International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 10 Oct. 2022, pp: 356-362 www.ijaem.net ISSN: 2395-5252

Answer selection model of interactional attention mechanism based on Twin BERT

Yulin Liao

[1]*Student, College of Communication Engineering, Chengdu University of Information Technology, Chengdu 610225, China; 2*

Date of Submission: 01-10-2022 Date of Acceptance: 12-10-2022

ABSTRACT:Answer selection techniques play a crucial role in the field of question answering. The technical task of answer selection aims to select the correct answer from the candidate answer set, so improving the model performance index is one of the important contents of this research. In recent years, with the rise and wide application of deep learning, a considerable number of problems have been solved. The current mainstream models still have problems that do not consider the interaction information between question-answer pairs and the existing semantic representation is insufficient. This paper proposes an answer selection model based on interactive attention and ESIM. The experimental results of the model in this paper on the Text Retrieval Conference Question Answering (TRECQA) and Wikipedia (WikiQA) datasets show that this model can effectively improve the answer selection task compared with other models.

Keywords: answer selection; interactive attention mechanism; ESIM; question answering system

KEYWORDS: Answerselection; Interactive attention mechanism; ESIM. Question answering system

I. INTRODUCTION

In recent years, with the rapid development of artificial intelligence, the technology of question answering system is also changing with the development of technology. The traditional question answering system based on keyword search can no longer meet the existing needs. Therefore, with the application of deep learning in the field of question answering, answer selection technology is widely used in intelligent question answering systems. Meanwhile, it is an important research direction in the field of natural language processing.

The answer selection task is a typical sentence matching task, that is, the matching relationship between questions and answers. The answer is technical

According to the given question and answer candidate set, the model is required to find the most matching answer [1] from the answer candidate set from the deeper semantic association, so as to return to the user. When choosing the best answer, researchers use a variety of methods such as cosine similarity, Manhattan distance, Jaccard similarity, etc., but its essence is to calculate semantic similarity to carry out semantic matching.

Answer selection technology has gone through many stages of development. In the early stage, the similarity between question and answer pairs was calculated based on lexical features . Due to the complexity of the semantic of text sequences, the method based on lexical features could not accurately capture semantic information. Subsequently, deep learning is widely applied, and the effect of answer selection task is improved compared with lexical features due to the application of deep learning. Subsequently, the representation of attention mechanism assigns different weights to different words, so as to enhance the weight of effective words and reduce the weight of invalid words. However, this method does not consider the influence of sentence context. Google[2] released BERT pre-training model, aiming to pre-train a deep bidirectional representation model from unlabeled text. The appearance of BERT greatly improves the effect of the answer selection task.

The following problems exist in today's answer selection task: 1. Questions and answers are semantically related, rather than simply semantically similar; 2. Incomplete expression of text sequence information; 3. The BERT model is directly used for training, which has the problem of complex network calculation. To solve these problems, this paper proposes an answer selection

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 10 Oct. 2022, pp: 356-362 www.ijaem.net ISSN: 2395-5252

model based on twin BERT's interactional attention. The main work is as follows:

A) On the basis of the pre-trained model BERT, the network structure based on twin BERT is adopted. While using the portable network of twin network, BERT is used to embed words in the input sequence.

B) In order to improve the performance of the answer interaction model and consider the connection between questions and answers, the interactive attention mechanism is proposed.

C) Finally, BiLSTM is used to strengthen the text sequence information and integrate the text semantics.

II. RELATED WORK

Based on traditional methods

For answer selection task, traditional answer selection methods mainly build text features based on feature engineering, such as edit distance, support vector machine, syntax tree, dependency tree and so on. Yih[3] et al. extracted semantic features of sentences based on wordNet. Joty[4] et al. used NLP toolkit to extract keyword matching features and named entity matching features in word layer. Surdeanu[5] et al. performed matching by extracting multiple features such as word frequency and similarity between words between questions and answers. Guzman[6] et al. conducted question and answer matching through translation model. Due to the inherent semantic complexity of natural languages, the methods based on feature engineering can be used to represent the information of question answering pairs to a certain extent, but such methods only represent the semantics superficially, and their model effects are still defective. Therefore, in order to solve the shortcomings of traditional answer selection methods, deep learning methods are constantly cited and selected.

Based on deep learning methods

As deep learning technology has achieved excellent results in image processing, computer vision, natural language processing and other fields. Therefore, many researchers use deep learning technology to make the model automatically learn the deep semantic features of the text, which can extract the text features at a deeper level compared with the traditional method while abandoning the disadvantages of manual feature extraction in the traditional method. Using deep learning to solve the answer selection task has become a mainstream way.

