Constructing Automated Scoring Model for Human Translation with Multidisciplinary Technologies

http://dx.doi.org/10.3991/ijet.v11i02.5137

Jinlin Jiang[1], Ying Qin[2] and Ya Sun[1]
[1] University of International Business and Economics, Beijing, China
[2] Beijing Foreign Studies University, Beijing, China

***Abstract*—This study developed a computer scoring model for Chinese EFL learners' English-to-Chinese translations using multidisciplinary techniques in corpus linguistics, natural language processing, information retrieval and statistics. The proposed model, once implemented as computer software, can score English-to-Chinese translations in large-scale examinations. This study built five scoring models with 50, 100, 130, 150 and 180 translations as the training set for 300 translations of an expository writing. The correlation coefficients between the computed scores of these models and human-assigned scores were above 0.8. The results further indicated that the computed scores with 130 training translations were closest to human-assigned scores. Therefore, it was concluded that the finalized model can produce reliable scores for Chinese EFL learners' English-to-Chinese expository translations.**

***Index Terms*—Automated scoring; English-to-Chinese translation; Multidisciplinary technologies; Text features**

I. INTRODUCTION

Since the 1960s several automated essay scoring systems have been developed and applied to GRE, GMAT and other large-scale examinations [1, 2, 3]. In China, Liang [4] developed an automated scoring system for Chinese EFL learners' English compositions and achieved good results. A few researchers studied the automatic scoring of Chinese writing as well and found that computer-calculated scores by the use of latent semantic analysis (LSA) were close to human-rated scores [5]. Some scholars also explored the automated scoring of short questions [6] and speech [7, 8, 9, 10] with multiple measures.

Compared with the maturity of automated essay scoring and the prosperity of automated scoring of short questions and speech, automated translation scoring is still confined to machine translation evaluation [11, 12, 13]. Only in recent years has automated scoring of human translation taken the first steps. Wang [14] constructed a computer-assisted scoring system for Chinese-to-English (C-E) translations of Chinese EFL learners, which consisted of diagnostic and selective scoring models. The diagnostic model was composed of four types of modules that can evaluate the form and meaning of both text translation and sentence translation. They can also provide learners with useful information about each module. The selective model can evaluate the semantic quality of text translation in large-scale tests. However, the conclusions of this study remain some uncertainties. First, it only used 300 translated texts of a narration to build scoring models, while different types of texts have remarkable differences in content and language, so it is hard to determine whether the quality predictors will be effective for other text types. Second, the study used a hold-out method, by which the training translations were only used for modeling and the validation translations were only utilized to test the models, so the results may be different if they switch roles [15].

The research dealing with English-to-Chinese (E-C) translations was even scarcer, as Chinese language has no morphological variety, fewer explicit connectives and more flexible syntactic rules and the natural language processing technologies of Chinese are far behind those of English. Wang [16] used 10-fold cross-validation to verify the models she constructed, but the source text in this study was an advertising paragraph, including only three sentences and 76 words. In addition, the human scoring rubric that the computer simulated was sketchy, the text features she quantified were basically on lexical level and were extracted against only four reference translations.

Focusing on the above shortcomings, this paper aims to develop a stable, reliable automated scoring model suitable for Chinese EFL learners' E-C translations in large-scale tests. The basic approach is to use technologies in multiple areas such as corpus linguistics, natural language processing and information retrieval to extract a variety of text features related to the quality of translations. Then through multiple linear regression analysis of text features (the independent variables) and human scoring (the dependent variable), the tentative computer scoring models can be constructed and validated. Finally, the regression equation which has the strongest predicting power of human scoring can be used to compute the scores of other translations of the same source text, and the similarity between machine-computed scores and human-rated scores will be analyzed.

This study is different from the existing literature in following aspects: human scoring process, text features extracted and text type. First, human scoring of semantic quality of translations takes "Translation Unit" (TU) as a single unit. TU identified in the source text is a monosemous multi-word unit which has complete meaning and high discriminating power, and conforms to grammatical rules. It can better evaluate translation quality in terms of semantic faithfulness, grammaticality, coherence and idiomaticity [17, 18]. Second, human scoring of the linguistic quality of translations considers "style closeness" as a supplementary component to “grammatical correctness” and “fluency”, the two traditional scoring criteria used for translations. Since the target language—Chinese is students’ mother tongue, the linguistic quality of translations needs to adopt higher evaluation standards, that is, the style of translations should be close to that of the original text. Third, by using multidisciplinary technologies, this study extracts a batch of new text features, such as the number of aligned key points. Fourth, this study constructs computer scoring models for translations of an expository writing, the most frequently used text type in large-scale English tests of China, which has special implications. Different text types may have substantial differences in content, language, style, etc. The exposition used in this study has a clear text structure, accurate and rigorous wording, and complex sentence structures. The scoring model built for translations of this typical writing helps improve the application scope of text features.

