US-China Education Review B 6 (2011) 767-777
Earlier title: US-China Education Review, ISSN 1548-6613

A Machine Learning Approach to Measurement of Text Readability for EFL Learners Using Various Linguistic Features

Katsunori Kotani
Kansai Gaidai University, Osaka, Japan

Takehiko Yoshimi
Ryukoku University, Shiga, Japan

Hitoshi Isahara
Toyohashi University of Technology, Aichi, Japan

The present paper introduces and evaluates a readability measurement method designed for learners of EFL (English as a foreign language). The proposed readability measurement method (a regression model) estimates the text readability based on linguistic features, such as lexical, syntactic and discourse features. Text readability refers to the comprehension rate of a text (0.0-1.0). The experimental results showed that the proposed readability measurement method yielded higher accuracy than a baseline method, which provides the mode value of the distribution of the comprehension rate data as the estimated value for any input.

Keywords: computer-assisted language learning, EFL (English as a foreign language), reading ability of language learners, natural language processing

Introduction

Automatic measurement of readability has been an important issue in the area of language learning. Classical research used statistical analyses to develop readability formulae, such as Flesch Reading Ease (Flesch, 1948) and Flesch-Kincaid Grade Level (Kincaid, Fishburne, Rogers, & Chissom, 1975), whereas more recent researches have employed machine learning algorithms to develop readability measurement methods (Nagata, Masui, Kawai, & Siino, 2004; Schwarm & Ostendorf, 2005; Heilman, Collins-Thompson, Callan, & Eskenazi, 2007).

The recently proposed readability measurement methods can be classified into three types. The first type is designed to indicate readability for native speakers. Schwarm and Ostendorf (2005) developed a readability measurement method for English texts using Support Vector Machines (Vapnik, 1998) to combine features of traditional readability indices, statistical language models and other language features. The second type of readability measurement method has been designed for second/foreign language learners (Nagata et al., 2004). The third type can measure readability for both native speakers and second language learners (Heilman et al., 2007).

Because the previously proposed methods compute the readability score by mainly examining how many specific grammatical constructions, such as relative clauses and participle constructions appear in a text, they are faced with the problem of technological errors made by natural language processing tools in identifying specific grammatical constructions. This problem has been noted and resolved by Kotani, Yoshimi, and Isahara

Katsunori Kotani, College of Foreign Studies, Kansai Gaidai University.
Takehiko Yoshimi, Department of Media Informatics, Ryukoku University.
Hitoshi Isahara, Information and Media Center, Toyohashi University of Technology.

(2010), who constructed a reading proficiency prediction model for learners of EFL (English as a foreign language), not a readability measurement method. They proposed to construct a reading proficiency prediction model with linguistic features, such as the number of branching nodes in a syntactic tree, because these linguistic features are supposed to be less likely to introduce technological errors. In an evaluation experiment, the proposed reading proficiency prediction model was compared with a model constructed with linguistic features of specific grammatical constructions. The experimental results showed that the linguistic features proposed by Kotani et al. (2010) were adequate for constructing a reading proficiency prediction model for EFL learners. However, it has not been clarified whether the linguistic features are also adequate for constructing a readability measurement method for EFL learners.

In the present paper, we introduce a readability measurement method for EFL learners based on the same linguistic features proposed by Kotani et al. (2010), and conduct an evaluation experiment to clarify the effectiveness of the readability measurement method. The proposed readability measurement method is constructed using regression. The independent variables of this regression are various linguistic (lexical, syntactic and discourse) features, and the dependent variable is the readability score for EFL learners. Here, the readability score of a text refers to the comprehension rate of a text, which is computed by dividing the number of correct answers by the number of comprehension questions in a text (range from 0.0 to 1.0). The proposed readability measurement method takes a text as input, extracts various linguistic features of the text and estimates readability scores based on the extracted linguistic features.

Related Studies

Recent research on second/foreign language learning has elicited reading models for second/foreign language learners (Nagata et al., 2004; Kotani et al., 2010; Heilman et al., 2007). We briefly review these studies below.

