Journal of Southeast University (English Edition) Vol. 28, No. 4, pp. 410 - 414 Dec. 2012 ISSN 1003—7985

A study on the feasibility of CET oral test based on automatic essay marking

Liu Jiangang[1] Zhen Yuqi[1] Chen Meihua[1] Jiang Hao[2] Zhao Li[3]

([1]School of Foreign Languages, Southeast University, Nanjing 210096, China)
([2]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China)
([3]School of Information Science and Engineering, Southeast University, Nanjing 210096, China)

Abstract: This paper discusses the approach to the CET (college English test) oral test based on automatic essay marking (AEM). The reliability and validity arouses a dispute on Internet-based CET (iB-CET) oral test in China, leading to the penetration into oral test contents in dialogue, description or comment, and question answering. Then, a probe into transformation from a spoken document into a textual document is touched on a conversion ratio of 72% adopted at the present time, being a necessary hypothesis for AEM to be carried forward. Afterwards, this paper focuses on the pipeline in AEM with the content features and the language itself, which is established in two different models in marking for extraction and detection. Through experiments, this paper reveals the fact that an iB-CET oral test can be performed perfectly in the pipeline of spoken documents transformed into textual documents which can be automatically marked. Hence the design of the iB-CET oral test reaches its reliability and validity.

Key words: reliability and validity; spoken document and textual document; automatic essay marking

doi: 10.3969/j.issn.1003 - 7985.2012.04.007

The college English test (CET) has been carried on for decades in China since 1987 with over one hundred thousand student examinees from the very start[1]. Influenced by the impact of the iB-TOEFL entrance into China, the Chinese Ministry of Education launched a project of developing the Internet-based college English test (iB-CET) in 2007, with a view of improving the validity and administration efficiency of the CET. The trial test of the iB-CET4 got started among 53 colleges and universities on the 20th of December, 2008 so that China made a historic milestone for high education[2]. Lots of experts and scholars testified to the test design with focuses on the purpose of the test, the authenticity of the test tasks, the ratios between constructed and selected response items, and the practicality of marking, convinced that the iB-CET would have a positive impact on the reform of college English teaching and on the implementation of the computer-and classroom-based college English teaching models.

The main feature for the iB-CET aims at scoring the oral test in the exam on line. The oral test takes up a proportion of 10% in the whole test marks, making a revolution in the paper-based test system; hence, the iB-CET comes to life, a wonder in Chinese education. The design and plan worked and became accepted by the Chinese authorities and faculties in educational fields.

However, the most challenging hit results in the credibility of the oral test, because nobody can find any explanation regarding the characteristics, performances and types of assessment on the official Website of CET (www.ccets.org). Doubts come from different corners that the iB-CET oral test only covers imitation performance rather than other performances like situational dialogue, self-introductive presentation concerned about the scientific criteria of reliability and validity.

This paper attempts to clear away illiteracy of linguistic perception, speech recognition technology and the automatic essay scoring method based on data mining through an illustration by our experiments on the oral test system in the network of Southeast University, convincing people involved that the iB-CET oral test works in some ways.

1 CET Oral Test Reliability and Validity

A scientific "truth" gets started as a hunch, which produces a hypothesis. When the ways of a test are devised, the truth of a hypothesis is established. Sometimes people often find that a theory is elegant and useful, but falsifiable. So reliability simply depends on validity in a sense.

The CET oral test requires reliability and validity, let alone the one in the iB-CET. To determine if a test is of

Received 2012-05-28.

Biography: Liu Jiangang (1959—), male, associate professor, jhonliunj@163.com.

Foundation items: The Humanities and Social Sciences Project granted from the Ministry of Education in China (No. 10YJA740061, 11YJA740121), the Teaching Reform Key Project under Educational Section of Jiangsu Province (No. 2011JSJG453), Southeast University Teaching Reform Project for Graduates (No. 2010-54), Southeast University Teaching Reform Project for Post-Graduates (No. KJGKT12-06), the Natural Science Foundation of Jiangsu Province (No. BK2008354), the National Natural Science Foundation of China (No. 6550181821, 60975017).

Citation: Liu Jiangang, Zhen Yuqi, Chen Meihua, et al. A study on the feasibility of CET oral test based on automatic essay marking. [J]. Journal of Southeast University (English Edition), 2012, 28(4): 410 - 414. [doi: 10.3969/j.issn.1003 - 7985.2012.04.007]

any good, a test designer must know what is going to be tested. The oral test includes imitation, spoken words in intonation, grammatical rules, correct answering and so on. All might be seen as indicative in testing one's communicative ability, which accounts for the validity of the oral test. This raises doubts for the present iB-CET oral test, which only covers imitation performance.

