Research on the Text Similarity Algorithms in Automatic Scoring of Subjective Questions

Juan Peng
Xi'an Eurasia University Xi'an, Shaanxi, China
pengjuan@eurasia.edu

Abstract. The theme revolves around the automatic scoring of subjective questions. The key technologies involved in the whole automatic scoring process are discussed in this paper, analyzes the advantages, disadvantages and application scope of different calculation methods for text similarity to understand the development trend of this field. It is helpful for further research on the automatic scoring method of subjective questions based on Word2vec improved model.

Keywords: Text Similarity, Automatic scoring, Neural network.

1. Introduction
With the development of computer technology and the enlargement of examination scale, the application of computer automatic evaluation technology in the examination system has become more and more popular. In the examination system, the objective problem of automatic marking technology already quite mature, but for the subjective questions such as short-answer questions and the discourse topic, because its answer is not unique, Chinese has complex semantic meaning, with synonyms and synonyms, etc. And limited by technologies such as artificial intelligence and natural language understanding, these make the automatic scoring of Chinese subjective questions more difficult. Therefore, it is of great practical significance to study how to use computer to realize automatic scoring of Chinese subjective questions.

2. Overseas and Domestic Research Status
2.1. Overseas Research Status
Abroad, it is relatively early to focus on the automatic scoring method of subjective questions, and a series of scientific research achievements have been made. The main automatic scoring systems include PEG, IEA, E-rater, etc.

2.1.1. Project Essay Grade (PEG). PEG scored the essays by analyzing the superficial linguistic features of the essays, without involving the right or wrong of the content. It uses proxy measurement criteria to measure the intrinsic quality of the composition to simulate the human rating of the composition. For example, the length of a composition represents the fluency of writing; prepositions, relative pronouns, and so on indicate the complexity of sentence structure; The word length indicates the wording of the article [1]. PEG pays more attention to surface structure than to semantics and content,
Because PEG focuses more on surface structure and less on semantics and content, the system does not provide instructive feedback to students.

2.1.2. Intelligent Essay Assessor (IEA). IEA uses the word frequency statistics method to calculate the similarity between the standard answer text and the student answer text by using the vector space model, and then obtains the score of the student text. IEA scoring system is able to calculate the similarity of large, word-rich texts, but there are many outliers in the vector space of short texts with fewer words [2].

2.1.3. Electronic Essay Rater (E-RATER). E-rater is mainly used for automatic scoring of English writing test. It uses both statistical methods and natural language processing methods. The overall scoring strategy adopted by the system can judge the examinee's essays from various aspects such as syntactic diversity and verbal ability, and through the verification of GMAT the evaluation accuracy is higher. However, the system also needs a large number of sample answers to build the scoring model and compare with the candidates' answers. It can only evaluate the candidates' writing level but cannot judge whether the candidates' answers are in good agreement with the questions [3].

2.1.4. IntelliMetric TM. IntelliMetric TM is a composition scoring system based on artificial intelligence. It mimics manual marking, grading essays on a scale of 1 to 4 or 1 to 6 for content, form, organization, and writing habits. In the scoring process, the system will first internalize the training set of the scores and constructs the model. Then the validity and generalization of the model are tested with a smaller test set. Once both items are confirmed, they can be used to judge the essay to be graded. In terms of performance, it achieves a high consistency with manual review.

2.2. Domestic research status
In China, since there is no natural separation between Chinese words, word segmentation should be involved, and Chinese semantics are relatively complex, which leads to a slow progress in the research on subjective automatic evaluation.

Han Yongguo et al. used the TF-IDF (Term Frequency-Inverse Document Frequency) algorithm to extract text keywords, and expressed the keywords through the vector space model. Finally, the text similarity was calculated according to the cosine of the included Angle. This algorithm was suitable for long texts, but did not consider the influence of word order [4].

Gao Sidan et al. studied in the way of DP (Dynamic Planning). At first, they needed to extract the keyword information of the text and calculate the similarity of the sentences with the extracted keywords through DP thought. However, this method ignored the word order of the keywords and did not conduct in-depth research on the semantic information [5].

