An Automatic Scoring System for Factual Subjective Questions Based on Language Element Extraction

Xudong Guo[*1], Yuan Li[*2], and Qinglin Wang[*3]

[*1] Beijing Institute of Technology, Beijing 100081, China
E-mail: 1264767553@qq.com
[*2] Beijing Institute of Technology, Beijing 100081, China
E-mail: liyuan@bit.edu.cn
[*3] Beijing Institute of Technology, Beijing 100081, China

Abstract. In the field of education, examinations are used more and more frequently as a mean to measure learning outcomes and summarize knowledge for both teachers and students. But the workload of scoring is also increasingly heavy, especially for subjective questions. In this paper, we designed an automatic question scoring system based on language element extraction, combining the deep learning related technology and the characteristics of objective fact-based subjective question. The system is mainly composed of entity recognition module and automatic scoring module. Attention mechanism is applied in entity recognition module to handle the different influences of word in a sentence. A double-layer network is designed considering the complexity of the entity. Experimental results verified the method, and the system achieved the expected design target.

Keywords: Named entity recognition, Attentional mechanism, Automatic grading of subjective questions

1. INTRODUCTION

There is an increasing demand for teachers to improve the efficiency of scoring under the premise of high accuracy in education field [1]. However, it is still difficult to understand natural language for modern tools and technologies, due to the widespread ambiguity and diversity at all levels of natural language. This problem is even more pronounced in Sino-Tibetan languages such as Chinese. Therefore, the current mainstream research on subjective subject automatic grading mainly focuses on the comparison and scoring with standard answers. However, such methods still need to solve the problems of feature extraction, similarity calculation at various levels, and the lack of structural characteristics of natural language.

At present, the research abroad on automatic scoring of subjective questions is relatively deep. There are also many relatively mature computer scoring systems [2]. However, intelligent scoring systems abroad still have a long way to go on complex subjective questions such as composition and reading comprehension [1].

There are also many domestic attempts to realize intelligent examination technology. For example, Anhui Provincial Education Admissions Examination Institute, in cooperation with IFLYTEK CO.LTD., conducted an intelligent evaluation of the answers to Chinese and English essays [3]. Beijing Eleventh Middle School and the China Institute of Science and Technology Information carry out an automatic scoring experiment in high school sophomore mathematics examination papers [4]. Yang Xian Yi Middle School in Zhongshan City conducted a comparative experiment of manual examination and artificial intelligence examination in the high school English subject test [5]. These experiments have brought certain technical and application experience to researchers. These methods in experiments calculate the text similarity between the answer and the standard answer. However, these methods can hardly replace manual, because the accuracy of the former is not so good. What is more, just comparing with the standard answer will make the innovative answer ignored or even depreciated.

Obviously, it is difficult to finish the scoring task just by calculating the similarity. But there is lack of competent technologies. Therefore, we narrow the scope of the question and design an automatic scoring system for objective factual subjective questions, to which the answers are relatively fixed and easy. After real implementation, this system could effectively reduce the burden of scoring.

The main work of this system is as follows:

i. *The entity recognition model relies on the attention mechanism to make different words have different influences, thereby improving the ability of neural networks to extract entities. In addition, considering the complexity of Chinese entity recognition, a double-layer network design is introduced. Compared with the basic model, this model achieved better experimental results in the entity recognition task.*
ii. *We designed and implemented an automatic scoring system for concept subjective questions based on named entity recognition. Considered and designed three scoring methods for*

objective factual subjective questions. Finally, an objective fact-based subjective question automatic scoring test was conducted, and the expected target was achieved.

The first part of this paper introduces the methods to improve the entity recognition model. Then the models and experiments are presented. Finally, the design of the automatic scoring system is shown.

