Maejo Int. J. Sci. Technol. **2015**, *9*(03), 355-369; doi: 10.14456/mijst.2015.27

Maejo International Journal of Science and Technology

ISSN 1905-7873
Available online at www.mijst.mju.ac.th

Full Paper

Assessing readability of Thai text using support vector machines

Yaw-Huei Chen and Patcharanut Daowadung*

Department of Computer Science and Information Engineering, National Chiayi University, Chiayi City 60004, Taiwan (R.O.C.)

* Corresponding author, e-mail: s0970384@mail.ncyu.edu.tw

Received: 13 February 2014 / Accepted: 3 November 2015 / Published: 16 November 2015

Abstract: The readability of a document is a measure of how easily the document can be read and understood. To select appropriate reading materials for children, techniques that can automatically assess readability are required. The objective of this study is to develop a machine-learning-based technique to assess the readability of Thai text. The experimental corpus, which was divided into training data and test data, consisted of articles selected from the textbooks of primary schools in Thailand. Documents in the corpus were first segmented into terms and then represented by feature vectors. Different combinations of feature sets including term frequencies of selected terms, shallow features and language model features were tested in the experiments. Classification and regression models were learned from the training data using support vector machines. Experimental results confirm that the proposed term-selection method can identify effective term frequency features for assessing the readability of Thai text.

Keywords: Thai readability, term frequency, feature selection, support vector machines

INTRODUCTION

Reading has an important role in learning for children because it can help them acquire knowledge and develop new ideas. However, articles with complex grammatical structure or difficult words may be overly complicated for children to comprehend. Children should read materials that are suitable to their reading ability. A task confronting schoolteachers is to choose appropriate reading materials for their students. The number of Thai articles available online is continuously increasing. This extensive number of digital Thai articles certainly improves the availability of reading materials for children, but it also increases the workload of the teacher in selecting suitable articles. The readability level of an article indicates how easily an article can be read and understood; therefore, it is reasonable to use the readability level as a major criterion to

Maejo Int. J. Sci. Technol. **2015**, *9*(03), 355-369; doi: 10.14456/mijst.2015.27

select appropriate reading materials for primary school students. Thus, we require an effective technique for assessing the readability of Thai text so that teachers can easily select appropriate reading materials.

Various techniques for assessing readability have been developed in the past, including both formula-based [1, 2] and machine-learning-based techniques [3, 4]. The majority of these techniques focus on English text. In Thai language the text is written without explicit word boundary delimiters, sentence endings or capital letters. For example, the word 'คนขับรถ' (kon-khup-rod) may refer to 'a driver' as a noun, 'a man drives a car' as a sentence, or a compound noun depending on the context where the word occurs. Word segmentation is a necessary pre-processing step in Thai text processing. Because of the fundamental difference between Thai and English, techniques developed for assessing readability of English text may not be effective for assessing the readability of Thai text.

In this paper a machine-learning-based technique is developed to assess the readability of Thai text. The proposed method predicts the reading levels of documents using support vector machines (SVM). Various features including term frequency features (TF), shallow features (SL) and language model features (LM) are extracted from the documents and are tested for their effectiveness for assessing readability. The documents are classified into six grade levels for students in primary school. A multiclass feature selection method is proposed to select the terms used for computing TF. In the experiments we selected 720 articles from the textbooks of primary schools in Thailand to form an experimental corpus. Feature selection methods based on mutual information and chi-square test were first evaluated and then different combinations of the feature sets were tested for assigning reading levels to the documents. The experimental results confirm that the proposed multiclass term-selection method can identify effective TF for assessing the readability of Thai text.

RELATED WORK

Readability formulas consisting of SL such as the average number of syllables per word and average number of words per sentence in a document were frequently used in early studies to predict the readability levels of the documents [2, 5, 6]. These formulas are usually simple and easy to calculate. However, complex words and long sentences do not always render a document difficult to read and therefore a simple readability formula cannot accurately predict the readability level of a document.

In addition to SL, word frequency is another common feature used for measuring readability. Chall and Dale [1] estimated the readability of a document using a combination of average sentence length and percentage of words occurring in a list of 3,000 familiar words identified manually. A document that contains fewer words in the common word list is likely to be more difficult. Stenner [7] combined the word frequency and sentence length features to generate a regression equation for predicting the difficulty of reading material. Heilman et al. [8] assessed reading difficulty using lexical features based on the frequencies of 5,000 common words in the training corpus and grammatical features derived from context-free grammar parses of sentences. They concluded that the combination of grammatical and lexical features was most effective. Chen et al. [9] calculated the term frequency and inverse document frequency of selected terms as features and applied SVM to assess the readability of Chinese text. The effectiveness of these methods depends primarily on the corpus from which the word list, the frequency information and the grammatical features are derived.

