Fuzzy Systems and Data Mining VII
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA210191

Semantic Repeatability Screening Mechanism of Intelligent Learning Platform Based on Bi-LSTM

Jianghui Liu[a,1] Bairu Xie[b] and Yuqing Shi[c]
[a]*Network and Information Center, Experimental Teaching Center, Guangdong University of Foreign Studies, Guangzhou, China*
[b]*School of Economics and Trade, Guangdong University of Foreign Studies, Guangzhou, China*
[c]*School of English Education, Guangdong University of Foreign Studies, Guangzhou, China*

Abstract. To solve the excessive utilization of back-end data caused by the sharp increase in the visit and consultation on the intelligent learning platform in the era of novel coronavirus epidemic, this study proposed to introduce Co-attention mechanism (Co-attention) into the Bidirectional Long Short Term Memory model (Bi-LSTM). The study employed Multi-layer Perception Network (MLP) for classification and screening to accurately judge the semantic repeatability. Lastly the study carried out contrast experiments for different models, using 1150 consultation posts about transposed determinant, using Newton's Leibniz formula to calculate definite integral, using Laplace's theorem to calculate determinant, how to do model analysis of STATA panel data, under what circumstances is the weighted least square method applicable, how to realize the Pareto optimality and finding the area of trapezoid with curve side from MOOC platform of Chinese universities. Results show that this model performs better than other existing models on the judgment accuracy and its accuracy is up to 89.42%.

Keywords. Intelligent learning platform, Bi-LSTM, Co-attention, Sentence vector, Semantic repeatability

1. Introduction

To meet the development needs of Chinese students in the new era, online education industry has gradually become a new trend in China's education. According to the survey of China Internet Network Information Center (CNNIC), the number of online education users in China has increased from 117 million in 2016 to 351 million in 2020, increasing by nearly three times. And the usage rate has increased from 16.6% to nearly 30%.

The rise of online smart teaching platforms has provided new ideas for solving the dilemmas faced by the traditional education industry under the new crown epidemic. Major online smart learning platforms should also begin to think about how to better improve the efficiency of platform operations, so as to efficiently support the large

[1] Corresponding Author: Jianghui Liu, Network and Information Center, Experimental Teaching Center, Guangdong University of Foreign Studies, China; Email: maggyliu1978@163.com

influx of students. [1] A survey of opinions on online learning conducted by undergraduates of Guangdong University of Foreign Studies revealed that while students generally recognized the safety and high participation of online learning platforms, they also raised questions about the knowledge of students on the platform by teachers. Questions are not answered in time. [2] Therefore, how to effectively improve the system operating efficiency of the online intelligent learning platform plays a crucial role in improving the efficiency and enthusiasm of learners in online learning.

In the process of learning on the online learning platform, students will also sent consultation posts to seek answers to the questions encountered during the online course learning. But among the countless consultation posts, there was inevitably a large number of repeated or similar questions that had been answered [3]. Learners may use completely different statements to ask the same or similar questions. The difficulties in detecting semantic similarity are as follows: First, similar or identical questions may contain completely different grammar, vocabulary or statement structures. Second, a single approach may be appropriate for different types of problems. Third, some teachers had answered the consulting contents before the repeatability screening of these uploaded contents, which caused the difficulty and burden of data processing. Fourth, it's hard to efficiently screen out the repeated statements from the numerous consultation posts.

Rather than expanding the capacity of back-end data from the aspect of information overload, it is better to improve the filtering system for duplicate values. In the processing of natural language, the detection of equivalent semantics has always been a difficult problem. In order to detect similar data in the smart medical platform, an innovative security mechanism called HealthGuard came into being. [3] In addition to the detection of semantic differences, an end-to-end multi-channel deep learning model SDCM is also applied to deal with the differences between image pairs. [4]

In recent years, various neural network models have also been widely used to process the semantic similarity of natural language. Existing research has tried to merge the CNN model and the RNN model to solve the problem of superimposing multiple convolutional layers when the CNN model recognizes the correlation between sentences. [5]

