Cross Language Text Categorization Using a Bilingual Lexicon

Ke Wu, Xiaolin Wang and **Bao-Liang Lu**[*]
Department of Computer Science and Engineering, Shanghai Jiao Tong University
800 Dong Chuan Rd., Shanghai 200240, China
{wuke,arthur_general,bllu}@sjtu.edu.cn

Abstract

With the popularity of the Internet at a phenomenal rate, an ever-increasing number of documents in languages other than English are available in the Internet. Cross language text categorization has attracted more and more attention for the organization of these heterogeneous document collections. In this paper, we focus on how to conduct effective cross language text categorization. To this end, we propose a cross language naive Bayes algorithm. The preliminary experiments on collected document collections show the effectiveness of the proposed method and verify the feasibility of achieving performance close to monolingual text categorization, using a bilingual lexicon alone. Also, our algorithm is more efficient than our baselines.

1 Introduction

Due to the popularity of the Internet, an ever-increasing number of documents in languages other than English are available in the Internet. The organization of these heterogeneous document collections increases cost of human labor significantly. On the one hand, experts who know different languages are required to organize these collections. On the other hand, maybe there exist a large amount of labelled documents in a language (e.g. English) which are in the same class structure as the unlabelled documents in another language. As a result, how to exploit the existing labelled documents in some language (e.g. English) to classify the unlabelled documents other than the language in multilingual scenario has attracted more and more attention (Bel et al., 2003; Rigutini et al., 2005; Olsson et al., 2005; Fortuna and Shawe-Taylor, 2005; Li and Shawe-Taylor, 2006; Gliozzo and Strapparava, 2006). We refer to this task as cross language text categorization. It aims to extend the existing automated text categorization system from one language to other languages without additional intervention of human experts. Formally, given two document collections $\{\mathcal{D}_e, \mathcal{D}_f\}$ from two different languages e and f respectively, we use the labelled document collection $\mathcal{D}_e$ in the language e to deduce the labels of the document collection $\mathcal{D}_f$ in the language f via an algorithm $\mathcal{A}$ and some external bilingual resources.

Typically, some external bilingual lexical resources, such as machine translation system (MT), large-scale parallel corpora and multilingual ontology etc., are used to alleviate cross language text categorization. However, it is hard to obtain them for many language pairs. In this paper, we focus on using a cheap bilingual resource, e.g. bilingual lexicon without any translation information, to conduct cross language text categorization. To my knowledge, there is little research on using a bilingual lexicon alone for cross language text categorization.

In this paper, we propose a novel approach for cross language text categorization via a bilingual lexicon alone. We call this approach as Cross Language Naive Bayes Classifier (CLNBC). The proposed approach consists of two main stages. The first stage is to acquire a probabilistic bilingual lex-

[*] Corresponding author.

icon. The second stage is to employ naive Bayes method combined with Expectation Maximization (EM) (Dempster et al., 1977) to conduct cross language text categorization via the probabilistic bilingual lexicon. For the first step, we propose two different methods. One is a naive and direct method, that is, we convert a bilingual lexicon into a probabilistic lexicon by simply assigning equal translation probabilities to all translations of a word. Accordingly, the approach in this case is named as CLNBC-D. The other method is to employ an EM algorithm to deduce the probabilistic lexicon. In this case, the approach is called as CLNBC-EM. Our preliminary experiments on our collected data have shown that the proposed approach (CLNBC) significantly outperforms the baselines in cross language case and is close to the performance of monolingual text categorization.

The remainder of this paper is organized as follows. In Section 2, we introduce the naive Bayes classifier briefly. In Section 3, we present our cross language naive Bayes algorithm. In Section 4, evaluation over our proposed algorithm is performed. Section 5 is conclusions and future work.