Jiang Zhenfeng proposed an evaluation method based on information extraction in his research on automatic scoring of subjective questions for computer-aided evaluation. Firstly, a knowledge base of standard answers is established. Each question contains several standard answers, and scoring is realized through synonym substitution and improved BLEU (Bilingual Evaluation Understudy) algorithm. However, this method does not consider the effect of word order on semantics [6].

Li Xuejun from Southwest University of Science and Technology put forward a subjective automatic scoring algorithm based on artificial intelligence. The algorithm aims to apply the research results (Text space vector model, Word segmentation algorithm) of artificial intelligence on the understanding of Chinese natural language to the understanding of subjective question answers in online exams. He put forward the text vector feature matching algorithm based on vector control model, and applied it to the automatic scoring of subjective questions. However, there are some defects in the algorithm, which can only accurately evaluate the topics with clear scoring points and few keywords. The algorithm needs to be further studied and expanded [7].
3. Relevant theories and technologies

In the process of manual marking, when teachers read short answer questions, discussion questions and other subjective questions, they will first divide the answers according to the grading standards. And then they assign the total score values of the test answers to each part according to the detailed standards or the understanding and mastery of the examination contents. Secondly, the teacher will make the same split for the students' answers, so as to compare with the standard answers and get the score of the students' answers. It is particularly noteworthy is that even if the student's answer is not exactly the same as the standard answer, but the overall meaning is consistent with the standard answer, except for the different ways of literal expression. Through the teacher's understanding, the student's answer is still considered to be correct.

The marking process described above is a natural one for teachers, but if the process of manual marking is simulated by computer and a similar result is obtained, there will be many problems to be solved, mainly including the following problems.

3.1. Sentence segmentation and Word segmentation.

Computers need to be able to divide entire paragraphs of text into sentences accurately and recognize all the words that make up a sentence. Chinese sentences can be divided in accordance with the punctuation to separate, but for Word segmentation, because no obvious spaces between Chinese word and the word, and Chinese has a large vocabulary, which leads to ambiguity when words are divided. For example, "What he said really makes sense" can be divided into "He/says/of/actually/in reason", can also be divided into "He/say/act/reality/reason". If the word is not properly segmented, it will affect the computer's understanding of the text. Therefore, efficient and accurate segmentation of Chinese text is one of the problems that need to be solved in automatic scoring of subjective questions by computer. For Chinese word segmentation, many experts have put forward the corresponding methods, and introduced a number of word segmentation tools. The existing segmentation methods mainly include the following categories.

- Word segmentation method based on lexicon. By matching strings of words to be segmented with words in a large-scale lexicon, word segmentation based on lexicon can only be matched against words already in the lexicon, and the Words that are not included cannot be matched.
- Word segmentation method based on statistics. This method builds a large corpus, in which the samples have been segmented, and then selects an appropriate statistical model for word segmentation of Chinese texts.
- Word segmentation method based on rules. This method applies syntactic information and grammatical information to word segmentation simultaneously, so the result of word segmentation is more accurate, but the complexity of this method is higher.

3.2. Semantic similarity calculation.

Due to the existence of synonyms in Chinese, there are often two sentences with different words but basically the same meaning. A computer should be able to determine whether two words are synonyms by calculating their semantic similarity.

There are two main methods to calculate the semantic similarity of words in natural language processing. One is to calculate the concepts of words in a tree structure by means of semantic dictionary. The other is mainly through the word context information, to use statistical methods to solve.

Dekang Lin believes that the similarity of any two words depends on their commonness and individualities, and then gives the definition formula from the perspective of information theory:

$$\text{Sim}(A, B) = \frac{\log p(\text{Common}(A, B))}{\log p(\text{Description}(A, B))}$$  (1)

Thereinto, numerator is the amount of information needed to describe the commonness of A and B; the denominator is the amount of information needed to complete the description of A and B.
According to Liu Qun and Li Sujian, if two words in different contexts can be replaced with each other without changing the syntactic and semantic structure of this paper, the more likely this is to happen, the higher the similarity between the two, otherwise the lower the similarity [8]. The similarity calculation formula is as follows:

\[ \text{Sim}(A, B) = \frac{\alpha}{\alpha + \text{Dis}(A, B)} \]  

(2)

Thereinto, \( \text{Dis}(A, B) \) represents the distance between the two words; \( \alpha \) is an adjustable parameter, it means the value of the word distance when the similarity is 0.5.

