A Comparative Study of Text Categorization Methods on a Multilingual Corpus

Nguyen GIANG SON† Shigeru OYANAGI† Katsuhiro YAMAZAKI†
Graduate School of Science and Engineering†
Ritsumeikan University

1 Introduction

Text categorization (TC) plays an important role in text processing tasks of information systems. Automatic TC techniques are mainly based on machine learning methods, which are mainly studied for English. Nowadays, multilingual information systems are very popular such as electronic news papers. Hence classification on multilingual corpora is needed. In this research, classification methods, which are Naive Bayes classifier and Support Vector Machine (SVM) classifier, are tested on a multilingual comparable corpus which has documents in both language Vietnamese and English to study their performance.

This paper is structured as follows. Section 2 describes briefly background information of classifiers and performance measures. Section 3 describes the approach to multilingual classification. Section 4 details the data of the experiment, results, and discussion. The last section is the conclusion and future work.

2 Classifiers and Performance Evaluation

2.1 Classifiers

This section describes a brief background of the classifiers Naive Bayes and SVMs used in this experiment.

Naive Bayes Classifier

Naive Bayes classifiers are based mainly on Bayes assumption, which is that words of a document are independent. Conditional probabilities of words with a given class are calculated by using specific event models. There are two event models, Bernoulli event model and multinomial event model. We chose the multinomial model for our experiment because it has better performance.

Support Vector Machines

Main idea of SVMs is to solve the binary linear separable problem by finding the linear separating hyper-plane which maximizes the margin, the optimal separating hyper-plane. For solving nonlinear separable problems, kernel functions are used to transform problem space to linear separable derived feature space. We applied the Linear SVM, because it has been proved Linear SVM is very efficient for solving TC problems.

2.2 Performance Measures

Our experiments adopt commonly used performance measures, including the micro-recall, micro-precision, and micro F1 measure to evaluate classifiers on multi-class problems.

3. The Approach to Multilingual Text Classification

There are three approaches for multilingual classification as in [4]. They are language-neutral document processing and a single classifier, language-specific document processing and single classifier, and language-specific document processing and independent classifiers. The third approach is chosen because Vietnamese has very different morphological structure in comparison with English. Vietnamese words can not be separated by spaces. A Vietnamese word may be combined of one or more syllables. The chosen approach also is more suitable in terms of performance, because feature vector size is smaller by specific language vocabulary.

4. Experiments and Evaluation

4.1 Multilingual Corpus

To make a comparison for classifiers, they must be trained and tested on a comparable corpus. Comparable corpus is a collection of texts in different language regarding similar topics in the same period such as collections of news published by an agency.

There is no standard datasets for the experiment. Therefore news articles were crawled from Vietnam News Agency website (http://www.vnagency.com.vn/) including English and Vietnamese. Articles' range is from August to December. The number of categories is 9, the same on both languages.

	Aug	Sept	Oct	Nov	Dec
Vietnamese	1436	1571	1728	1682	1474
English	1227	1079	1383	1299	1159

Table 1. News articles by month and language.

In the crawled data, some topics are combined of two topics organized by the agency. The corpus is not well balanced by topics. World (about 30%), Society & Education (16%), and Business & Finance (18%) have considerable amount of news. Vietnamese has more news articles than English and Vietnamese news articles are not mainly translated into English. Articles are separated by their language. Therefore the language identification step for classification is omitted.

4.2 Feature Extraction and Selection

Steps for preprocessing English documents are: tokenization (extracting words to build feature vectors), removing stopwords, and stemming using Porter stemming algorithm.

For Vietnamese documents, word segmentation is executed by a segment tool as in [3] which has performance of F1 measure 0.9409. Stop words are also removed by a stopword list. There is no stemming step for Vietnamese.

For both languages, standard term weight TFxIDF is used for Support Vector Machine classifier. Term weights are normalized to the scale [0, 1] by cosine normalization. The Naive Bayes classifier uses term frequency.