2 The Naive Bayes Classifier

The naive Bayes classifier is an effective known algorithm for text categorization (Domingos and Pazzani, 1997). When it is used for text categorization task, each document $d \in \mathcal{D}$ corresponds to an example. The naive Bayes classifier estimates the probability of assigning a class $c \in \mathcal{C}$ to a document d based on the following Bayes' theorem.

$$P(c|d) \propto P(d|c)P(c) \tag{1}$$

Then the naive Bayes classifier makes two assumptions for text categorization. Firstly, each word in a document occurs independently. Secondly, there is no linear ordering of the word occurrences.

Therefore, the naive Bayes classifier can be further formalized as follows:

$$P(c|d) \propto P(c) \prod_{w \in d} P(w|c) \tag{2}$$

The estimates of $P(c)$ and $P(w|c)$ can be referred to (McCallum and Nigam, 1998)

Some extensions to the naive Bayes classifier with EM algorithm have been proposed for various text categorization tasks. The naive Bayes classifier was combined with EM algorithm to learn the class label of the unlabelled documents by maximizing the likelihood of both labelled and unlabelled documents (Nigam et al., 2000). In addition, the similar way was adopted to handle the problem with the positive samples alone (Liu et al., 2002). Recently, transfer learning problem was tackled by applying EM algorithm along with the naive Bayes classifier (Dai et al., 2007). However, they all are monolingual text categorization tasks. In this paper, we apply a similar method to cope with cross language text categorization using bilingual lexicon alone.

3 Cross Language Naive Bayes Classifier Algorithm

In this section, a novel cross language naive Bayes classifier algorithm is presented. The algorithm contains two main steps below. First, generate a probabilistic bilingual lexicon; second, apply an EM-based naive Bayes learning algorithm to deduce the labels of documents in another language via the probabilistic lexicon.

Table 1: Notations and explanations.

Notations	**Explanations**
e	Language of training set
f	Language of test set
d	Document
$\mathcal{D}_e$	Document collection in language e
$\mathcal{D}_f$	Document collection in language f
$\mathcal{V}_e$	Vocabulary of language e
$\mathcal{V}_f$	Vocabulary of language f
$\mathcal{L}$	Bilingual lexicon
$\mathcal{T} \subseteq \mathcal{V}_e \times \mathcal{V}_f$	Set of links in $\mathcal{L}$
λ_γ	Set of words whose translation is γ in $\mathcal{L}$
$E \subseteq \mathcal{V}_e$	Set of words of language e in $\mathcal{L}$
$w_e \in E$	Word in E
$F \subseteq \mathcal{V}_f$	Set of words of language f in $\mathcal{L}$
$w_f \in F$	Word in F
$\lvert E\rvert$	Number of distinct words in set E
$\lvert F\rvert$	Number of distinct words in set F
$N(w_e)$	Word frequency in $\mathcal{D}_e$
$N(w_f, d)$	Word frequency in d in language f
$\mathfrak{D}_e$	Data distribution in language e

For ease of description, we first define some notations in Table 1. In the next two sections, we detail the mentioned-above two steps separately.

3.1 Generation of a probabilistic bilingual lexicon

To fill the gap between different languages, there are two different ways. One is to construct the multilingual semantic space, and the other is to transform documents in one language into ones in another language. Since we concentrate on use of a bilingual lexicon, we adopt the latter method. In this paper, we focus on the probabilistic model instead of selecting the best translation. That is, we need to calculate the probability of the occurrence of word w_e in language e given a document d in language f, i.e. $P(w_e|d)$. The estimation can be calculated as follows:

$$P(w_e|d) = \sum_{w_f \in d} P(w_e|w_f, d)P(w_f|d) \quad (3)$$

Ignoring the context information in a document d, the above probability can be approximately estimated as follows:

$$P(w_e|d) \simeq \sum_{w_f \in d} P(w_e|w_f)P(w_f|d) \quad (4)$$

where $P(w_f|d)$ denotes the probability of occurrence of w_f in d, which can be estimated by relative frequency of w_f in d.

