Journal of Artificial Intelligence 5 (1): 56-63, 2012
ISSN 1994-5450 / DOI: 10.3923/jai.2012.56.63

Automated Scoring Research over 40 Years: Looking Back and Ahead

[1,2]Jinlin Jiang and [3]Wei Wei

[1]Research Center for Business Engligh and Cross-cultural Studies, University of International Business and Economics, Bejing 100029, China

[2]Research Centre of Applied Linguistics, School of International Studies, University of International Business and Economics, Beijing 100029, China

[3]School of Computer and Engineering, Xi'an University of Technology, Xi'an 710048, China

Corresponding Author: Jinlin Jiang, Room 1316, Chengxin Building, School of International Studies, University of International Business and Economics, No. 10, Huixin Dongjie, Chaoyang District, Beijing 100029, China

ABSTRACT

This study reviews the developments of automated scoring systems and techniques over the past 40 years and the strengths and weaknesses of each technique are analyzed in detail. Results of the study showed that automated scoring systems are becoming more complicated and advanced by the integrated use of knowledge in natural language processing, statistics, information retrieval, corpus linguistics, etc. but they still have drawbacks in one or two aspects. The study provides some valuable insights into further studies in this line.

Key words: Automated scoring, essay scoring, translation scoring

INTRODUCTION

Performance assessment is an effective way of evaluating language skills and has been widely used in various types of English examination. In the field of language testing, automated scoring of performance assessment has been a focus of attention, among which automated scoring of English writing has been the most studied area. Since the 1960s a number of automated essay scoring systems have been developed and applied to GRE, GMAT and other large-scale tests as well as classroom teaching of writing (Landauer *et al*., 2001; Shermis and Burstein, 2003; Dikli, 2006). In China, an automated scoring system of Chinese EFL learners' English writing has been developed and yielded good results. In addition to these developments in essay scoring, automated scoring of machine translation has been prosperous and scoring of human translation has made the first rudimentary steps, but further studies need to be carried out. This study will review the characteristics of the existing automated scoring systems and provides some implications for future studies.

AUTOMATED ESSAY SCORING SYSTEMS

Automated essay scoring refers to the evaluation or rating of writing based on computer technology (Shermis and Barrera, 2002; Shermis and Burstein, 2003). The first automated essay scoring system PEG was developed in 1966 (Page, 2003). Since the 1990s, IEA, E-rater, IntelliMetric, MY Access and other essay scoring systems have emerged one after another (Landauer and Dumais, 1997; Burstein, 2003). A scoring system was developed in China as well

Table 1: Main characteristics of automated essay scoring system

	PEG	IEA	E-rater	Liang's system
Measurement object	Language	Content	Language; content; organization	Language; content; organization
Scoring method	Variable extraction; multiple regression; score computation	Essay scoring based on text similarity	Variable extraction; multiple regression; score computation	Variable extraction; score multiple regression; computation
Main techniques	Statistical techniques; natural language processing	Information retrieval (latent semantic analysis)	Statistical techniques; natural language processing; information retrieval (vector space model)	Statistical techniques; natural language processing; information retrieval (latent semantic analysis)
Major variables	Superficial features of language, such as text length and the number of prepositions	Semantic similarity	Syntactic structure; cohesive devices; semantic similarity;	Variables indicating fluencey, idiomaticity and complexity; sementic similarity; cohesive devices
Validation method	Correlation between machine /human scoring	Correlation between machine/human scoring	Correlation, exact agreement and adjacent agreement between machine /human scoring	Correlation, exact agreement and adjacent agreement between machine/human scoring

From table 1 we can find that the four systems have five major characteristics

to rate Chinese EFL learners' essay. Automated scoring was even extended to medicine, architecture, art, computer science and other subjects; scoring objects covered short essay, drawing and other subjective questions (Bejar, 1991; Clauser *et al*., 2000; Mislevy *et al*., 1999). The researcher only summarizes four major automated essay scoring systems: PEG, IEA, E-rater and Liang's system. Their main characteristics are shown in Table 1.

Measurement object: Table 1 shows that the measurement object of automated essay scoring systems expands from the initial measurement of language to three aspects: language, content and organization. Liang in 2005 study, the content module examines whether the composition focuses on the theme, the language module measures the accuracy of linguistic form and the structure module judges whether the essay is a complete and independent piece of writing. These three modules can directly traces back to the constructs of writing ability, thus has good validity and is more in line with the evaluation criteria of writing assessment (Chung and Baker, 2003).