In answer selection, deep learning mainly uses neural network models such as convolutional neural network or bidirectional long short-term memory network to vectorize sentences and calculate the similarity between questions and answers. For example, Yu et al. [7] used CNN to extract Q&A features, and used deep learning to solve the answer selection problem. Wang[8] used bidirectional long-short memory network to represent sentence pairs. Feng et al. proposed the CNN-QA method, which uses the structure of twin networks to share CNNS to encode questions and answers and learn the overall representation of question sentences and answer sentences. Tan et al. [9] proposed the QA-LSTM method, which used bidirectional LSTM to encode question and answer sentences respectively. With the proposal of attention mechanism, attention mechanism shows good performance in sentence representation [10]. With the emergence of ELMO, GPT, BERT and other pre-trained models, it has achieved good results in many natural language processing tasks. BERT has better feature extraction ability than traditional neural network. BERT can not only be regarded as the embedding layer to obtain the context embedding of text, but also as the encoding layer to achieve the feature extraction of text . Therefore, based on the pre-trained model BERT, this paper proposes an interactive attention mechanism answer selection model based on twin BERT.

In this paper, methods

This chapter will introduce the interactive attention mechanism model based on Twin BERT in detail. The overall framework of the model is shown in Figure 1. Data first, after pretreatment, coding layer of coding sequence of questions and answers after pretreatment, the interaction of coding layer the questions and answers for interaction of information extraction, and interactive processing sequence of questions and answers for more fully to extract semantic features, to the question and answer information into the semantic feature extraction. So the feature fusion layer can fuse and concatenate the semantic features and interactive information features. Finally, the similarity between the question and the answer was calculated by the softmax layer.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 10 Oct. 2022, pp: 356-362 www.ijaem.net ISSN: 2395-5252

Fig. 1 Interactive attention answer selection model based on Twin BERT

BERT pre-training model

This layer uses BERT pre-trained model to encode the input question sequence $X = \{x_1, x_2 ..... x_n\}$ and the answer sequence $Y = \{y_1, y_2 ..... y_m\}$. The pre-training model BERT uses the Encoder structure in the bi-directional Transformer language model to solve the long-term dependence and unidirectional limitations of the traditional language model. There are two tasks in the BERT language model: first, the mask language model task, compared with the traditional language model, directly adopts the random mask of some words in the whole sentence, and uses bidirectional coding to predict instead of borrowing the words and predicting the next word; Second, context prediction: by using a large number of corpora to train the model, it can learn the context logical relationship between phrases and sentences. BERT aims to build a bidirectional language model to better capture the context semantics between statements and make it more generalized in downstream tasks.

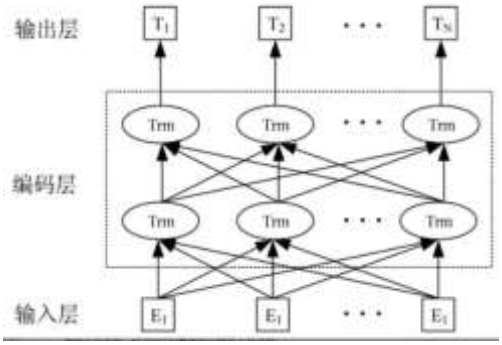

Fig. 2 BERT model diagram

To cope with downstream tasks. BERT gives a sentence-level representation. In this paper, the input of the BERT pre-trained model is two sentences, so it is necessary to add identifiers [CLS] at the head of the sentence, and use separators [SEP] between the two sentences and at the end of the sentence. The input part of BERT model consists of three parts: word embedding vector, piecewise embedding vector and position encoding vector. The word embedding vector is no different from the traditional word embedding vector; The piecewise embedding vector is actually a 0-1 representation used to distinguish the input context. The position coding vector indicates that the words in different positions in the sentence represent different semantics.

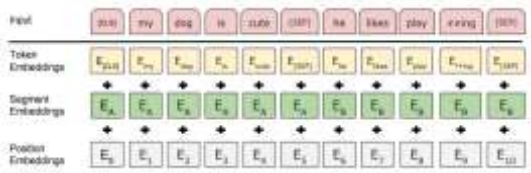

Fig. 3 Example of BERT input

Shared Neural Network

After preprocessing the text dataset to get the word vector that can represent the text, the word vector needs to be processed by the language model. In the answer selection task, the essence is to calculate the similarity of two texts in the data set, so the twin network is used as the main model structure.