In order to induce $P(w_e|d)$, we have to know the estimation of $P(w_e|w_f)$. Typically, we can obtain a probabilistic lexicon from a parallel corpus. In this paper, we concentrate on using a bilingual lexicon alone as our external bilingual resource. Therefore, we propose two different methods for cross language text categorization.

First, a naive and direct method is that we assume a uniform distribution on a word's distribution. Formally, $P(w_e|w_f) = \frac{1}{\lambda_{w_f}}$, where $(w_e, w_f) \in \mathcal{T}$; otherwise $P(w_e|w_f) = 0$.

Second, we can apply EM algorithm to deduce the probabilistic bilingual lexicon via the bilingual lexicon $\mathcal{L}$ and the training document collection at hand. This idea is motivated by the work (Li and Li, 2002).

We can assume that each word w_e in language e is independently generated by a finite mixture model as follows:

$$P(w_e) = \sum_{w_f \in F} P(w_f)P(w_e|w_f) \quad (5)$$

Therefore we can use EM algorithm to estimate the parameters of the model. Specifically speaking, we can iterate the following two step for the purpose above.

- E-step

$$P(w_f|w_e) = \frac{P(w_f)P(w_e|w_f)}{\sum_{w \in F} P(w)P(w_e|w)} \quad (6)$$

- M-step

$$P(w_e|w_f) = \frac{(N(w_e)+1)P(w_f|w_e)}{\sum_{w \in E}(N(w)+1)P(w_f|w)} \quad (7)$$

$$\begin{aligned} P(w_f) &= \lambda \cdot \sum_{w_e \in E} P(w_e)P(w_f|w_e) \\ &+ (1-\lambda) \cdot P^{'}(w_f) \end{aligned} \quad (8)$$

where $0 \le \lambda \le 1$, and

$$P^{'}(w_f) = \frac{\sum_{d \in \mathcal{D}_f} N(w_f, d) + 1}{\sum_{w_f \in F}\sum_{d \in \mathcal{D}_f} N(w_f, d) + |F|} \quad (9)$$

The detailed algorithm can be referred to Algorithm 1. Furthermore, the probability that each word in language e occurs in a document d in language f, $P(w_e|d)$, can be calculated according to Equation (4).

3.2 EM-based Naive Bayes Algorithm for Labelling Documents

In this sub-section, we present an EM-based semi-supervised learning method for labelling documents in different language from the language of training document collection. Its basic model is naive Bayes model. This idea is motivated by the transfer learning work (Dai et al., 2007). For simplicity of description, we first formalize the problem. Given the labelled document set $\mathcal{D}_e$ in the source language and the unlabelled document set $\mathcal{D}_f$, the objective is to find the maximum a posteriori hypothesis h_{MAP}

Algorithm 1 EM-based Word Translation Probability Algorithm

```
Input: Training document collection $\mathcal{D}_e^{(l)}$, bilingual lexicon $\mathcal{L}$ and maximum times of iterations $T$
Output: Probabilistic bilingual lexicon $P(w_e|w_f)$
Initialize $P^{(0)}(w_e|w_f) = \frac{1}{|\lambda_{w_f}|}$, where $(w_e, w_f) \in \mathcal{T}$; otherwise $P^{(0)}(w_e|w_f) = 0$
Initialize $P^{(0)}(w_f) = \frac{1}{|F|}$
for t =1 to $T$ do
    Calculate $P^{(t)}(w_f|w_e)$ based on $P^{(t-1)}(w_e|w_f)$ and $P^{(t-1)}(w_f)$ according to Equation (6)
    Calculate $P^{(t)}(w_e|w_f)$ and $P^{(t)}(w_f)$ based on $P^{(t)}(w_f|w_e)$ according to Equation (7) and Equation (8)
end for
return $P^{(T)}(w_e|w_f)$
```

from the hypothesis space H under the data distribution of the language e, $\mathfrak{D}_e$, according to the following formula.