2. FEATURES

2.1. Attention

In the task of named entity recognition, sometimes the content related to the recognition result is just a small part of a long sentence. In traditional methods, recurrent networks are usually used to encode sentences into fixed-length intermediate semantic vectors, which are then used to guide the output of each step. However, it will cause information overload, limited shape of models, reduced operating efficiency with this method [6]. What is more, the optimization algorithm is complex for this method. Although optimization operations can simplify the neural network model and alleviate the contradiction between the complexity of the model and the ability to express, such as weight sharing, local connection and pooling, too much data will lead to poor memory of information [7].

The attention mechanism was originally used in machine translation tasks. Its principle is like that of human selective visual attention mechanism. The core theory is to choose the most important information for the current task and goal from complex data. The attention mechanism model is successfully applied to various fields of deep learning, such as image processing, speech recognition and natural language processing tasks [8].

An attention mechanism was added to the original BiLSTM-CRF(Bidirectional Long Short Term Memory - Conditional Random Field) network to improve it. Instead of just using location information processing to deal with the relationship between different words, the new entity recognition model can make good use of the information of words, too.

2.2. Multi-Layer Neural Network

Although the attention mechanism can improve the ability of neural network entity recognition, single-layer recurrent neural networks can hardly recognize the entities with multiple meanings. Embeddings from Language Models (ELMo) uses a multi-layer bidirectional LSTM when training language models. Different layers of ELMo can extract different granularity and level of information. At the same time, a typical two-stage training process is used to solve the problem of ambiguity [9]. We draw on the ideas of the ELMo method and further improve the attention model into a two-layer neural network, which to a certain extent improves the model's ability to extract polysemous words.

3. MODEL

The architectures of basic BiLSTM-CRF model, Att-BiLSTM-CRF model and Multi-layer model are illustrated in Figure 1-3, respectively.

3.1. BiLSTM-CRF

The basic neural network consists of four parts, from bottom to top are the word embedding layer, BiLSTM layer, full connection layer and conditional random field layer.

3.1.1 word embedding layer

To understand the deep meaning of the word as much as possible, we introduced the word embedding method. Popularly and specifically, this method is to map a word as a point into a word vector with a certain length. For a sentence, it is to convert a word sequence composed of words into a matrix of word vectors.

To reduce the phenomenon of overfitting of neural networks during training, we introduce a random inactivation (Dropout) method [10], which prevents some neural network units from updating weights according to a certain probability while training. It is equivalent to randomly finding a thinner network from the original network for neural network training from the perspective of network structure. Practice shows that this method can significantly reduce the overfitting of large network models.

3.1.2 BiLSTM layer

The BiLSTM layer is used to obtain the features in the sentence. The input is the output of the word embedding layer through dropout. After the word vector of each word is input into the bidirectional LSTM, the semantic vector obtained by the forward LSTM and the semantic vector obtained by the reverse LSTM are spliced to a complete semantic vector $H = (h^1, h^2, \cdots, h^i \cdots, h^t) \in R^{n*m}$. Parameter m represents the dimension of the hidden state h_i, parameter t represents the length of the text sequence, and h^i represents the semantic vector of the i-th word in the sequence after going through the bidirectional LSTM layer.

3.1.3 full connection layer

After the bidirectional long-short neural network, the fully connected network is used to process the data. Then the data is sent to the conditional random field layer for annotation processing.

The input of the fully connected layer comes from the output of the bidirectional LSTM layer. The specific calculation of this layer is as follows:

$$a = \tanh(w^T H + b) \quad (1)$$

Vector a is the output of this layer, vector b and w consists of parameter will be updated while training.

3.1.4 CRF layer

The conditional random field (CRF) is a Markov random field that meets a certain discriminant. Although LSTM neural networks can directly output the labeled results, they have relatively weak learning ability for specific

laws because of their strong nonlinear fitting ability. The conditional state transition probability matrix is obtained by CRF training. The entity recognition model with CRF can learn the limited features of the data more accurately, such as the transfer rule of the label to be marked, etc.

Therefore, instead of modeling tagging decisions independently, the CRF layer is used to decode the best tag path in all possible tag paths.