3.3. Text similarity calculation.

The computer should be able to simulate the process of manually marking subjective questions and calculating final scores. That is, obtaining the score of the student's answer according to the degree of similarity between the student's answer and the standard answer. The more similar a student's answer is to a standard answer, the higher his score should be. Therefore, the computer should have a better ability to calculate text similarity to improve the accuracy of subjective questions scoring.

In the calculation of text similarity, there are two key elements: one is the representation model, which represents the text as a numerical vector that can be calculated by the computer, namely the eigenvalue. The other is the algorithm, which calculates the similarity between texts based on the eigenvalues. Here are some common methods.

3.3.1. Based on string.

The method starts from the string matching degree and takes the degree of string co-occurrence and repetition as the criterion of similarity. According to the granularity of calculation, the methods can be divided into methods based on character and methods based on word. One is to consider the similarity algorithm based on the composition of characters or words, such as Levenshtein Distance, Hamming distance, Cosine Similarity, Dice Index, Euclidean distance, etc. In another kind of methods, character order is added. That is, having the same character composition and character order is a requirement for string similarity. Such as longest common subsequence (LCS), Jaro-Winkler. In another kind of methods, the idea of set is adopted, and a string is regarded as a set composed of words, and words co-occurrence can be calculated by the intersection of sets[9]. Such as N-gram, Jaccard, Overlap Coefficient.

Among them, the principle of cosine similarity algorithm is to calculate the frequency of all words in the two comparison texts respectively, so as to obtain the corresponding vector of the two texts. And then we calculate the included angle cosine value of the two vectors by using the cosine theorem. The smaller the similarity, the greater the distance. The greater the similarity, the smaller the distance. The calculation formula of cosine is as follows.

\[ \cos \alpha = \frac{\sum_{i=1}^{n} f_{A_i} \times f_{B_i}}{\sqrt{\sum_{i=1}^{n} (f_{A_i})^2} \times \sqrt{\sum_{i=1}^{n} (f_{B_i})^2}} \]  

(3)

3.3.2. Based on corpus.

The corpus-based method uses the information obtained from the corpus to calculate the text similarity. The methods based on corpus can be divided into: the method based on Bag-of-words model, the method based on neural network and the method based on search engine. The first two sets of documents to be compared with each other are corpus, while the latter takes Web as corpus.

The method based on Bag-of-words model uses a set of word sequences to represent a piece of text. This set of word sequences is Bag-of-words model, also known as vocabulary. Bag-of-words model holds that the occurrence probability of one word has nothing to do with the occurrence probability of other words. And the occurrence probability of words is independent of each other. The biggest defect of Bag-of-words model is the high dimension of vectors, which leads to a large amount of computation for subsequent similarity or text classification, and the sparse data also leads to the lack of obvious similarity distinction. According to the different degree of semantics considered, the methods based on
Bag-of-words model mainly include Vector Space Model (VSM), Latent Semantic Analysis (LSA), Probability-based Latent Semantic Analysis (PLSA), and Latent Directray Distribution (LDA).

The method based on neural network is to use machine learning to generate word vector through neural network language model (NNLM). The word vector is the incidental output of the language model. The basic idea behind NNLM is to predict the words that appear in the context, and this prediction of the context is essentially a study of the statistical characteristics of co-occurrence. The commonly used methods to generate word vectors include: Skip-gram, CBOV, GloVe, LBL, C&W, etc. The word vector generated based on the neural network model is a trained low-dimensional real vector. The dimension can be artificially limited and the real value can be adjusted according to the text distance. This text representation conforms to the way people understand the text.

