Question Similarity Model
Term Paper Report

SUBMITTED BY-
Enrollment No - 18803033
Name - Yashi Agarwal

SEPTEMBER 2022

Submitted in partial fulfillment of the Degree of
5 Year Dual Degree Programme B. Tech
in
Computer Science Engineering

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING & INFORMATION TECHNOLOGY
JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, NOIDA

INTRODUCTION

In the field of natural language processing (NLP), a core problem is to determine if two sentences have approximately the same meaning. That is, we need to know “Thou art mine” and “You are mine” share the same meaning, and “How old are you?” and “What is your age?” express the same question. One of the most typical applications of question similarity is the community Question Answering (cQA) system. Most of the traditional search is based on keyword matching, or text surface meaning of the query, so the problem is that it is difficult to get the content we users need to get, not dig out the deep information resources, not semantic information.

One of the most vital steps in automatic Question Answer systems is Question classification. The Question classification is also known as Answer type classification, identification, or prediction. The precise and accurate identification of answer types can lead to the elimination of irrelevant candidate answers from the pool of answers available for the question. High accuracy of Question Classification phase means highly accurate answer for the given question.

Determining question similarity is difficult because of the complication of semantics in human languages. The same word may have multiple meanings, and the same meaning can be expressed in many different ways. At sentence level, the syntactical structures that contribute to paraphrases are even more complicated. Particularly, in community question answering, similarities over the surface text do not always imply that the questions share the same answer. For instance, for the query “What must not I feed a dog?”, similar questions would be “What can’t dogs eat?”, and “What food may make dogs sick?”. On the other hand, the seemingly similar question “What can I feed a dog?” has just the opposite meanings. Simply taking its answer may lead to tragic consequences.

Upon receiving a query from the user, a cQA system first retrieves a list of possible candidate questions from a large database (typically via an indexing service), resulting in only a few hundred ones. Then, it selects the most similar question set from the candidate list with a sentence similarity algorithm. The answers to these existing questions are most likely the correct responses to the original query. In such a system, the key issue is to determine the query-question similarity, i.e., find a question that is most similar to the user’s query. More formally, the problem is defined as follows:

Given a query Q and a set of relevant question candidates {C1,C2,...,Cn} retrieved from an indexing service, determine whether or not each candidate Ci is similar to Q, and rank them by their similarities to the original query Q.

In addition, Chinese language processing has its unique difficulties. For instance, unlike English, Chinese text is written continuously. Therefore, word segmentation is typically the first step of analyzing one sentence. A poor-performing word segmenter could reduce the accuracy of all succeeding processes

In our model, the goal of the training is to maximize the probability of the first sentence given a similar one, while minimizing the probability of the first sentence given an irrelevant one. The model solves the similarity problems in three major ways:

- The semantic relatedness between words is captured by the pre-trained word vectors.

- The syntactic properties of the sentences are implicitly learned by the RNN. During its sequential processing, the RNN adaptively learns to update information according to current and previous words using reset and update gates.

- In order to deal with the lack of labelled data, we develop a two-step learning scheme. In the first step (pretraining), we use some basic features to automatically label a large amount of question pairs. Then, we train the final model using a smaller amount of finely labelled data. Experiments show that the pre-training boosts the performance of our model.

This proposes using RNN encoderdecoder to solve the problem of sentence similarity. This novel approach is able to capture both textual similarity and semantic similarity between two sentences. Our two-step training framework effectively tackles the lack of training data. Another contribution of the paper is the building of a Chinese question similarity dataset, which will be available upon request.

II. RELATED WORK

QSE (Question Sentence Embedding)

Among many models and algorithms being used, one of them is Question Sentence Embedding(QSE).

for question classification by utilizing semantic features. Extracting a large number of features does not solve the problem every time. Our proposed approach simplifies the feature extraction stage by not extracting features such as named entities which are present in fewer questions because of their short length and features such as hypernyms and hyponyms of a word which requires WordNet extension and hence makes the system more external sources dependent. We encourage the use of Universal Sentence Embedding with Transformer Encoder for obtaining sentence level embedding vector of fixed size and then calculate semantic similarity among these vectors to classify questions in their predefined categories.

This model has been tested over COVID dataset as people are more curious to ask questions about COVID. So, the experimental dataset is a publicly available COVID-Q dataset. The acquired result highlights an accuracy of 69% on COVID questions. The approach outperforms the baseline method manifesting the efficacy of the QSE method.