The specific calculation of CRF layer calculation is as follows:

Step1: Calculate the scoring function $score(x, y)$ with input x and output y. P_{i,y_i} represents the probability that the corresponding output of the i-th word is y_i. Each item in the transfer matrix A represents the transfer probability between each label. for example, A_{y_{i-1},y_i} means the transition probability to the output of the i-th word is y_i, when the output of the $(i-1)$ word is y_{i-1}. n indicates the length of the sentence.

$$score(x, y) = \sum_{i=1}^{n} P_{i,y_i} + \sum_{i=1}^{n+1} A_{y_{i-1},y_i} \quad (2)$$

Step2: Normalize the score. The purpose is to make the probability of each label in the range of [0,1].

$$P(Y|X) = \frac{exp(score(x,y))}{\sum_{y'} exp(score(x,y'))} \quad (3)$$

Step4:The calculation formulas of the final model training and testing process are shown as follows. y^x represents the output annotation result corresponding to each input in the training set. y' represents the output annotation result of the model f or each input sequence during training. y^* represents the model annotation result during testing.

$$log\ P(y^*|x) = score(x, y^*) - log\left(\sum_{y'} exp(x, y')\right) \quad (4)$$

$$y^* = argmax_{y'} score(x, y') \quad (5)$$

The automatic output of the recognition model are the labels of the input words. For example, o means that the word is not an entity. S means the word is part of a place name. B or E means the word is at the beginning or the end of the entity name.

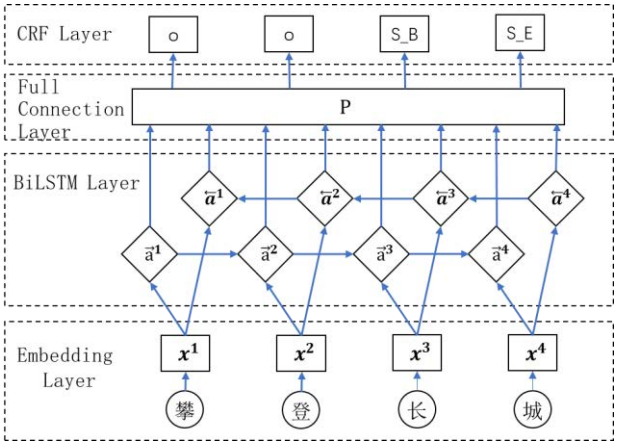

Fig. 1 Structure of BiLSTM-CRF model.

3.2. Att-BiLSTM-CRF

Compared with the model above, this model replaces the fully connected layer with an attention layer to process the output of the BiLSTM layer. The attention layer introduces an attention mechanism to calculate the weight of the influence of each word in the sentence on the output of the current position. At the same time, it automatically updates the weight during model training. Therefore, the words that have a great impact on the entity recognition task will get more attention of this system. The input of the attention layer comes from the output of the BiLSTM layer.

The specific calculation of this layer is as follows:

Step1: The semantic vector matrix H is processed with the tanh activation function to obtain the processed semantic vector matrix M.

$$M = \tanh(H) \quad (6)$$

Step2: Calculate the attention weight a corresponding to the semantic vector, where w represents the weight matrix of the attention mechanism layer

$$a = w^T M \quad (7)$$

Step3: Compute the semantic vector r with the attention weight matrix added.

$$r = [H; a] \quad (8)$$

Step4: Calculate the output of the attention mechanism layer. w_1 is the weight matrix used for the output of this layer.

$$P = tanh(w_1 r) \quad (9)$$

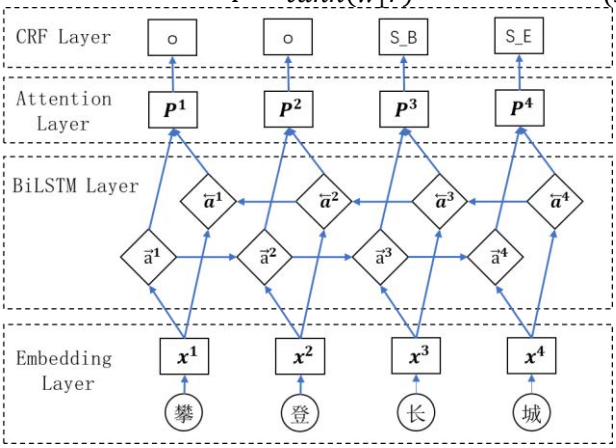

Fig. 2 Structure of Att-BiLSTM-CRF model.

