Corpus-Based Automatic Integrated Scoring Algorithm for English Composition

Miao Wang[1(✉)], Desheng Zhu[2], and Sufang Wang[2]

[1] Modern College of Northwest University, Xi'an 710130, Shaanxi, China
57510514@qq.com
[2] Dongying Vocational Institute, Dongying 257091, Shandong, China

Abstract. In recent years, with the popularization and application of artificial intelligence technology in various fields, the field of automatic scoring of English composition has also received great attention and development. Using corpus technology to score English writing automatically, this paper shows the operation method of corpus technology in automatic scoring of English writing, and demonstrates the feasibility of this scoring method. This paper develops an automatic scoring method for large-scale automatic scoring of English writing. By comparing it with the traditional manual marking method of English composition, this paper evaluates the advantages and disadvantages of corpus based automatic scoring method of English composition. Automatic composition grading has the advantages of high working efficiency, strong objectivity and fairness, detailed and comprehensive marking results, etc., but it also has the disadvantages of mechanical and blindness, errors in error analysis, and the reliability and validity still need to be improved. Therefore, college English teachers can't regard this automatic grading method as a shortcut to correct English compositions and rely too much on it.

Keyword: Corpus · Automatic scoring · Efficiency

1 Introduction

With the development of computer technology, especially the development of artificial intelligence technology in recent years, computer technology is more and more applied to people's study, work and life. Automatic scoring of English composition, abbreviated as AES, is one of them. There are English college entrance examination and CET-4 and CET-6 for college students in China. All these examinations have English composition and writing assessment, which is used to judge students' logical thinking ability and language mastery. However, due to the subjective expression of English writing, the artificial scoring of English compositions is subjective. Due to the lack of clear and clear evaluation criteria, and each rater may have subjective preferences, different raters of the same English composition will also have deviations in scoring. In addition, grading English compositions requires a lot of teachers' energy. In addition to these drawbacks,

Y. Zhang and N. Shah (Eds.): BigIoT-EDU 2023, LNICST 580, pp. 538–544, 2024.
https://doi.org/10.1007/978-3-031-63130-6_60

manual grading also has the problem of untimely feedback. Students need to wait a long time to know the scores and the teacher's feedback after writing the composition, which is undoubtedly inefficient for improving students' English writing level. If the English composition automatic scoring system can be introduced to score students' compositions, the above problems can be effectively solved.

Corpus refers to "written and oral language materials processed and stored by computer for the study of language". In the construction and development of corpus, some principles and methods are gradually put forward, theoretically discussed and summarized, and corpus linguistics is formed. Corpus linguistics is defined as "taking corpus as the starting point of language description or using corpus to verify relevant language hypotheses". It is also defined as "a discipline that uses corpus to carry out linguistic research". Automatic composition scoring refers to the process of using computer technology to evaluate and score the composition. Its essence is to classify the composition to be scored according to the accurate manual scoring of the characteristic information of the composition. Since the automatic scoring system entered the application stage, there have been different comments on it. Supporters believe that the automatic composition scoring system is highly operable, economical and practical, objective and accurate, and has a high correlation with manual scoring. From this point of view, we can use computer automatic scoring instead of manual scoring. However, there are not a few dissenters. Among them, the most criticized one is that the automatic scoring system relies on computer programs, which are generally flawed. As long as the examinee has the relevant vulnerability information, he can get high scores to cater to the computer's liking, but his writing may be meaningless.

2 Corpus-Based English Composition Scoring

2.1 Advantage

There are five main advantages of corpus based automatic scoring of English composition (Fig. 1). First, high work efficiency. The efficiency of English composition automatic scoring method is 100 times higher than that of teachers' correction. This is also its biggest advantage. When students submit their English compositions to the network system, they will soon be able to see the scores and comments of their compositions. Because the feedback is timely and in line with students' expectations, it is a positive feedback stimulus for students, which improves students' interest in English composition and their enthusiasm for English writing. Secondly, it has strong objectivity and fairness. To judge the quality of an English composition, we need to comprehensively measure the language, content and writing skills of the composition. However, in the traditional way of manually marking English compositions, teachers tend to pay more attention to them. Some teachers attach importance to the comprehensive application ability of English, others focus on the composition content, while others attach importance to students' flexible application of the previously trained writing skills. Coupled with other subjective factors, every composition review is not objective and fair enough. This is often a psychological blow to students and a negative stimulus feedback. However, every time an English composition is evaluated, the automatic scoring method will make a comprehensive score from the language, content and writing skills of the composition according