Scoring method: There are three steps in PEG, E-rater and Liang in 2005, system to complete essay scoring: variable extraction, multiple regression and score computation. First, a series of text features are extracted from a number of essays which have been previously scored by human raters. These essays should cover students with different levels of writing proficiency so that they can provide enough information for automated scoring. After that, multiple linear regression analysis is conducted taking these features as independent variables and human-assigned scores as dependent variable. In this way the regression equation that can best predict human scoring is produced. Finally, the text variables of new compositions are put into the equation to obtain the machine-assigned scores to them. These three steps provide important hints for future automated scoring.

Main techniques: Four scoring systems have adopted a variety of techniques to extract variables and conduct relevant analysis. Among them, IEA and Liang in 2005, system used Latent Semantic

Analysis (LSA) of information retrieval. The basic assumption is that there exists a hidden semantic space in each text which is the accumulation of all words' meaning. Since in each language there are a large number of synonyms and lexicons with polysemy, a lot of noises emerge in the semantic space. It usually takes three steps to compress the semantic space: filtering, choosing and extracting features. First, use a stoplist to filter the lexicons with every little information. Then, select a number of texts related to the topic (such as expert writing, subject knowledge materials) to build a word frequency matrix and give different weights to each word in accordance with its frequency. The more a word appears, the smaller amount of information it contains and thus the lower its weight is. After that, use Singular Value Decomposition (SVD) technique to reduce the dimensionality of the matrix. This technique is similar to principal component analysis. The compressed matrix retains the important information of the original matrix and eliminates interference information, so it becomes a typical representative of the latent semantic space of the subject essay (Landauer and Dumais, 1997; Landauer *et al*., 1998). Latent semantic analysis has the advantage of extracting semantic content and it can even handle creative narrative writing (Landauer *et al*., 2001). However, it ignores word order, syntax, logic and other information and cannot reflect all the knowledge of a student (Landauer *et al*., 2001). Therefore, it would be better if it can be used in combination with variables that reflect the linguistic form of essay.

Similar to IEA and Liang's system, E-rater uses information retrieval technology, but it adopts Vector Space Model (VSM) of it to determine the relevance of text content (Burstein, 2003). Vector Space Model is based on keyword analysis technique, so it cannot achieve dimension reduction, elimination of noises and other effects of latent semantic analysis.

Major variables: Major variables used in each system correspond to the object it measures. For example, Liang's system used variables of linguistic form to measure the fluency, idiomaticity and complexity of language, used semantic similarity to evaluate quality of content and used such features as cohesive devices to assess the quality of essay structure.

Validation method: The above systems mainly test the correlation and consistency of machine- and human-assigned scores to validate their proximity. Correlation reflects the similarity of sorting made by machine and human raters (Chung and Baker, 2003). The existing systems analyzed not only the correlation between machine and a single rater's scoring, but also the correlation between machine scoring and the average score of multiple human raters. The first type of correlation is not necessarily reliable, as there may be bias within a rater's scoring and internal consistency cannot be guaranteed (Yang *et al*., 2002). The second type of correlation is more valuable, as the average score assigned to the same student by many raters is close to its true score (Page, 2003).

Consistency reflects the degree of accordance between specific ratings, including percentage of exact agreement and adjacent agreement (Chung and Baker, 2003). The former refers to the proportion of texts that are assigned the same rating by machine and human raters to the total number of texts. The latter refers to the proportion of texts that are assigned ratings next to each other by machine and human raters to the total number of texts. They both have their own strengths. When the score results are discrete data and there are a very small number of rating grades, the percentage of exact agreement is usually adopted; when the number of rating scale is larger, the percentage of adjacent agreement is more suitable (Yang *et al*., 2002). E-rater and Liang's system analyzed both two types of statistics.

当分数跨度很大时，邻近一致性被使用，否则用精确一致性。

Research results to date have indicated that regardless of examinee age and essay topic, computer assigned scoring has a correlation of 0.7 to 0.9 with human assigned scoring, mostly from 0.8 to 0.85 which provides strong evidence that computer can replace one human rater when scoring essay (Burstein and Chodorow, 1999; Landauer *et al*., 2001; Nichols, 2004; Page, 2003).

