2011 11th IEEE International Conference on Advanced Learning Technologies

A Robust Estimation Scheme of Reading Difficulty for Second Language Learners

Yi-Ting Huang, Hsiao-Pei Chang,Yeali Sun
Department of Information Management
National Taiwan University
Taipei, Taiwan
{d97725008, r98725037, sunny}@ntu.edu.tw

Meng Chang Chen
Institute of Information Science
Academic Sinica
Taipei, Taiwan
mcc@iis.sinica.edu.tw

***Abstract*—Reading difficulty is a measurement for estimating the appropriate reading level of a document. Almost all prior studies are designed for first language learners, but not for second language learners. In this study, we propose a robust estimation scheme, including features such as word frequency, official word grade and grammar patterns, to train a linear model to estimate the difficulty of the document for second language learners. The experiment results show that the proposed estimation scheme outperforms other reading difficulty estimations.**

Keywords-readability; reading difficulty; second language learning; linear regression model

I. INTRODUCTION

Reading difficulty (also called readability) is often used to estimate the reading level of a document, so that learners can choose appropriate learning materials. Heilman et al. [4] (denoted as the Heilman method hereafter in this paper) described reading difficulty as a function that maps a document to a numerical value corresponding to a difficulty or grade level. A list of lexical and grammatical features extracted from a document usually act as the inputs of this function, and one of the ordered difficulty grade levels is the output corresponding to a learner's reading skill.

Numerous researchers have studied the issue of reading difficulty, applying various lexical and grammatical features in statistic models. Early studies such as the Dale-Chall model [1][3], the Flesch-Kincaid measures [6], and the Lexile [8] only adopted simple lexical and grammatical features and designed a regression model to predict difficulty levels. Schwarm and Ostendorf [7] proposed more complicated features that significantly increased the performance of readability prediction. Furthermore, some researchers also took language models into consideration [2][4][5], in order to obtain a probability distribution for each grade.

In bilingual education environments, it is necessary for second language (also called L2) learners to choose suitable documents to improve their language skill through intensive reading. However, it is not suitable to apply prior work directly to second language learners. As Heilman et al. [4] pointed out, the learning timeline and processing between first language (called L1) learners and second language learners is different: first language learners learn all grammar rules before formal education, whereas second language learners learn grammatical structures and vocabulary simultaneously and incrementally. In their work, they used a linear model to combine and weigh the unigram language model and the interpolation of grammar level.

Most past literature was designated for first language learners, and consulted word frequency from general corpora composed of articles for native speakers. But for second language learners, the word difficulty depends on the structure of the material they study, not its popularity in the real world. In this paper, we design a reading difficulty estimation scheme for second language learners that adopts several word frequency features from corpora, official word grades from language authorities, and grammar patterns collected from textbooks. We also conduct two experiments: the first compares the proposed estimation with the other methods, while the second compares the estimation to the results of three English teachers. The corpus contains 175 documents in six grade levels, which were gathered from high school English textbooks designed for Chinese students. The results show that the performance of the proposed estimation is better than other methods and is similar to human experts.

The remainder of this paper is organized as follows. Section 2 describes related work on reading difficulty. In Section 3, we define the research problem and present lexical and grammatical features of the task. Section 4 contains the experimental method and results. Finally, Section 5 provides discussion and Section 6 summaries a conclusion and the future work.

II. RELATED WORK

Early related works only used a few simple features to measure lexical complexity, such as word frequency or number of syllables per word. Since they took fewer features into account, most studies made assumptions on what variables affected readability, and then based their difficulty metrics on these assumptions. One example is the Dale-Chall model [1] [3] which determined a list of 3,000 commonly known words and then used the percentage of rare words to measure the lexical difficulty. Another example is the Lexile Framework [8] which used the mean log word frequency as a feature to measure lexical complexity. And then the researchers entered the parameters into a logistic regression analysis to obtain a logit difficulty level, which helps determine if the learner can comprehend 75% of a given document. Using word frequency to measure lexical difficulty assumes that a more frequent word is easier

978-0-7695-4346-8/11 $26.00 © 2011 IEEE
DOI 10.1109/ICALT.2011.25

for learners. This assumption seems fair, since a widely used word has a stronger chance to be seen and absorbed by learners, but this method is susceptible to the diverse word frequency rates found in various corpora.