Nagata et al. (2004) proposed a readability measurement method using a neural network learning algorithm. This method examines the number of specific grammatical constructions, such as post-nominal modifiers (e.g., relative clauses and participle constructions), appearing in a text. In this method, the readability score is weighted for these constructions, because, according to Nagata et al. (2004), it is difficult for Japanese EFL learners to comprehend these constructions.

Heilman et al. (2007) developed a readability measurement method for both native speakers and second language learners and compared the vocabulary-based and the grammar-based readability measurement methods. While the vocabulary-based method outperformed the grammar-based method, syntactic features were found to play an important role in second language readability in the grammar-based method.

The vocabulary-based readability measurement method is based on a unigram language model. Although Heilman et al. (2007) considered the unigram language model to be a weak model, it could be more effectively trained than more complex bi- or tri- gram models.

The grammar-based readability measurement method uses grammatical constructions, such as passive voice, past participles and relative clauses. These grammatical constructions were extracted from grammar textbooks for EFL learners and were implemented as syntactic patterns for a parsing tool.

Kotani et al. (2010) proposed a reading proficiency prediction model, not a readability measurement method. However, they did not use the number of specific grammatical constructions as syntactic features. According to Kotani et al. (2010), although the number of specific grammatical constructions undeniably

affects reading proficiency, a reading model using these features is also affected by technological errors made by natural language processing tools used for extracting linguistic features. When using a syntactic parser, we must consider the presence of technological errors, such as the incorrect labeling of syntactic nodes. For instance, a non-relative clause might be incorrectly labeled as a relative clause. Given this possibility, it is crucial to minimize the effects of such errors as much as possible. Kotani et al. (2010) solved this problem by using syntactic features that are available without labeling—specifically, the size of a sentence in terms of the number of syntactic branching nodes, which is believed to affect the reading proficiency from a psycholinguistic perspective, such as through the garden-path effect (Frazier & Rayner, 1982). In addition to syntactic features, Kotani et al.'s (2010) reading proficiency prediction model used lexical and discourse features and their reading model showed an 8.0% lower prediction error than a conventional model.

Linguistic Features of the Proposed Readability Measurement Method

Following the previous model (Kotani et al., 2010), we developed a readability measurement method that estimates readability scores for texts intended for EFL learners by examining the various linguistic features of the texts.

In the present paper, "linguistic features" refers to lexical, syntactic and discourse features. Of these, we selected those features that can be automatically derived with state-of-the-art natural language processing tools, as the goal of this study is to implement a readability measurement method into a computer-assisted language learning system. In the rest of this section, we review the features used to develop the proposed readability measurement method.

Lexical Features

Lexical features represent the vocabulary-related difficulties faced by EFL learners. As noted by Sano and Ino (2000), reading comprehension can be difficult for EFL learners even when only short words are used. Consequently, Kotani et al. (2010) assigned vocabulary difficulty scores based on heuristically determined vocabulary difficulty, which is summarized in the JACET (The Japan Association of College English Teachers) 4,000 Basic Words list (JACET, 1983). Vocabulary difficulty was determined by teachers of English working with Japanese EFL learners. The JACET list provides difficulty scores for 11 levels (Someya, 2000). The vocabulary difficulty of a given text is determined by summing the difficulty scores of all the words in the text.

The vocabulary difficulty list contains more than 35,000 words. However, authentic texts may contain words that are not registered in this list. Therefore, the reading model of Kotani et al. (2010) takes into account the fact that the model cannot estimate the difficulty of words that are not registered in the list.

Since the vocabulary difficulty list is compiled mainly for EFL learners, it is intended to cover words that EFL learners should study. As a result, the vocabulary difficulty of unregistered words is assumed to be higher than that of registered words. Following this assumption, the problem of unregistered words can be solved by either regarding unregistered words as more difficult than registered words or considering the number of unregistered words in a text to be a lexical feature. The former solution is hardly feasible, as it is difficult to precisely determine the vocabulary difficulty of unregistered words. Thus, following Kotani et al. (2010), we employed the latter strategy in this paper.