II. RESEARCH DESIGN

A. *Research questions*

This study addresses the following questions:

(1) How much predicting power do the computer scoring models built based on different sizes of training sets have? How reliable are the predicted scores?

(2) How many translations does the training set should contain in order to meet the need of automated scoring in large-scale tests?

B. *Research tools*

This study uses a large number of text analysis and data analysis tools:

(1) Text pre-processing tools. Most of them are perl programs utilized to organize irregular input in the translations, to randomly number the translations, and to separate and integrate sentences.

(2) Text analysis tools. They are used to extract text features that are closely related to the semantic quality of translations, including R software and perl programs. R is statistical analysis software. In this study it is used to carry out LSA. Perl program is used first to extract the number of aligned n-grams, ranging from unigrams to fourgrams. In addition, it is used to realize key-point alignment. The reference of these text features is the best translation set composed of 30 expert translations. The closer a student translation to the best set, the higher the quality is.

(3) Data analysis tools. SPSS is used to calculate the correlation coefficients between text features and human scoring of translations. “Stepwise” multiple linear regression analysis is conducted as well to build and verify the scoring models.

C. *Research procedures*

This study can be divided into five stages: translation collection, human scoring, feature extraction, model construction and model validation. The first three steps are preparation before modeling. They will be introduced respectively in the following part.

(1) Translation collection. The source text of translations in this research is from “The significant Americans” written by Cuber, J. F. and Haroff, P. B. in 1966. It is an expository writing of 235 words, including 15 sentences. The study uses 300 E-C translations from Parallel Corpus of Chinese EFL Learners [19], which were originally translated by junior and senior students in three Chinese universities of different levels. In classroom, the whole text was shown to students first in order to facilitate their overall understanding; then each sentence was presented to them and they were asked to write down the corresponding translation below each sentence within 45 minutes. In this way both the translation of each sentence and the complete translation of the text can be collected by combining them together.

(2) Human scoring. In automated scoring studies, human scoring must be highly reliable to ensure the effectiveness of computer scoring. There are three raters in this study, two male and a female, two of whom are associate professors and one is a lecturer in three Chinese universities. They are all doctoral candidates majoring in applied linguistics and have large-scale test scoring experience.

There are 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 linguistic quality based on the criteria of faithfulness, expressiveness and closeness. For semantic scoring, each source sentence was divided into two or three translation units and each unit was assigned a full score of 5 points. The fidelity of the equivalent for each translation unit was evaluated compared with the correct and half-correct reference translations for each unit provided. For linguistic scoring, grammaticality, fluency and style closeness of each sentence translation was evaluated. The scoring process lasted about 240 hours. The full score of “form” and “meaning” for each sentence were 10 points respectively, and the total full score of “form” and “meaning” for the whole text were both 150 points.

Since the first round grading is time-consuming and not suited to large-scale examination as the lack of efficiency, this study carried out another round of scoring one year later. In this simplified scoring process, only highly discriminative points were semantically evaluated. Two experts of translation research in China were invited to decide on the key points of the source text. After discussion, 35 key points were determined, which made up one eighth of the lexicons in the whole text. The scoring took about 32 hours.

Table 1 shows that in detailed scoring, the mean correlation coefficient of three raters when evaluating the semantic quality is 0.891** and Cronbach’s alpha is 0.957; the mean correlation coefficient of three raters when evaluating the linguistic quality is 0.857** and alpha value is 0.946, demonstrating that the three raters have good consistency. In simplified scoring, the raters display satisfactory correlation coefficient and alpha coefficient as well.

TABLE I.
RATER RELIABILITY OF TEXT TRANSLATION SCORING[1]

	Module	Mean of correlation coefficients	Cronbach’s alpha
First round scoring	Semantic scoring	0.891**	0.957
	Linguistic scoring	0.857**	0.946
Second round scoring		0.944**	0.980

As the first round scoring gives an exhaustive evaluation of the semantic quality of each translation and the

[1] In this paper ** indicates that the correlation coefficient is significant at the 0.01 level (2-tailed).

second round scoring is greatly simplified, the effectiveness of the latter depends on its similarity with the first scoring. Statistical analysis shows that the correlation coefficient between the two rounds of average semantic scoring is as high as 0.924, so the key-point-based scoring method is very effective. In addition, human scoring is analyzed based on Many-facet Rasch Measurement Model. Research results indicate that the severity of raters has significant difference, which is not expected in a good exam [20]. This research takes a remedy of using the average score of three raters. In this way the severity difference will be reduced.