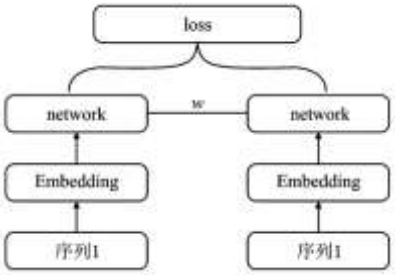

Fig. 4 Structure of Siamease network

As shown in the twin network structure diagram above, Network1 and Network2 are two neural networks with the same weight. In the Siamease network structure, the same neural network can be used or different neural networks can be adopted. Therefore, the same neural network is used to reduce the parameter setting and lightweight network structure. In this paper, the question sequence $X = \{x_1, x_2 ..... x_n\}$ and answer sequence $Y = \{y_1, y_2 ..... y_m\}$ of the twin BERT network structure are used for coding. After

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 10 Oct. 2022, pp: 356-362 www.ijaem.net ISSN: 2395-5252

the twin BERT coding, the question vector and answer vector are obtained.

$$Q = BERT(X_i)$$
$$A = BERT(Y_j)$$

Interactive attention mechanism

Aiming at the question vector and answer vector encoded by the pre-trained model, in order to consider the interaction information between question and answer pairs.

First, the attention matrix E is obtained by multiplying the question vector Q with the answer vector A, where e_{ij} represents the attention score of the ith word q_i in the question vector and the jth word a_j in the answer vector.

$$E = QA^T$$

Then the score of Q against A is softmax:

$$Attn_{qa} = \sum_{j=1}^{a\_len} \frac{\exp(e_{ij})}{\sum_{k=1}^{a\_len} \exp(e_{kj})}$$

In the same way, softmax the score of A against Q:

$$Attn_{aq} = \sum_{j=1}^{q\_len} \frac{\exp(e_{ij})}{\sum_{k=1}^{q\_len} \exp(e_{ik})}$$

Matrix multiplication of the attention of Q to A with respect to the A vector will give you

$$V_q = Attn_{qa} A$$

Similarly, A takes the matrix multiplication of Q's attention with the Q vector

$$V_a = Attn_{aq} Q$$

Aiming at the interactive attention vector between the obtained question and the answer V_q, V_a, in order to further enhance the interactive attention mechanism obtained. In this paper, the difference and dot product between $V_{i\in(q,a)}$ and Q/A are also calculated. This operation helps to sharpen the local inference information between the question vector and the answer vector, and capture more other inference information. The sequence model is enhanced in this way, so as to obtain the feature vector $Comb_q$、$Comb_a$ based on the interactive attention mechanism.

$$Comb_q = cat(V_q, Q, Q - V_q, QV_q)$$

$$Comb_a = cat(V_a, A, A - V_a, AV_a)$$

The fused feature vectors are trained into the language model, so as to strengthen the text sequence information and integrate the text semantics.

In this paper, LSTM is used to train Q and A vectors after attention mechanism interaction. LSTM (Long short-term Memory Neural Network), as a kind of RNN (recurrent neural network), can well process the data input as time series. At the same time, the LSTM model is proposed to solve the problems of the traditional RNN model, such as gradient dispersion, long training time and slow updating of network weights. To solve the above problems, LSTM adds a memory unit to store memory. As shown in the picture below:

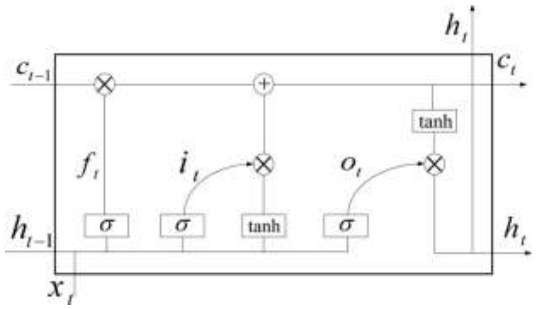

Fig. 5 LSTM structure diagram

LSTM consists of input gate, output gate and memory gate. After the vector representation of the sentence is obtained, the word vector passes through the LSTM layer, and the encoding information of the sentence vector can be obtained through the LSTM layer. When the input sequence is $X = \{x_1, x_2 ..... x_t\}$, x_t is a D-dimensional word vector, and the number of nodes in the hidden layer of the LSTM recurrent neural network model is H, the input, output and forgetting gate at time T can be expressed as: i_t, o_t, f_t, LSTM unit update method is:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$
$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$$
$$c_t = f_t \otimes c_{t-1} + i_t \otimes \tanh(W_c x_t + U_c h_{t-1} + b_c)$$
$$h_t = o_t \otimes \tanh(c_t)$$

The state vector of memory unit at time t is CT, which is calculated by the input gate, input vector and forgetting gate. Ht is the final output vector of the LSTM unit and the output vector of

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 10 Oct. 2022, pp: 356-362 www.ijaem.net ISSN: 2395-5252

the shadow layer calculated by the LSTM recurrent neural network model at the next time, and its dimension is equal to the node number H of the hidden layer. Where, σ is the Sigmoid function. $W \in R^{H\times E}, U \in R^{H\times H}, b \in R^{H\times 1}$ are the weight matrix and bias quantity of the model.