Although the vocabulary difficulty list covers basic vocabulary for EFL learners, some basic words might be more difficult than expected. For instance, words classed among the least difficult in the list, such as "get" and "make", may have various and complex usages and the difficulty of these words may depend on the

context in which they appear. Kotani et al. (2010) attempted to solve this problem by including the number of word meanings as another lexical feature. The number of word meanings was measured using Word Net 2.0 (Fellbaum, 1998), a large lexical database of the English language. The number of word meanings in a text was determined by summing the word meanings of each word in the text.

Syntactic Features

Syntactic features comprise two types, following Kotani et al. (2010): One is the number of all the branching nodes constituting a syntactic tree; and the other is the number of branching nodes stored in short-term memory under human language processing.

Since the number of syntactic nodes explains the size of a syntactic tree, we decided to use this quantificational information of syntactic nodes as a syntactic feature, following Kotani et al. (2010). In addition, Kotani et al. (2010) have suggested that the number of branching nodes is highly correlated with readability for EFL learners. The garden-path effect is a similar branching node effect (Frazier & Rayner, 1982). Syntactic parsing was performed using the Apple Pie Parser (Sekine & Grishman, 1995). Kotani et al. (2010) considered that the number of syntactic nodes could take into account the presence or absence of specific grammatical constructions that affected the reading comprehension of EFL learners. The number of syntactic nodes in a text is determined by summing all of the syntactic nodes in each sentence in the text.

Since a syntactic tree represents a result of syntactic parsing, it does not explain memory load during psychological syntactic parsing. Thus, following Kotani et al. (2010), we used the number of syntactic nodes stored in short-term memory as a syntactic feature representing short-term memory load. Syntactic nodes stored in short-term memory refer to those stored in a stack when analyzing a sentence in a top-down fashion using a push-down automaton. The number of nodes stored in a stack when parsing a text is determined by summing the numbers of nodes stored when parsing all the sentences in a text.

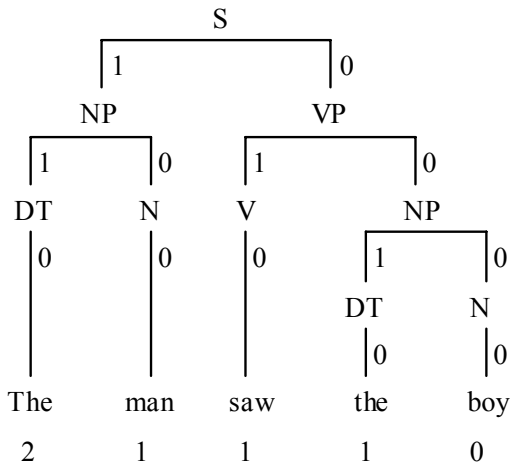

Figure 1. Number of nodes stored in short-term memory.

Figure 1 shows how the number of non-terminal symbols stored in a stack is determined for the sentence “The man saw the boy” in a push-down automaton (Yngve, 1960). When the first word “the” is inserted, theterminal symbol “S” (sentence) is transformed into (NP (noun phrase); VP (verb phrase)) and “VP” is memorized, that is, the symbol “VP” is stored in a stack. Next, the terminal symbol “NP” is transformed into (DT (determiner); N (noun)) and “N” is stored in a stack. Then, “DT” is rewritten as “the”. Therefore, the two

non-terminal symbols “N” and “VP” are stored in a stack, while “the” is processed.

The number of nodes stored in a stack is measured as follows (Yngve, 1960). As shown in Figure 1, beginning from zero, a number is assigned to each branch from right to left. The sum of the numbers in the path from “S” to a word indicates the number of symbols stored in a stack for that word. The following numbers are assigned to each word as the number of nodes stored in a stack for the sentence “The man saw the boy”: 2, 1, 1, 1 and 0.