3.3. Multi-layer model

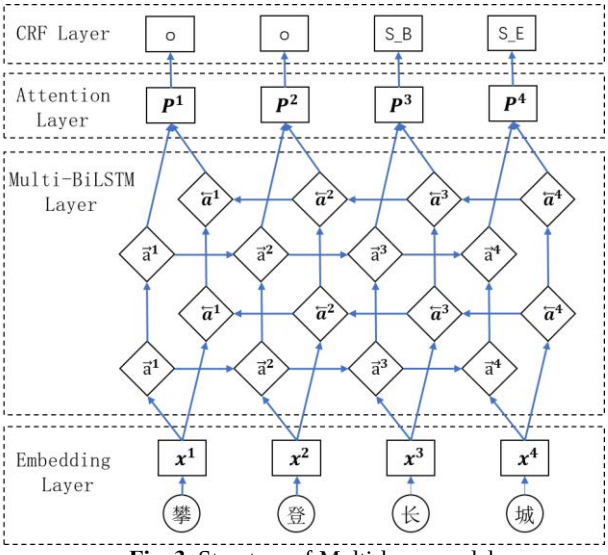

Fig. 3 Structure of Multi-layer model.

The realization method of this attention mechanism can improve the influence ability of the context of the word on the result to the same degree as the influence ability of the word itself, thereby enhancing the ability of the entity recognition model to deal with long sentences and long dependence.

This model changes the BiLSTM layer from single layer to a double layer based on the previous model. The forward LSTM cells of these two layers are connected to each other, so do the backward LSTM cell. Finally, the semantic vector obtained by the forward LSTM and the semantic vector obtained by the reverse LSTM are spliced to a complete semantic vector H.

4. RESULT

4.1. Experimental datasets and settings

The corpus data used for model training and testing is MSRA Microsoft Asia Research Open Source data and 1998 People's Daily annotated data, which only marked three entity types: person name, place name, and organization name. The corpus data includes 79213 training sets and 8802 test sets.

All the parameters are adjusted by the Adam algorithm while training. Model performance is measured by accuracy, recall and F1 values. The hyperparameters adjusted by experiments are shown in the following table.

Table. 1 Hyperparameters of model.

name	meaning	value
Embedding_dim	Embedding dim	100
Sen_len	Max length of sentence	60
Batch_size	Length of Batch	32
Epochs	Number of epochs	31
Learning_rate	Learning rate	0.001
Dropout	Dropout rate	0.5
β_1	Parameter descent weight	0.9
β_2	Parameter descent weight	0.999
ε	Very small constant	10^{-8}

4.2. Comparison of three models

The data in the experimental results are averaged from the multiple training results in the training set and the test set. The performance of these three models in the training set and test set is shown in table 2.

Table. 2 Performance comparison of three models.

Training set	Precision	Recall	F1-score
Fully connection network	92.42%	91.31%	91.86%
Attention network	92.86%	91.95%	92.40%
Multi-layer network	93.59%	92.67%	93.13%
Test set	**Precision**	**Recall**	**F1-score**
Fully connection network	86.77%	84.91%	85.83%
Attention network	86.80%	85.07%	85.93%
Multi-layer network	87.47%	85.33%	86.38%

The result shows that the Att-BiLSTM-CRF model obtains better performance than basic model in precision, recall and F1-score, which proves the effectiveness of the attention mechanism for entity recognition tasks. The performance improvement on the training set indicates that the model with the attention mechanism can better fit the training data and can achieve the training effect faster.

