2021 International Conference on Information Science, Parallel and Distributed Systems (ISPDS) | 978-1-6654-1738-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISPDS54097.2021.00018

Research on Quality Evaluation of Requirement Analysis Specification Based on Text Similarity Calculation

Yuyi Peng
School of Information Science and engineering
Hunan Normal University
Changsha, China
18173396080@163.com

Jin Zhang*
School of Information Science and engineering
Hunan Normal University
Changsha, China
* Corresponding author mail_zhangjin@163.com

Yiqi Huang
School of Information Science and engineering
Hunan Normal University
Changsha, China
1240065133@qq.com

Jie Tang
School of Information Science and engineering
Hunan Normal University
Changsha, China
justontang@qq.com

***Abstract*—Aiming at the problem of automatic quality evaluation of requirement analysis specification, an evaluation method based on text similarity is proposed. This method uses the typical process of natural language processing. Firstly, the evaluation document is preprocessed to complete word segmentation and stop word processing; Then, word2vec vector is constructed to replace the original document to be evaluated; Finally, the quality of standard documents is evaluated. Simulation experiments show that, compared with manual evaluation, its efficiency can be greatly improved, but considering the personalized problem of requirement analysis document writing, its evaluation accuracy needs to be further improved.**

Keywords-component; Natural language processing; Text similarity; Word2vec model; Requirements specification

I. INTRODUCTION

Chinese economic and social development has led to the rapid development of various industries in the software field, so that the software engineering industry has a strong demand for talents. The society is also trying to cultivate all kinds of high-level talents of software engineering to promote the overall development of the software industry. College students, especially the students majoring in software, should cultivate all kinds of software abilities to meet the social demand for talents in the software industry.

The development process of large-scale software products is a complex process that requires the participation of many developers. The traditional software development process has the main links of software demand analysis, software design and coding, software testing and maintenance, which is a process of understanding, creation and confirmation. In order to complete a software product that can make users satisfied and correct, software developers must first understand the user's requirements correctly, such as functional requirements, data resources and restrictions on function, data or behavior. Writing software requirements specification is regarded as a rigorous method to help solve software development problems. Software requirements specification can help developers understand users' needs and purposes correctly.

From the above two conditions, it is necessary to cultivate the ability of software engineering students to analyze and write software requirements specification. In daily teaching, teachers train students to analyze and write software requirements specification, and make evaluation to improve students' ability of analyzing and writing software requirements specification. However, due to the long document length of software requirements specification, low efficiency and high cost of manual evaluation, it causes a great waste of teachers' resources. Therefore, this paper proposes the application of text similarity calculation in the quality evaluation of needs analysis, which reduces the burden of teachers.

II. PRE-TREATMENT

A. Chinese word segmentation

Phrase is the basic unit of a sentence, whether it is to find keywords in a sentence or to calculate the similarity in understanding the meaning of a sentence, it is necessary to obtain the information of phrase, so word segmentation is the preparation of text similarity calculation[1]. At present, Chinese word segmentation technology has been more mature.

B. Stop words processing

It means that in order to save storage space and improve the search efficiency, some words will be filtered out automatically before processing natural language data or text[2]. For example, 'the', 'is',' at ',' which ',' on ', etc. These words or words are called "stop words", only the words with the specified part of speech, such as nouns, verbs and adjectives,are reserved.

This paper collects and sorts out a comprehensive stop words.txt from the list of "stop words in Harbin University", "disabled words in machine learning intelligent laboratory of Sichuan University" and "Baidu stop words list".

978-1-6654-1738-9/21/$31.00 ©2021 IEEE
DOI 10.1109/ISPDS54097.2021.00018

III. Word2Vec model

Word2vec is an open source model tool which is proposed by Google in recent years. It is based on neural network model to extract information from text[3]. It is one of the deep learning models. Neural network model plays an important role in the research of word vectorization. CBOW and skip gram models are mainly used for training, including input layer, projection layer and output layer. Hierarchical softmax and negative sampling technology are used to speed up the training process[4]. The final requirement is the vector representation of the projection layer. As shown in Figure 1.