Murata et al. (2001) modified this number assignment procedure in certain aspects; for instance, NP that has no postmodifier will not be transformed. Thus, as NPs in the sentence “The man saw the boy” have no postmodifier, the numbers of nodes in a stack is 1, 1 and 0. Murata et al. (2001) determined the number of nodes in a stack following this revised procedure. The number of nodes stored in a stack in a text is determined by summing all of the numbers of nodes in a stack in each sentence.

Discourse Features

The discourse feature of the proposed readability measurement method is the number of pronouns, following Kotani et al. (2010). While reading a text, referents of pronouns must be identified and this requires comprehension of the discourse structure. Thus, the number of pronouns can be used as an indicator of the complexity of the discourse structure of a text.

Although a text may include other anaphoric expressions, such as definite expressions, these are not included as a discourse feature due to the technological error effect. Kotani et al. (2010) considered that the detection of pronouns involves fewer technical problems.

Comprehension Rate Data Collection

In the proposed readability measurement method, readability scores are assessed based on comprehension rate. In the present study, comprehension rate was defined as the correct answer rate for comprehension questions about the texts (range from 0.0 to 1.0). In addition to the linguistic features reviewed in above, comprehension rate data were used as training data in order to develop the proposed method.

Comprehension rate data were collected as follows. Participants were recruited from a job information Website and were chosen on the basis of the following criteria: Those who had taken the TOEIC (Test of English for International Communication) (Website, http://www.ets.org/toeic, a test of English language skills used in the workplace), those who could submit a TOEIC score sheet, and those who lived near the data collection site. Among the respondents, 64 took part in the data collection process. All the participants had taken the TOEIC within the previous one-year period and their native language was Japanese.

We prepared test sets based on 84 texts extracted from TOEIC preparation textbooks (Arbogast et al., 2001; Lougheed, 2003). Each test set consisted of seven texts and every test set contained different texts. Each text was accompanied by two to five multiple-choice comprehension questions. We randomly provided participants with one or two test sets. Thirty-one participants took one test set and 33 participants took two test sets.

Comprehension rate data were collected using a reading process recording tool (Yoshimi, Kotani, Kutsumi, Sata, & Isahara, 2005). This tool displays one sentence at a time (see Figure 2). A sentence appears on the computer screen when the cursor is positioned over a reading icon and it disappears when the cursor is moved away from the icon.

Participants used this tool while reading the text and answering the comprehension questions. When the

cursor was positioned over a question icon, a comprehension question appeared. Participants answered the question by clicking on one of the answer icons.

Figure 2. Screenshot of the reading process recording tool.

After receiving instructions about the tool, participants practiced by reading several texts and answering comprehension questions. The participants were instructed first to read the text and then answer the comprehension questions. We also directed participants to attempt to understand the text well enough to correctly answer the comprehension questions. Since we did not impose time constraints, the participants could take as much time as they needed. In order to reduce the pressure on the participants, we did not inform them that the tool would be measuring their reading times.

We excluded comprehension rate data of four participants whose reading speed (WPM (words per minute)) was extremely fast or slow (> 200 WPM or < 70 WPM), as slow reading speed might have been the result of unnecessarily careful reading and excessively fast reading speed could indicate that participants did not properly read the materials (average reading speed of native English speakers is reported to be in the range of 200 to 300 WPM (Carver, 1982)). The comprehension rate data we obtained consisted of 451 instances. An instance consists of the linguistic features of a text and the comprehension rate when an EFL learner reads the text. The mean age of the participants whose comprehension rate data were included in analysis was 29.8 years (*SD* (standard deviation) = 9.5). Nine participants were males and 51 were females.

The distribution of the comprehension rate data is shown in Figure 3. The comprehension rate data comprises ten values from 0.0 to 1.0, and each value refers to the comprehension rate calculated by dividing the number of correct answers (one to five) by the number of comprehension questions (two to five). The comprehension rate data showed a skewed distribution plotted with a dotted line, because the comprehension rate was 1.0 in 59.6% of the instances (269 out of 451). The reason why so many instances of comprehension rate 1.0 were observed could be due to the fact that, as there was no time restriction in this experiment, the participants could spend as much time as they wanted to complete each question.