Maejo Int. J. Sci. Technol. **2015**, *9*(03), 355-369; doi: 10.14456/mijst.2015.27

Statistical language models compute the probability of the next word from the previous n - 1 words. Si and Callan [10] used unigram models to measure the reading difficulty of science web pages. Collins-Thompson and Callan [11] focused on using smoothed unigram-language models to predict the grade level of web documents; this approach demonstrated superior performance to traditional methods. Language modelling techniques have also been used to assess the readability of non-English texts. For example, Sato et al. [12] devised a character unigram model to measure the readability of Japanese text because each Kanji character in Japanese can be considered a single term.

With the ever-increasing computing power of modern computers, researchers have acquired the ability to use machine-learning-based techniques to assess the readability of documents. For example, Schwarm and Ostendorf [13] utilised SVM to combine features from traditional reading-level measures, parse trees and statistical language models for assessing the reading level. Vajjala and Meurers [14] tested various syntactic and lexical features on a corpus created from two web sources: Weekly Reader and BBC Bitesize. They found that a combination of development measures from second language acquisition research and traditional readability features significantly improved the performance of the classifiers. Francois and Miltsakaki [15] compared readability formulas with machine-learning-based methods for assessing the readability of French text. The best result was obtained when they used a combination of traditional readability features and new features derived from languages models, parse tree-based predictors and other measures.

PROPOSED METHOD

The problem of readability assessment has been studied for several languages. However, research on Thai readability remains in its initial stage. Preliminary experiments in our previous study indicated that using SVM to analyse the term frequency and inverse document frequency values of selected terms in Thai text is promising in classifying documents for primary school students [16]. Because the feature set used in a machine-learning-based approach is critical to the performance of a learned-text classifier, we propose to compare the effectiveness of prediction models with different feature sets which are derived from SL, LM and TF of selected terms. A machine-learning-based approach is applied to produce the prediction models for assessing the readability of Thai text. The proposed method consists of the tasks of Thai word segmentation and pre-processing, feature selection, feature value computation, and prediction model generation.

Thai Word Segmentation and Pre-processing

Owing to the lack of explicit word boundaries in Thai written text, word segmentation has a significant role in extracting terms for Thai language processing. Dictionary-based techniques and machine-learning-based algorithms are two well-known approaches for Thai word segmentation [17]. The former segments input text strings into words based on terms defined in a dictionary, which must contain an extensive number of terms for this approach to perform well. The latter learns a classification model from a training corpus to predict whether a character in the input text string is a word beginning. The performance of the classification model depends on the quality and size of the training corpus, where word boundaries are clearly identified.

In this study LexTo (Thai Lexeme Tokenizer) [18] was applied to segment text strings into words using both the longest matching and dictionary-based techniques; the former solves the ambiguity problem by selecting the longest matched word in the dictionary. LexTo provides a source code of the program and a dictionary (i.e. Lexitron [19]) containing approximately 40,000

Maejo Int. J. Sci. Technol. **2015**, *9*(03), 355-369; doi: 10.14456/mijst.2015.27

words. Because the performance of the word segmentation program can be improved by increasing the size of the dictionary, we added 5,000 proper names, organisations and places as new words into the dictionary. We also removed punctuation, numeric characters, special symbols and Thai number characters from the corpus.

Feature Selection

Feature selection is the process of selecting a subset of terms in the training data such that values computed from these terms can be used as features more effectively in the text classification. Several term-selection methods have been studied for text categorisation [20]. We compared mutual information and chi-square test as the feature selection methods for this study. Mutual information measures the importance of the presence or absence of term t in a document for prediction on class c. In a two-class classification problem mutual information is computed using equation (1), where N is the total number of documents and N_{ij} is the number of documents that contain term t ($i = 1$) or do not contain term t ($i = 0$) and are in class c ($j = 1$) or are not in class c ($j = 0$) [21]. For example, N_{11} is the number of documents that contain term t and are in class c. Note that '.' represents both '0' and '1', and therefore $N_{.1} = N_{01} + N_{11}$ is the total number of documents in class c.

$$MI = \frac{N_{11}}{N}\log_2\frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N}\log_2\frac{NN_{01}}{N_{0.}N_{.1}} + \frac{N_{10}}{N}\log_2\frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N}\log_2\frac{NN_{00}}{N_{0.}N_{.0}} \quad (1)$$

The chi-square test measures the independence of two events: the occurrence of the term and the occurrence of the class. In a two-class classification problem, the chi-square test value is computed using Equation (2), where N and N_{ij} are defined as in Equation (1) [21].

$$\chi^2 = \frac{N(N_{11}N_{00} - N_{10}N_{01})^2}{(N_{11}+N_{01})(N_{11}+N_{10})(N_{10}+N_{00})(N_{01}+N_{00})} \quad (2)$$