to certain standards, write a comprehensive comment on writing, and point out and correct the mistakes in all aspects. This way of marking ensures objectivity and fairness for every student. Once again, the results of marking are detailed and comprehensive. In recent years, English teachers no longer point out and correct the mistakes in students' compositions as in the past. In order to encourage students, teachers often give general comments with praise. The students don't know that there are many mistakes in their compositions, so they don't bother to correct the language mistakes one by one. In this way, it is difficult to improve students' Comprehensive English application ability. In the corpus based English composition scoring method, by comparing the students' articles with the English articles in the corpus, the system will find the language errors that do not conform to the grammatical rules or pragmatic habits in the students' compositions, point out and correct them one by one, and evaluate the content and organizational structure of the articles. Detailed correction information can enable students to fully understand their English composition, so as to correct their mistakes in time and improve their comprehensive English application ability. Thirdly, promote the effective transformation and repositioning of teachers' role. The way of English composition grading frees English teachers from the task of composition correction which was overwhelmed before, and concentrates on the effective guidance and supervision of students' writing process. Before writing a composition, teachers can train students in writing skills and guide them to pay attention to the content and structure of the article. Teachers can also find out the submission time and times of each student's composition in the class through the correction network, and urge those students who are lagging behind and lazy in time. The traditional teacher-led writing teaching method has changed into a teaching method centered on students' independent writing. English teachers have also changed from former "teachers" and "critics" to "organizers" and "instructors". Finally, reduce students' pressure and improve students' interest. In the traditional manual marking method of English composition, due to time and energy reasons, teachers rarely require students to modify the composition according to the comments and carry out follow-up inspection; On the other hand, students first pay attention to the registered composition scores rather than the teachers' comments, so many students turn a blind eye to the teachers' comments. The automatic scoring method of English composition based on corpus provides students with the opportunity to modify the composition content and language in time, and promotes students' in-depth thinking of the composition content and the continuous improvement of their comprehensive English application ability. According to the timely feedback of composition errors, students can continuously improve their composition scores through repeated revision, which is a positive feedback stimulation for students. It encourages students to think deeply, write seriously, revise repeatedly and strive for perfection. This is an effective way to improve students' English writing ability.

2.2 Disadvantages

There are four main disadvantages of corpus based automatic scoring of English composition (Fig. 2). First, mechanical and blindness. After all, the automatic scoring method of English composition based on corpus is machine operation. Like other information processed by computer, it carries out mechanical operation according to the input program, and can not be handled creatively and flexibly in case of special situations and

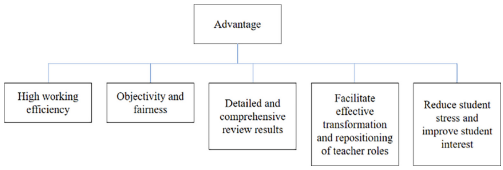

Fig. 1. Advantage

new problems. For example, it may turn a blind eye to the creative thinking shown in students' compositions, and haggle over every penny for students' temporary clerical errors, which will inevitably hurt students' creative passion. In addition, it also has the blindness unique to the machine. For example, it will not appropriately improve or reduce the scoring standard according to the difficulty of each composition, and will not take into account the subjective feelings of students when writing comments. It can not exchange ideas with students and form good interaction like teachers. Secondly, there are errors in error analysis. In the process of using the automatic scoring system, careful teachers and students will find that this kind of automatic English composition scoring often makes some wrong judgments when marking language errors in compositions. Thirdly, it is easy for teachers to have dependence psychology and slack psychology. The quickness and convenience of automatic English composition grading can easily make teachers feel completely dependent on the marking network and slack off. Although the scoring system provides teachers with the opportunity to revise online comments, when they see comprehensive comments and detailed error analysis on the Internet, teachers may stop manually marking them. Over time, they are prone to slack off, leaving the arduous task of marking compositions to the marking network. In this way, they can not correct the correction errors of the correction network in time, and lack an in-depth understanding of the students' composition, so they can not effectively guide the students to correct some errors. Finally, reliability and validity still need to be improved. Automatic composition scoring system is a very complex technology. It needs to make rational use of multi-disciplinary technology and language testing theory in order to achieve the ideal effect. At present, various automatic online evaluation software for English composition developed at home and abroad have their own advantages and disadvantages. Their reliability and validity in evaluating English composition, which is a highly subjective topic, has been questioned.