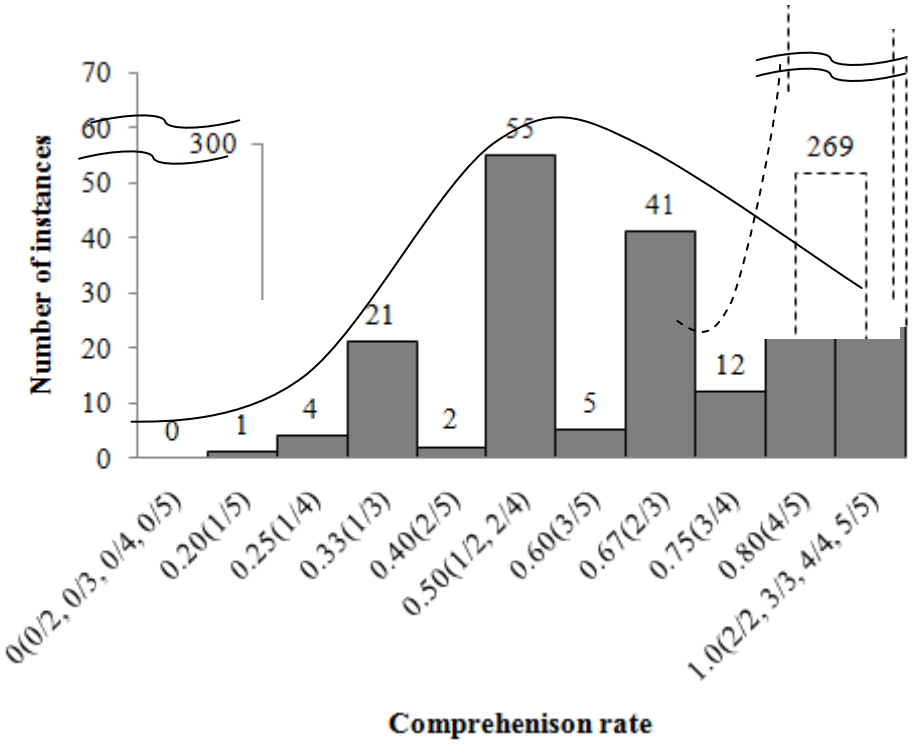

Figure 3. Histogram of comprehension rate data.

As the target of this readability measurement method was texts intended for use in norm-referenced tests, the readability of the texts should follow a normal distribution. Since a readability measurement method trained with skewed data estimates skewed values, it is highly likely that a method trained with skewed data cannot properly measure the readability of texts intended for norm-referenced tests. To address this problem, we corrected the distribution of the comprehension rate data by randomly selecting 31 instances of comprehension rate 1.0. The modified comprehension rate (plotted with an actual line) included 194 instances. We considered this to be a roughly normal distribution.

Evaluation Experiment

In this section, we describe experiments for the evaluation of the proposed readability measurement method. First, we describe the experimental methods, and then, we report the experimental results.

Experimental Method

We developed a readability measurement method using comprehension rate as a dependent variable and linguistic features as independent variables. This readability measurement method was evaluated using the 194 instances of comprehension rate data described above. The evaluation was performed using five-fold cross-validation tests.

Support vector regression (Vapnik, 1998) was carried out using an algorithm implemented in the author's SVM (support vector machine) software (Website, http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/index.html). The d-th polynomial kernel function (d = 2, 3 and 4) and soft margin parameter C (C = 1, 0.1 and 0.01) were selected, and the other settings remained at the default values.

The performance of the proposed readability measurement method was examined in terms of the absolute error (the absolute value of the difference between the estimated value and the observed value). The estimated

values refer to the readability scores calculated with the readability measurement method and the observed value indicates the learner's actual comprehension rate obtained in the data collection described above. The absolute error shows the degree to which the readability measurement method correctly indicates readability scores for EFL learners.