$$h_{MAP} = \arg\max_{h \in H} P_{\mathfrak{D}_e}(h|\mathcal{D}_e, \mathcal{D}_f) \quad (10)$$

Instead of trying to maximize $P_{\mathfrak{D}_e}(h|\mathcal{D}_e, \mathcal{D}_f)$ in Equation (10), we can work with $\ell(h|\mathcal{D}_e, \mathcal{D}_f)$, that is, $\log\left(P_{\mathfrak{D}_e}(h)P(\mathcal{D}_e, \mathcal{D}_f|h)\right)$. Then, using Equation (10), we can deduce the following equation.

$$\begin{aligned} \ell(h|\mathcal{D}_e, \mathcal{D}_f) &\propto \log P_{\mathfrak{D}_\mathfrak{e}}(h) \\ &+ \sum_{d \in \mathcal{D}_e} \log \sum_{c \in \mathcal{C}} P_{\mathfrak{D}_e}(d|c) P_{\mathfrak{D}_e}(c|h) \\ &+ \sum_{d \in \mathcal{D}_f} \log \sum_{c \in \mathcal{C}} P_{\mathfrak{D}_e}(d|c) P_{\mathfrak{D}_e}(c|h) \end{aligned} \quad (11)$$

EM algorithm is applied to find a local maximum of $\ell(h|\mathcal{D}_e, \mathcal{D}_f)$ by iterating the following two steps:

- E-step:

$$P_{\mathfrak{D}_e}(c|d) \propto P_{\mathfrak{D}_e}(c) P_{\mathfrak{D}_e}(d|c) \quad (12)$$

- M-step:

$$P_{\mathfrak{D}_e}(c) = \sum_{k \in \{e,f\}} P_{\mathfrak{D}_e}(\mathcal{D}_k) P_{\mathfrak{D}_e}(c|\mathcal{D}_k) \quad (13)$$

$$P_{\mathfrak{D}_e}(w_e|c) = \sum_{k \in \{e,f\}} P_{\mathfrak{D}_e}(\mathcal{D}_k) P_{\mathfrak{D}_e}(w_e|c, \mathcal{D}_k) \quad (14)$$

Algorithm 2 Cross Language Naive Bayes Algorithm

```
Input: Labelled document collection $\mathcal{D}_e$, unlabelled document collection $\mathcal{D}_f$, a bilingual lexicon $\mathcal{L}$ from language $e$ to language $f$ and maximum times of iterations $T$.
Output: the class label of each document in $\mathcal{D}_f$
Generate a probabilistic bilingual lexicon;
Calculate $P(w_e|d)$ according to Equation (4).
Initialize $P_{\mathfrak{D}_e}^{(0)}(c|d)$ via the traditional naive Bayes model trained from the labelled collection $\mathcal{D}_e^{(l)}$.
for t =1 to $T$ do
    for all $c \in \mathcal{C}$ do
        Calculate $P_{\mathfrak{D}_e}^{(t)}(c)$ based on $P_{\mathfrak{D}_e}^{(t-1)}(c|d)$ according to Equation (13)
    end for
    for all $w_e \in E$ do
        Calculate $P_{\mathfrak{D}_e}^{(t)}(w_e|c)$ based on $P_{\mathfrak{D}_e}^{(t-1)}(c|d)$ and $P(w_e|d)$ according to Equation (14)
    end for
    for all $d \in \mathcal{D}_f$ do
        Calculate $P_{\mathfrak{D}_e}^{(t)}(c|d)$ based on $P_{\mathfrak{D}_e}^{(t)}(c)$ and $P_{\mathfrak{D}_e}^{(t)}(w_e|c)$ according to Equation (12)
    end for
end for
for all $d \in \mathcal{D}_f$ do
    $c = \arg\max_{c \in \mathcal{C}} P_{\mathfrak{D}_e}^{(T)}(c|d)$
end for
```

For the ease of understanding, we directly put the details of the algorithm in cross-language text categorization algorithmin which we ignore the detail of the generation algorithm of a probabilistic lexicon.

