

Development of a Multi-Classifer Approach for Multilingual Text Categorization

Chung-Hong Lee, Hsin-Chang Yang, Ting-Chung Chen and Sheng-Min Ma

Abstract - Research work related to applying text categorization methods to a monolingual corpus such as English text collections has been well established by several research teams in recent years. However, little attention has been paid to applying the techniques to classify the documents in multiple languages such as English and Chinese by means of a unified model. In this paper we propose a multi-classifier system platform to enable multilingual documents be effectively categorized. First, we utilized a number of selected corpora in multiple languages collected from internet to train several text classifiers based on the Support Vector Machines (SVM) model. Subsequently, the multilingual texts of unknown category were classified by the trained classifiers. Finally, we evaluated our experimental results by accuracy, recall, precision, and F1 measures. The preliminary results show that our platform model has the potential for multilingual text categorization.

I. INTRODUCTION

THE information available in languages more than one single language (like English) in the global information systems is increasing significantly. Users of internationally distributed information networks need tools and methods that will enable them to discover, retrieve and understand relevant information, in whatever language and form it may have been stored. This drives a convergence of numerous interests from diverse research communities focusing on the issues related to multilingual information discovery. Multilingual text categorization is a novel topic that was, in some sense, larger than the problem of cross-language information retrieval and classic text categorization. Nowadays, the issues of multilingual information access and discovery have been studied by researchers from several different disciplines, including: digital library, information retrieval, and computational linguistics. Research work related to applying text categorization methods to a monolingual corpus such as English text collections has been well established by several research teams. However, little attention has been paid to applying the techniques to handle the documents in multiple languages such as English and Chinese, by means of a unified model to support classifying multilingual documents. One

major reason is associated with the nature of language used in the texts. Chinese differs from English in various ways that impact upon text mining techniques. Words, for example, in written text commonly consist of one, two, three, or four characters, or sometimes more, but there are no word spaces. Thus, words cannot be identified in the trivial manner applicable in English. Previous text categorization techniques discussed in the literature appear to lack significant capabilities required to handle the multilingual information sources. Therefore, in this research we set out to explore whether there is a method for text categorization approach applicable to multilingual (English and Chinese) information sources. We developed a unified technique to tackle the language difficulties in discovering the implicit classification knowledge from multilingual text collections. It is also applicable to the development of multimedia medical databases, through combining text and content-based methods for retrieval, and tackling the domain-specific medical terminology as well as notes of varying quality in mixed target languages.

In this work, the primary thrust is to make the text categorization techniques fully multilingual, classifying open-source texts in different languages. However our goal in these experiments is not to establish the ideal full-functional multilingual information categorization system, as the time and resources required for this task are considerable. Rather, we restrict our attention to bilingual corpora and try to understand the basic requirements for effective multilingual information categorization and the problems that arise from a simple implementation of such a system. This research work is mainly carried out by experimenting with text extraction from parallel corpora, a variation of bilingual corpora. Bilingual corpora can be used in many ways: For multilingual information access, bilingual corpora make it possible to investigate syntactic, semantic and lexical relationships between languages and are also important sources of contrastive evidence of language in usage. Also, for multilingual text categorization, in this work we expect it to be acted as a starting point for exploring the impacts on linguistics issues with the machine learning approach to discovering sensible linguistics elements from multilingual text collections. It is believed that, if the corpus is large enough, then statistical or other algorithm-based techniques can be used to produce bilingual text or term equivalents as well as associations by comparing which strings co-occur in the same sentences in the resulting categorized texts over the whole corpus.

Chung-Hong Lee is with the Department of Electrical Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan (e-mail: leechung@mail.ee.kuas.edu.tw).

Hsin-Chang Yang is with the Department of Information Management, Chang Jung University, Tainan, Taiwan (e-mail: hcyang@mail.cju.edu.tw).

Ting-Chung Chen is with the Department of Electrical Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan (e-mail: tinchung@dml.ee.kuas.edu.tw).

Sheng-Min Ma is with the Department of Electrical Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan (e-mail: smma@dml.ee.kuas.edu.tw).