More recent approaches have started to take n-gram language models into consideration to assess lexical complexity, which can more accurately measure difficulty. Collins-Thompson and Callan [2] used the smoothed unigram language model to measure the lexical difficulty of a given document. For each document, they generated language models by levels of readability, and then calculated the likelihood ratios to assign the level of difficulty; in other words, the prediction is the level that has the highest likelihood ratio of the document. Schwarm and Ostendorf [7] also utilized statistical language models to classify documents based on reading difficulty level, and they found that trigram models are more accurate than bigrams and unigrams.

Prior studies [1] [3] [6] [8] only calculated the mean number of words per sentence to estimate grammatical readability. Using sentence length to measure grammatical difficulty assumes that a shorter sentence is syntactically more simple than a longer one, but it does not mean long sentences are always more difficult than shorter sentences. More recent approaches have started to consider the structure of sentences when measuring grammatical complexity because of the increasing parser accuracy rate. This research usually considered more grammatical features such as parsing features per sentence in order to make a more accurate difficulty prediction. Schwarm and Ostendorf [7] employed four grammatical features derived from syntactic parsers. These features included the mean parsing tree height, mean number of noun phrases, mean number of verb phrases, and mean number of SBARs to assess a document's readability. Heilman et al. [5] used grammatical features extracted from automatic context-free grammar parsing trees of sentences, and then computed the relative frequencies of partial syntactic derivations. The more frequent subtrees will be viewed as less difficult for learners.

Most prior work has proposed some critical features to assess reading difficulty; however, these methods are not designed for second language learners. It is inappropriate to use these methods directly to predict the recommended document grade level for second language learners. For example, the unigram language model was used in Heilman et al. [4] to estimate the probability distribution of words in each grade level. When the model predicts the reading difficulty for L2 learners, it will recommend unsuitable documents, because the model is based on a L1 corpus. In this paper, we consult some meaningful lexical and grammatical features in early work, and then further consider several word frequency features from corpora, official grading indexes of vocabulary from language experts, and grammar patterns collected from textbooks—those which represent words and grammar patterns that L2 learners have learned at various grade levels.

III. The L2 Reading Difficulty Estimation Scheme

Existing approaches of reading difficulty estimations are insufficient for L2 learners, because they only rely on vocabulary lists or some superficial representation of syntax. The proposed approach selects some representative lexical and grammatical features to compose a function to predict readability. The inputs of this function are a list of lexical and grammatical features of a document, and the output is a document difficulty score. The scores can also correspond to one of the ordered difficulty levels.

A. Notations

Let D represent the document that a learner reads, while S represents the sentences in D. Suppose that D has n sentences, $S_1, S_2, \ldots, S_n$, so that $D = \{ S_1, S_2, \ldots, S_n \}$. Let W be the set of words in D. Suppose D has m distinct words, $W_1, W_2, \ldots, W_m$, so that a document as $D = \{ W_1, W_2, \ldots, W_m \}$. We further suppose that the sentence S has k words, $W_1, W_2, \ldots, W_k$, so that $S = \{ W_1, W_2, \ldots, W_k \}$, $m > k$. Let L represent the lexical features, so that $L = \{L_1, L_2 \ldots, L_n\}$. The grammatical features are represented as G, $G = \{G_1, G_2 \ldots, G_n\}$. More detailed descriptions and definitions of each feature are in the next section.

B. Features

For a given training data set, the features are extracted and sent to a linear regression process to obtain a linear model that includes the weight of each feature. The linear model is then applied to a document to estimate the difficulty level. Here we explain and define lexical and grammatical features used in the proposed scheme.