And the adoption of Siamese neural network to detect the semantic similarity of statements had become the focus of research in this field.[6] Many scholars have used neural network models to improve the repeatability detection mechanism of Quora and Stack Overflow platforms.[7,8] There are also studies that have improved related technologies from the perspective of models, such as DupPredictorRep and DupeRep models based on DupPredictor and Dupe models.[9] Some scholars have further proposed that introducing the attention mechanism into the deep learning model to screen the similarity of sentences.[10,11] Then, the ESIM model came into being. This model further improves the ability of the Bi-LSTM model to obtain the semantic information of sentences at different time steps. [12]

According to the literature review, the current research on the use of deep learning models to repetitively screen online learning platform consulting issues still has the following shortcomings. First, the optimization of the intelligent learning platform system is mainly focused on the recommendation system of learning paths and learning resources, and there is a lack of research on the optimization of the processing system of student consulting data. Second, the attention mechanism introduced in the current deep learning model is limited to obtaining semantic information from a single

direction of the sentence, and the inference of the similarity of sentence meaning is prone to large errors. Third, the current deep learning model mainly uses the Word2Vec model or the GloVe model to convert sentences into word vector representations, but the word vector representation cannot efficiently process large-scale sentence data and lacks the ability to identify the connection between sentence context information Ability to effectively improve the prediction rate.

To solve the above problems, this study introduced the Siamese sub-neural network into the twin growth short-term memory model Bi-LSTM, and used the DM model in the sentence vector technology (doc2vec) to improve the prediction rate of the model, sentence vector technology and two-way mutual attention. The introduction of the mechanism enabled the model to effectively obtain the context information of the sentence, thereby achieving accurate semantic recognition. The study also optimized the hidden layer, so that the model can have a higher operating efficiency than the existing model. Finally, the Euclidean distance function was used to measure the similarity between the question sentences, and the characteristic information was obtained, so as to judge whether the sentences had repeated values.

2. System design

In this system, the input question statement will firstly pass through the word vector layer DM to produce the corresponding word vector feature representation. And it will be transformed from the discrete word vector into continuous one-dimensional vector by Embedding.

Then, the produced characteristic vector will be inputted into the neural network of Bi-LSTM. And then the result of the Bi-LSTM output will be introduced into Co-attention mechanism to generate corresponding statement's Co-attention representation. Then the system will use the weighted Euclidean Distance function to measure the similarity degree between characteristic vectors. Lastly, the system will employ Multi-layer Perception Network (MLP) for classification and screening to judge whether the problem statement has duplicate values or not. The process is shown in Figure 1:

2.1. Assume that there is a set of problem statements $\boldsymbol{W_1}$ and $\boldsymbol{W_2}$ in the consultation posts

Denote as: $S(W_1,\ W_2)$ (1)

Meanwhile when $S(W_1,\ W_2) \to 1$, it means that the semantics of the two sentences are similar or the same, that is, the two sentences W_1 and W_2 may have different grammatical structures, but their meanings are similar or the same. In this case, there are duplicate values between this set of statements.

When $S(W_1,\ W_2) \to 0$, it means that the semantics of the two sentences are not the same, that is, whether the two sentences W_1 and W_2 have similar grammatical structures, the meanings expressed are not similar or the same. In this case, there are no duplicate values between this group of statements.