The proposed readability measurement method was also compared with a baseline method, which provides the mode value of the distribution of the comprehension rate data as the estimated value for any input data. As shown in Figure 3, the mode value was 0.5. Thus, the absolute error of the proposed method should be smaller than the absolute value of the difference between 0.5 and the observed value.

Experimental Results

Figure 4 shows the distribution and the cumulative relative frequency of the absolute error of the proposed method. This distribution shows the absolute error when the order of the kernel function d is set to 2 and the soft margin parameter C is set to 0.1 among the nine combinations of parameter settings of $d = 2, 3, 4$ and $C = 1, 0.1, 0.001$. The distribution of absolute error indicates that the absolute error appears mostly in the lower range (0.1 to 0.2). The absolute error less than 0.4 has a cumulative relative frequency of 93.8%. Moreover, the distribution of the absolute error is positively skewed. The median absolute error of the proposed readability measurement method was 0.13 (range from 0.00 to 0.60).

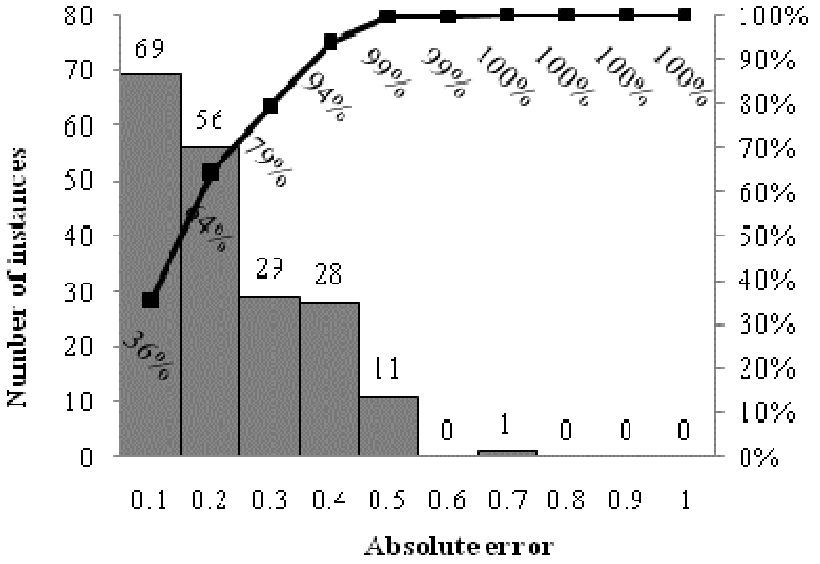

Figure 4. Histogram of the absolute error of the proposed method.

In order to further examine the appropriateness of the readability measurement method, we compared the absolute error of the proposed readability measurement method with that of the baseline method. Figure 5 shows the distribution and the cumulative relative frequency of the absolute error of the baseline method. The distribution of absolute error indicates that the absolute error of the baseline method appears mostly in the lower range (range from 0.1 to 0.2), similar to the absolute error of the proposed readability measurement method. Absolute error less than 0.4 has a cumulative relative frequency of 84.0%. The cumulative relative frequency of the proposed readability measurement method is higher than that of the baseline method. The median absolute error was 0.17 (range from 0.00 to 0.50).

The significance of the difference in median absolute error values was examined using the Wilcoxon pair

matched rank sum test. A significant difference ($p < 0.05$) was found between the absolute error of the proposed readability measurement method and that of the baseline method.