Because these feature selection methods are only suitable for two-class classification problems, we devised a new method to select the features for multiclass-classification problems. For each term, a one-against-all approach was employed to calculate the goodness measures of the term as a feature for k binary classifiers, where k is the number of classes in the training data. The ith binary classifier is trained with documents in the ith class as positive and all other documents as negative. For each classifier, we sorted the goodness measures of all the terms and assigned order numbers to the terms. Therefore, every term received an order number for each classifier. We chose the best order number in the k classifiers as the representing order number of the term. Then we sorted the representing order numbers and used the sorted list of terms for feature selection. Note that ties were broken arbitrarily when sorting the terms. A fixed number of terms were selected as features from the top of the sorted list. We conducted experiments to determine the feature selection method for calculating the goodness measures to be used in the experiments.

Feature Value Computation

Three different sets of features were used: SL, LM and TF. Because the Thai language has no explicit word boundaries, sentence endings or capital letters, many SL such as the average number of words per sentence, average number of sentences per paragraph and average number of syllables per word are difficult to extract. Therefore, as indicated in Table 1, we adopted features from Coh-

Maejo Int. J. Sci. Technol. **2015**, *9*(03), 355-369; doi: 10.14456/mijst.2015.27

Metrix-Port [3] as our SL; these include the average length per word, percentage of some connectives and percentage of words in word lists for different grade levels defined by the Office of the Basic Education Commission of Thailand.

Table 1. SL used in the proposed method

Number	Feature
1	Average word length
2	Ratio of 'และ' (and)
3	Ratio of 'หรือ' (or)
4	Ratio of 'ถ้า' (if)
5	Ratio of other connectives (i.e. 'แต่' (but), 'แต่กระนั้น' (yet), 'แม้' (although), 'แม้ว่า' (even), 'แต่ทว่า' (whereas), 'ทว่า' (albeit) and 'แต่ว่า' (however))
6	Ratio of words in word list of Grade 1
7	Ratio of words in word list of Grade 2
8	Ratio of words in word list of Grade 3
9	Ratio of words in word list of Grade 4
10	Ratio of words in word list of Grade 5
11	Ratio of words in word list of Grade 6

To compare the different aspects of Thai readability assessment, we used *n*-gram language models to assign probability measures to the word strings at each reading level. For any given text, each language model was evaluated through its perplexity defined by Equation (3), where $P(t \mid c)$ is the conditional probability of a word sequence of length m: $t = w_1, \ldots, w_m$ relative to class c. Because a lower perplexity indicates a higher probability, we can use the perplexity as a feature in the readability assessment task [13].

$$perplexity = P(t \mid c)^{-\frac{1}{m}} \tag{3}$$

Unigram, bigram and trigram were the language models that we used for capturing the term frequency and collocation information in the text. For a test document, we generated a perplexity value from each language model that was trained on documents in one of the six grade levels. Therefore, as shown in Table 2, there were eighteen perplexity values in the language model feature set.

Maejo Int. J. Sci. Technol. **2015**, *9*(03), 355-369; doi: 10.14456/mijst.2015.27

Table 2. Language model feature set

Number	Perplexity
1	Generated by unigram model trained on Grade 1 text
2	Generated by unigram model trained on Grade 2 text
3	Generated by unigram model trained on Grade 3 text
4	Generated by unigram model trained on Grade 4 text
5	Generated by unigram model trained on Grade 5 text
6	Generated by unigram model trained on Grade 6 text
7	Generated by bigram model trained on Grade 1 text
8	Generated by bigram model trained on Grade 2 text
9	Generated by bigram model trained on Grade 3 text
10	Generated by bigram model trained on Grade 4 text
11	Generated by bigram model trained on Grade 5 text
12	Generated by bigram model trained on Grade 6 text
13	Generated by trigram model trained on Grade 1 text
14	Generated by trigram model trained on Grade 2 text
15	Generated by trigram model trained on Grade 3 text
16	Generated by trigram model trained on Grade 4 text
17	Generated by trigram model trained on Grade 5 text
18	Generated by trigram model trained on Grade 6 text

The term frequency measure, which is proportional to the number of occurrences of a term in a document, can be used to evaluate the importance of the term to the document in a collection. We used a sublinear term frequency scaling method to compute the term frequency measure in Equation (4), where $wf_{t,d}$ is the modified term frequency value of term t in document d and $tf_{t,d}$ is the number of occurrences of term t in document d [21].

$$wf_{t,d} = \begin{cases} 1 + \log tf_{t,d} & \text{if } tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Prediction Model Generation