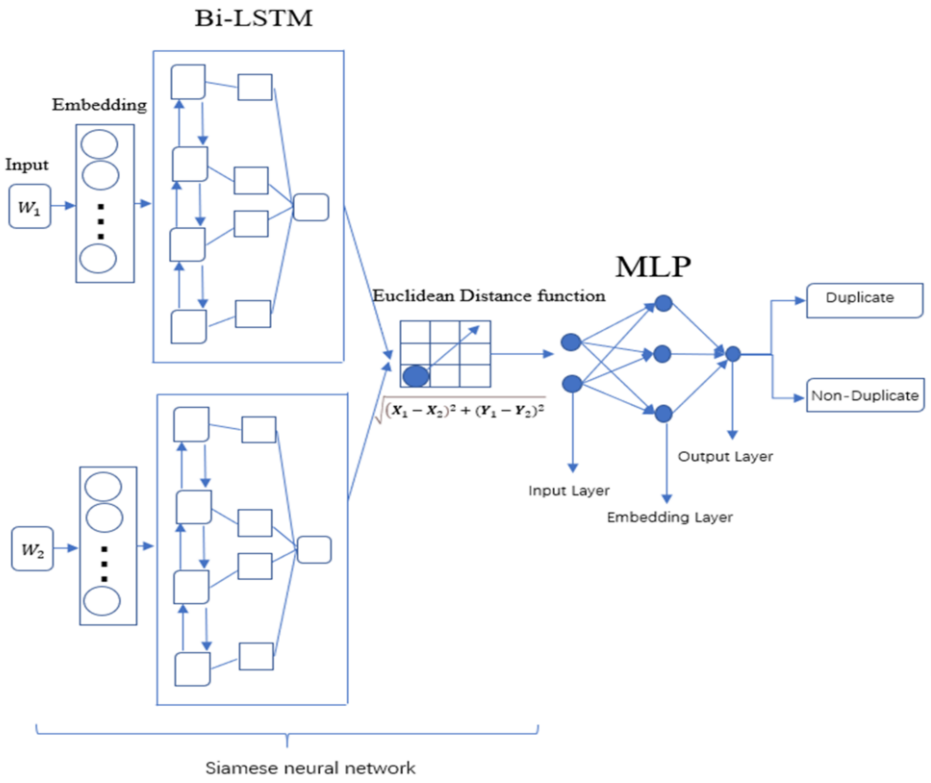

Figure 1. Question-and-answer system based on Bi-LSTM and Co-attention model.

2.2. Input layer

Each question statement is separately put into the sub-network of Siamese neural network, and the parameter and weight in each sub-network is the same. Siamese neural network is mainly used to measure the similarity between the two inputs of the model. The two inputs correspond to two sub-networks, which map the input statements into the new space, thus forming the sentence feature vectors in the new space. In Siamese neural network, the contrastive loss function it used can achieve the matching degree between samples well and also be effectively used for the training model. Because the problem statements are character-level text, statements in this layer are placed in the form of individual characters or symbols into the sub-networks.

2.3. sentence vector layer

This study adopts the DM model of doc2Vec technology, and its framework is shown in Figure 2:

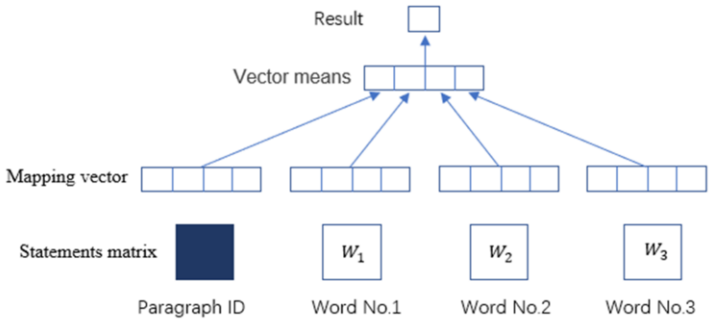

Figure 2. DM model.

In the training process of DM model, Paragraph ID will generate a statement recognition vector with the same dimension as the word vector at first. And then all the sentence vectors and word vectors will be accumulated to obtain the corresponding sentence vector. Finally the corresponding sentence vector will be transferred to the next coding layer. Moreover, during the training of a sentence, its Paragraph ID will remain fixed and share the same sentence vector with other sentences. So in each probability prediction of the words in the statements, the model can make complete use of the full meaning of the statements.

2.4 Coding layer

The system will transfer the sentence vectors obtained by DM model in the sentence vector layer to the Bidirectional Long Short Term Memory model (Bi-LSTM). And the semantic information of the bidirectional tense of the sentence vector is read and encoded to obtain it more completely. Then, the feature vector output from the Bi-LSTM model will be introduced into the Co-attention model to further obtain the body information of the statements.