1) Lexical Features: For every word in a document, we find its word frequency from the BNC corpus and also use a Google search result count as an alternative frequency. The use of word frequency is based on the assumption that if a word is more frequent, it tends to be easier.

a) Word frequency in BNC corpus: The British National Corpus (BNC) (http://www.natcorp.ox.ac.uk/) is a 100 million word collection of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later 20th century. For each word in a document, we calculate the distinct word frequency (*wf*) that refers to the times it appears in the BNC corpus. Word frequency is defined as follows:

$$wf_i = \frac{n_i}{|dj|} \quad (1)$$

where n_i is the number of occurrences of the considered distinct word w_i in document d_j, and the denominator is the sum of the number of occurrences of all distinct words in document d_j, that is, the size of the document $| d_j |$. The document's difficulty value of word frequency in the BNC corpus is defined as follows:

$$BNC = log \frac{\sum_{i=1}^{m} wf_i}{m} \quad (2)$$

b) Google search result count: For a given query, Google will return a list of documents containing the queried words and a search result count. We use the search

result count as a measure of word frequency, like the word frequency from a corpus. The document's difficulty value based on word frequency from Google is defined as follows:

$$Google = log \frac{\sum_{i=1}^{m} google\_count_i}{m} \quad (3)$$

where $google\_count_i$ is the search result count of a word i from Google.

Due to differences in learning time between L1 learners and L2 learners, we build two dictionaries from formal and educational grading indices for L2 learners made by human experts. This helps to better catch the vocabulary difficulty level of L2 learners.

c) CEEC – High School English Reference Vocabulary: The "High School English Reference vocabulary" text made by the College Entrance Examination Center (CEEC) (http://www.ceec.edu.tw/Research/paper_doc/ce37/ce37.htm) of Taiwan contains 6,480 words in English, divided into six levels, which represent the specified range of the Department Required Test. For each word from the document, we identify its difficulty by first referencing its difficulty level from within the CEEC word lists, counting the number of distinct words in each level, and then normalizing by the total number of distinct words in each level. The CEEC difficulty of a document is defined as follows:

$$CEEC = \{ceec_1, ceec_2, \ldots, ceec_6\} \quad (4)$$

d) GEPT word lists: The General English Proficiency Test (GEPT) provides a reference vocabulary list with 8000 words (http://www.lttc.ntu.edu.tw/academics/wordlist.htm), divided into three levels: elementary, intermediate and high-intermediate. For each word from the article, we identify its vocabulary difficulty by searching the word's level from the GEPT word lists, counting the number of distinct words in each level, and finally normalizing by the total number of distinct words in each level. The GEPT difficulty of a document is defined as follows:

$$GEPT = \{gept_1, gept_2, gept_3\} \quad (5)$$

e) Number of syllables per word: A syllable is a unit of organization for a sequence of speech sounds. For example, the word *water* is composed of two syllables: *wa* and *ter*. We find the number of syllables of every word in a document to measure the reading difficulty. This is based on the assumption that a mono- or bi-syllabic word is more familiar to learner than polysyllabic words. The syllable difficulty of a document is defined as follows:

$$Syllables = \frac{\sum_{i=1}^{m} word\_syllables_i}{m} \quad (6)$$

where $word\_syllables_i$ is the number of syllables within a word i.

2) Grammatical Features: Grammatical features include word count, sentence length, grading index of grammar, and parsing features.

a) Word count: The number of words in a document is used as one of the features to estimate reading difficulty. We assume that a longer document is more difficult than a shorter one. The word count difficulty is defined as follows:

$$Word\_Count = log(|D|) \quad (7)$$

b) Sentence length: For each document, we consider the average sentence length as a feature of syntactic complexity. This assumes that a shorter sentence is easier than a longer one. The sentence length difficulty is defined as follows:

$$Sentence\_Length = \frac{Word\_Count}{n} \quad (8)$$

c) Grading index of grammar: We consider the grammatical difficulty as a linguistic processing factor for second language learning in estimating reading difficulty. We first collected sentences from six English textbooks and parsed the sentences to find their grammar patterns. We assigned each grammar pattern to the textbook in which it first appears. In other words, the grammatical difficulty level of a grammar is the assigned textbook grade.