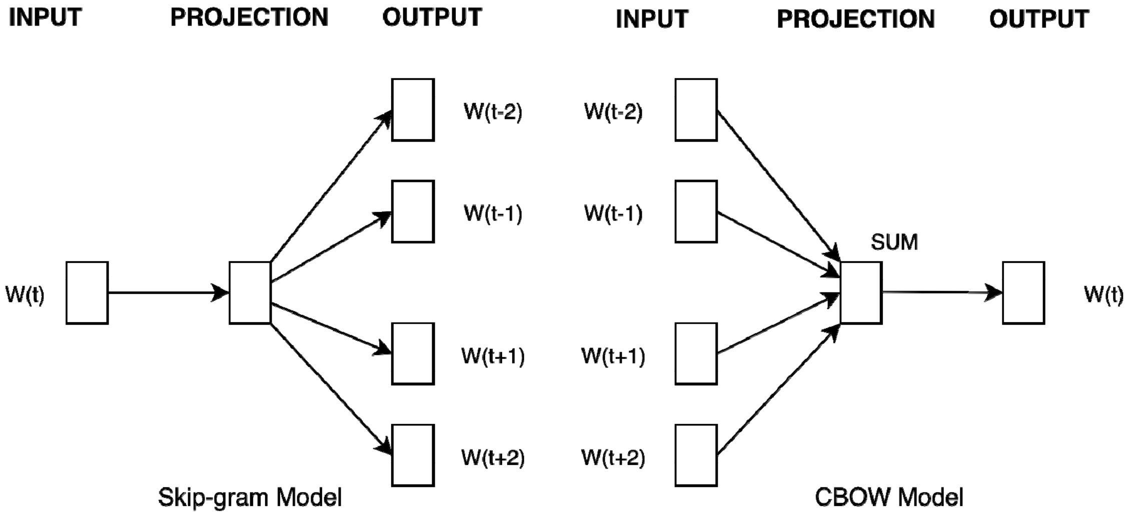

Figure 1. Word2Vec model.

TABLE I. Word2Vec model application example

father	
mother	0.9374490976333618
son	0.9025691151618958
daughter	0.8766654133796692
wife	0.8749578595161438
brother	0.8715890645980835
grandpa	0.8632363677024841
parents	0.8598790764808655
husband	0.8520638942718506

W(T) denotes the t-th word in W. It can be seen from the figure that when the head word is used as input in skip gram model, the context of its occurrence can be predicted. CBOW model, on the contrary, is the prediction head word for a given context.

A. *Word2vec model training*

Before training word2vec model, we need to do text pre-processing, including word segmentation, stop words and so on[5]. This paper selects 1.54g Sogou news corpus, adopts CBOW model, and uses genius module to train on word2vec[6].

The training parameters of word2vec are set as follows:

1) Size: the dimension of the output word vector, that is, the number of units in the hidden layer of the neural network. If the value is too small, the result of word mapping will be affected due to conflict. If the value is too large, the memory will be consumed and the algorithm calculation will be slow. A large size needs more training data, but the effect will be better[7]. The size value is set to 300 dimensions.

2) Window: the maximum distance between the current word and the target word in the sentence, that is, the window. The size of the window is set to 5.

3) min_ Count: filter words. Words whose frequency is less than min count will be ignored[8]. The default value is 4. Worker: the number of threads used to train the model. The number of threads is set to 4.

B. *Application of Word2Vec model*

After using corpus to train the model, we can use the model to view the vector of words, calculate the Related words of words and the similarity between words. Table 1 shows the Related words of "father". (The original Chinese words have been translated into English)

IV. Text similarity calculation

1) Pre-treatment. The student documents and reference documents are processed by text processing, namely word segmentation and stop word processing.

2) TF-IDF keyword extraction. TF-IDF algorithm extracts keywords, which measures the frequency of a word in a document. When a word appears repeatedly in a document, it means that the word may have a certain meaning in the article.

But not all words appear more frequently, the more meaningful they are. For example, if a word appears very frequently in all documents, it proves that the word is not of great value.

TF-IDF algorithm is a good measure of these problems, where: TF = The number of times the word appears in the document / Total number of articles, IDF = log (Total articles / Number of articles in which the word appears), TF-IDF = TFIDF. The higher the TF-IDF value, the higher the probability that the word will become a keyword.

3) Generate word vectors. Word vector is a feature extraction technology in natural language processing, which is trained on large-scale corpus by neural network language model. It maps words into dense, low dimensional real vectors, which has better expression effect than traditional feature extraction methods, and effectively avoids the dimension disaster of feature vectors[9]. After training the model, the keywords obtained in the previous step are put into the model to make the keywords vectorized.

4) Cosine similarity calculation. The cosine similarity is used to calculate the similarity of the vector after the word vector quantization is completed. The result of the calculation can determine the similarity between the student document and the reference answer[10].

Vector space model is a common similarity calculation model in the field of natural language processing, which has think of the word frequency vectors of two articles as two line segments in the space from the origin of coordinates and pointing to different directions[11]. The smaller the angle is, been used in the current research and development widely[9]. In the vector space model, we can calculate a word frequency vector according to the word frequency of each article, and the closer the cosine is to 1, that is, the smaller the distance between the two vectors is, that is, the more similar the words are.

Therefore, this paper uses cosine similarity to calculate the similarity of the vector for the word set that has completed word vectorization, and the result of the calculation can determine the similarity between the student document and the reference answer.

The specific process of text similarity calculation is shown in Figure 2.