II. RELATED WORK

Automatic text categorization is normally performed by supervised learning techniques, where pre-defined category labels are assigned to documents based on the likelihood suggested by a training set of labeled documents. Many learning algorithms such as k-nearest neighbor (k-NN)[9][14], Rocchio [2][12], Support Vector Machines (SVM) [6][12], neural networks [11][13], linear least squares fit (LLSF) [15], Winnow[2], and Naive Bayes (NB)[7][10] have been applied to text classification. A comparison of these techniques is addressed by Yang and Liu [14]. They conclude that all these approaches perform comparably when each category contains over 300 documents. However, when the number of positive training documents per category is less than 10, SVM, k-NN, and LLSF outperform significantly neural networks and NB. These techniques did provide state-of-the-art learning approaches to represent a viable and well-performing solution for monolingual categorization problems. Unfortunately, little attention has been paid for practical applications which need applying text categorization approaches to multilingual information sources in the real-world scenario.

Multilingual text categorization is a relatively new research topic, about which not much previous work in the literature appears to be available. Still, it concerns a practical problem, which is increasingly felt in some application fields, such as the documentation departments of international organizations as they come to rely on automatic text classification. It is also manifest in many news sites on the web, which rely on a quick classification of multinational news information. Adeva [1] provided a review of methods related to multilingual text categorization (i.e. Spanish and Basque) . They compared different feature extraction strategies such as n-gram-based stemming and classic stemming in preprocessing of multilingual documents. On the other hand, they also compared performance of different classification methods in multilingual text categorization such as Naïve Bayes, Rocchio, and k-nearest neighbor. Jalam [5] proposed an original framework for multilingual text categorization (i.e. English, French, and German) . Their framework contains two new steps including language identify and language translation. They applied their framework to classifying news articles which were written in different languages. Chou [3] proposed a concept-based approach on multilingual text categorization using fuzzy techniques. They integrated fuzzy c-means and fuzzy k-nearest neighbor algorithms to implement their system. The goal of their system is developed to solve language-independent problem, and all multilingual documents will be mapped into a common semantic space. Also, Chou [4] provided a method for multilingual text categorization by self-organizing map (SOM) and hierarchical clustering algorithms. First, they developed a universal concept space, in which the relationships of multilingual terms can be discovered by self-organizing map

(SOM) and hierarchical clustering algorithms. Second, they implemented a concept-based classifier using a 3-layer feed-forward neural network to carry out a concept-based multilingual text categorization task.

III. AUTOMATIC TEXT CATEGORIZATION BASED ON SUPPORT VECTOR MACHINES

In this work we developed a multi-classifier technique to enable multilingual documents be effectively categorized. We utilized a number of selected corpora in various languages to train the classifiers based on the *Support Vector Machines* (SVM) model. SVM is a relatively new learning technique for data classification. The goal of SVM is to find a decision surface to separate the training data samples into two classes and make decisions based on the *support vectors* that are selected as the only effective elements from the training set. For text classification, SVM makes decision based on the globally optimized separating hyper-plane. It simply finds out on which side of the hyper-plane the test pattern is located (see Figure 1). This characteristic makes SVM highly competitive, compared with other pattern classification methods, in terms of predictive accuracy and efficiency. In particular, Joachims has done much research on the applying SVM to text categorization [6].

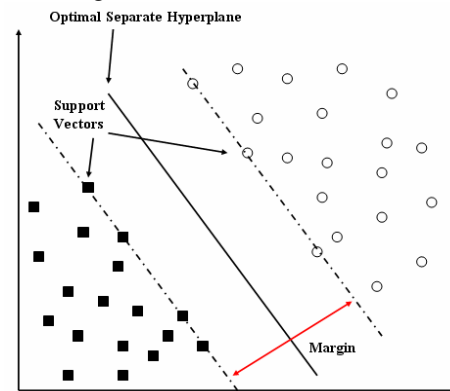


Figure 1. Structure of SVM classifier

On the other hand, if training samples can't be separated into two classes, SVM maps the samples into a high dimension feature space and find a decision surface to separate the training samples into two classes (see Figure 2). Such mappings can be performed by several kernel functions, such as Radial Basic Function (RBF) and polynomial kernels, as shown in Equation (1) and Equation (2).

