Stop Word in Readability Assessment of Thai Text

Patcharanut Daowadung and Yaw-Huei Chen
Department of Computer Science and Information Engineering
National Chiayi University
Chiayi, Taiwan
{s0970384, ychen}@mail.ncyu.edu.tw

***Abstract*—Teachers and parents may use readability to select appropriate learning materials for primary school students. This research constructs Thai stop word list and evaluates the impact of eliminating stop words on readability assessment of Thai text. The corpus contains 1,188 textbook articles used by students from grade 1 to grade 6. Word segmentation, stop word list extraction, and feature selection are the preprocessing tasks performed on the articles in the corpus. Then, term frequency and inverse document frequency (TF-IDF) of the selected terms are used as features for support vector machines (SVMs) to generate classification models. Experimental results show that F-measure can reach 0.87 when identifying Thai articles suitable for middle grades primary school students.**

Keywords-readability; stop word list; mutual information; TF-IDF; SVM

I. INTRODUCTION

Reading is important for learning because children may obtain knowledge and develop new ideas from reading textbooks and Websites. However, it is not easy for children to understand articles that contain complex grammatical structure and/or difficult vocabulary (e.g., Pali and Sanskrit languages in Buddhist text [6]). Readability of an article indicates how easy the article can be read and understood. Readability levels are frequently used as a main criterion for teachers and parents to select appropriate learning materials for primary school students.

A variety of prediction techniques for readability of English articles have been developed during the past decades, such as SMOG grading formula [10], language models [4], and machine learning techniques [11]. Thai language is very different from English. For example, there are no explicit word boundary delimiters and capital letters in Thai language. Since the number of articles in Thai available online is continuously growing, it is necessary to have readability assessment tools specifically designed for Thai text. A machine learning based approach has been proposed for assessing readability of Thai text [4].

Stop words, which frequently occur but do not convey significant semantics in documents, can be eliminated in many fields on natural language processing. In bag-of-visual-word model, removing stop words can improve the performance of semantic concept detection in large-scale multimedia corpus [8]. In text classification, eliminating of stop words can improve the classification results [2]. Moreover, it has been shown that removing tailored stop words in latent semantic indexing based information retrieval systems can enhance retrieval performance [12]. Stop words are domain dependent. For example, common stop words "from" and "to" may have important semantics in travel business [7].

This research generates stop word lists and evaluates the impact of stop words elimination on readability assessment of Thai text. The proposed method consists of the following steps: using Ling CU word segmentation program to identify terms [1]; generating stop word list to eliminate terms that do not convey significant semantics; computing mutual information to select terms [9]; combining TF-IDF values of selected terms to produce a feature vector for each document [9]; establishing SVM prediction models in the training phase; classifying articles in the testing phase. Experimental results show that eliminating stop words may improve the performance of middle grades classifier but decrease the performance of lower grades classifier.

The remainder of the paper is organized as follows. Section 2 describes the proposed method. Section 3 presents the experimental results. Finally, we conclude the paper in Section 4.

II. READABILITY ASSESSMENT

Stop word elimination is incorporated into the process of readability assessment of Thai text. All articles must be preprocessed by a word segmentation program so that individual terms can be identified for further processing. We find stop words in the training data and use the stop word list to remove the terms that usually do not convey important semantics. After that, mutual information is used as a feature selection method to select a subset of terms that can most likely discriminate articles between classes. TF-IDF values of the selected terms are calculated as feature vectors to represent both training and testing articles.

We use LIBSVM [3] machine learning toolkit to train the classification models, where cross-validation is used to tune the parameters. When building the classifier for lower grades articles, we use articles from textbooks of grade 1 and grade 2 as positive training data and those of grade 3 to grade 6 as negative training data. The classifier for middle grades articles uses articles for grade 1 to grade 4 as positive training data and those for grade 5 and grade 6 as negative

978-0-7695-4702-2/12 $26.00 © 2012 IEEE
DOI 10.1109/ICALT.2012.9

training data. The classification results are obtained by processing the testing articles in the same way and feeding the feature vectors to the classifiers generated in the training phase.

A. Thai Word Segmentation and Preprocessing

Because there is no explicit word boundary in Thai language, word segmentation plays an important role in extracting terms for advanced language processing. In this research, we use program Ling CU to segment both training and testing articles [1]. The accuracy of this word segmentation step can reach 98%.