The oral test should be a test assessment based on the contents of the performance designed for encouraging students to talk[3]. According to the design, the CET oral test intends to detect the communicative ability in three ways: self-introduction, description and comment on a picture given, and answering questions from examiners[4]. Since there is no evidence given on speech recognition rate, it is more likely to reject the present iB-CET oral test in imitation performance covering a small portion of the imitation ability of the examinees. The test is based on the hypothesis that machine scoring of repetition questions can work perfectly well in large-scaled English oral test with a satisfactory result in a matching ratio of 82.7% between man-powered scoring and machine scoring (see Fig. 1)[5].

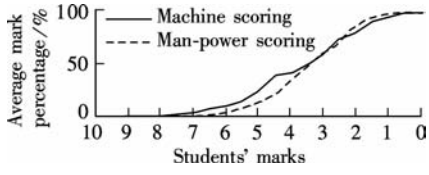

Fig. 1 Matching valid results on repetition scoring hypothesis

With disproof and doubts regarding the validity of the iB-CET oral test performance on imitation-based repetition matching technology, is there any other hypothesis of communicative ability testing methods based on the present advanced technology? The answer is definitely "YES" in speech recognition and data-mining technology, which is the main focus of this paper.

2 Speech Recognition Technology Applied

Accurate transcription of speech into text (STT for short) is a prerequisite for virtually all other natural language applications operating on audio sources which can be practically obtained through the iB-CET oral test because the speech recognition technology has reached a recognition rate of 95% in the lab and 70% to 80% in real customer application[6]. Advances in speech processing make the application of information extraction possible on spoken documents (such as iB-CET audio records), beyond the traditional textual documents[7]. Different detection systems are designed to reach perfect speech recognition. This can be well illustrated in the STT integration recognition pipeline (see Fig. 2)[8].

According to Professor Feng Jiali, the conversion ratio can be 100% with a matching sample corpus, and 72% without a matching sample corpus in the random model[9]. This conversion ratio can also be evidenced by using a method based on a short-time spectrum and prosodies such as pitch contour, duration and stress[10]. This sounds like great news for the iB-CET oral test in validity, since audio samples can be easily collected before the oral test properly takes place. The Chinese education is somewhat compulsory when the CET certificate is required before leaving school, iB-CET designers can put instructions to read a short passage before matching corpus. Under the circumstances, we can apply the transferred spoken version (now in textual version) to match the communicative ability so as to reach the demands of oral test validity by, say, 72% out of the converted textual version. Out of this 72% converted textual version, we seek out the self-introduction, picture description or comment and question answering any matching points required by test designers so as to mark the oral test.

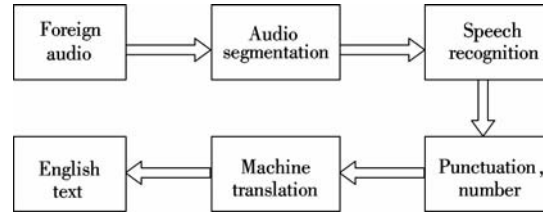

Fig. 2 STT integration pipeline illustration

3 AEM Applying for iB-CET Oral Test

Automatic essay marking (AEM) has been a hot study in China, which has been applied in Chinese education, such as in Chinese essay marking. In large-scale CET exams, English writing marking is an obstacle in English teaching; hence, lots of projects have been launched by the authorities to solve this problem so as to raise the efficiency of English teaching.

As a practice, Southeast University also sets up a small foundation for essay marking under the project of data mining. English teachers are required to accomplish article proofreading once a week for a student, which seems a heavy task since there are hundreds of students assigned to a teacher. Supported by the School of Computer Science and Engineering, the study on automatic essay marking got started two years ago, and it has worked well up to the present.

The study focuses on four aspects: content, organization, sentence structure and diction.

3.1 Target of essay marking

Linguistically, a good article consists of four parts: coherent content, clear organization, grammatical sentences and precise diction[11]. In real practice, there is a discussion among examiners on how to give correct marks in large-scale exams like the CET, concentrating on the percentage of content, organization, sentence structure and diction. Take content for example: if the examinee deviates from the topic, even if he/she well organizes his/her

essay with excellent grammar in perfect diction, he/she might get a zero mark in the final result. So, we arrange content at the top, organization in the second place, sentence in the third place and diction at the bottom. This gives good guidance for the curve of marking in the computer.

In computer-assisted essay marking, relying on the principles guided by linguists, the marking target aims at the content features and the language itself [12]. Content features cover semantic analysis such as key words and topic sentences by setting up three content recognition models: the vector space model (VSM), the latent semantic analysis model (LSAM) and the model with vector space model based on wordnet semantic dictionary (VSMBWSD). Language itself covers vocabulary, sentence structure and grammar, in which, some variable characteristics are selected, making a comprehensive analysis to measure the quality of the composition language, and score according to each feature item.