The method based on search engine. With the continuous development of network technology, Web has become the corpus with the most abundant content and the largest amount of data. At the same time, the progress of search engine related algorithms enables users with any needs to find the answers through search[9]. The famous one is the normalized Google distance, which is based on the number of keyword search results provided by search engines and measures the degree of semantic similarity. Keywords with the same or similar meaning in natural language tend to be "close" in search engines, and words with different meanings tend to be more distant. The biggest disadvantage of this method is that the results are completely dependent on the query effect of the search engine and the similarity varies with the search engine.

3.3.3. Based on world knowledge. The method based on world knowledge is to use the knowledge base with normative organization system to calculate text similarity, which is generally divided into two types: ontology-based knowledge and network-based knowledge. The former is generally based on the upper and lower and isotopic relationships between concepts in the ontology architecture. If concepts are semantically similar, then there is only one path between the two concepts [8,10]. In the Internet knowledge, entries are structured and hyperlinks are used to show the relationship between the upper and lower positions, which is closer to the computer's understanding of information organization. Paths between concepts or links between terms become the basis of text similarity calculation.

3.3.4. Other methods. In addition to the above three methods, there are some other methods for text similarity calculation, such as syntactic analysis and hybrid method [9]. The main task of syntactic analysis is to identify the syntactic components contained in a sentence and the relationship between these components. When the similarity is calculated, word similarity and relationship similarity are taken into account. Syntactic trees are generally used to represent the results of syntactic analysis. However, the complexity of sentence itself brings a certain degree of difficulty and a great amount of work for frame analysis. Hybrid method is to make up for the deficiencies of a single algorithm. It is the comprehensive use of two or more methods to calculate text similarity. To some extent, it can improve the effect of text similarity calculation.

4. Summary
Many achievements have been made in the research of text similarity methods. This paper mainly focuses on the topic of automatic scoring of text subjective questions, discusses the key technologies involved in the whole automatic scoring process, and lays a theoretical foundation for the later research of automatic scoring of text subjective questions based on improved model of Word2vec. The research is of great significance to ensure the fairness of the examination questions, improve the teachers' work efficiency, reduce the teachers' burden, and give feedback to the students' knowledge in time.

Acknowledgments
University-level scientific research Fund project of Xi'an Eurasia University (2020XJZK12).
References

[1] Valenti S, Neri F, Cucchiarelli A, An Overview of the Current Research on Automated Essay Grading [J]. Journal of Information Technology Education, 2003.

[2] Foltz PW, Streeter LA, Lochbaum K E, Implementation and applications of the Intelligent Essay Assessor [M]. Handbook of automated essay evaluation. Routledge, 2013, pp.90-110.

[3] Dikli s, An overview of automated scoring of essays [J]. The Journal of Technology, Learning and Assessment, 2006, 5(1).

[4] Zhao Zhihong, HAN Yongguo, Research on Subjective Topic Automatic Evaluation System based on NLP-TF Technology [J]. Software Guide, 2013(1), pp. 82-83.

[5] Gao Sidan, Yuan Chunfeng, Preliminary Application of Sentence Similarity Calculation in Automatic scoring of Subjective Questions [J]. Computer engineering and applications, 2004, 40(14), pp. 132-135.

[6] Jiang Zhenfeng, Liu Li, Research on Automatic Evaluation of Subjective Questions based on Computer-aided Evaluation [J]. Journal of Chongqing Normal University: natural science, 2013, 30(2), pp. 74-78.

[7] Li Xuejun, Realization of automatic scoring algorithm for subjective questions based on artificial intelligence [J]. Journal of Jiangnan University: Natural Science edition, 2009, 8(3), pp. 292-295. (in Chinese)

[8] Liu Qun, Li Sujian, Lexical Semantic similarity calculation based on CNKI [J]. Chinese Computational Linguistics, 2002, 7(2), pp. 59-76.

[9] Chen Erjing, Jiang Enbo, Review of text Similarity Calculation Methods [J]. Data analysis and knowledge discovery, 2017,1(6), pp. 1-11.

[10] Sun Haixia, Qian Qing, Cheng Ying. A Review of Ontology-based Semantic Similarity Computing Methods [J]. Modern Library and Information Technology, 2010(1), pp. 51-56.