Question Answer systems(QAs) are an advanced field of Natural language processing and Information retrieval and are different from search engines. Among the five steps in QAs,

1. Question classification is the first and vital step

The main purpose of question classification which is also known as Answer type identification/ classification/ prediction is to classify questions into one of the predefined categories/classes. The set of categories are defined according to the domain. Classifying a question in a specific category helps in narrowing down the answer selection process and removing the unwanted answer. 36.4% of the wrong answers in QAs are due to incorrect classification of questions.

As the pandemic has created panic among people and put people in confusion with so many questions that were never asked before. To better answer, the COVID-19 related questions asked by people on multiple sites such as Yahoo, Quora, Center for Disease Control and Prevention (CDC), etc

have built a question classifier system that helps people answer these questions more accurately by classifying the questions into one of the 15 predefined categories. They define how NLP can be used to help gather and process information about this pandemic.

They presented a BERT baseline system achieving the highest accuracy of 58.8% when trained on augmented data and tested on generated questions from the dataset. Classifying COVID-19 questions is a new research area and demands extensive research as well as methods to gather, pre-process and handle the questions asked by common people.

2. Question Understanding

Question Understanding is the first step of any QA system and also determines the strategy used for answer extraction. Question understanding comprises of multiple stages.

(1) question keywords extraction (or question labelling), identifying the focus of the question.

(2) question classification, determining the category or semantic type of the question.

(3) question extension, generating similar questions

In this paper, the focus is only on question classification using machine learning techniques along with word embedding features. Question classification can use a rule-based or linguistic template-based/statistical/machine learning-based approach. Rule-based methods are time-consuming and require human efforts as compared to other methods. Previous research on question classification focuses mostly on ontological approaches in designing domain-based question classification systems while ignoring machine learning or NLP techniques. Performance of question classification depends on features extracted and classifiers used.

Designing an effective strategy for classifying COVID-19 questions is a challenging task due to the short length of questions as well as new words and information being asked in the questions. This paper presents a universal approach for answer type identification and classification method, named Question Sentence Embedding (QSE), based on sentence embedding feature extraction with transformer encoder.

Feature Extraction-

Universal Sentence Encoder (USE) is used to encode the input before it is passed on to the transformer encoder module. The transformer encoder is based on the original transformer architecture. This architecture consists of six layers of transformers stacked onto each other. The output generated by this module is feature vectors that are context-aware word embedding. These context-aware word embedding are placed elementwise and then a module divides them with the square root of the length.

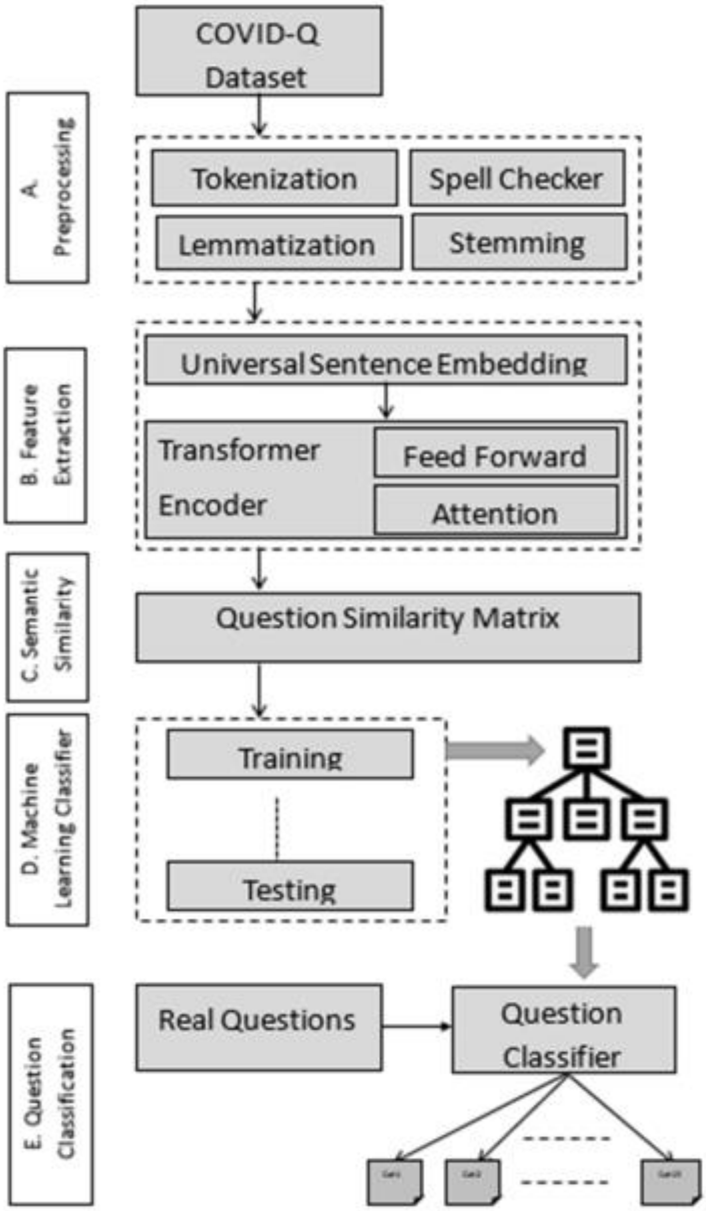
This results in size 512-dimensional vectors. Apart from the questions, the categories of the questions are also labeled by implementing the label encoder module.