(3) Feature extraction. This study extracted the following semantic features:

a) The number of matched n-grams with reference to those in the best translation set. Since there is no space within Chinese words, the translations were first segmented word by word. Then word-based n-grams were extracted correspondingly in student translations and the best translations. A Chinese stoplist was used to filter out the words without specific meaning in the best unigrams. As it is difficult to find a match for the best fivegrams in student translations, the researchers extracted one-to-four grams. N-grams occurring more than twice in the best translation set formed a dictionary. The frequency of each n-gram in the dictionary was searched one by one in each student translation, and the number of matched one-to-four grams was used as a text feature. It should be noted that opposed to "word-based" approach, there exists "character-based" approach in Chinese language, according to which the translations were segmented character by character and the corresponding character-based n-grams were extracted. Experiments showed that word-based n-grams had better performance than character-based ones when predicting translation quality, so word-based n-grams were chosen at last. However, n-gram is a mechanic sequence of words which is not necessarily in line with grammatical rules. In some cases it's not a unit of meaning, so this feature does not fully consider the contextual factor.

b) The number of aligned words with reference to an E-C dictionary and the expanded version of Synonymy Thesaurus. During word alignment process, two key points of E-C equivalence need to be considered. First, the matching of one English word with one or more Chinese words that are separated by other words is not uncommon. For example, within "xiang…yi yang Kuai (像……一样快)", a translation equivalent of "as quickly as", one or more words can be inverted. In this study, such a phenomenon was taken into consideration and properly dealt with. Second, there are a lot of synonyms in Chinese language. For example, "in radiant bloom" can be translated into "shengkai (盛开)", "kaihua (开花)", "kaifang (开放)", "zhanfang (绽放)", "nufang (怒放)", and so on. In this study, not only were the translation equivalents of an English word in the E-C dictionary counted, but other synonyms of the correct translation equivalents were taken into consideration. Word alignment took the following steps: first, dictionary-based word alignment was conducted using fuzzy matching method; second, the expanded version of Synonymy Thesaurus was adopted for further word alignment; third, the matching of one English word with more than one Chinese words, and more than one English words with one Chinese word were carried out. Through these procedures the multiple alignment situations between English and Chinese language were considered comprehensively.

c) The number of aligned key points with reference to an E-C dictionary of correct translations of key points. This is an imitation of the human scoring practice in large-scale tests of China. Highly discriminating language points in the source text are chosen and only the translation equivalents of these points are scored by human raters. In order to save time and draw machine scoring closer to human scoring, the alignment of key point translation was performed in this study. The number of aligned key points with reference to a list of correct translations was counted as a text feature. This variable can measure such cases as leakage of translation and mistranslation. It was expected to have high predicting power of translation quality.

d) Semantic similarity between student translation and the best translation set through LSA. The basic assumption of LSA 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 very little information. Then, select a number of texts related to the topic (in this study, the best translation set) 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 interfering information, so it becomes a typical representative of the latent semantic space of the subject text [21]. LSA has the advantage of extracting semantic content and can even handle creative narrative writing [22].

The above variables have their own strengths and weaknesses, and the ones with a significant correlation coefficient with the human scoring will become a quality predictor of translations and enter the model building phase.

III. RESULTS AND DISCUSSION

In the model building phase, the researchers first randomly chose certain translations as the training set and the remaining translations as the validation set. In this study, 50, 100, 130, 150 and 180 translations were chosen respectively in order to determine the appropriate size of the training set. Then the researchers calculated the correlation coefficients between the extracted variables and simplified 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 the simplified 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.

There are three standards used to evaluate the tentative models: First, in order to avoid collinearity and ensure the reliability of models, the independent variables whose correlation coefficients were higher than 0.8 were prevented from entering regression analysis simultaneously. Collinearity refers to the phenomenon that two independent variables in the regression equation overlap a lot or the variance an independent variable explains can be basically covered by a number of other independent variables [23]. Second, the coefficient's direction of an independent variable (negative or positive) should be the same as the direction of correlation coefficient between this variable and the dependent variable. Otherwise this variable is a negative suppressor and is usually associated with collinearity problem [23]. Third, the model with the largest coefficient of determination (R Square) and the most reasonable collinearity data was selected. Frequently used collinearity statistical standards include tolerance, variance inflation factor and condition index [24].