SVM is a supervised learning classification method that attempts to identify a maximum-margin hyper plane between two classes [22]. SVM works well on a large feature space in terms of both the accuracy of the classification results and the efficiency of training and classification algorithms. Moreover, SVM has generated good classifiers for many different types of datasets [23]. Because SVM is a binary classifier, it must be extended for solving multiclass-classification problems. In a one-against-one approach all possible $k(k - 1)/2$ two-class classifiers are first

Maejo Int. J. Sci. Technol. **2015**, *9*(03), 355-369; doi: 10.14456/mijst.2015.27

generated from a training set of k classes; then the class label of a test document can be determined using a voting strategy [24]. We used LIBSVM [25] to build the multiclass classifiers based on different combinations of feature sets. However, because the readability was divided into six reading levels corresponding to the six grades in primary school, the readability levels were continuous and exhibited a natural order. This ordinal-regression problem can be solved using new SVM formulations that are modified from two-class classification approaches [26, 28]. In the experiments we evaluated both multiclass-classification and ordinal-regression approaches.

EXPERIMENTS AND RESULTS

This section describes the environment, corpus and results of the experiments on the feature selection methods and feature sets. Both multiclass-classification and ordinal-regression learning techniques were evaluated in the experiments.

Experimental Environment

We developed JAVA-based programs to assess the readability of Thai articles. The applications were installed on a Windows-based PC with an Intel Core 2 Quad CPU and 8 GB of RAM. Thai words were segmented using the LexTo application. The corpus was stored in a MySQL database. LIBSVM, a machine-learning toolkit, was used for learning and testing the SVM prediction models [25]. A program modified from LIBSVM was used for performing ordinal regression in the experiments [27]. SRILM, a language modelling toolkit, was used for building and applying the n-gram models to the generation of LM [29].

Corpus

The corpus was selected from textbooks in six core subjects, namely 'occupations and technology', 'social studies, religion and culture', 'health and physical education', 'Thai language', 'arts', and 'science', which are mandatory courses for primary school students in Thailand. The articles were retrieved from textbooks in paper and digital formats. The *trueplookpanya* website provided content for all six subjects. *Max Education* provided content for Grades 4-6 of all six subjects, whereas another website provided articles in the subject of the Thai language. The articles were captured either by entering the texts manually or by copying the texts electronically.

One of the challenges in training and testing models for assessing readability is to correctly assign a reading level to the documents in the corpus. Originally, there were 1,080 articles retrieved from the textbooks. We invited five primary school teachers, each with more than ten years of teaching experience, to assess the reading levels of the articles. Only articles that were labelled with the same reading level by all five teachers were selected for inclusion in the corpus. We selected 720 articles to form the final corpus, where all the grade levels were equally represented with the same number of documents. The distribution of the 720 documents in the corpus is presented in Table 3.

Maejo Int. J. Sci. Technol. **2015**, *9*(03), 355-369; doi: 10.14456/mijst.2015.27

Table 3. Distribution of articles in the corpus

Grade	Number of documents	Number of words
1	120	22,844
2	120	39,298
3	120	40,973
4	120	44,866
5	120	91,540
6	120	93,553

Experiments on TF

Because even a small corpus may contain many thousands of unique terms, the learning algorithms must evaluate an enormous number of feature values when using TF. The dimensionality of the feature space can be reduced by employing feature selection methods to determine the most discriminative terms for classification and regression. To determine the number of terms to be used as features, we compared two feature selection criteria: mutual information and chi-square. For each term, we applied the one-against-all approach to determine the order of the term in each classifier according to the feature selection criterion. Then we selected the best order number in the six classifiers as the representing order number of the term. Finally, we sorted the representing order numbers of all terms and selected the terms from the top of the sorted list. The modified term frequency of the selected terms calculated from Equation (4) was used as the feature for training and testing the prediction models. Note that the feature values were scaled to the interval of numbers between 0 and 1 for both the training and test data in all the experiments.

The experimental results of the SVM-based approaches with linear kernel functions for multiclass classification and ordinal regression are presented in Tables 4 and 5 respectively. The performance of randomly selected terms is also included as baseline comparison. In the experiments the mutual information method generated the highest accuracy in all cases except for the '5,000 terms' in the ordinal-regression approach. Therefore, we selected mutual information as the feature selection criterion for the TF in the remaining experiments. The experimental results also showed that the multiclass-classification model provided superior performance compared to the ordinal-regression model. A reason for the inferior performance of the latter may lie in the reading levels assigned to the documents. Because these levels were manually rated, the evaluation may be imprecise and the difference in reading difficulty among levels may vary, which can cause performance degradation in the ordinal-regression model. Another possible reason for the resulting performance may be due to the lack of parameter adjustment. However, the experiments indicated that both multiclass-classification and ordinal-regression models had the highest or near-highest accuracy values when the number of terms was 4,000. Therefore, we used the 4,000-term frequency features in the experiments on feature set combination.