AUTOMATED TRANSLATION SCORING SYSTEMS

There are two types of automated translation scoring systems, one is adopted to score machine translation and the other is to score human translation. In the following part, these two types of systems will be discussed, respectively.

Automated scoring systems of machine translation: There are mainly two ways to evaluate machine translation.

Ngram-based evaluation: The underlying idea of Ngram-based evaluation is that compared with bad translation, high-quality machine translation should have more identical words or language fragments with human translation. BLEU (Bilingual Evaluation Understudy) and NIST (National Institute of Standards and Technology) are the main representatives of this method. BLEU examines the quality of machine translation by analyzing its similarity with a set of reference translations, or the proportion of identical Ngrams in machine translation with reference translations to the total number of Ngrams in machine translation. As it's quite difficult to find a match for a fivegram of human translation in machine translation, five-grams make little contribution to the similarity of machine and human translation. So the average proportion of unigram to four-gram matching is used. If a machine translation is shorter than the reference translation which bears the highest similarity with it, the similarity result should be multiplied by Brevity Penalty (BP) in order to get a certain punishment (Papineni *et al*., 2001). On the basis of BLEU, NIST give different weights to Ngrams according to their frequency in the reference translations. The lower the frequency, the more information it contains and the greater its weight is (Brew and Thompson, 1994; NIST, 2002). The techniques BLEU and NIST used are easy to operate and the machine scoring based on them is highly correlated with human rating (Brew and Thompson, 1994; NIST, 2002; Papineni *et al*., 2001).

Test point-based evaluation: The fundamental ideas of test point-based evaluation are the following: first, simulate the human rating method in standardized tests. In large-scale tests, raters usually do not evaluate the whole sentence or passage; the test target is simplified into test points. Test points can be divided into six groups: vocabulary test, test of fixed phrases, lexical test as well as test of preliminary, intermediate and advanced syntax. Second, describe the test point in each sentence with the use of Test Description Language. In this way the evaluation can be fully automatic. Procedures evaluate the quality of machine translation based on each test point and the weighted average grade will be the final result of machine translation evaluation. Third, build a large-scale test set of over 9,000 sentences and evaluate the translation quality of the entire sentence by pooling a large number of isolated test points. The advantage of this method is self-evident: some language points in the source text can highly differentiate students' translation ability while other points cannot, so to evaluate the whole piece of translation based on the quality of these test points can shorten evaluation time.

Automated scoring systems of human translation: In addition to automated scoring systems of machine translation, automated scoring systems of human translation has taken the first steps as well. The existing automated scoring systems and constructed a computer-assisted scoring system of Chinese EFL learners' Chinese-to-English translation. This study will be reviewed from six aspects: translation source, model design, human scoring, variable mining, model building and model validation.

Translation source: The study used 300 Chinese-to-English translated texts of third and fourth grade students from three universities in China. The source text is a narrative writing Letters from Home of Qin Shutong published in the supplement of Yangzi Evening News. It includes 3 paragraphs and 9 sentences in total. The translation was required to be finished in 60 min. To meet the research needs, the source text was first presented to the students as a whole for them to get a thorough understanding and then 9 sentences were shown to them one by one for translation. In this way the researchers can collect the translation of each sentence and the complete translation of the text as well by combining them together. All the translated texts are a part of Parallel Corpus of Chinese EFL Learners (Wen and Wang, 2008).

Model design: Computer-assisted scoring system of Chinese EFL learners' Chinese-to-English translation was divided into diagnostic and selective scoring models. Diagnostic model consists of modules that can be used to evaluate the form and meaning of both text translation and sentence translation. By extracting the features corresponding to each module, scoring models for each module can be built, respectively and they can provide targeted diagnostic information to learners. Selective models can only evaluate the semantic quality of text translation in large-scale tests.

Human scoring: There were three raters in this study, two male and a female, two of whom were associate professors and one was a lecturer. They were all Ph.D candidates in Applied Linguistics and had some university teaching experience and large-scale test scoring experience. There were two rounds of human scoring in this study with a one-year interval in between. In the first round of scoring, very detailed evaluation was made for the semantic and language quality, each following the standard of faithfulness and expressiveness, respectively. The scores were used to construct scoring models serving diagnostic purpose. For semantic scoring, each source sentence was divided into two or three semantic units and each unit was assigned a full score of 5 points. For form scoring, each sentence has a full score of 10 points. After scoring, the total score was calculated as 60% of semantic score plus 40% of the form score.