Semantic Similarity between questions-

In this step, we find the questions that are semantically similar and assign them categories. Cosine similarity is one of the powerful measures to find similarity between two vectors by calculating the cosine of the angle between them

Accuracy = Number of correctly classified questions/ Total number of classified questions

$$\cos\theta = \frac{A.B}{||A||\,||B||} = \frac{\sum A_i B_i}{\sum A_i{}^2 \sum B_i{}^2}$$



Multi Model Fine-Grained Nonlinear Fusion

To alleviate the problem, the combination of the traditional statistical method and deep learning model as well as a novel model based on multi model nonlinear fusion are proposed in this paper. The model uses the Jaccard coefficient based on part of speech, Term Frequency-Inverse Document Frequency (TF-IDF) and word2vec-CNN algorithm to measure the similarity of sentences respectively. According to the calculation accuracy of each model, the normalized weight coefficient is obtained and the calculation results are compared. The weighted vector is input into the fully connected neural network to give the final classification results. As a result, the statistical sentence similarity evaluation algorithm reduces the granularity of feature extraction, so it can grasp the sentence features globally. Experimental results show that the matching of sentence similarity calculation method based on multi model nonlinear fusion is 84%, and the F1 value of the model is 75%.

The semantic similarity between the corresponding words is calculate respectively for the sentences after word segmentation. If the similarity of words exceeds the set threshold, the location information will be compared. Finally, the weight vector is multiplied with the word vector mapping file to get the final feature matrix of the sentence. In this feature matrix, the initial coarse-grained extraction of sentence features is realized, but the influence of high-dimensional word vector weight on sentence feature matrix is too mild. This method propose a sentence similarity calculation model based on the fusion of deep learning model and statistical method. The model calculates TF-IDF vector through co-occurrence words and other information, and preprocesses sentences twice before calculation.

In this paper, a sentence similarity calculation model based on multi model nonlinear fusion is proposed. The model combines the traditional sentence similarity calculation method based on statistics, and completes the coarse-grained extraction of sentence. The main contributions and innovations of this article are summarized as follows:

1. A concept of text extraction based on granularity is proposed. The feature extraction based on statistical method is defined as coarse-grained feature extraction. We combine the coarse-grained features with the fine-grained features through the fusion of various computational methods, so as to grasp the semantics of sentences from a global perspective.

2. An improved Jaccard coefficient calculation method is proposed in this paper. The traditional Jaccard coefficient only calculates the amount of words in the intersection of sentence and word segmentation results, which ignores the influence of part of speech on sentence semantics.

Nevertheless, for complex parts of speech, the number of words in the intersection set cannot accurately reflect the degree of semantic similarity. Therefore, improve the Jaccard algorithm by adding weighting the part of speech.

3. Compared with the traditional deep learning model, a multi feature weighting mechanism is added to the word2vec-CNN model based on multiple features. We weight the initial feature matrix by measuring the semantic similarity between words. Compared with the direct extraction of sentence feature matrix, this weighting mechanism can highlight the key points of extraction.