Compared to the attention model, the accuracy, recall and F1 value of the multi-layer network on the data set also improves. The improvement in accuracy is much higher than the recall rate. The Multi-layer model increases the number of non-linear nodes and thus improves the ability to fit complex systems.

The improvement of performance in the test set is not obvious. The reason should be that there is less training data and the length of a single training sentence is short, making it difficult to play the ability of the attention mechanism to integrate the context. What is more, the performance gap between the training set and the test set is large. The reason should be the lack of training data, the large difference in data between the training set and the test set, and the high degree of fitting of the model to the training set.

5. SYSTEM DESIGN

5.1. Architecture

The overall architecture of the objective fact-based subjective question automatic scoring system based on language element extraction designed in this paper is shown in Figure 4. It consists of three parts: data layer, processing layer and application layer from the bottom to the top.

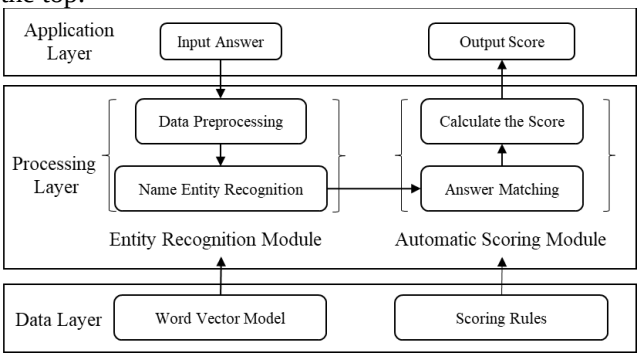

Fig. 4 The overall architecture of the factual subjective question automatic scoring system based on language element extraction.

Data layer: The data layer is located at the bottom of the system's overall architecture and provides data for the entire system. It consists of a word vector model and scoring rules. The word vector model is used in the entity recognition model to provide a word vector that converts each word of the input problem into a fixed dimension, and then recognizes the entity and its type. Scoring rules are used to assist the automatic scoring module to output correct scores.

Processing layer: The processing layer is the middle layer of the overall architecture of the system. It is the core of the factual subjective question automatic scoring

system, including entity recognition module and automatic scoring module.

Application layer: The application layer is located at the top of the overall system architecture and is used to provide a window for users to interact with the system. In the system interaction interface, the user will get the automatically evaluated score after entering the answer to the system.

5.2. Block Diagram

The overall program framework of the subjective question automatic scoring system based on language element extraction is shown in Figure 5:

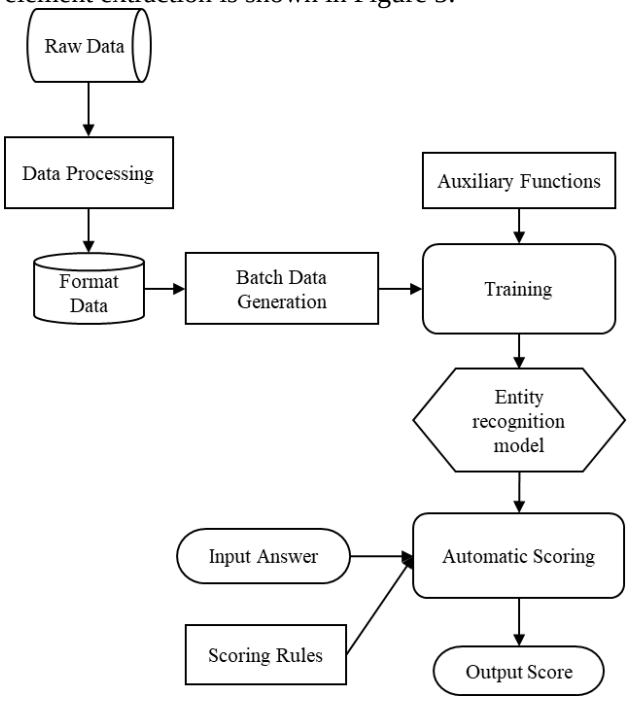

Fig. 5 The block diagram of the factual subjective question automatic scoring system based on language element extraction.