The second round of human scoring was more simplified. Only part of the semantic point needs to be scored. The scores were used to construct selective model.

Variable mining: The study used corpus linguistics tools, natural language processing, information retrieval techniques and statistical methods to extract a number of text features. Table 2 summarized some important variables.

Table 2 shows that the study extracted three types of variables that can reflect the semantic quality of translation: Ngram, semantic similarity and number of aligned semantic points. Among them, semantic point alignment technique matched a list of correct translations of highly discriminating language points in student translation which is similar to the evaluation method. The study also extracted three levels of variables relevant to linguistic form: lexical level, sentence level and text level.

Table 2: Sample variables in computer-assisted scoring system of Chinese EFL learners' Chinese-to-English translation

Category	Variables	Extraction method
Content	Matched unigrams, bigrams, trigrams	Match unigrams, bigrams and trigrams from the set of 25 best translations in student translation
	SVD	Similarity between student translation and 25 best translations using latent semantic analysis
	number of aligned semantic points	Match a list of correct translations of semantic points in student translation
Form		
Lexical level	Type, token and type/token ratio	Absolute value of type, token and type/token ratio minus the corresponding average value of 25 best translations
	Word length	Number of letters in each word
	Percentage of noun, verb, adjective, etc.,	Number of noun, verb, adjective, etc. divided by token
Sentence level	Number of sentences, average sentence length	Absolute value of sentences, average sentence length minus the corresponding average value of 25 best translations
Text level	Transitional lexicons	Match a list of self-made transitional lexicons in student translation

Model building: In the study, first, human raters scored half of the translated texts, that is, the training set. Then the researcher extracted variables from the training set and calculated the correlation between these variables and the corresponding human scoring. The variables that significantly correlated with human scoring would become predictors of the translation quality. After that, multiple linear regression analysis was conducted, in which the predictors were independent variables and human scoring was the dependent variable. Lastly, the model with the best performance would be chosen which was in essence an equation indicating the relationship between human scoring and effective predictors. Data showed that the correlation coefficient R in diagnostic semantic and form scoring model were 0.891 and 0.740, respectively and coefficient of determination R^2 were correspondingly 0.794 and 0.547. The study further constructed selective scoring models based on four types of training set (30, 50, 100 and 150 translated texts, respectively). Results indicated that the correlation coefficient R in four types of models were all above 0.8.

Model validation: The study used the regression equation to compute the scores of another half of translated texts. These translated texts were scored by human raters as well. Then the correlation coefficients and consistency between machine scoring and human scoring were analyzed. The results showed that the correlation between machine and human scoring of semantic quality for text translation was 0.842 and correlation for form quality was 0.741. For selective purpose, the scoring model built on the training set of 100 translated texts had the best performance, with the computed scoring bearing a correlation of over 0.8 with human scoring. Therefore, 100 translated texts meet the needs of automated training in large-scale test.

In short, the study constructed a system to evaluate Chinese students' Chinese-to-English translation accurately and efficiently. This research is important in several aspects: first, it focused on Chinese learners' Chinese-to-English writing which have their own characteristics and need to be dealt with accordingly, such as human scoring based on semantic points. Second, it explored diagnostic and selective scoring models and satisfied different purposes. However, the study has some weaknesses as well. First, the study only used 300 translated texts of a narrative writing to build scoring models, while different style of texts and translation of them may have significant

differences in content, language, etc., so it is difficult to determine whether the quality predictors can be effective in other text types. Second, the study used a hold-out method to build and validate models which means the training set was only used for modeling and the validation set was only adopted to test models, so the results may be different if they switch their roles.

CONCLUSIONS

This study analyzed the existing automated scoring systems and techniques and pointed out their strong points and drawbacks. From the paper it can be concluded that automated essay scoring has been quite mature. The existing systems conform to the construct of writing ability and exhibit satisfying performance in high-risk large-scale tests. In addition, with the integration of interdisciplinary knowledge and technologies, automated scoring of translation has made preliminary progress, especially the scoring of Chinese EFL learners' Chinese-to-English translation. Future study can make more improvements by constructing models for different text types, different topics and validating computer scoring with comprehensive validation evidence.