1. SENTENCE SIMILARITY CALCULATION BASED ON TF-IDF-

Term Frequency (TF) and Inverse Document Frequency (IDF) are indicators to measure words in text. TF represents the number of times a word appears in the text, which reflects the importance of the word to the file. The IDF represents the frequency of words appearing in a single document.

$$TF-IDF(w_i)=\frac{trem(w_i)}{Num(Sen_A \cup Sen_B)}\times\log(\frac{|T|}{1+w_i\ :\ w_i \subset T}),$$

2. JACCARD SIMILARITY COEFFICIENT BASED ON SENTENCE COMPONENTS

Jaccard coefficient is weighted based on the word components. After word segmentation analysis, a sentence tree is obtained.

$$Jaccard\_Sim = \frac{\alpha(Sen_A \cap Sen_B)}{Sen_A \cup Sen_B},$$

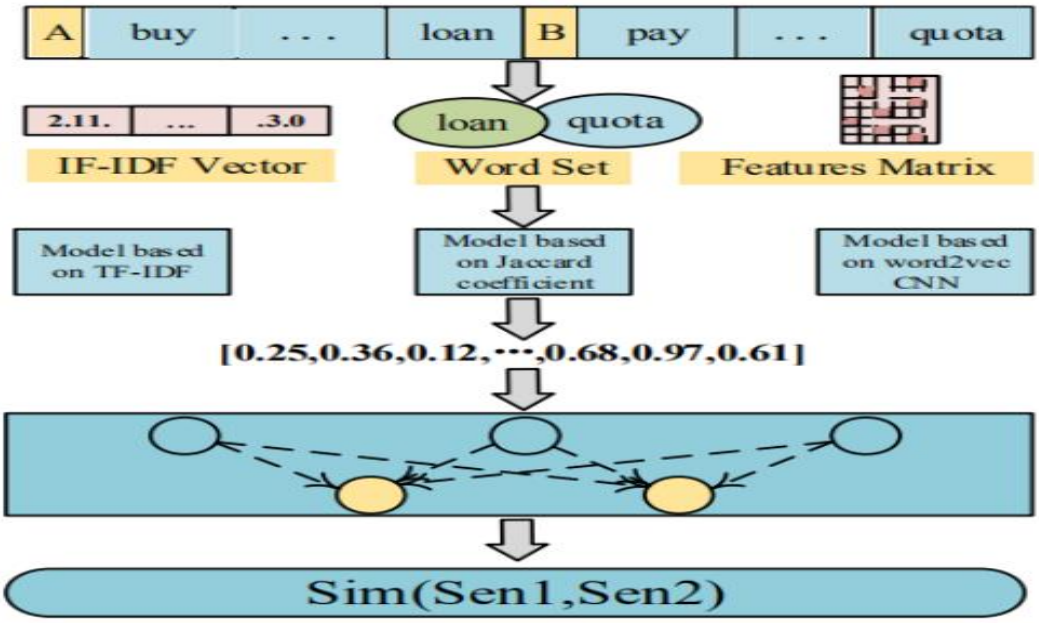
3. WORD2VEC-CNN ALGORITHM BASED ON MULTIPLE FEATURES

word2vec-CNN algorithm we present takes multiple features into account, which is a sentence feature extraction model based on attention mechanism. The model constructs a multi feature attention matrix based on the co-occurrence word information, location information and semantic information between texts to realize the weighting of the original feature matrix

$$Sim_{word} = \sum_{i}^{n}\sum_{j}^{m} COS(w_i, w_j),\ w_i \in Sen_A,\ w_j \in Sen_B,$$

$$w2v\_matrix = \sum_{i}^{n} \sum_{j}^{m} COS(w_i, w'_j)$$

$$= \begin{bmatrix} COS(w_1, w'_1), & ... & , COS(w_1, w'_j) \\ COS(w_2, w'_1), & ... & , COS(w_2, w'_j) \\ COS(w_3, w'_1), & ... & , COS(w_3, w'_j) \\ ...., & ... & , ... \\ COS(w_i, w'_1), & ... & , COS(w_i, w'_j) \end{bmatrix}$$



BERT

The semantically based retrieval algorithm can effectively improve the retrieval rate by calculating the similarity between two sentences. This paper proposes a method to obtain the similarity between sentences based on Bert model, and compares the traditional ALBERT, ESIM and BIMPM models. Experimental results show that the accuracy of **BERT** model in calculating text similarity reaches 87%, which is obviously better than other models. At the same time, the synonym model is trained based on Word2Vec to obtain synonyms related to the target word. Therefore, the algorithm adopted in this paper can effectively improve the retrieval efficiency.

We propose a FAQ retrieval system that considers the similarity between a user's query and a question as well as the relevance between the query and an answer. Although a common approach to FAQ retrieval is to construct labelled data for training, it takes annotation costs. Therefore, we use a traditional unsupervised information retrieval system to calculate the similarity between the query and question. On the other hand, the relevance between the query and answer can be learned by using QA pairs in a FAQ database. The recentlyproposed BERT model is used for the relevance calculation. Since the number of QA pairs in FAQ page is not enough to train a model, we cope with this issue by leveraging FAQ sets that are similar to the one in question. We evaluate our approach on two datasets. The first one is local gov FAQ, a dataset we construct in a Japanese administrative municipality domain. The second is Stack Exchange dataset, which is the public dataset in English. We demonstrate that our proposed method outperforms baseline methods on these datasets.