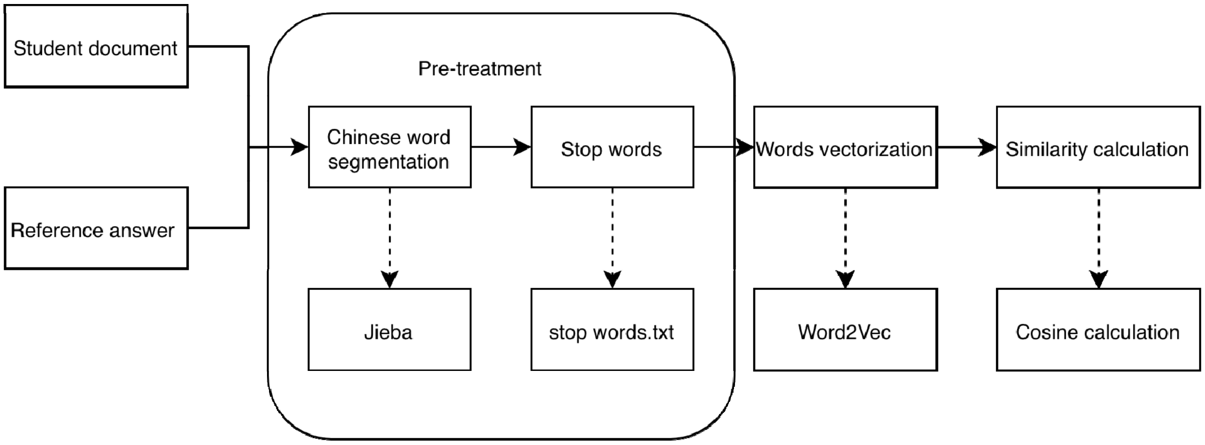

Figure 2. Flow chart of text similarity calculation.

V. Conclusion

In the text similarity calculation test evaluation index, the test data content comes from the content of the product manual, and the existing manual content is used as the reference answer for scoring. Select 10 content points from the manual to make questions, and count the score points of each question according to the manual marking method. Then let 10 students to answer the questions, according to the score points to 10 students to answer the content of scoring, and then use the evaluation system to score the answers of 10 students, and finally compare the two ways of scoring. According to the results of manual scoring and model scoring, as shown in the Figure 3.

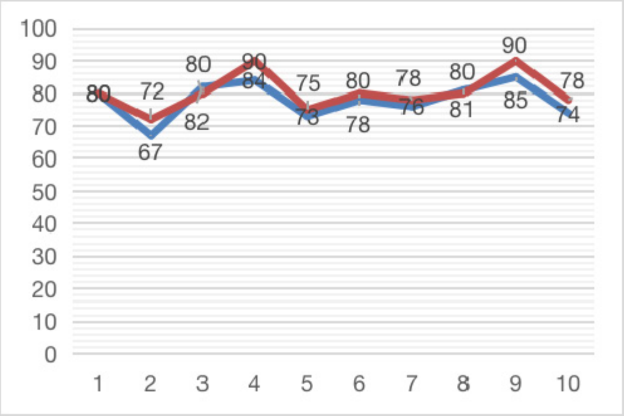

Figure 3. Comparison chart of manual scoring and model scoring.

Aiming at the application research of text similarity in requirement analysis document quality evaluation, this paper uses word2vec word vector to complete the similarity calculation. After pre-processing student documents and reference answers, word vectors are generated, and the quality of student documents is evaluated by cosine similarity calculation.

The red line is the result of manual scoring, and the blue line is the result of model scoring. It can be seen from the figure that the result of model scoring is very close to that of manual scoring. There is only a little deviation between the two, which can prove the reliability of the model evaluation results.

REFERENCES

[1] S. Omid, A. Lugowski, K. Younge, “Text similarity in vector space models: a comparative study.” In 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA) , pp. 659-666.

[2] L. Yubei, C. Feng, et al, “Design and Implementation of Intelligent Scoring System for Handwritten Short Answer Based on Deep

[3] Learning.”2020 IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS), pp. 184-189.

[4] W. Yongliang, et al, “Text classification method based on TF-IDF and cosine similarity,” J. Journal of Chinese Information Processing, 2017, pp. 138-145.

[5] C. Wang, Y. Yang, L. Deng, et al, “Review of text similarity calculation methods,” J. Inf. Sci, 2019, pp. 158-168.

[6] L. Zhifang, Z. Guoen, L. Junfeng, et al, “Semantic similarity algorithm for Chinese short texts,” Journal of Hunan University (Natural Science Edition),2016, pp. 35-140.

[7] A. Rokade, B. Patil, S. Rajani, et al, “Automated grading system using natural language processing,” In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), pp. 1123-1127, April 2018.

[8] W. Yongqiang, “Comprehensive application of text similarity detection in campus management system,” J. Computer programming skills and maintenance, 2014, pp. 55-56 + 63.

[9] P. Sravanthi, B. Srinivasu, “Semantic similarity between sentences,” International Research Journal of Engineering and Technology (IRJET), 2017, pp. 156-161.

[10] J. Oliva, J. Serrano, D. Castillo, et al, “SyMSS: A syntax-based measure for short-text semantic similarity,” J. Data & Knowledge Engineering, 2011, pp. 390-405.

[11] G Chongyang, X Haoyu, Z Han, et al, “Text similarity calculation based on lexical semantic information,” J. Computer application research, 2018, pp. 391-395.