Maejo Int. J. Sci. Technol. **2015**, *9*(03), 355-369; doi: 10.14456/mijst.2015.27

Table 4. Accuracy of different feature selection criteria for multiclass classification

Number of terms	Mutual information	Chi-square	Random
100	41.389%	40.000%	22.083%
200	44.722%	39.722%	27.639%
300	46.111%	41.944%	29.583%
400	45.278%	44.167%	30.972%
500	46.111%	42.083%	33.056%
600	45.694%	43.889%	35.556%
700	46.528%	45.417%	32.222%
800	46.528%	45.833%	33.750%
900	47.222%	45.417%	34.583%
1000	46.528%	45.694%	36.111%
2000	47.361%	47.083%	38.056%
3000	49.167%	47.500%	42.222%
4000	50.278%	47.917%	44.028%
5000	49.444%	48.472%	45.000%

Table 5. Accuracy of different feature selection criteria for ordinal regression

Number of terms	Mutual information	Chi-square	Random
100	37.778%	34.861%	20.694%
200	37.361%	35.972%	24.583%
300	37.500%	35.139%	25.417%
400	37.083%	35.278%	28.472%
500	36.944%	34.444%	29.722%
600	39.167%	36.944%	30.000%
700	36.528%	36.111%	28.472%
800	36.250%	37.778%	29.861%
900	36.528%	36.111%	30.000%
1000	38.472%	35.694%	31.806%
2000	40.000%	37.778%	32.639%
3000	42.778%	39.028%	34.722%
4000	41.667%	39.861%	36.389%
5000	41.111%	42.361%	38.611%

Maejo Int. J. Sci. Technol. **2015**, *9*(03), 355-369; doi: 10.14456/mijst.2015.27

Experiments on SL

SL have been used in many traditional readability formulas [1-3, 5, 6]. In these experiments we investigated the performance of shallow features for assessing Thai text readability. The SL listed in Table 1, which include average word length, ratio of some connectives and ratio of words in word lists for different grades, were used as features for the SVM-based learning algorithms for multiclass classification and ordinal regression. Table 6 presents the average results of the experiments which used a 5-fold cross validation in terms of accuracy, mean absolute error, and squared correlation coefficient. The accuracy values of both the multiclass-classification and ordinal-regression models using SL were inferior to those of the models using TF as shown in Tables 4 and 5. The number of SL used in the experiments was very limited, which may have caused the inferior performance. However, it is difficult to extract additional SL owing to the characteristics of Thai text.

Table 6. Performance of 11 SL

	Accuracy	Mean absolute error	Squared correlation coefficient
Multiclass classification	29.306%	1.20694	0.31503
Ordinal regression	32.080%	1.09167	0.46422

Experiments on LM

To compare the different aspects of textual properties on the performance of Thai text readability assessment, we used *n*-gram language models to predict the probability that a particular word sequence would occur in an article. The language models were created with language modelling toolkit SRILM [29]. To avoid over-fitting by less informative terms in the corpus, we retained only the 400 terms with the highest mutual information and replaced the remaining terms with 'unknown' tags. Table 7 presents the average accuracy obtained from the 5-fold experiments. The performance of the LM was even poorer than that of the SL in these experiments.

Table 7. Performance of the 18 LM

	Accuracy	Mean absolute error	Squared correlation coefficient
Multiclass classification	27.361%	1.53056	0.21621
Ordinal regression	17.222%	1.95555	0.10219

Experiments on Combination of Feature Sets

We evaluated SL, LM and TF for both SVM-based multiclass-classification and ordinal-regression approaches. In these experiments various combinations of feature sets were also tested, viz. SL+LM, TF+SL, TF+LM, TF+SL+LM, and the set of TF of all terms. For these tests, the TF set contained 4,000 terms as suggested by the aforementioned experiments.

The experimental results are presented in Table 8. Among the eight feature sets, the LM had the lowest accuracy values for multiclass classification and ordinary regression. This result indicates

Maejo Int. J. Sci. Technol. **2015**, *9*(03), 355-369; doi: 10.14456/mijst.2015.27

that LM alone are not sufficient for assessing Thai text readability. However, the performance of the combined feature set SL+LM was superior to either the single feature set SL or LM. This outcome is similar to the conclusions drawn from other studies [10, 13]. As indicated in Table 8, prediction models with TF outperformed those without these features and the multiclass-classification model outperformed the ordinal-regression model. The addition of SL and/or LM did not improve the accuracy of the model, while the TF exhibited effective predictive capabilities for assessing the readability of Thai text. We also noticed that using all 14,205 terms as features (all terms) produced the best performance with lowest mean absolute error and highest squared correlation coefficient in the ordinal-regression model and good performance in the multiclass-classification model.