After several rounds of model diagnostic analysis and comparison, the models finally decided on were shown in table II. All the variables in these models were effective and the collinearity statistics were within acceptable range. Limited by paper length, specific data were not reported here.

Table II shows that the correlation coefficients of five models constructed with different sizes of training set are above 0.8, indicating that the variables in the models can explain a large proportion of translation quality. When there are 50 translations in the training set, the model has the highest correlation coefficient. With the gradual increase of the training set, the correlation coefficient displays an overall downward trend. However, the smaller the training set translation is, the greater the impact of specific translations is and the more unstable the model is, and thus it cannot be concluded that 50 translations meet the need of large-scale scoring. Since fitness statistics are not enough for determining the appropriate size of the training set, further evidence from the scoring performance of the models is needed for a sound judgment.

This research used the validation set to testify the efficacy of the above models. First, the equations in table II and the text features were used to calculate the scores of meaning for the validation set. Then the reliability of these scores was computed compared with human scoring. The results were shown in tables III.

Table III shows that the correlation coefficients and the alpha values between the predicted scores of different models and human-assigned scores are above 0.8, which indicates that the models can effectively predict the quality of the validation set. These results are better than those of the existing automated scoring studies of oral performance, in which the correlations between machine and human scorings are between 0.5-0.7. [8]. The results are also better than the essay scoring model for Chinese students' English writing, in which the correlations between machine and human scorings of which are between 0.65-0.74. [4]. But the performance of the current models is not as good as that of the models built for Chinese students' C-E translations [14]. In that study, when there are 50, 100 and 150 translations in the training set, the correlations between machine and human scorings are 0.870, 0.878 and 0.897 respectively, about 0.03 higher than the results in table III. As the object of this study is Chinese translation and Chinese language is parataxis, it's not easy to obtain the results in table III. These results are also considerably better than the existing automated scoring study of students' E-C translations, in which the correlation between machine and human scorings is 0.75 [16].

TABLE II.
MODELS WITH DIFFERENT SIZES OF TRAINING SET

Training translations	R	R Square	Adjusted R Square
50	0.908	0.824	0.816
100	0.852	0.726	0.717
130	0.835	0.698	0.690
150	0.834	0.696	0.690
180	0.848	0.720	0.715

TABLE III.
RELIABILITY OF MACHINE SCORING

Training translations	Correlation between machine/human scorings	Cronbach's alpha of machine/human scorings
50	0.832**	0.909
100	0.837**	0.908
130	0.860**	0.917
150	0.862**	0.919
180	0.851**	0.910

TABLE IV.
PAIRED-SAMPLES T TESTS OF MACHINE/HUMAN SCORINGS

Training translations	Paired differences		t	Sig. (2-tailed)
	Mean	Std. Deviation		
50	1.085	4.293	4.076	0.000
100	1.288	4.045	4.614	0.000
130	0.471	3.873	1.631	0.105
150	0.434	3.886	1.412	0.160
180	0.184	4.077	0.514	0.608

Table III further shows that when there are 50, 130 and 150 translations in the training set, the correlation coefficients between machine scoring and human scoring gradually increases. Comparing with table II, it can be found that the variance the models explains for the training set is not consistent with the performance of the models when predicting the quality of the validation set. The smaller the training set is, the higher predicting power the models have for the training set, the less stable and effective they are when predicting the quality of the validation set. Therefore, it can be seen that the training translations need to reach a certain number to ensure the validity of the model. But still there is very small difference in the correlation coefficients produced by different models, so the final conclusion cannot be reached yet.

This study further compared machine scoring with human scoring with paired-samples t test. The results were in table IV.