6. CONCLUSIONS

In this paper, we do research on the entity recognition technology and improve the performance of entity recognition by changing the original neural network model structure. What is more, a factual subjective automatic scoring system based on named entity recognition was designed and implemented. The main research contents and innovations of this paper are summarized as follows:

i. *Since different words of the sentence in the entity recognition task have different effects on the entity labeling results, we introduced attention mechanism to improve the basic model and highlight the role of keywords. The experiment shows that the effect of entity recognition is improved.*

ii. *We designed a two-layer network to further improve the entity recognition model, aiming at the problem of polysemy. The performance of the model is better than the attention model.*

iii. *Three scoring rules are considered and designed considering the diversity of objective fact-based subjective question scoring task. Input and output were also optimized to make the automatic scoring system more convenient to use.*

iv. *We combined the entity recognition module and the automatic scoring module to design and implement an objective fact-based subjective question automatic scoring system based on named entity recognition. The automatic scoring test was carried out with news materials. The expected results verified that the design scheme has certain feasibility.*

However, due to the limitation of research time and experimental conditions, there are still many areas for improvement in the research process. In the future, the performance of the automatic scoring system can be further improved by solving the following issues.

i. The entity recognition module is weak in resolving words with multiple combinations. For example, "香山" and "檀香山" are two different place names, and the entity recognition module will only recognize them as "香山". The entity recognition module in this paper can only recognize the three major entities of person name, place name and organization name. This problem can be solved by increasing corpus and extending tags of data.

ii. *Design document-level automatic scoring function. This function is obviously very important in the actual scoring work. The automatic scoring of multiple subjective questions can reduce manual influence on the system and improve scoring efficiency.*

iii. *Combined with machine vision, Internet and other technologies, this system can be more user-friendly and intelligent. For example, the system can search the answer itself with the help of internet technology.*

Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 61472037).

REFERENCES:

[1] Beck J, Stern M, Haugsjaa E. Applications of AI in education. 1996, 3(1):11-15.

[2] Li B. Review of the Application of Computer Technology in the Automatic Examination of Subjective Questions. Jiangsu Science & Technology Information, 2019 (8), pp.39-43+54.

[3] He, Y., Sun, Y., Wang, Z. and Zhu, l. Application of Artificial Intelligence Evaluation Technology in Large-scale Chinese and English Essay Review. China Examinations, 2018 (6), pp.63-71.

[4] Liu, Y., Lu, Y., Ding, L. and Wang, X. Bi-LSTM-based Method for Automatic Examination of Subjective Questions in Mathematics. Management Observer, 2020 (2), pp.109-113.

[5] Li, J. Application of Artificial Intelligence in the Evaluation of English Subjects in High Schools. Educational Information Technology, 2018 (12), pp.41-43.

[6] Qin, M. Research On Key Technologies Of Knowledge-Based Question Answering System Based On Deep Learning. M.D. Beijing Institute of Technology, China, 2019.

[7] Ubale R, Qian Y, Evanini K. Exploring End-To-End Attention-Based Neural Networks For Native Language Identification. In Proceedings

of the 2018 IEEE Spoken Language Technology Workshop, 2018, pp.84-91.

[8] Kingma D P , Ba J . Adam: A Method for Stochastic Optimization. Computer ence, 2014, pp.1412: 6980.

[9] Yu, C., Wang, M., Lin, H., Zhu, X., Huang, T. and An, L. Contrastive Research on Vocabulary Representation Model Based on Deep Learning. Data Analysis and Knowledge Discovery, 2020(7), pp.1-19.

[10] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 2014, 15(1), pp.1929-1958.