Table 8. Comparison of feature sets

Feature set	No. of features	Multiclass classification			Ordinal regression		
		Accuracy	Mean absolute error	Squared correlation coefficient	Accuracy	Mean absolute error	Squared correlation coefficient
SL	11	29.306%	1.20694	0.31503	32.080%	1.09167	0.46422
LM	18	27.361%	1.53056	0.21621	17.222%	1.95555	0.10219
TF	4000	50.278%	0.73611	0.60536	41.667%	0.79028	0.60575
SL+LM	29	31.528%	1.12917	0.33831	37.083%	0.99305	0.49733
TF+SL	4011	50.417%	0.72222	0.61741	42.500%	0.77917	0.61873
TF+LM	4018	50.139%	0.72639	0.61408	42.083%	0.78056	0.61551
TF+SL+LM	4029	50.972%	0.70972	0.62303	40.972%	0.79028	0.62355
All terms*	14205	49.306%	0.74028	0.60628	44.861%	0.70694	0.64837

* Set of TF of all terms

Experiments on Feature Sets with Different Numbers of Terms

Because the feature sets with a smaller number of features could generate comparative performance in less time, we tested the models with different numbers of terms as features in the subsequent experiments. Figures 1 and 2 illustrate the accuracy of the multiclass-classification and ordinal-regression models respectively with different numbers of terms. It can be observed that the four tested feature sets, i.e. TF, TF+SL, TF+LM and TF+SL+LM, generate similar accuracies in both Figures. Therefore, TF alone seem to be sufficient for assessing the readability of Thai text.

Maejo Int. J. Sci. Technol. **2015**, *9*(03), 355-369; doi: 10.14456/mijst.2015.27