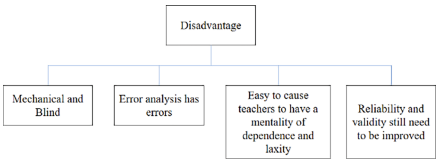

Fig. 2. Disadvantages

3 English Composition Automatic Grading

3.1 Research Status

Corpus refers to "written and oral language materials processed and stored by computer for the study of language". In the construction and development of corpus, some principles and methods are gradually put forward, theoretically discussed and summarized, and corpus linguistics is formed. Corpus linguistics is defined as "taking corpus as the starting point of language description or using corpus to verify relevant language hypotheses". The study of Chinese corpus linguistics has developed rapidly in the past decade. There are more than 300 articles on corpus research in more than 20 core foreign language journals, such as modern foreign languages, foreign language teaching and research, and foreign languages (Fig. 3).

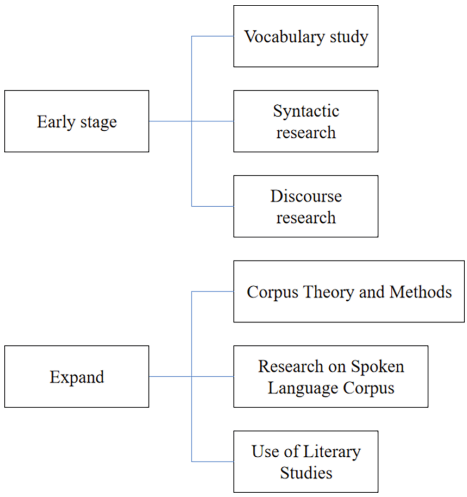

Fig. 3. Corpus research direction

At present, the level and depth of corpus in language ontology research continue to extend. But generally speaking, corpus is rarely used in language teaching. Therefore, how to integrate the resources and means of corpus into English teaching practice has become a new challenge for corpus research. First, collect corpus; Secondly, build a corpus and process the collected corpus; Secondly, corpus labeling and processing, unmarked corpora sometimes can not fully provide the information needed for research; Thirdly, data extraction (Fig. 4); Finally, the data are compared and analyzed.

No matter what stage of students, English has always been a very important subject. Students usually have English writing tasks. In the final or entrance examination, they have to take a large-scale English examination, in which the scoring of English writing is a heavy task. If we can use the English composition automatic scoring system for automatic scoring, it has the following significance. First of all, computer automatic scoring is very objective. Compared with the subjectivity of manual scoring, computer

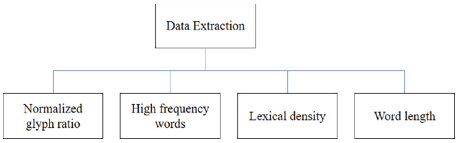

Fig. 4. Data extraction

scoring will be more fair to students. Secondly, computer automatic scoring can quickly obtain scoring results and relevant feedback, which is conducive to students' timely feedback and help students save valuable time. Thirdly, computer scoring can save labor costs.

3.2 Language Model

Language model plays an important role in language processing. The following model is a classic language model. When the statistical language model is used for language modeling, the prediction of the **n**th item is based on the previous $\mathbf{n-1}$ items. In information theory, it is mentioned that if a string sequence is given, the next character of the string sequence can obtain the N probability distributions of the next character of the string sequence by training corpus data and using maximum likelihood estimation method.