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad (1)$$

$$K(x_i, x_j) = (x_i \cdot x_j + 1)^d \quad (2)$$

Equation (1) is Radial Basic Function (RBF) kernel, γ is one parameter of the kernel. Equation (2) is polynomial kernel function, d is degree of the kernel.

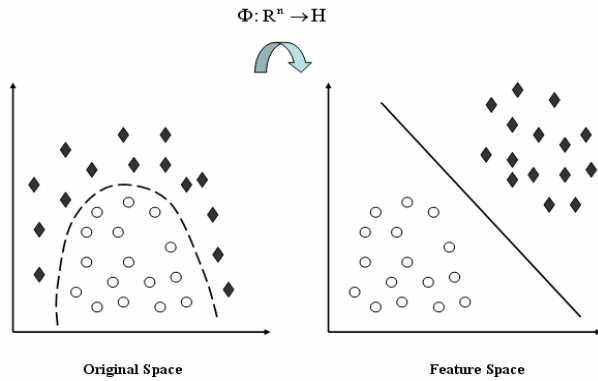


Figure 2. Kernel Mapping of SVM classifier

IV. SYSTEM FRAMEWORK

In this section we describe our system framework. As shown in Figure 2, in this work we employ Chinese and English corpora as training sources for constructing the multi-classifier system. After training process, we used several unlabeled texts written in Chinese and English to evaluate the performance of categorization.

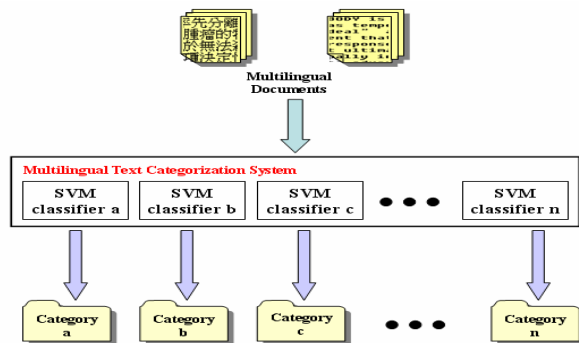


Figure 3. System framework

The original idea of our framework is to assist people who try to categorize specific types of multilingual documents related to certain known fields. It is not designed to the general public who try to find information without much knowledge apriori most of the time. As a framework prototype, therefore, we used only six selected categories of texts in English and Chinese to train six Support Vector Machines (SVM) classifiers, which are believed to be sufficient for defining a fundamental system model for testifying the theory of multilingual text categorization. This method works well if the original number of classes are limited, however, we allow the number of classes to be increased by the use of our system to a certain amount. We constructed SVM classifiers with a one-against-all (OAA) learning strategy to implement our multi-classifier system, as shown in Figure 4. Although there were still several learning strategies such as one-against-one (OAO) and DAGSVM methods applicable, the OAA technique selected for our implementation was based on the tradeoffs of the total costs of the amounts of classifiers and system performance. The OAA technique allows the deployment of fewer classifiers to

achieve the functionalities of multi-class categorization and also obtain a reasonable performance in our application.

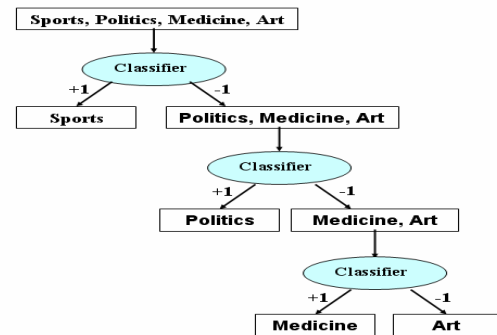


Figure 4. An example of One-Against-All

V. EXPERIMENTAL RESULTS

A. Preparation of the multilingual corpora

In this work we collected text corpora covering six different categories from various websites as shown in Table 1. In each category we collected 300 articles in Chinese and 300 articles in English respectively. The total amounts of texts in the corpora are 3,600. These articles collected in categories containing Politics, Art, Astronomy, Finance, Medicine, and Physics. Each category in the corpora includes a number of articles in either Chinese or English languages.