2.4.1. Introduce the Co-attention mechanism into the Bi-LSTM model

The traditional one-direction LSTM model cannot go forward and backward from the statement at the same time to get the semantic information. We cannot fully use the coded information of a statement if we only use time step information of the statements on a single direction. [11] Based on the above shortcomings, this study decided to use multi-layer and bidirectional Bi-LSTM model. This model has two sub-networks, allowing it to get the information both forward and backward. So the model can obtain the statement semantics on two voice directions, that is, the past and the future. And it can also connect state of hidden layers of the time step in both directions, so as to gain more accurate semantic information.

$$q_t = \sigma(p_{q_y} y_t + p_{q_h} h_{t-1} + B_q) \tag{2}$$

$$d_t = \sigma(p_{d_y} y_t + p_{d_h} h_{t-1} + B_d) \tag{3}$$

$$o_t = \sigma(p_{o_y} y_t + p_{o_h} h_{t-1} + B_o) \tag{4}$$

$$s_t = \tan h\,(p_{s_y} y_t + p_{s_h} h_{t-1} + B_s) \tag{5}$$

$$B_t = d_t \odot B_{t-1} + q_t \odot s_t \tag{6}$$

$$h_t = \tan h(b_t) \odot o_t \tag{7}$$

$\odot$: corresponding product of elements

$p_q p_d p_o p_s$: Weight factors in the hidden layer

$B_q B_d B_o B_s$: Deviation vector

σ : S-shaped function as an activation function

$\tan h$: Hyperbolic tangent function

$X_t = [h_{t-n}, h_{t-(n-1)}, \dots, h_{t-2}, h_{t-1}]$ is the output of the LSTM layer, which represents the vector of all output results.

Then, we use the LSTM model to connect the state of hidden layer of each time step backward and forward. And through (8), we can calculate each vector output by Bi-LSTM.

$$X_t = ((h_t)^{\rightarrow} \odot (h_t)^{\leftarrow}) \tag{8}$$

$\odot$: Connection function between two output results

$h^{\rightarrow}, h^{\leftarrow}$: It respectively represents the calculation output results of the forward sequence from the t-n time period to the t-1 time period and the reverse input of the forward layer

Therefore, the output vector of Bi-LSTM can be expressed as:

$$X_t = [x_{t-n}, x_{t-(n-1)}, \dots, x_{t-2}, x_{t-1}] \tag{9}$$

2.4.2. Co-Attention Mechanism

In this part, the system will encode the input statement to form the intermediate content vector, so as to distribute different weights to different parts of the statement and then effectively obtain the topic information of it. The Co-attention mechanism is used to optimize the Bi-LSTM model, and the semantic similarity matrix is constructed as follows:

$$s_{ij} = E(\overline{n_i})^T . E(\overline{m_j})^T, s \in R^{l_n \times l_m} \tag{10}$$

$\overline{n_i}$,m_i: Respectively represent the i-th and j-th words in statements n and m.

$E(.)$: Single-hidden Layer Feedforward Neural Network and meanwhile $e(x) = Relu(p(x) + g)$.

Maximum pooling is carried out for row and column s to extract the features of the pooling layer. The process is as follows:

$$n' = S(\max_{col}(s))^T n \tag{11}$$

$$m' = S(\max_{row}(s))^T m \tag{12}$$

s_{ij}: The similarity matrix of n and m statements, and $s \in R^{l_n \times l_m}$。

$S(.)$ is the softmax function.

n' , m' : Respectively the Co-attention representations of n and m statements

2.4.3. Measurement of semantic similarity

The weighted Euclidean Distance function is used to measure the similarity distance between feature vectors. The formula is as follows:

$$u(x,y) = \sqrt{\sum_{i-0}^{n-1}(x(i) - y(i))^2} \tag{13}$$

x, y: vectors in K-dimensional space

The Euclidean Distance of x and y is defined as the real distance between x and y in space. If x and y are vectors in two-dimensional space, the value of $u(x,y)$can be measured using the Pythagorean theorem. Even if the statements are not similar, when the Euclidean distance between the feature vectors is small, the loss value of the Contrastive loss function in the neural network will still increase, so that the degree of similarity between the statements can still be well judged.