In this layer, BiLSTM is mainly used to capture the information of COMB and comA vectors and their context after interactive attention information, so as to facilitate subsequent processing.

$$V_q = BiLSTM(Comb_q)$$

$$V_a = BiLSTM(Comb_a)$$

Finally, the vectors $Comb_q$、$Comb_a$ obtained by BiLSTM are averaged and pooled, and then the question answering sequence is maximally pooled. And put the pooled pieces together again.

$$V_{q,ave} = \sum_{i=1}^{q\_len} \frac{V_q}{q\_len}$$

$$V_{a,ave} = \sum_{j=1}^{a\_len} \frac{V_a}{a\_len}$$

$$V_{q,\max} = \max_{i=1}^{q\_len} Q$$

$$V_{a,\max} = \max_{j=1}^{a\_len} A$$

$$V = Cat(V_{q,ave}, V_{q,\max}, V_{a,ave}, V_{a,\max})$$

Finally, the results are put into multi-layer perceptron (MLP) classifier. In this paper, a hidden layer with TANH activation function and softmax output layer is adopted.

III. EXPERIMENTAL RESULTS AND ANALYSIS

The data set

In order to verify the effectiveness of the model, this experiment is conducted on the open Q&A dataset WikiQA dataset and TrecQA dataset. As an open question-and-answer dataset, WikiQA uses Bing query logs as the question source to reflect the real information needs of ordinary users. As part of the original data set did not have corresponding answers, the unanswered questions were deleted. Its dataset contains more than 1250 questions and 13,000 question pairs. TrecQA is from Text Retrieval Conference QA Track. TrecQA is the most widely evaluated and long-term standard dataset for QA. It contains more than 13k questions and 57K question and answer pairs. WikiQA and TrecQA related data are shown in the following table:

Table. 1 Distribution table of question answers in experimental data sets.

Dataset	Number of questions			Number of answers		
	Train	Valid	Test	Train	Valid	Test
WikiQA	875	132	245	8697	1126	3251
TrecQA	1219	82	97	55360	1136	1563

Evaluation Index

In the answer selection task, MAP and MRR are usually used as the criteria to measure the model performance and method. MRR(mean reciprocal ranking) is used to measure the ranking of the first relevant answer. MAP(Average accuracy) measures the order in which the answers are related.

$$MAP = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\left(\frac{1}{m}\sum_{j=1}^{|m_i|}\frac{j}{p_j}\right)$$

Where represents the number of candidate answers to the first question, and represents the ranking position of the correct answer to the first question.

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{p_i}$$

Where represents the position of the correct answer in the ranking and represents the size of the dataset.

Experimental parameter setting

There are mainly model parameters and hyperparameters in the training process of deep learning network model, so the selection of parameters is closely related to the performance evaluation of the overall model. In this paper, the English model "Bert-base (uncased)" in pre-trained BERT published by Googel is used in the public dataset. The model cased The number of

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 10 Oct. 2022, pp: 356-362 www.ijaem.net ISSN: 2395-5252

Transformer layers is 12, the dimension of hidden layers is 768, the number of attention heads is 12, and the dropout is 0.2 to avoid overfitting.

Table. 2 Parameter Settings

Parameter reference	value
Optimization	Adam
Learning-rate	1e-5
epoch	8
Hidden-size	768
Batch-size	32
Dropout	0.2
BiLSTMdim	768

The cross entropy loss function is used to calculate the loss in baseband model

$$L = -[y \log \hat{y} + (1-y) \log(1-\hat{y})]$$

Analysis of experimental result

To verify the effectiveness of the model in the answer selection task. Experimental comparison was conducted on WikiQA and TrecQA respectively. In addition, in order to highlight the efficiency of the experimental model, this experiment was compared with multiple baseline models, so as to compare the performance differences between the models.