Many FAQ retrieval models use the dataset with the relevance label between q and a QA pair. However, it costs a lot to construct such labelled data. To cope with this problem, we adopt an unsupervised method for calculating the similarity between a query and a question. Another promising approach is to check the q-A relevance trained by QA pairs, which shows the plausibility of the FAQ answer for the given q. Studies of community QA use a large number of QA pairs for learning the q-A relevance . However, these methods do not apply to FAQ retrieval task, because the size of QA entries in FAQ is not enough to train a model generally. We address this problem by collecting other similar FAQ sets to increase the size of available QA data. In this study, we propose a method that combines the q-Q similarity obtained by unsupervised model and the q-A relevance learned from the collected QA pairs.

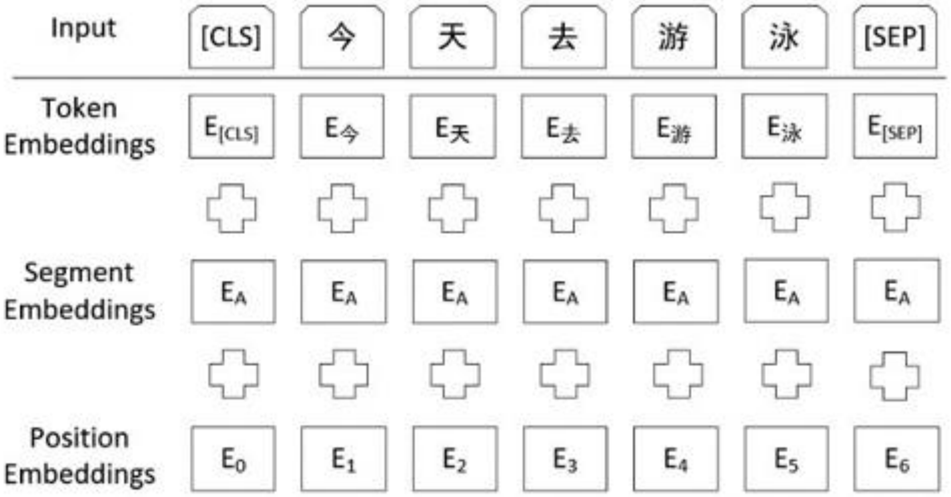
Based on natural language processing technology, semantically based sentence similarity algorithm and synonym model algorithm are introduced into the retrieval field, and the similarity algorithm is used to find the sentence with the highest similarity, and synonyms can also be used for more accurate retrieval.

Model Structure -

BERT Model is a multi-layer bidirectional structure, based on the Transformers, which is not based on the previous RNN and CNN structure, but uses Mask Language Model and Next Senence Prediction (NSP) multi-task training target. Can be used to train more corpus. Compared with the previous GPT models, although they are based on Transformers, BERT is a two-way structure

Embedding- input representation -

The input representation of the Bert model is summed by the three Embedding types shown below. The Token Embedding represents word vector. The first flag of each sentence is CLS and the end is ESP. The Segment Embeddings distinguish the two kinds of sentence.



Pre-training –

This step is the most important step of BERT model, which is divided into two stages

(1) Masked Language Model(MLM)

The role of MLM is to randomly mask some of the tokens in the input and then predict the tokens in the covered portion.

(2) Next Sentence Prediction(NSP)

The task of this section is to get the model to fully understand how thetwo sentences are related. Enter two sentences 1 and 2, and the model predicts whether 2 is the next sentence of 1.