According to table IV, when there are 50 and 100 translations in the training set, the mean differences between machine and human scorings of the validation set are 1.085 and 1.288 respectively, which are statistically significant. When the training set increases to 130 translations, the mean difference drops to 0.471 and has no statistical significance. As the training set becomes larger, the mean difference between machine and human scorings lowers. Therefore, it can be seen that 130 training translations can meet the need of machine scoring. However, this conclusion requires further investigation in large-scale translations. The finalized computer scoring model is as follows:

Semantic scores = -10.988 + 0.983 × the number of aligned key points + 0.098 × the number of aligned unigrams + 24.163 × semantic similarity

In this equation, the number of aligned key points is a very effective predictor. The standard coefficient of this feature is 0.549 (confined by paper length, detailed statistics are not reported), the largest among the three variables in the equation, indicating that it has the strongest predicting power of the semantic translation quality. Key points are a simplified representative of translation units. They are smaller language units and there are fewer of them in the source text. They grasp the essence of the source text and have distinctively more discriminating power than translation units. Their effectiveness verifies the researchers' assumption before the study.

In addition, the contribution of the number of aligned unigrams is second only to the number of aligned key points. The standard coefficient of this feature is 0.279, the second largest among the three variables. This proves the effectiveness of n-gram in automated scoring. N-gram has been used as a text quality predictor in machine translation evaluation. The BLEU (Bilingual Evaluation Understudy) and NIST (National Institute of Standards and Technology) are two representatives. BLEU examines the quality of machine translation by analyzing its similarity with a set of reference translations, or the proportion of identical n-grams in machine translation with reference translations to the total number of n-grams in machine translation. NIST give different weights to n-grams according to their frequency in the reference translations. The lower the frequency, the more information it contains and the greater its weight is. The machine scoring based on BLEU and NIST is highly correlated with human scoring [11, 12]. In this study, matched unigrams reflects the consistency of one Chinese character between student translation and the best translations, which is complementary to the highly discriminating multi-word-unit key points.

Semantic similarity is another predictor in the equation. With a standard coefficient of 0.146, it has certain predicting power of translation quality as well. This variable is the degree of closeness calculated by the use of LSA. Through filtering, choosing and extracting features, LSA can effectively screen out noises and compress the semantic before making the comparison. This technology has been adopted in the existing studies and plays an important role in automated essay scoring models [4, 22] and Chinese students' C-E scoring model [14, 25]. This study demonstrates its effectiveness when the scoring object shifts to Chinese translations.

IV. CONCLUSION

In this study, multidisciplinary knowledge and technologies are used to construct automated scoring model for Chinese students' E-C translations in large-scale tests. The results show that the scoring model performs well. First, the simplified human scoring saves four fifths of the time and the scores are highly correlated to and consistent with detailed human scoring, indicating that the human evaluation method based on key points are effective and feasible. Second, the computer scoring models built with 50, 100, 130, 150 and 180 translations as the training set have high predicting power of the translation quality. In addition, when there are 130 translations in the training set, the scores produced by the automated scoring model are the closest to human-assigned scores. Using this model will not only save cost of human rating but also meet the need of automatic scoring in large-scale tests.

This study also has some limitations. First, large-scale translations are needed to test whether 130 training translations can produce the same effect in other styles, topics and sizes of translations. Second, some text features need to be further improved. For example, the list of correct translation equivalents of key points is not exhaustive. Moreover, it's difficult for the automatic scoring model to judge creative translation. This study has made some efforts in this regard. The variables it extracts take 30 expert translations as a reference, but these translations cannot cover all creative ones, so human intervention is inevitable when there is a big difference between machine- and human-assigned scores.

REFERENCES

[1] S. Dikli, "An Overview of Automated Scoring of Essays", Journal of Technology, Learning, and Assessment, no. 1, pp. 3-35, 2006.

[2] E. S. Quellmalz, J. W. Pellegrino, "Technology and Testing", Science, no. 2, pp. 75-79, 2009. http://dx.doi.org/10.1126/science.1168046

[3] D. M. Williamson, "A Framework for Implementing Automated Scoring", Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, San Diego, USA, April 13-17, 2009.

[4] M. C. Liang, "Constructing a Model for the Computer-assisted Scoring of Chinese EFL Learners' Argumentative Essays", Foreign Language Teaching and Research Press, Beijing, China, 2011.

[5] Y. W. Cao, C. Yang, "Automated Chinese Essay Scoring with Latent Semantic Analysis", Examinations Research, no. 1, pp. 63-71, 2007.

[6] N. T. Carr, X. Xi, "Automated Scoring of Short-answer Reading Items: Implications for Constructs". Language Assessment Quarterly, no. 3, pp. 205-218, 2010. http://dx.doi.org/10.1080/15434300903443958

[7] J. Bernstein, J. Cheng, "Logic and Validation of a Fully Automatic Spoken English Test". In V. M. Holland, F. P. Fisher, (Eds.), The Path of Speech Technologies in Computer Assisted Language Learning: From Research Toward Practice. New York: Routledge, 2008.