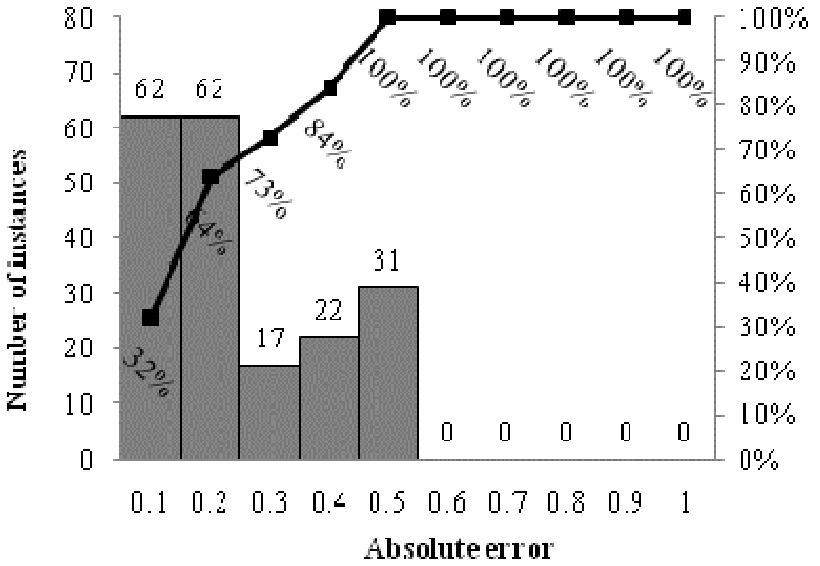

Figure 5. Histogram of the absolute error of the baseline method.

In addition, we also calculated the absolute values of the difference between the absolute error of the proposed measurement method AER_P (absolute error rate of the proposed measurement) and that of the baseline method AER_B (absolute error rate of the baseline measurement).

The absolute differences (AER_P -AER_B) are plotted in descending order in Figure 6. The horizontal axis represents the number of instances, and the vertical axis represents the absolute values of the differences of the two readability measurement methods. In Figure 6, "proposed method" represents the cases in which $AER_P < AER_B$. The "baseline method" represents the cases in which $AER_P \geq AER_B$. Note that the differences are normalized by dividing each absolute difference by the maximum absolute difference.

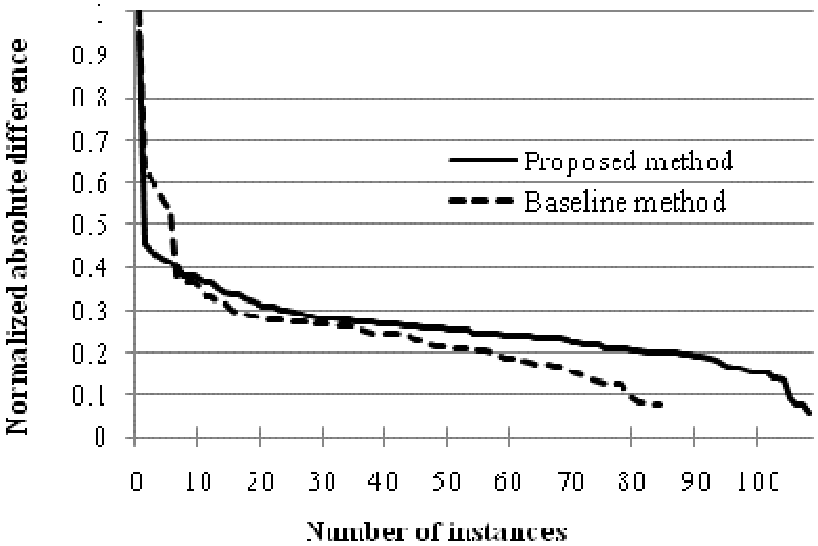

Figure 6. Normalized absolute differences of absolute error between the proposed readability measurement method and the baseline method.

Figure 6 shows that the proposed readability measurement method resulted in larger error than the baseline method when the normalized absolute differences fell between 0.4 and 1.0 (six instances). However, the proposed method resulted in smaller error in most cases, as most plots of the proposed method appear above those of the baseline method. In addition, as the distribution for the proposed method is longer than that of the baseline method, there were more instances in which the proposed method resulted in smaller errors (109 instances, 54%).