B. Building Thai stop word list

A stop word is a word that has frequent occurrences in the text but does not bear significant information. Stop words have little effect on identifying classes of articles, and therefore may be removed in advance. A combination of a statistical model and an information model are applied to identify stop words, where the statistical model calculates the probability and distribution of word occurrences and the information model measures the significance of a word using entropy. The mean of probability (*MP*), variance probability (*VP*), and a ratio of the two values of a single word are defined in (1), (2), and (3), respectively [13].

$$MP(w_j) = \frac{\sum_{1 \le i \le N} P_{i,j}}{N} \tag{1}$$

$$VP(w_j) = \frac{\sum_{1 \le i \le N} (P_{i,j} - \overline{P}_{i,j})^2}{N} \tag{2}$$

$$SAT(w_j) = \frac{MP(w_j)}{VP(w_j)} \tag{3}$$

Equation (1) measures the mean of probability of one word in individual document, where $P_{i,j}$ is the number of occurrences of word w_j in document i divided by the total number of words in document i and N is the total number of documents. Equation (2) calculates the variance of the probability of a word. Equation (3) combines the two results into the statistical value. Information model measures the entropy of a word using (4) [13].

$$H(w_j) = \sum_{1 \le i \le N} P_{i,j} \times \log(\frac{1}{P_{i,j}}) \tag{4}$$

The aggregation step combines the statistical model with the information model. The words are first sorted by *SAT* value in ascending order. Then, each word in the sorted list is assigned with an *S* weight, which is the order of the word in the sorted list. Similarly, an *H* weight is also assigned to each word to indicate the order of the word in the list sorted by ascending *H* value. Finally, we sort the words by the summation of *S* and *H* weights in ascending order. The resulting list is the stop word list used in this research.

C. Feature Selection

Because there are more than 11,000 terms in the corpus, it is time-consuming to encompass all of them as features for the machine learning algorithm. More important, using all terms as features does not guarantee satisfactory results. Therefore, we need a feature selection method to reduce the number of features. Mutual information reveals the importance of the presence or absence of a term in distinguishing a class, so we can use it to analyze the terms in the training data and find the most important subset of terms for classification. The appropriate size of the subset will be decided in the experiments.

D. TF-IDF Calculation

The TF-IDF measure can be used to evaluate the importance of a term to a document in a collection. It is proportional to the number of occurrences of the term in the document, and inversely proportional to the number of documents that contain the term in the collection.

III. Experiment and Results

This research evaluates the effect of eliminating stop words on the assessment of readability. We generate classifiers for articles suitable for lower grades and middle grades students. Precision, Recall, and F-measure are the measures of performance used in the experiments.

A. Corpus

The corpus used in the experiments consists of textbooks in both printed and digital formats. We collect articles in seven subjects from textbooks used by primary school students: Mathematics; Occupations and Technology; Social Studies, Religion and Culture; Health and Physical Education; Thai Language; Arts; and Science. We randomly choose 80% of the articles as the training set and the rest 20% to be the testing set, resulting in 950 articles for training and 238 articles for testing.

B. Stop word Generation

The above mentioned method for generating stop words is applied on the preprocessed articles to get the statistical model and the information model. The two models are aggregated together to construct Thai stop word list. It shows that most of the Thai stop words have comparable stop word in English, such as ที่ (at), มี (have), และ (and), and เป็น (be).

C. Classifier for Lower Grades

To create a classifier for articles suitable for lower grades students, we use articles of grade 1 and grade 2 in the training set as positive data, a total of 298 articles. The other 652 articles of grade 3 to grade 6 in the training set are used as negative data. Using the same criteria, we get 78 positive articles and 160 negative articles from the testing set. We remove various numbers of stop words and then test different numbers of terms selected to investigate their effects on classification performance. As shown in Fig. 1, the highest

F-measure 0.76 in the testing result appears when no stop word being removed using 600 terms. The results indicate that we should not remove stop word for lower grades classifier.

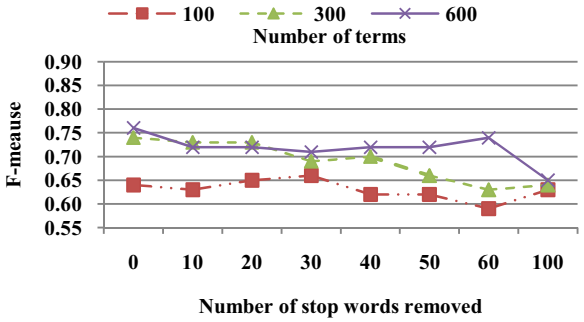

Figure 1. Testing result of predicting readability for lower grades

D. Classifier for Middle Grades

When building the classifier for articles suitable for middle grades students, we use articles of grade 1 to grade 4 in the corpus as positive data and those of grade 5 and grade 6 as negative ones. There are 646 positive articles and 304 negative articles for training, and 167 positive articles and 71 negative articles for testing. In testing result, as shown in Fig. 2, F-measures reach their peaks at 0.87 when using 600 terms and removing 30, 40, and 50 stop words. The overall performance of the classifier for the testing data set is above 0.82 in F-measure. These findings suggest that we can remove 30 to 50 stop words for middle grades classifier.