2.5. Output layer

Multi-layer perception network (MLP) is used for classification and screening to achieve prediction. The hybrid Siamese neural network and multi-layer perception network (MLP) are generated, which enables the output of the Siamese encoder to be passed to the MLP and thus simulate the interaction between the two problem statements. Thus, the MLP model can obtain the vector representation of the problem statement and the connection between the similarity distance of its semantics as the input of the model, namely:

$$V = [f(w_1); f(w_2); d(w_1, w_2)] \tag{14}$$

Then the matching probability between the two question statements is printed as the output. Finally a single 1×2 vector is printed to determine whether there are duplicate values or not.

3. Model experiment

3.1. Experimental environment

This experiment adopted the TensorFlow system, which realizes the programming of machine learning algorithm through the programming of data flow and symbol teaching. And C ++ was chosen to speed up its operation efficiency, and advanced machine learning application program interface (API) was used to maximize the efficiency of model training.

3.2. Collection and processing of experimental data

The experiment selected 1150 consultation posts sent in the MOOC platform of Chinese universities about "transposed determinant", "using Newton's Leibniz formula to calculate definite integral", "using Laplace's theorem to calculate determinant", "how to do model analysis of STATA panel data", "under what circumstances is the weighted least square method applicable", "how to realize the Pareto optimality" and "finding the area of trapezoid with curve side" as the experimental data, and integrated the

statements from the question-and-answer results of the posts. And this resulted in two text documents, respectively named question text and answer text.

A total of about 6400 pieces of consulting data, 4800 pieces of data were selected for training, and 1600 pieces of data were tested for about 6400 iterations of data. Use the jieba tokenizer to segment the sentences in the text, and then delete punctuation marks, etc., to prepare for the subsequent generation of sentence vectors.

Then the Euclidean Distance function was used to measure the similarity of the output results of the Bi-LSTM model based on the Co-attention mechanism. And the MLP filter was used to judge whether there were duplicate statements.

3.3. Model training results and analysis

The TensorFlow technology was used to train the model. And when the network convergence was achieved, the model data was saved. Parameter settings of the model are shown in Table 1:

Table 1. Parameter of model

Parameter	Value
Initial learning rate	0.001
Iteration	6400
Dimension	256
Dimension of sentence vector	256

In addition, in order to better evaluate the performance of the model proposed above, the experimental results are shown in Table 2:

Table 2. Comparison with other models

Model	Vector	Result
Bi-LSTM	Word vector	76.26%
Attention-Bi-LSTM	Word vector	81.13%
Co-attention-Bi-LSTM	Word vector	86.76%
Bi-LSTM	sentence vector	80.61%
Attention-Bi-LSTM	sentence vector	85.86%
Co-attention-Bi-LSTM	sentence vector	89.42%

Moreover, according to Table 2, Bi-LSTM model introduced with the Co-attention mechanism had higher accuracy compared with the other two models no matter word vectors or sentence vectors were used in the same way. What's more, for the above three models, when sentence vectors were used, they all had a higher judgment accuracy of semantic repeatability than when word vectors were used.

Then use Stata to test the significance of the experimental results. The test results show that, under the control of variables such as corpus size, number of iterations, and number of query sentences, the accuracy of repeatability prediction using sentence vectors and Bi-LSTM is significantly positive at the 2.5% level. It shows that the use of word vectors can improve the prediction accuracy of Bi-LSTM based on the mutual attention mechanism. And, according to whether the attention mechanism is introduced

for different models to perform sub-sample regression, the regression results show that after the attention mechanism is introduced, the prediction accuracy of the model is significantly positive at the 1% level. The use of sentence vectors can effectively improve the prediction accuracy of models based on the attention mechanism. For the model without the attention mechanism, its prediction accuracy is significantly positive at the level of 1% when compared with the sentence vector, indicating that the use of the sentence vector can also effectively improve the prediction accuracy of the non-attention mechanism model. The results of the significance test fully illustrate that the core conclusion that the use of word vectors and the introduction of the attention mechanism has a positive effect on the model's prediction accuracy is robust.