Conclusions

We proposed a readability measurement method for EFL learners based on various linguistic features that consist of lexical, syntactic and discourse features. The median absolute error of the proposed method was relatively low at 0.13 (range from 0.00 to 0.60), and this was lower than the absolute error of the baseline method (0.17 (range from 0.00 to 0.50)). Also, the proposed method had a higher cumulative relative frequency of error below 0.4 than the baseline method. Finally, the distribution of the absolute error of the proposed method was significantly different from that of the baseline method ($p < 0.05$). From these experimental results, we concluded that the proposed method can effectively assess the readability of texts intended for EFL learners.

The present paper leaves several problems unresolved. First, we must improve the accuracy of the proposed method. Second, we should examine the other possible independent variables. By using learner features, such as reading time, we may be able to develop a more effective readability measurement method on a learner-by-learner basis. Finally, we must examine other possible dependent variables, and we may use reading time data and a complex measure of comprehension and reading time (known as effective reading speed (Jackson & McClelland, 1979)).

References

Arbogast, B. (Ed.). (2001). *TOEIC official test-preparation guide: Test of English for international communication*. N. J.: Peterson's, Lawrenceville.

Carver, R. P. (1982). Optimal rate of reading prose. *Reading Research Quarterly, 18*(1), 56-88.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32,* 221-233.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology, 14,* 178-210.

Fellbaum, C. (1998). *Word net: An electronic lexical database*. Cambridge, M. A.: The MIT Press.

Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. Proceedings of *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)* (pp. 460-467). Rochester, New York, April 2-22.

JACET (The Japan Association of College English Teachers.). (1993). *JACET 4000 basic words*. The Japan Association of College English Teachers, Tokyo.

Jackson, M. D., & McClelland, J. L. (1979). Processing determinants of reading speed. *Journal of Experimental Psychology, 108,* 151-181.

Kincaid, J. P., Fishburne, R. P. Jr., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Research Branch Report,* 8-75. US: Naval Air Station, Memphis.

Kotani, K., Yoshimi, T., & Isahara, H. (2010). A prediction model of foreign language reading proficiency based on reading time and text complexity. *US-China Education Review, 7*(10), 1-9.

Lougheed, L. (2003). *How to prepare for the TOEIC Test: Test of English for international communication*. Hauppanuge, N. Y.: Barron's Educational Series, Inc..

Murata, M., Uchimoto, K., Ma, Q., & Isahara, H. (2001). Magical number seven plus or minus two: Syntactic structure recognition in Japanese and English sentences. In *Computational linguistics and intelligent text processing* (Lecture Notes in Computer Science, 2004, pp. 43-52). Springer Berlin: Heidelberg.

Nagata, R., Masui, F., Kawai, A., & Siino, T. (2004). A method of rating English texts by reading level for Japanese learners of English. *The Transactions of the Institute of Electronics, Information and Communication Engineers, D-II*(6), 1329-1338.

Sano, H., & Ino, M. (2000). Measurement of difficulty on English grammar and automatic analysis. *IPSJ SIG Notes, 2000*(117), 5-12.

Schwarm, S. E., & Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. Proceedings of *the 43rd Annual Meeting of the Association for Computational Linguistics* (pp. 523-530). Ann Arbor, Michigan, June 25-30.

Sekine, S., & Grishman, A. (1995). A corpus based probabilistic grammar with only two non-terminals. Proceedings of *the 4th International Workshop on Parsing Technologies* (pp. 216-223). Prague, Czech Republic, September 20-23.

Someya, Y. (2000). *Word level checker: Vocabulary profiling program by AWK, 1.5.* Retrieved from http://www1.kamakuranet.ne.jp/someya/wlc/wlc_manual.html

Vapnik, V. (1998). *Statistical learning theory*. Wiley-Interscience, N. Y..

Yngve, V. H. (1960). A model and a hypothesis for language structure. *The American Philosophical Society, 104*(5), 444-466.

Yoshimi, T., Kotani, K., Kutsumi, T., Sata, I., & Isahara, H. (2005). A method of measuring reading time for assessing EFL learners' reading ability. *Transactions of Japanese Society for Information and Systems in Education, 22*(1), 24-29.