[8] M. Chen, K. Zechner, "Computing and Evaluating Syntactic Complexity Features for Automated Scoring of Spontaneous Nonnative Speech". In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, pp. 722-731, 2011.

[9] X. M. Xi, et al., "Automated Scoring of Spontaneous Speech Using SpeechRater v1.0" (ETS Research Report No. RR-08-62). Princeton, NJ: Educational Testing Service, 2008.

[10] X. M. Xi, et al., A Comparison of Two Scoring Methods for An Automated Speech Scoring System. Language Testing, no. 3, pp. 371-394, 2012. http://dx.doi.org/10.1177/0265532211425673

[11] G. Doddington. Automated Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. In Proceedings of the Second International Conference on Human Language Technology, pp. 138-145. San Diego, CA, 2002. http://dx.doi.org/10.3115/1289189.1289273

[12] K. Papineni, et al., "Bleu: A Method for Automatic Evaluation of Machine Translation", In Proceedings of the 40th Annual Meeting

of the Association for Computational Linguistics, Philadelphia, PA, pp. 311-318, 2002.

[13] W. Wu, L. Li. Automated Chinese-English Translation Scoring Based on Answer Knowledge Base. In Proceedings of 12th IEEE International Conference on Cognitive Informatics & Cognitive Computing, pp. 341-346, New York, USA, 2013. http://dx.doi.org/10.1109/icci-cc.2013.6622264

[14] J. Q. Wang, "Computer-assisted Scoring Models of Chinese Learners' Chinese-English Translation: Construction and Research", Foreign Language Teaching and Research Press, Beijing, China, 2010.

[15] J. L. Jiang, W. Wei, "Automated Scoring Research over 40 Years: Looking Back and Ahead", Journal of Artificial Intelligence, no. 5, pp. 56-63, 2012.

[16] L. X. Wang, "Research on Automatic Quantification of Translation Criterion", Unpublished Ph.D dissertation, Shanghai International Studies University, Shanghai, China, 2007.

[17] J. L. Jiang, Q. F. Wen, "A Comparative Study of n-gram and Translation Unit Alignment in Automated Scoring of Students' English-Chinese Translation", Modern Foreign Languages, no. 2, pp. 177-184, 2010.

[18] W. Teubert, "The Role of Parallel Corpora in Translation and Multilingual Lexicography", In B. Altenberg, S. Granger, (Eds.), Lexis in Contrast: Corpus-Based Approaches, Benjamins, Amsterdam and Philadelphia, pp. 189-214, 2002. http://dx.doi.org/10.1075/scl.7.14teu

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

[20] J. L. Jiang, W. Wei, "Many-facet Rasch Model's Application in the Evaluation of Test Validity", International Journal of Digital Content Technology and Its Applications, no. 11, pp. 52-59, 2011.

[21] T. K. Landauer, P. W. Foltz, D. Laham, "Introduction to Latent Semantic Analysis", Discourse Processes, no. 2, pp. 259-284, 1998. http://dx.doi.org/10.1080/01638539809545028

[22] T. K. Landauer, D. Laham, P. W. Foltz, "Automated Essay Scoring and Annotation of Essays with the Intelligent Essay Assessor". In M. D. Shermis, J. C. Burstein, (Eds.), Automated Essay Scoring: A Cross-Disciplinary Perspective, Lawrence Erlbaum Associates, Mahwah, NJ, pp. 87-112, 2003.

[23] T. P. Ryan, Modern Regression Methods, John Wiley and Sons, New York, 2009.

[24] X. Q. Qin, Quantitative Analysis in Foreign Language Teaching Research, Huazhong University of Science and Technology Press, Wuhan, China, 2003.

[25] J. Q. Wang, Q. F. Wen, "A Model for the Computer-assisted Scoring of Chinese EFL Learners' Chinese-English Translation", Modern Foreign Languages, no. 4, pp. 415-420, 2009.

AUTHORS

Jinlin Jiang is with School of International Studies, University of International Business and Economics, Beijing, 100029 China (jiangjinlin2014@163.com).

Ying Qin is with Department of Computer Science, Beijing Foreign Studies University, Beijing, 100089 China (qinyingmail@163.com).

Ya Sun is with School of International Studies, University of International Business and Economics, Beijing, 100029 China (sawyersun@126.com).

Submitted 11 October 2015. Published as resubmitted by the authors 28 January 2016.