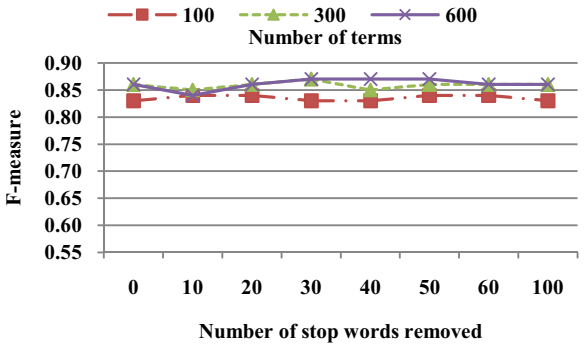

Figure 2. Testing result of predicting readability for middle grades

IV. Conclusions

In this research, we evaluate the impact of removing stop words on the performance of readability assessment of Thai text. The texts are first segmented into individual terms, and then a stop word list is constructed for eliminating stop words from the corpus. Mutual information is the method used to select top-ranking terms and a TF-IDF value vector of the selected terms is computed for each document. Finally, SVM is used to generate prediction models for assessing the readability of Thai text using these feature vectors. Experiments on a corpus of textbooks in seven subjects for primary school have been conducted. The results show that the F-measure of the classifiers using stop word elimination can reach 0.87 for middle grades. However, as shown in the experimental results, stop words should not be removed for the lower grades classifier.

Acknowledgment

This work was supported in part by the National Science Council under the Grant NSC99-2511-S-415-007-MY2.

References

[1] W. Aroonmanakun, "Collocation and Thai Word Segmentation," *the Fifth Symposium on Natural Language Processing & The Fifth Oriental COCOSDA Workshop*, Thailand, 2002, pp. 68–75.

[2] H. Ayral and S. Yavuz, "An automated domain specific stop word generation method for natural language text classification," *2011 International Symposium on Innovations in Intelligent Systems and Applications (INISTA)*, Istanbul 2011.

[3] C. C. Chang and C.-J. Lin, "LIBSVM - a library for support vector machines," http://www.csie.ntu.edu.tw/~cjlin/libsvm /index.html, 2011.

[4] K. Collins-Thompson and J. Callan, "Predicting reading difficulty with statistical language models," *J. Am. Soc. Inf. Sci. Technol.,* vol. 56, pp. 1448-1462, 2005.

[5] P. Daowadung and Y.-H. Chen, "Using word segmentation and SVM to assess readability of Thai text for primary school students," in *2011 Eighth International Joint Conference on Computer Science and Software Engineering (JCSSE),* Nakhon Pathom ,Thailand, 2011, pp. 170-174.

[6] E-book: Office of the Basic Education Commission, http://www.obec.go.th

[7] E. Dragut, F. Fang, P. Sistla, C. Yu, and W. Meng, "Stop word and related problems in web interface integration," *Proc. VLDB Endow.*, *vol. 2*, pp. 349-360, 2009.

[8] Y. G. Jiang, J. Yang, C.W. Ngo, and A.G. Hauptmann, "Representations of keypoint-based semantic concept detection: A comprehensive study," *IEEE Transactions on Multimedia*, *vol. 12*, pp. 42-53, 2010.

[9] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press, 2008.

[10] G. H. McLaughlin, "SMOG grading - a new readability formula," *Journal of Reading*, *vol. 12*, no. 8, pp. 639-646, 1969.

[11] S. E. Petersen and M. Ostendorf, "A machine learning approach to reading level assessment," *Computer Speech & Language*, *vol. 23*, pp. 89-106, 2009.

[12] A. Zaman, P. Matsakis, and C. Brown, "Evaluation of stop word lists in text retrieval using Latent Semantic Indexing," *2011 Sixth International Conference on Digital Information Management (ICDIM)*, Melbourne, Australia, 2011, pp. 133-136.

[13] F. Zou, F. L. Wang, X. Deng, S. Han, and L. S. Wang, "Automatic construction of Chinese stop word list," *Proceedings of the 5th WSEAS international conference on Applied computer science*, Hangzhou, China, 2006.