3.2 Pipeline of essay marking

The essay marking in the computer is designed in two aspects: content features abstraction and language itself detection. The two aspects can be simply illustrated as follows[13].

Content features abstraction adopts statistical methods to analyze the potential text semantic structure between words so as to extract the semantics of the words. Singular value decomposition is based on the matrix having different entities in the ranks, which will be broken down into three specific forms (see Fig. 3).

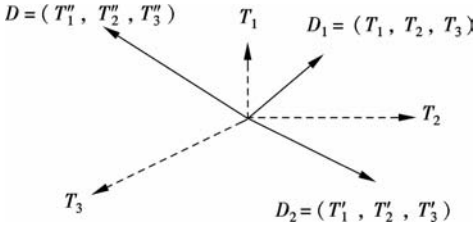

Fig. 3 Decomposition matrix in three main specific forms

Language itself detection tries to accomplish the task of sorting out vocabulary, sentence structure and grammar. Take the sentence structure for example, the sentence "Then I offered to wager a hundred dollars that he could get an answer by return mail" can be matched in grammar tree (see Fig. 4) for a perfect illustration.

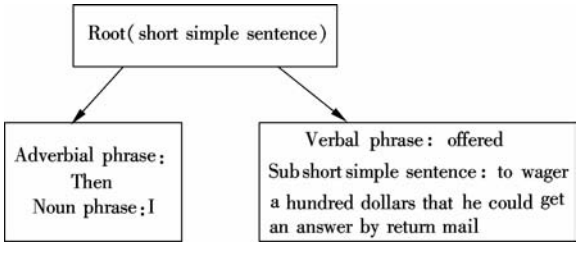

Fig. 4 Stanford Parser in the sentence matching

3.3 Experiments in essay marking

The essay marking experiment started with a composition writing task in iB-CET 2010, entitled "cash incentive program". Five English teachers were involved in manpower marking in different professional titles: one professor, two associate professors and two lecturers. 300 hundred students in classes were asked to write an essay in argumentation for the experiment. The composition task was designed as follows:

1) Write a summary of student's responses to the cash incentive programs and teachers' concerns about their effects.

2) Comment on the cash incentive programs and state your views on the best way to motivate students at school.

The five teachers were asked to mark the same essay written by the same student. After that, a collection of the marks were gathered and waited for the experiment. We designed software to see if the marks given by teachers can meet an average so as to prove the credibility of the software.

The software was designed into the following essay marking system.

1) Platform: Intel® Core (TM)2, CPU6320, 1.86 GHz with 2.5 GB Memory;

2) Operational system: Windows XP professional;

3) Tool: My Eclipse 6.5;

4) Programming language: Java;

5) Open source software: Stanford Parser;

6) Corpus: Spoken and Written English Corpus of Chinese Learners(2.0)[14].

3.4 Results in machine essay marking experiment

Based on the corpus of "Spoken and Written English Corpus of Chinese students", and the essay topic "cash incentive program", we put about 300 essays into the self-designed system, and arrives at the following result in content features abstraction marking (see Fig. 5) and language itself marking.

Then we made a comparison between automatic machine

Fig. 5 Content features abstraction marking illustration (SLA model)

marking and teacher man-power marking, resulting in a surprising finding: machine marking is close to manpower marking with a narrow closure to average marks in 10 points (see Fig. 6).

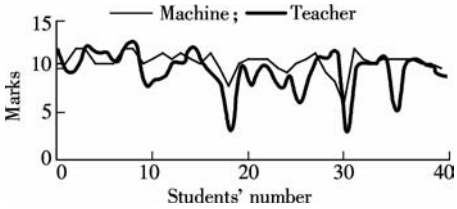

Fig. 6 Comparison between machine and teachers in essay marking

In conclusion, AEM can perform well in marking in textual documents.

3.5 Hypothesis for iB-CET oral test with AEM

Since spoken documents in audio can be transferred into textual documents in writing, it makes people imagine that the oral test can be tested in traditional ways with dialogues, description, comments and answering questions. In this case, the iB-CET can add more traditional items to be tested, such as translation and interpretation. Only in this way can reliability and validity be returned to the genuine language proper.

4 Conclusion

In this paper, we discuss the reliability and validity of oral test, leading to a dispute on the narrow validity of the iB-CET oral test based on imitation, which results in the hypothesis of converting spoken documents in audio into textual documents in writing, thus moving into automatic essay marking, establishing an imaginative in practice system that the iB-CET oral test can be tested in dialogues, description and comments, and answering questions.

When we say "imaginative in practice", we mean that it only gets started in the lab, but not in the iB-CET oral test, because nobody would like to risk the responsibility of failure in large-scaled tests (18 million examinees annually). In another sense, it can work well in practice, since there are compulsory requirements from the authorities in Chinese universities to pass the oral test in the course of experiencing English.