Assuming a sentence $s = w_1 w_2 \cdots w_n$, the probability of the sentence can be calculated by the following formula:

$$p(s) = p(w_1)p(w_2/w_1)p(w_3/w_1 w_2)\cdots p(w_n/w_1 \cdots w_{n-1}) \tag{1}$$

In order to calculate the probability of the occurrence of this sentence, we need to estimate the probability of n parameters. With the increase of the length of the sentence, more and more parameters are required for the calculation of the probability of the occurrence of words behind the sentence. At the same time, a large amount of training data is needed to reasonably estimate it. In addition, because many sentence sequences may not appear in the training data, the longer the sentence sequence, the more serious the problem of data sparsity. In order to solve this problem, Markov hypothesis is introduced, that is, the current word is only related to the previous words. According to the above description, the sentence probability is calculated as:

$$p(s) = \prod_{i=1}^{n} p(w_n/w_1 w_2 \cdots w_{n-1}) \approx \prod_{i=1}^{n} p(w_i/w_{i-2}w_{i-1}) \tag{2}$$

The value of $p(w_i/w_{i-1}w_{i-2})$ can be obtained by the following formula:

$$p(w_i/w_{i-2}w_{i-1}) = \frac{c(w_{i-2}w_{i-1}w_i)}{c(w_{i-2}w_{i-1})} \tag{3}$$

In Formula (3), $c(w_{i-2}w_{i-1}w_i)$ represents the number of times $(w_{i-2}w_{i-1}w_i)$ triples appear in the corresponding training corpus, and $c(w_{i-2}w_{i-1})$ represents the number of times $(w_{i-2}w_{i-1})$ triples appear in the corresponding training corpus.

4 Conclusion

This paper makes a preliminary comparative analysis of the automatic scoring method of English composition based on corpus. As for whether this new method can effectively improve students' writing ability and promote college English writing teaching, as well as the reliability and validity of this new method, it still needs to be tested in the long-term use process and detailed and in-depth empirical research. This paper applies corpus text analysis technology to the evaluation of English writing quality, and explains the operation steps of corpus quantitative evaluation. Through the investigation of five corpus evaluation parameters: standardized aspect ratio, vocabulary density, high-frequency words, word length and average sentence length, it is demonstrated that corpus evaluation parameters can be used as reliable parameters to quantitatively evaluate English writing. Therefore, we have successfully explored a new scoring method for English writing corpus. This method is scientific, systematic, efficient and fair. It has wide application value and practical applicability, and can be popularized and applied to all units to automatically evaluate the quality of English writing. English teachers can't regard this automatic grading method as a shortcut to correct English compositions and rely too much on it. Admittedly, this method frees English teachers from the task of correcting compositions that were overwhelmed before, but it also puts forward new requirements and challenges for teachers. Teachers should change their roles in English writing in time and reposition themselves.

References

1. Zhou, M., Jia, Y., Zhou, C., et al.: An automatic scoring method for English composition based on text structure. Comput. Sci. **046**(003), 234–241 (2019)
2. Han, H.: Problems and countermeasures in the research on automatic scoring of college English compositions. Lit. Youth **2019**(12), 1
3. Wang, H., Li, F.: The application of corpus in English writing teaching. College English Teach. Res. **2016**(3), 9
4. Du, S., Chang, R.: Corpus-based research on the relationship between English writing level and lexical chunk usage. J. Heilongjiang Inst. Educ. **37**(7), 3 (2018)
5. Zhou, M., Jia, Y., Zhou, C., et al.: An automatic scoring method for English composition based on text structure. Comput. Sci. **46**(3), 8 (2019)
6. Liu, G.: Automatic evaluation of local coherence in English learners' composition based on WordNet semantic knowledge base. J. Henan Normal University: Nat. Sci. Ed. **44**(6), 10 (2016)
7. Zhong, X.: Research on higher vocational college students' English writing self-efficacy based on the composition automatic scoring system. J. Changsha Vocat. Tech. College Commun. **016**(004), 64–67 (2017)
8. Li, X., Zhong, L.: An Empirical study of the automatic composition evaluation system in college English writing teaching—taking Juku correction network as an example. Teach. Res. **40**(1), 5 (2017)
9. Bai, Y.: Research on the characteristics of online composition self-modification based on automatic scoring system. Modern Commun. **2020**(15), 2
10. Yang, L.: Category-based English composition scoring standards and its operability research. J. Zhengzhou Railw. Vocat. Tech. College **31**(3), 4 (2019)