In feature selection step, number, date, symbols are excluded from feature vectors. Feature selection by Document Frequency is chosen because its computation is not expensive like X^2 and has quite good competitive result as in [2]. The size of the feature vector is 3000, which is medium.

4.3 Experimental Results and Discussion

Figure 1 and 2 describes the classifiers' experimental results with F1 measure. The horizontal axis of graphs presents size of training datasets increasing by months. Articles from August to November were used for building training datasets. The test dataset is articles in December.

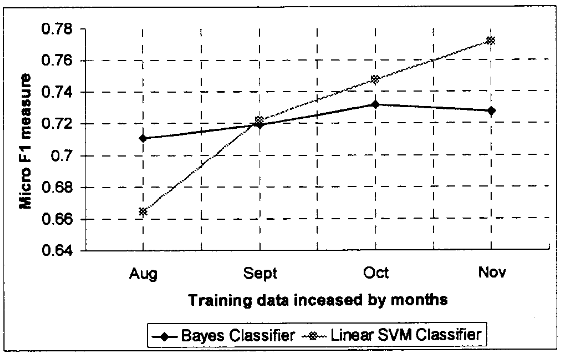

Figure 1. Learning curve for English

From the figure 1, the test of English documents shows that Linear SVM has lower F1 measure on the first training dataset of August. Then it outperformed the Naive Bayes classifier and get maximum at 0.7715. While the Naive Bayes Classifier slow down a little to 0.7273 for the training data range from August to November.

Figure 2 shows the test results of Vietnamese documents. SVM always outperformed Bayes classifier and gets maximum value at 0.7558 F1 measure.

In both cases, the Naïve Bayes classifier's performance did not change much only about 3 percent when the training data set size increased. In contrast, SVM classifier performance increased with average 10 percent.

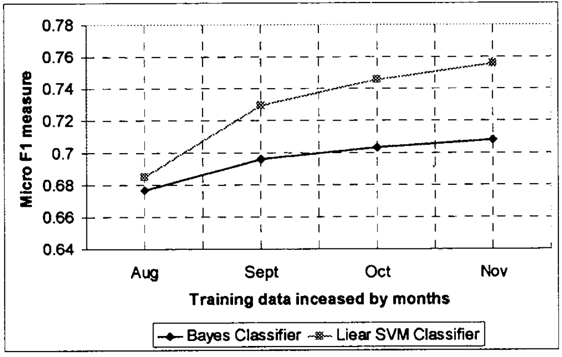

Figure 2. Learning curve for Vietnamese

The results of both languages were not so high mainly possibly due to ambiguity between topics in the corpus. The ambiguity is that documents contain phrases express both topics. For example, region topic and world topic can easily share a certain amount of documents. There also are some other minor factors which could affect to the accuracy such as word spelling in both languages and normalization in Vietnamese (different tone mark position, foreign proper nouns, and so on). One example of Vietnamese normalization issues is the translation of foreign proper noun by different ways due to different phonetics translations or remaining the origin words.

5 Conclusions and Future Work

The performances of both Naive Bayes classifier and Support Vector Machine classifier were investigated on a multilingual comparable corpus. The performances were not so high in both languages because of the ambiguity of topics. The performances were slightly different by languages. In future, to invest further, the experiment should be tested with a multilingual corpus which has more than two languages like having additional languages such as Japanese.

References

[1] Fabrizio Sebastiani. Machine learning in automated text categorization. In *ACM Computing Surveys*. 2002

[2] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 412-420, Nashville, US, 1997.

[3] Cam-Tu Nguyen and Xuan-Hieu Phan, "JVnSegmenter: A Java-based Vietnamese Word Segmentation Tool", *http://jvnsegmenter.sourceforge.net/*, 2007.

[4] J. J. García Adeva, D. López de Ipiña, R. Calvo, Multilingual Approaches to Text Categorisation, The European Journal for the Informatics Professional, 2005.