The feasibility-proof is very important, not only for the iB-CET oral test, but also for the realization of oral interpretation in the future. The audio sample collection can only be obtained in such a large-scaled test without payment in a genuine and wholesome way. Only when the lab or certain university oral tests are justified in scientific papers with approvals from engineering scholars and linguistic experts, can the iB-CET be really recommended for oral testing in dialogues, description and comments, and answering questions. And the day will come when serious attention is paid to this issue.

References

[1] Yang Huizhong. The 15 years of CET and its impact on teaching [J]. *Journal of Foreign Languages,* 2003, **145** (3): 21 – 29. (in Chinese)

[2] Liu Jiangang, Dong Jing. A study on language identification in call English oral test[J]. *Chinese Journal of Electron Devices,* 2011, **34**(4): 482 – 484. (in Chinese)

[3] Bachman L, Paulmer A. *In language testing in practice* [M]. Oxford: Oxford University Press, 1996: 25 – 26.

[4] Yang Jingpeiyi. A critical review of college English spoken test (CET-SET)[J]. *Journal of Kunming University,* 2010, **32**(1): 135 – 137. (in Chinese)

[5] Gong Li, Liang Weiqian, Ding Yuguo. Feasibility study and practice of machine scoring of repetition question in large-scaled English oral test[J]. *Computer-Assisted Foreign Language Education,* 2009(02): 10 – 15. (in Chinese)

[6] Chen Xiaoya, Li Aijun, Sun Guohua. Phonetic research orientating speech technology[C]//*The 6th National Conference on Man-Machine Speech Communication* (NCMMSC6). Shenzheng, China, 2001: 65 – 72.

[7] Favre B, Grishman R, Hillard D, et al. Punctuating Speech for Information Extraction[C]//*IEEE International Conference on Acoustic, Speech and Signal Processing*. Las Vegas, NV, USA, 2008: 5013 – 5016.

[8] Matsoukas S, Bulyko I, Xiang B, et al. Integrating speech recognition and machine translation [C]//*IEEE International Conference on Acoustic, Speech and Signal Processing*. Honolulu, HI, USA, 2007: 1281 – 1284.

[9] Fang Hongfeng, Feng Jiali, Wei Mengyun, et al. English pronunciation transform English character software actualize[J]. *Journal of Haibin Engineering University*, 2006, **27**(Sup): 584 – 586. (in Chinese)

[10] Li li, Yu Yibao. Voice conversion using spectrum with super-segment prosody features [J]. *Signal Processing,* 2012, **28**(2): 289 – 294. (in Chinese)

[11] Ding Wangdao. *A handbook of writing—revised English edition* [M]. Beijing: Foreign Languages Learning and Research Publisher, 1994: 151 – 153.

[12] Jiang Hao, Huang Guoqiang, Liu Jiangang. The research on CET automated essay scoring based on data mining [C]//*International Conference on Advances in Computer and Education Applications*(CSE 2011). Qiangdao, China, 2011: 100 – 105.

[13] Tao Huang. The research on CET automated essay scoring [D]. Nanjing: School of Computer Science and Engineering, Southeast University, 2011. (in Chinese)

[14] Wen Qiufang. *Spoken and written English corpus of Chinese learners* [M]. Beijing: Foreign Languages Learning and Research Press, 2009. (in Chinese)

基于作文自动评分的大学英语口语评估系统可行性

刘健刚[1] 郑玉琪[1] 陈美华[1] 姜 浩[2] 赵 力[3]

([1] 东南大学外国语学院, 南京 210096)
([2] 东南大学计算机科学与工程学院, 南京 210096)
([3] 东南大学信息科学与工程学院, 南京 210096)

摘要:论述了作文自动评分系统平台上大学英语口语测试的实现途径. 英语口语测试的可信度和有效度向中国当今的大学英语口语机考提出了挑战,涉及机考英语口语测试是否需要还原传统的对话、叙述或评述、问答问题三大评估内容. 首先简述了口头言语转化文字言语实现概率(72%)的现况,并以此为自动评分的前提设想. 其次,设想了大学英语口语测试的实现途径,即在口头言语转化为文字后,应用作文自动平分对话语进行口语测试系统的2种建模:言语内容检测模型和语言本身检测模型. 通过实验,论证了运用作文自动评分系统平台进行大学英语口语测试的可实现性,即实现了首先借助语音识别技术将口头言语转化为文字言语,然后进行文字言语评估的自动评分. 这种基于作文自动评分的大学英语口语测试系统建模研究和试验揭示了我国大学英语口语机考的可信度和有效度.

关键词:可信度和有效度;口头言语记录和文字言语记录;作文自动评分
中图分类号:H319.3