The grammatical difficulty level of a document is represented by a set of grammatical difficulty values for each level. To decide the grammatical difficulty level of a document, firstly we find the grammatical difficulty level of each sentence. Then for each grammatical difficulty level, we count the number of sentences in this level and normalize by the total number of sentences, denoted as G_i. Then the grammatical difficulty is defined as follows:

$$Grammar = \{G_1, G_2, \ldots, G_6\} \quad (9)$$

We consider the following syntactic features from parsing results generated by a parser (http://nlp.stanford.edu/software/lex-parser.shtml), average parsing tree height, average number of noun phrases, average number of verb phrases and average number of SBARs.

d) Average parsing tree height: Suppose the height of a parsing tree of a sentence is h. The average parsing tree height difficulty of a document is defined as follows:

$$Parse\_tree\_height = \frac{\sum_{h=1}^{n} h}{n} \quad (10)$$

e) Average number of noun phrases: Suppose a sentence has NP_i noun phrases. The average NP difficulty of a document is defined as follows:

$$NP = \frac{\sum_{i=1}^{n} NP_i}{n} \quad (11)$$

f) Average number of verb phrases: Suppose a sentence has VP_i verb phrases. The average VP difficulty of a document is defined as follows:

$$VP = \frac{\sum_{i=1}^{n} VP_i}{n} \quad (12)$$

g) Average number of SBARs: Subsidiary conjunctions (SBAR), for example, "because," "unless," "even though," and "until," are placed at the beginning of a subordinate clause that links the subordinate clause and the dominant clause. SBAR is an indicator to measure sentence complexity. The SBAR difficulty of a document is defined as follows:

$$SBAR = \frac{\sum_{i=1}^{n} SBAR_i}{n} \quad (13)$$

C. Statistical models

Linear regression is an approach to modeling the relationship between a scalar variable Y and variables denoted X. A prediction of a given document is the inner product of a vector of feature values for the document and a vector of regression coefficients estimated from the training data.

$$Y = \alpha + \sum_{i=1}^{n}(\beta_i X_i + \varepsilon_i) \quad , i = 1, 2, , n \quad (14)$$

where Y is the difficulty value of document, α is the intercept parameter, $X = \{X_1 \quad X_2 \quad \dots \quad X_n\}$ represent for the lexical and grammatical feature values, $\beta = \{\beta_1 \quad \beta_2 \quad \dots \quad \beta_n\}$ refers to the regression coefficient for each feature value i, ε is an unobserved random variable that represents noise of the linear relationship between the dependent variable and regressors. Each level l has its own threshold θ_l value as the basis for the document reading level classification:

$$\theta_l = \frac{\bar{X_l} + \bar{X_{l+1}}}{2}, 1 \le l \le 6 \quad (15)$$

$$\bar{X_l} = \alpha + \frac{1}{|X_l|} \sum_{li} \sum_i \beta_i X_{li} \ , X_l \in X \ , X_{l+1} \in X \quad (16)$$

IV. Experiments

This section presents two experiments to exemplify the merits of the proposed method. The first experiment compares the proposed scheme with other methods, and the second experiment compares the scheme with the results generated by human experts. First we describe the experimental setup, which consists of two evaluation corpora and metrics. Two compared estimations are also introduced.

A. Experiment Settings

Two corpora were used in the experiments. The first corpus was from high school English textbooks designed for Chinese students to learn English as a second language. It gathered 175 documents in six grades from three different publishers (including Far East Book Company, Lungteng Cultural Company, and San Min Book Company). The first experiment was designed to compare the estimation correctness within the proposed estimation scheme, Lexile [4] (http://lexile.com/) and the Heilman method [6] (http://boston.lti.cs.cmu.edu/demos/readability/). The second corpus contained 12 documents extracted from online news websites and reading difficulty level labels annotated by three high school teachers. Kappa statistics was used to evaluate the inter-rater agreement between three annotators. The Kappa value of annotators on the second corpus was *0.08*, which implies that the annotations were not consistent and subjective. While annotators did not agree with each other on the level of many documents, most of their differences were only 1 level apart.

Three measurements, accuracy, the root mean squared error (RMSE) and the Pearson's correlation coefficient, were used in the experiments in order to evaluate the effectiveness of the estimated reading difficulty. Accuracy is defined as the proportion of the correctness of generated results within the ground truth. RMSE shows the average distance between the ground truth and the generated results. The Pearson's correlation coefficient measures the trends between the ground truth and the generated results. A ten-fold cross-validation was employed in the first experiment.