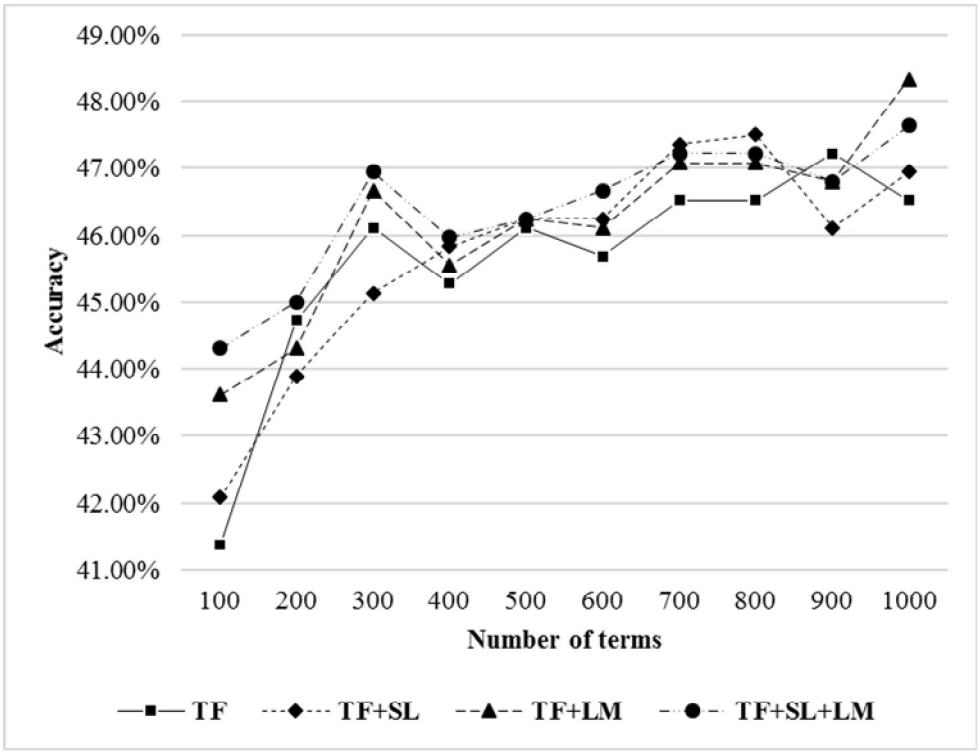

Figure 1. Accuracy for multiclass classification

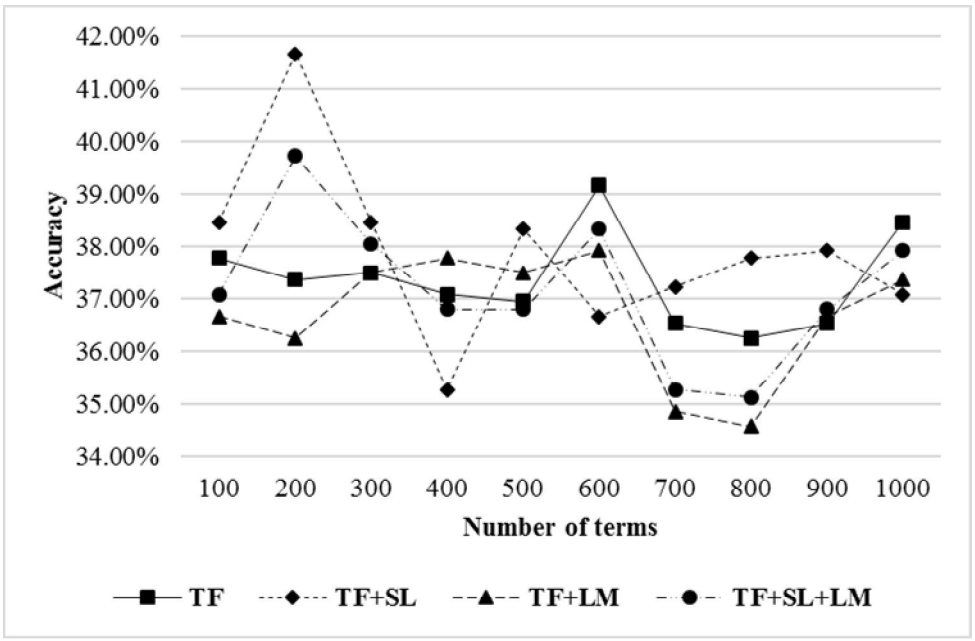

Figure 2. Accuracy for ordinal regression

DISCUSSION

SL have been used successfully for assessing the readability of English text, but their performance is not good for Thai text because we could extract only eleven SL. Although LM are expensive to compute, it did not perform well in the experiments. The combined feature set SL+LM provided an improvement in performance compared to either single feature set SL or LM, which is consistent with the results of previous studies [10, 13]. The experimental results suggested that the TF set performed well for all the models and the combined feature set TF+SL+LM did not provide

Maejo Int. J. Sci. Technol. **2015**, *9*(03), 355-369; doi: 10.14456/mijst.2015.27

noticeable improvement in performance. Because the terms were sorted by the proposed feature selection method, a relatively small number of terms were enough for generating the TF set for the classifier and we could process these TF very efficiently. In addition, it was demonstrated that the multiclass classification provided a higher accuracy compared to the ordinal regression in every experiment. Therefore, the multiclass-classification model with TF selected by the proposed feature selection method should be capable of effectively assessing the readability of Thai text.

There are still areas in the proposed method that require further investigation. For example, the quantity and quality of the training data which are crucial to the success of the supervised learning can be improved by increasing the number of documents from other sources (e.g. annotated web pages). Moreover, other types of features such as topic information may be capable of improving the performance. We expect to achieve superior accuracy in the future using new features such as name entity, discourse features and topic information. Based on the proposed method, a software tool using a larger training data could be developed for accessing the readability level of Thai text and teachers could easily use this tool to select suitable reading materials for primary school students.

CONCLUSIONS

The machine-learning-based method introduced in this study can effectively assess the readability of Thai text. Term frequency values generated from a small number of terms which are selected by the proposed feature selection method can form the term frequency feature set of a good classification model for accessing the readability. The performance of the term frequency feature set is superior to either a single shallow feature set or language model feature set and is comparable to combinations of feature sets. Additionally, the classification model with the term frequency feature set is computationally more efficient in comparison to other models.

ACKNOWLEDGEMENT

This work was supported in part by the National Science Council under Grant No. NSC102-2511-S-415-009.

REFERENCES

1. J. S. Chall and E. Dale, “Readability Revisited: The New Dale-Chall Readability Formula”, Brookline Books, Cambridge (MA), **1995**.
2. R. Flesch, “A new readability yardstick”, *J. Appl. Psychol.*, **1948**, *32*, 221-233.
3. S. Aluisio, L. Specia, C. Gasperin and C. Scarton, “Readability assessment for text simplification”, Proceedings of NAACL HLT 2010 5th Workshop on Innovative Use of NLP for Building Educational Applications, **2010**, Los Angeles, USA, pp.1-9.
4. S. E. Petersen and M. Ostendorf, “A machine learning approach to reading level assessment”, *Comput. Speech Lang.*, **2009**, *23*, 89-106.
5. J. P. Kincaid, R. P. Fishburne, Jr., R. L. Rogers and B. S. Chissom, “Derivation of new readability formulas (automated readability index, Fog count and Flesch reading ease formula) for navy enlisted personnel”, Research Report, **1975**, Institute for Simulation and Training, University of Central Florida, USA.
6. G. H. McLaughlin, “SMOG grading: A new readability formula”, *J. Reading*, **1969**, *12*, 639-646.

