. A study of the knowledge base requirements for passing an elementary science test. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction.(AKBC-13)*, pages 37–42.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Peter Clark, Oren Etzioni, and Dan Roth. 2016. Question answering via integer programming over semi-structured knowledge. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI-16)*, pages 1145–1152.

Tushar Khot, Niranjan Balasubramanian, Eric Gribkoff, Ashish Sabharwal, Peter Clark, and Oren Etzioni. 2015. Exploring markov logic networks for question answering. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 685–694.

Pengfei Liu, Shafiq Joty, and Helen Meng. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 1433–1443.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Mrinmaya Sachan, Avinava Dubey, and Eric P. Xing. 2016. Science question answering using instructional materials. *CoRR abs/1602.04375.*

T. N Sainath, O Vinyals, A Senior, and H Sak. 2015. Convolutional, long short-term memory, fully connected deep neural networks. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4580–4584.

Carissa Schoenick, Peter Clark, Oyvind Tafjord, Peter Turney, and Oren Etzioni. 2017. Moving beyond the turing test with the allen ai science challenge. *arXiv preprint arXiv:1604.04315.*

Jin Wang, Liang Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016. Dimensional sentiment analysis using a regional cnn-lstm model. In *Proceedings of the Association for Computational Linguistics*, pages 225–230.

Yang Yi, Wen Tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018.