B. Baseline Methods

There are two baseline estimations available online for first language learners, Lexile and the Heilman method. Because the grades of the two baseline methods are divided into different scales, we have to find the threshold between each level of the two estimations. In the training phase, we collected the reading difficulty prediction generated from each document by Lexile, and found the threshold in each grade, shown as Table □. In the testing phase, the level of a testing document was determined by examining its Lexile estimation with the thresholds. In the Heilman method, training documents with different grades were estimated into the same Heilman grade. As a result, it performed poorly in the first experiment.

TABLE I. THE RESPONDING LEVELS IN LEXILE ESTIMATION AND THE HEILMAN METHOD.

Grade	Lexile value	Heilman method
1	L<817.944	L < 6.264
2	817.944≦L< 852.379	6.264≦L< 6.795
3	852.379≦L< 935.713	6.795≦L< 7.619
4	935.713≦L< 1020.495	7.619≦L< 8.402
5	1020.495≦L< 1062.490	8.402≦L< 8.782
6	1062.490 ≦L	8.782 ≦L

C. Results

In the first experiment, Table Ⅱ shows the results between the proposed estimation scheme, Lexile and the Heilman method. For both the accuracy measurement and RMSE, the proposed estimation produces more accurate reading difficulty predictions than both Lexile and the Heilman method. When the proposed estimation fails to predict the correct reading difficulty, its error ranges are almost within one grade; in contrast, the error ranges of Lexile estimations are between one to two grades, and the Heilman method has an even wider error range. Table Ⅲ reports the Person's correlation coefficient among the three estimations. All three estimations are positively correlated, however the proposed estimation reports a particularly high correlation at *0.897 ($p<0.001$)*. This suggests that the relationship between the proposed estimation and the ground truth is stronger than the others.

In the second experiment, Table Ⅳ shows the results between the estimation predicted by the proposed scheme and three human experts' annotations. The average accuracy of the proposed scheme is lower than the first experiment, and the average RMSE indicates that the error ranges are almost within one grade; in other words, the proposed estimation was close to the annotators' judgments. Furthermore, some annotators revealed that they estimated the document difficulty by the document's length, word difficulty, sentence complexity, and grammar patterns, which are similar to the features used in the proposed estimation.

The results indicate that the proposed method still performs stably with previously unknown documents.

From the experiments, we found that the existing difficulty estimation methods did not perform well for second language learners due to the different and insufficient features used. The proposed estimation prediction is consistent, although it tends to predict easy documents with a lower grade and difficult documents with higher grades. In contrast, the results of Lexile and the Heilman method are fluctuant.

TABLE II. THE RESULTS OF ACCURACY AND RMSE AMONG THE PROPOSED ESTIMATION, LEXILE AND HEILMAN METHOD.

Level	Proposed		Lexile		Heilman	
	Accuracy	RMSE	Accuracy	RMSE	Accuracy	RMSE
1	0.767	0.587	0.667	0.899	0.833	0.847
2	0.400	0.811	0.033	1.529	0.000	1.529
3	0.450	0.911	0.300	1.436	0.000	1.738
4	0.533	0.712	0.167	1.870	0.167	2.169
5	0.350	0.770	0.100	1.648	0.000	1.958
6	0.500	0.865	0.400	1.796	0.450	1.405
Avg.	0.500	0.776	0.278	1.530	0.242	1.607

TABLE III. THE RESULTS OF PEARSON'S CORRELATION COEFFICIENT AMONG THE PROPOSED ESTIMATION, LEXILE AND HEILMAN METHOD

	Proposed Estimation	Lexile	Heilman Method
Correlation	0.897***	0.539***	0.547***

*(*** = p<0.001)*

TABLE IV. THE RESULTS OF ACCURACY, RMSE AND PEARSON'S CORRELATION COEFFICIENT AMONG THE THREE EXPERTS.