TABLE I
SOURCES OF THE CORPORA

Source	Website
CAN News	http://www.cna.com.tw/
UDN News	http://udn.com/NEWS/main.html
Yahoo!	http://www.yahoo.com
Science American	http://www.sciam.com.tw/
Grolier online	http://go-passport.grolier.com/
Taiwan Panorama	http://www.taiwan-panorama.com/

B. Document Preprocessing

In this work, each training document in either Chinese or English in the corpus is deconstructed into a bag of Chinese and English terms respectively. The Chinese-text part of the corpus was first preprocessed by the word segmentation program as in previous experiments. For English documents, the document preprocessing still begins with term indexing, establishing stop lists, stemming, and then encoding documents with binary vectors, in which each component corresponds to a different word, and the value of the component reflects the presence of word in the document. For preprocessing of Chinese documents, we employed CKIP Chinese Word Segmentation System [8] to process the Chinese documents.

C. Evaluation

In this section, the evaluation results of system performance are addressed. We used 1,000 labeled documents (i.e., 500 texts in Chinese and 500 ones in English) to train our classifier system, and 200 unlabeled documents (i.e., 100 texts in Chinese and 100 ones in English) to evaluate performance of system. We compare performance of the six classifier results by Accuracy, Recall, Precision, and F1 measures. As shown in Figure 5-8, the measures of the developed SVM classifiers with Linear SVM, Gaussian Radial Basis Function (RBF) and Polynomial kernels are illustrated. In the final paper, we will further discuss the implications of the experimental results mentioned above and compare the results with ones of other learning approaches to multilingual text categorization.