Table. 3 Comparison of experimental results based on WikiQA/TrecQA dataset

Models	WikiQA		TrecQA	
	MAP	MRR	MAP	MRR
AP-CNN	0.632	0.641	0.632	0.641
AP-LSTM	0.670	0.684	0.713	0.803
AP-BiLSTM	0.683	0.690	0.751	0.824
KABLSTM	0.699	0.732	0.767	0.841
MP-CNN	0.709	0.723	0.777	0.836
ESIM	0.709	0.731	0.791	0.839
Sia-BERT	0.736	0.753	0.796	0.855
本文模型	**0.748**	**0.764**	**0.875**	**0.920**

According to the above experiments, all experiments in Table 3 are completed in WikiQA and TrecQA datasets. Compared with the classical neural network, the text model has a good effect on the index MAP and MRR in answer selection task. In the comparison experiment with the baseband model SIA-BERT, the evaluation index MAP/MRR of the proposed model on WikiQA and TrecQA is improved by 1.2%/1.1% and 7.9%/6.5% respectively.

Model optimization

In order to improve the overall performance of the model and the imbalance phenomenon in the data set, this paper proposes three methods in the model optimization and conducts experiments. 1. Focal loss loss function was used to replace the cross-entropy loss function. 2. Use label smoothing to increase the generalization capability of models.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 10 Oct. 2022, pp: 356-362 www.ijaem.net ISSN: 2395-5252

Table.4 Comparison of experimental results of model optimization WikiQA/TrecQA dataset

Models	WikiQA		TrecQA	
	MAP	MRR	MAP	MRR
Our（logistic loss）	0.748	0.764	0.875	0.920
Our+labelsmoothing	0.755	0.767	**0.882**	**0.930**
Our+focal loss	**0.757**	**0.773**	0.875	0.929

According to the above experiments, in this paper, two optimization methods, Focal loss function and Label Smoothing, are used in the answer selection task of WikiQA dataset and TrecQA dataset. Compared with Logistics loss, which is not used in optimization, both of them are obviously optimized. By comparing three experimental methods in WikiQA data set, Focal Loss is better than Label Smoothing method. For TrecQA data set, Label Smoothing method is more effective than Focal Loss. The main reason is that the WikiQA dataset is not balanced. Compared with Logistics loss, the optimal utility method in WikiQA and TrecQA datasets improves by 0.9%/0.9% and 0.7%/0.09% respectively in MAP/MRR index.

IV. CONCLUSION

This experiment improves the extraction of semantic information while the network structure islightweight, so as to propose an interactive attention answer selection model based on Twin BERT. By utilizing the structure of twin networks, the network structure is lightweight and the complexity of network computation is reduced. The BERT pre-trained model is used to extract features, and then the extraction of interactive information is proposed in the attention layer, which helps the model to accurately and immediately connect the question with the answer. The effectiveness of this model in answer selection task is proved by experimental comparison. In the future work, we can try to enhance the data set and improve the model.

REFERENCES

[1]. LASKAR M T R, HUANG J, HOQUE E. Contextualized Embeddings based Transformer Encoder for Sentence Similarity Modeling in Answer Selection Task [M]. 2020.

[2]. DEVLIN J, CHANG M-W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [J]. ArXiv, 2019, abs/1810.04805(YIH W-T, CHANG M-W, MEEK C, et al. Question Answering Using Enhanced Lexical Semantic Models [M]. 2013.

[3]. BARRóN-CEDEñO A, CEDEñO C, BONADIMAN D, et al. ConvKN at SemEval-2016 Task 3: Answer and Question Selection for Question Answering on Arabic and English Fora [M]. 2016.

[4]. SURDEANU M, CIARAMITA M, ZARAGOZA H. Learning to Rank Answers to Non-Factoid Questions from Web Collections [J]. Computational Linguistics, 2011, 37(351-83.

[5]. GUZMAN F, MàRQUEZ L, NAKOV P. Machine Translation Evaluation Meets Community Question Answering [M]. 2019.

[6]. YU L, HERMANN K, BLUNSOM P, et al. Deep Learning for Answer Sentence Selection [J]. Proceedings of the Deep Learning and Representation Learning Workshop: NIPS-2014, 2014,

[7]. WANG J, LI X, LI J, et al. NGCU: A New RNN Model for Time-Series Data Prediction [J]. Big Data Research, 2021, 27(100296.

[8]. TAN M, DOS SANTOS C, XIANG B, et al. Improved Representation Learning for Question Answer Matching [M]. 2016.

[9]. DENG Y, XIE Y, LI Y, et al. Contextualized Knowledge-aware Attentive Neural Network: Enhancing Answer Selection with Knowledge [J]. ACM Transactions on Information Systems, 2022, 40(1-33.