	Expert 1	Expert 2	Expert 3	Average
Accuracy	0.500	0.250	0.333	0.361
RMSE	0.707	2.566	3.304	2.192
Correlation	0.924***	0.603*	0.679*	0.735

*(*** = p<0.001, *=p<0.05)*

V. DISCUSSION

The coefficient of a feature indicates the importance of the feature in determining the difficulty grade. Source (17) lists the linear model trained in experiment 1. The coefficient value of the official word grade (e.g., CEEC and GEPT word list), no matter positive or negative, is far larger than the word frequency from BNC or Google. In other words, it supports the assumption that the difficulty of the structure of L2 materials is different from L1.

$$Y = -64.01 + 0.24BNC + 1.73Google + 16.79CEEC_1 + 20.49CEEC_2 + 25.12CEEC_3 + 21.50CEEC_4 + 24.88CEEC_5 + 0.00CEEC_6 - 18.14GEPT_1 - 8.97GEPT_2 + 0.00GEPT_3 - 1.86Syllables + 0.08Word\_Count + 1.17Sentence\_Length - 2.32G_1 - 0.73G_2 - 1.63G_3 - 0.66G_4 - 0.06G_5 - 0.02G_6 - 0.59Parse\_tree\_height + 0.11NP + 0.64VP + 0.69SBAR \quad (17)$$

In order to examine the proposed scheme also works in other second language environments, we removed the official grading indexes of vocabulary, CEEC and GEPT. In Table Ⅴ, the first row is the average performance of the estimation with all the features, the second row is without the GEPT word lists feature, and the third row is without the CEEC word lists feature. It is clear that with official grading indexes of vocabulary, the estimation performance degrades, although it still remains better than Lexile and the Heilman method. It shows that the proposed scheme can be employed in other second language learning environments when the reading materials are incrementally constructed.

TABLE V. THE ESTIMATION PERFORMANCE WITH FULL FEATURES AND WITHOUT OFFICIAL WORD GRADE FEATURES.

	Accuracy	RMSE	Correlation
Full features	0.500	0.776	0.897***
- GEPT	0.483	0.780	0.893***
- CEEC	0.451	0.782	0.859***
- CEEC & GEPT	0.428	0.945	0.841***

VI. CONCLUSION AND FUTURE WORKS

In this paper, we proposed a state-of-the-art approach to estimating a learning document's difficulty level for second language learners. We also identified features such as word frequency from corpora, official grading indexes of vocabulary and grammar that can effectively improve performance. Two experimental results have shown that the proposed estimation outperforms the other estimations, and is close to the annotation of human experts. In the future, we will integrate the proposed scheme into the AutoQuiz (http://autoquiz.iis.sinica.edu.tw/) project and provide second language learners with personalized learning materials.

ACKNOWLEDGMENT

This work was partially supported by National Science Council, Taiwan, with Grant No. NSC99-2631-H-008-004 and NSC97-2221-E-001-014-MY3.

REFERENCES

[1] J. S. Chall and E. Dale, "Readability Revisited: The New Dale-Chall Readability Formula," Brookline Books, Cambridge, MA. 1995.

[2] K. Collins-Thompson and J. Callan, "Predicting reading difficulty with statistical language models," Journal of the American Society for Information Science and Technology, Vol.56, No. 13, pp. 1448-1462, 2005.

[3] E. Dale and J. S. Chall, "A Formula for Predicting Readability," Educational Research Bulletin, Vol. 27, No. 1, 1948.

[4] M. Heilman, K. Collins-Thompson, J. Callan, and M. Eskenazi, "Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts," Proceedings of the Human Language Technology Conference. Rochester, 2007.

[5] M. Heilman, K. Collins-Thompson and M. Eskenazi, "An Analysis of Statistical Models and Features for Reading Difficulty Prediction," Proceedings of the Third ACL Workshop on Innovative Use of NLP for Building Educational Applications, pp. 71–79, 2008.

[6] J. Kincaid, R. Fishburne, R. Rodgers, and B. Chissom, "Derivation of new readability formulas for navy enlisted personnel," Branch Report, pp. 8-75. 1975.

[7] S. Schwarm and M. Ostendorf, "Reading level assessment using support vector machines and statistical language models," Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, 2005.

[8] J. Stenner, "Measuring reading comprehension with the Lexile framework," In Fourth North American Conference on Adolescent/Adult Literacy, 1996.