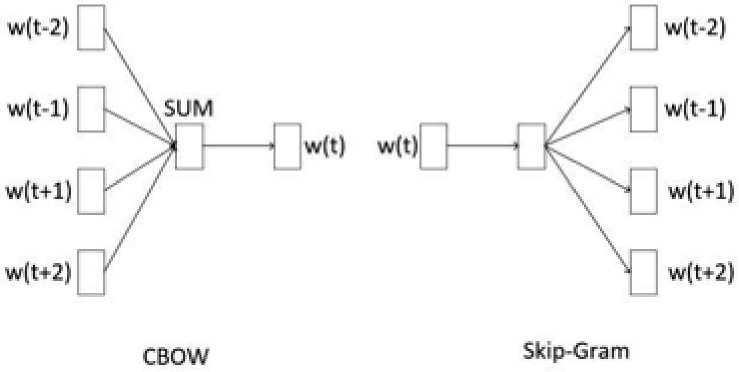
Dataset

The BERT model adopted in this paper is the Bert-Base- Uncased model released by Google to obtain sentence vector for text similarity calculation. It is composed of 12 heads of attention mechanism, hidden layer has 768 dimensions and 110M parameters in total. The data used is the Chinese semantic matching dataset LCQMC released by the Intelligent Computing Research Center of Harbin Institute of Technology, which includes training set, verification set and test set.

Word2Vec model

The traditional language model is facing a string, but the string itself is unable to store the semantic information, the traditional model can’t use no tagging corpus, intelligent rely on manual design algorithm to solve, is laborious and the effect is not very good.

Word2Vec includes CBOW model and Skip-Gram model and have achieved good results in many natural language processing tasks.
The CBOW model predicts the probability of the occurrence of the target word according to the words in the context, as shown in the figure below, and estimates the language model of the current word through the given context.
The Skip-Gram model uses the current target word to predict its context, as shown in the figure below. Meanwhile, in order to improve the training, the negative sampling algorithm is applied to the Skip-Gram model.



As shown in the figure above, although skip-Gram model is a little more complex than CBOW model, its effect is better than CBOW model. paper trains word vector based on Skip-Gram model in Word2VEc, and adds negative sampling algorithm to obtain synonym model, and then identifies synonyms of current words and other words related to them. Prepare for subsequent tasks.

Dataset –

When Word2Vec trains the synonym model, the data set used is the latest data of Wikipedia corpus, including
396,842 documents (a total of 3757530 articles).

Conclusion

In view of the sentence similarity problem encountered in the current intelligent retrieval, this paper proposes sentence similarity calculation based on BERT and combines Word2Vec model to train synonyms to solve the problem that the traditional algorithm model can only understand the surface meaning of text and thus affect the retrieval efficiency. By comparing with different algorithms, the argument of this article is solved by using the good sentences and words between the efficiency of search, this paper puts forward the sentence similarity computing and synonyms matching method can effectively improve the performance of information retrieval, but the effect of synonyms model here also is not very good, the subsequent will continue through parameter optimization, language training to improve the efficiency of model.

References

[1] Zhang Jianwei. Research on full-text information Retrieval technology for wechat content [D]. East China Normal University,2018.

[2] Li Nan, Zhang Fauna. A Semantic Search Engine Algorithm and its implementation [J].Journal of Shangluo university,2018,32(06):l-5+29.

[3] C. Bo and L. Yang-Mei, "Design and Development of Semantic-Based Search Engine Model," 2014 7th International Conference on Intelligent Computation Technology and Automation, Changsha, 2014, pp. 145-148, doi:10.1l09/ICICTA.2014.43.

[4] Yuan Fang. Research on text retrieval model technology based on semantic analysis [D]. Central China Normal University,2016.

[5] Guo Weiwei, Liu Feng. Journal of lanzhou university of arts and sciences (natural science edition),2016,30(01):51-55.

[6] Wu Yan, Wang Rujing. Application of Semantic Matching Algorithm based on BERT in Question answering System [J]. Instrument Technology, 2020, No.374(06):23-26+34.

[7] N. Arif, S. Latif and R. Latif, "Question Classification Using Universal Sentence Encoder and Deep Contextualized Transformer," 2021 14th International Conference on Developments in eSystems Engineering (DeSE), 2021, pp. 206-211, doi: 10.1109/DeSE54285.2021.9719473.

[8] B. Ye, G. Feng, A. Cui and M. Li, "Learning Question Similarity with Recurrent Neural Networks," 2017 IEEE International Conference on Big Knowledge (ICBK), 2017, pp.

111-118, doi: 10.1109/ICBK.2017.46.

[9] P. Zhang, X. Huang, Y. Wang, C. Jiang, S. He and H. Wang, "Semantic Similarity Computing Model Based on Multi Model Fine-Grained Nonlinear Fusion," in IEEE Access, vol. 9, pp. 8433-8443, 2021, doi: 10.1109/ACCESS.2021.3049378.

[10] Sakata, Wataru, Tomohide Shibata, Ribeka Tanaka, and Sadao Kurohashi. "FAQ retrieval using query-question similarity and BERT-based query-answer relevance." In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1113-1116. 2019.