Maejo Int. J. Sci. Technol. **2015**, *9*(03), 355-369; doi: 10.14456/mijst.2015.27

7. A. J. Stenner, "Measuring reading comprehension with the Lexile framework", Proceedings of 4th North American Conference on Adolescent/Adult Literacy, **1996**, Washington, DC, USA.
8. M. Heilman, K. Collins-Thompson and M. Eskenazi, "An analysis of statistical models and features for reading difficulty prediction", Proceedings of 3rd Workshop on Innovative Use of NLP for Building Educational Applications, **2008**, Columbus, USA, pp.71-79.
9. Y.-H. Chen, Y.-H. Tsai and Y.-T. Chen, "Chinese readability assessment using TF-IDF and SVM", Proceedings of International Conference on Machine Learning and Cybernetics, **2011**, Guilin, China, pp.705-710.
10. L. Si and J. Callan, "A statistical model for scientific readability", Proceedings of 10th International Conference on Information and Knowledge Management, **2001**, Atlanta, USA, pp.574-576.
11. K. Collins-Thompson and J. Callan, "Predicting reading difficulty with statistical language models", *J. Am. Soc. Inform. Sci. Technol.*, **2005**, *56*, 1448-1462.
12. S. Sato, S. Matsuyoshi and Y. Kondoh, "Automatic assessment of Japanese text readability based on a textbook corpus", Proceedings of 6th International Conference on Language Resources and Evaluation, **2008**, Marrakech, Morocco, pp.654-660.
13. S. E. Schwarm and M. Ostendorf, "Reading level assessment using support vector machines and statistical language models", Proceedings of 43rd Annual Meeting of Association for Computational Linguistics, **2005**, Ann Arbor, USA, pp.523-530.
14. S. Vajjala and D. Meurers, "On improving the accuracy of readability classification using insights from second language acquisition", Proceedings of 7th Workshop on Building Educational Applications Using NLP, **2012**, Montreal, Canada, pp.163-173.
15. T. Francois and E. Miltsakaki, "Do NLP and machine learning improve traditional readability formulas?", Proceedings of 1st Workshop on Predicting and Improving Text Readability for Target Reader Populations, **2012**, Montreal, Canada, pp.49-57.
16. P. Daowadung and Y.-H. Chen, "Using word segmentation and SVM to assess readability of Thai text for primary school students", Proceedings of 8th International Joint Conference on Computer Science and Software Engineering, **2011**, Nakhon Pathom, Thailand, pp.170-174.
17. C. Haruechaiyasak, S. Kongyoung and M. Dailey "A comparative study on Thai word segmentation approaches", Proceedings of 5th International Conference on Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology, **2008**, Krabi, Thailand, Vol. 1, pp.125-128.
18. NECTEC, "LexTo: Thai lexeme tokenizer", **2007**, http://www.sansarn.com/lexto/ (Accessed: December 2013).
19. S. Klaithin, P. Chootrakool and K. Kosawat, "LEXiTRON-Pro editor: An integrated tool for developing Thai pronunciation dictionary", Proceedings of International Multiconference on Computer Science and Information Technology, **2010**, Wisla, Poland, pp.429-433.
20. Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization", Proceedings of 14th International Conference on Machine Learning, **1997**, Nashville, USA, pp.412-420.
21. C. D. Manning, P. Raghavan and H. Schutze, "Introduction to Information Retrieval", Cambridge University Press, Cambridge, **2008**.
22. C. Cortes and V. Vapnik, "Support-vector networks", *Mach. Learn.*, **1995**, *20*, 273-297.
23. M. Reif, F. Shafait, M. Goldstein, T. Breuel and A. Dengel, "Automatic classifier selection for non-experts", *Pattern Anal. Appl.*, **2014**, *17*, 83-96.

Maejo Int. J. Sci. Technol. **2015**, *9*(03), 355-369; doi: 10.14456/mijst.2015.27

24. C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines", *IEEE Trans. Neural Netw.*, **2002**, *13*, 415-425.
25. C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines", *ACM Trans. Intell. Syst. Technol.*, **2011**, *2*, no. 27.
26. W. Chu and S. S. Keerthi, "Support vector ordinal regression", *Neural Comput.*, **2007**, *19*, 792-815.
27. H.-T. Lin and L. Li, "Reduction from cost-sensitive ordinal ranking to weighted binary classification", *Neural Comput.*, **2012**, *24*, 1329-1367.
28. F. Xia, L. Zhou, Y. Yang and W. Zhang, "Ordinal regression as multiclass classification," *Int. J. Intell. Ctrl. Syst.*, **2007**, *12*, 230-236.
29. A. Stolcke, "SRILM—An extensible language modeling toolkit", Proceedings of 7th International Conference on Spoken Language Processing, **2002**, Denver, USA, pp.901-904.