In addition, the prediction accuracy rate of Bi-LSTM based on the mutual attention mechanism and the more widely used neural network model are compared and tested. The experimental results are shown in Table 2. Comparative experiments show that, compared with other neural network models, the Bi-LSTM model has a better accuracy in the detection of sentence repeatability.

4. Conclusion

This study carried out a control experiment for the designed model to obtain the accuracy of judging statement duplicate values. The results showed that the use of sentence vector and the introduction of bidirectional Co-attention mechanism both significantly improved the judgment accuracy of the model. This model is expected to be used in the online intelligent learning platform for semantic repetitive screening, thereby improving the operating efficiency of the background system and efficiently assisting students in personalized learning. However, the training results of this model in a large-scale corpus are not ideal. As a result, its efficiency in large-scale repetitive screening tasks needs to be improved, and its applicable consulting platform is too single to efficiently identify consulting sentences in different fields. In future work, we will focus on introducing the dynamic attention mechanism and the attention mechanism based on the capsule network into the Bi-LSTM model, so as to better improve the efficiency of model detection. It will also increase the size of the corpus used for model training and try to apply it to different intelligent consulting platform fields.

Acknowledgement

This study was financially supported by the Undergraduate Innovation Training Project of Guangdong University of Foreign Studies in 2021 (NO. S202111846022).

References

[1] Hussein E , Daoud S , Alrabaiah H , et al. Exploring Undergraduate Students' Attitudes towards Online Learning during COVID-19: A Case from the UAE[J]. Children and Youth Services Review, 2020, vol. 119, 105699

[2] Li Ruiqian. Research on the Evaluation of Online Education Platform by University Students: Based on Technology Acceptance Model [J]. Continue Education Research. 2021, No. 257(01):116-119.

[3] A.I. Newaz, A.K. Sikder, M.A. Rahman, A.S. Uluagac. Healthguard: a machine learning-based security framework for smart healthcare systems, in: 2019 Sixth International Conference on Social Networks Analysis, Management and Security, SNAMS, IEEE, 2019, pp. 389–396.
[4] Oluwasanmi A , Aftab M U , Alabdulkreem E , et al. CaptionNet: Automatic End-to-End Siamese Difference Captioning Model With Attention[J]. IEEE Access, 2019, vol. 7, pp.106773-106783.
[5] Hassan A , Mahmood A . Convolutional Recurrent Deep Learning Model for Sentence Classification [J]. IEEE Access, 2018, vol.6, pp13949-13957.
[6] Zhu P , Tan Y , Zhang L , et al. Deep Learning for Multilabel Remote Sensing Image Annotation With Dual-Level Semantic Concepts[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, (99) pp.1-14.
[7] Imtiaz Z , Umer M , Ahmad M , et al. Duplicate Questions Pair Detection Using Siamese MaLSTM[J]. IEEE Access, 2020, PP (99):1-1.
[8] M. Ahasanuzzaman, M. Asaduzzaman, C.K. Roy, K.A. Schneider, Mining duplicate questions in stack overflow, in: Proceedings of the 13th International Conference on Mining Software Repositories, ACM, 2016, pp. 402–412.
[9] K. R.F. Silva, M. de Almeida Maia, Duplicate question detection in stack overflow: A reproducibility study, in: IEEE 25th International Conference on Software Analysis, Evolution and Reengineering (SANER), 2018, pp. 572–581.
[10] Ning Chunmei. Research and Application of Text Similarity Algorithm Based on Deep Learning [D].Chongqing University.2019.
[11] Wenpeng Yin, Hinrich Schütze, Bing Xiang, et al. ABCNN: Attention-Based Convolutional Neural Network for Modeling Sentence Pairs. 2016, 4:259-272.
[12] CHEN Q, ZHU X D, LING Z H, et al. Enhanced LSTM for natural language inference[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2016: 1657-1668.