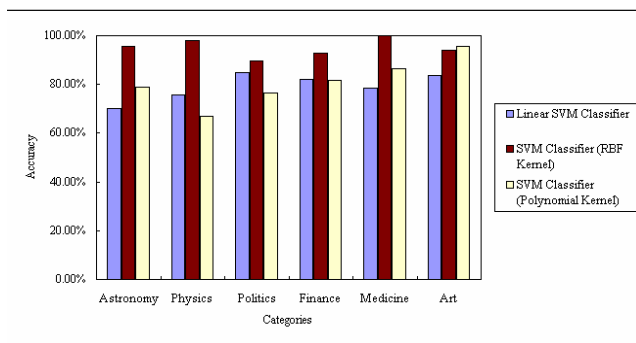


Figure5. Results of accuracy rate of developed SVM classifiers

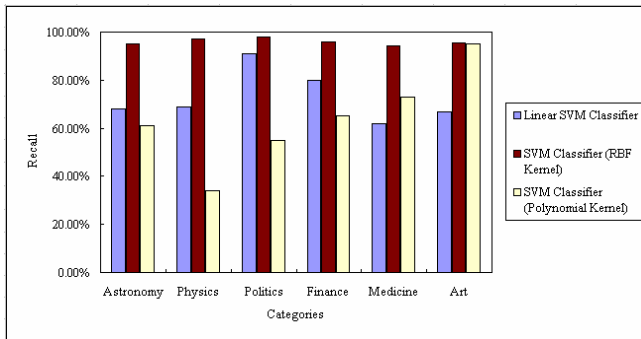


Figure6. Results of Recall measures of developed SVM classifiers

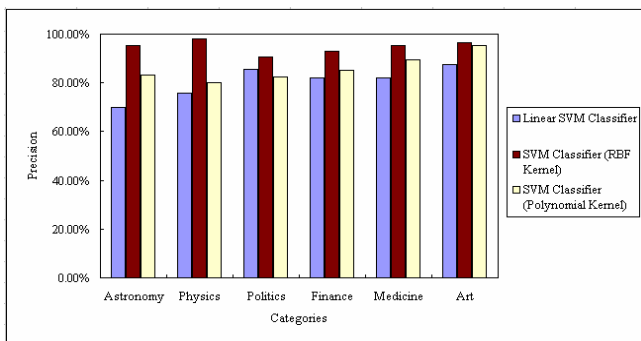


Figure7. Results of Precision measures of developed SVM classifiers

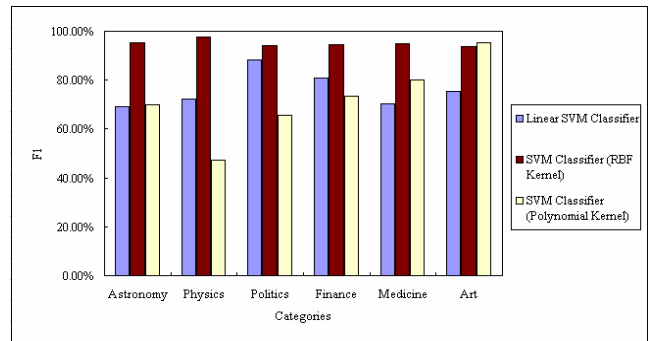


Figure8. Results of F1 measures of developed SVM classifiers

VI. CONCLUSION

In this paper we propose a multi-classifier system platform to enable multilingual documents be effectively categorized. First, we utilized a number of selected corpora in multiple languages collected from internet to train several text classifiers based on the Support Vector Machines (SVM) model. Subsequently, the multilingual texts of unknown category were classified by the trained classifiers. Finally, we evaluated our experimental results by accuracy, recall, precision, and F1 measures. The preliminary results show that our platform model has the potential for multilingual text categorization.

REFERENCES

- [1] Adeva, J. J.G., Calvo, R. A., and Ipiña, D. L. d. "Multilingual Approaches to Text Categorisation." *UPGRADE: The European Journal for the Informatics Professional*, Vol. 6, No.3, (2005) 43 - 51
- [2] Bel, N., Koster, C.H., and Villegas M.: "Cross-Lingual Text Categorization." In *Proceedings of Research and Advanced Technology for Digital Libraries, 7th European Conference, ECDL 2003*, Trondheim Norway (2003) 126-139
- [3] Chau, R., and Yeh, C.H. "Multilingual Text Categorization for Global Knowledge Discovery Using Fuzzy Techniques" *2002 IEEE International Conference on Artificial Intelligence Systems, 2002. (ICAIS 2002)*. 5-10 Sept. 2002. Page(s):82 - 86
- [4] Chau, R., Yeh, C.H. and Smith, K.A. "A Neural Network Model for Hierarchical Multilingual Text Categorization." *ISNN 2005, Second International Symposium on Neural Networks*, Chongqing, China, May 30 - June 1, 2005, *Proceedings, Part II. Lecture Notes in Computer Science 3497 Springer 2005, ISBN 3-540-25913-9*
- [5] Jalam, R., Clech, J., and Rakotomalala, R. "Cadre Pour la Catégorisation de Textes Multilingues." In C. Fairon, G. Prunelle, and A. Dister, editors, *7èmes Journées internationales d'Analyse statistique des Données Textuelles*, pages 650-660, Louvain-la-Neuve, Belgique, Marsh 2004. PUL.
- [6] Joachims, T., "Text Categorization with Support Vector Machines: Learning with Many Relevant Features", *European Conference on Machine Learning (ECML)*, 1998.
- [7] Koller, D. and Sahami, M. "Hierarchically Classifying Documents Using Very Few Words", *Proceedings of 14th International Conference on Machine Learning*, pp.170-178, Nashville, US, 1997.
- [8] Ma, W.Y. and Chen, K.J. "Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff", *Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing*, pp168-171.
- [9] Masand, B., Linoff, G. and Waltz, D. "Classifying News Stories Using Memory Based Reasoning", *15th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 59-64, 1992.

- [10] McCallum, A. and Nigam, K. "A Comparison of Event Models for Naive Bayes Text Classification", *AAAI-98 Workshop on Learning for Text Categorization*, 1998.
- [11] Ng, H. T., Goh, W. B. and Low, K. L. "Feature Selection, Perceptron Learning, and a Usability Case Study for Text Categorization", *20th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 67–73, 1997.
- [12] Sebastiani, F.: "Machine Learning in Automated Text Categorization." *ACM Computing Surveys*, Vol. 34, No. 1(2002) 1-47
- [13] Wiener, E., Pedersen, J. O. and Weigend, A. S. "A neural network approach to topic spotting", *Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval*, 1995.
- [14] Yang, Y. and Liu, X. "A Re-examination of Text Categorization Methods", *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, Berkeley, US, 1999.
- [15] Yang, Y. and Pedersen, J. P. "A Comparative Study on Feature Selection in Text Categorization", *The Fourteenth International Conference on Machine Learning*, pp. 412–420, 1997.