ACKNOWLEDGMENTS

This study is supported by National Social Science Found Project of China, Construction of Computer-Assisted Scoring System for English-Chinese and Chinese-English Translations of Specific-Purposed Texts (No. 11CYY007). 中国社会科学基金项目支持的研究成果
特定文本英汉互译结果的自动评分系统建设

REFERENCES

Bejar, I.I., 1991. A methodology for scoring open-ended architectural design problems. J. Applied Psychol., 76: 522-532.

Brew, C. and H. S. Thompson, 1994. Automated evaluation of computer generated text: A progress report on the text eval project. Proceedings of the Human Language Technology Workshop, March 8-11, 1994, Plainsboro, New Jerey, USA., pp: 108-113.

Burstein, J. and M. Chodorow, 1999. Automated essay scoring for nonnative English speakers. Proceedings Joint Symposium of the Association of Computational Linguistics and the International Association of Language Learning Technologies, June 1999, College Park, Maryland, pp: 68-75.

Burstein, J., 2003. The E-rater Scoring Engine: Automated Essay Scoring with Natural Language Processing. In: Automated Essay Scoring: A Cross-Disciplinary Perspective, Shermis, M.D. and J. Burstein (Eds.). Lawrence Erlbaum Associates, NJ., pp: 113-121.

Chung, G. K. and E.L. Baker, 2003. Issues in the Reliability and Validity of Automated Scoring of Constructed Responses. In: Automated Essay Scoring: A Cross-Disciplinary Perspective, Shermis, M.D. and J. Burstein (Eds.). Lawrence Erlbaum Associates, NJ., pp: 23-40.

Clauser, B.E., P. Harik and S.G. Clyman, 2000. The generalizability of scores for a performance assessment scored with a computer-automated scoring system. J. Educ. Measure., 37: 245-262.

Dikli, S., 2006. An overview of automated scoring of essays. J. Technol. Learn. Assess., 5: 3-35.

Landauer, T.K. and S.T. Dumais, 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychol. Rev., 104: 211-240.

Landauer, T.K., D. Laham and P.W. Foltz, 2001. The intelligent essay assessor: Putting knowledge to the test. Proceedings of the Association of Test Publishers Conference on Computer-Based Testing: Emerging Technologies and Opportunities for Diverse Applications, February 26-28, 2001, Tucson, AZ., USA.

Landauer, T.K., P.W. Foltz and D. Laham, 1998. An introduction to latent semantic analysis. Discourse Process., 25: 259-284.

Mislevy, R.J., L.S. Steinberg, F.J. Breyer, R.G. Almond and L. Johnson, 1999. Making sense of data from complex assessments. Applied Measure. Educ., 15: 363-389.

NIST, 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. http://www.itl.nist.gov/iad/894.01/tests/mt/doc/ngram-study.pdf.

Nichols, P.D., 2004. Evidence for the interpretation and use of scores from an automated essay scorer. Proceedings of the Annual Meeting of the American Educational Research Association (AERA), April 12-16, 2004, San Diego, CA.

Page, E.B., 2003. Project Essay Grade: PEG. In: Automated Essay Scoring: A Cross-Disciplinary Perspective, Shermis, M.D. and J. Burstein (Eds.). Lawrence Erlbaum Associates, NJ., pp: 43-54.

Papineni, K., S. Roukos, T. Ward and W. Zhu, 2001. BLEU: A method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center.

Shermis, M.D. and F.D. Barrera, 2002. Exit assessments: Evaluating writing ability through automated essay scoring. Proceedings of the Annual Meetings of the American + Educational Research Association, April 1-5, 2002, New Orleans, LA., pp: 1-30.

Shermis, M.D. and J. Burstein, 2003. Automated Essay Scoring: A Cross-Disciplinary Perspective. Lawrence Erlbaum Associates, NJ.

Wen, Q.F. and J.Q. Wang, 2008. Parallel Corpus of Chinese EFL Learners. Foreign Language Teaching and Research Press, Beijing.

Yang, Y., C.W. Buckendahl, P.J. Juszkiewicz and D.S. Bhola, 2002. A review of strategies for validating computer-automated scoring. Applied Measure. Educ., 15: 391-412.