In Equation (12), $P_{\mathfrak{D}_e}(d|c)$ can be calculated by

$$P_{\mathfrak{D}_e}(d|c) = \prod_{\{w_e | w_e \in \lambda_{w_f} \wedge w_f \in d\}} P_{\mathfrak{D}_e}(w_e|c)^{N_{\mathfrak{D}_e}(w_e, d)} \quad (15)$$

where $N_{\mathfrak{D}_e}(w_e, d) = |d| P_{\mathfrak{D}_e}(w_e|d)$.

In Equation (13), $P_{\mathfrak{D}_e}(c|\mathcal{D}_k)$ can be estimated as follows:

$$P_{\mathfrak{D}_e}(c|\mathcal{D}_k) = \sum_{d \in \mathcal{D}_k} P_{\mathfrak{D}_e}(c|d) P_{\mathfrak{D}_e}(d|\mathcal{D}_k) \qquad (16)$$

In Equation (14), similar to section 2, we can estimate $P_{\mathfrak{D}_e}(w_e|c, \mathcal{D}_k)$ through Laplacian smoothing as follows:

$$P_{\mathfrak{D}_e}(w_e|c, \mathcal{D}_k) = \frac{1 + N_{\mathfrak{D}_e}(w_e, c, \mathcal{D}_k)}{|\mathcal{V}_k| + N_{\mathfrak{D}_e}(c, \mathcal{D}_k)} \qquad (17)$$

where

$$N_{\mathfrak{D}_e}(w_e, c, \mathcal{D}_k) = \sum_{d \in \mathcal{D}_k} |d| P_{\mathfrak{D}_e}(w_e|d) P_{\mathfrak{D}_e}(c|d) \qquad (18)$$

$$N_{\mathfrak{D}_e}(c, \mathcal{D}_k) = \sum_{d \in \mathcal{D}_k} |d| P_{\mathfrak{D}_e}(c|d) \qquad (19)$$

In addition, in Equation (13) and (14), $P_{\mathfrak{D}_e}(\mathcal{D}_k)$ can be actually viewed as the trade-off parameter modulating the degree to which EM algorithm weights the unlabelled documents translated from the language f to the language e via a bilingual lexicon. In our experiments, we assume that the constraints are satisfied, i.e. $P_{\mathfrak{D}_e}(\mathcal{D}_e) + P_{\mathfrak{D}_e}(\mathcal{D}_f) = 1$ and $P_{\mathfrak{D}_e}(d|\mathcal{D}_k) = \frac{1}{|\mathcal{D}_k|}$.

4 Experiments

4.1 Data Preparation

We chose English and Chinese as our experimental languages, since we can easily setup our experiments and they are rather different languages so that we can easily extend our algorithm to other language pairs. In addition, to evaluate the performance of our algorithm, experiments were performed over the collected data set. Standard evaluation benchmark is not available and thus we developed a test data from the Internet, containing Chinese Web pages and English Web pages. Specifically, we applied RSS reader[1] to acquire the links to the needed content and then downloaded the Web pages. Although category information of the content can be obtained by RSS reader, we still used three Chinese-English bilingual speakers to organize these Web pages into the predefined categories. As a result, the test data containing Chinese Web pages and English Web pages from various Web sites are created. The data consists of news during December 2005. Also, 5462 English Web pages are from 18 different news Web sites and 6011 Chinese Web pages are from 8 different news Web sites. Data distribution over categories is shown in Table 2. They fall into five categories: *Business*, *Education*, *Entertainment*, *Science* and *Sports*.

Some preprocessing steps are applied to Web pages. First we extract the pure texts of all Web pages, excluding anchor texts which introduce much noise. Then for Chinese corpus, all Chinese characters with BIG5 encoding first were converted into ones with GB2312 encoding, applied a Chinese segmenter tool[2] by Zhibiao Wu from LDC to our Chinese corpus and removed stop words and words with one character and less than 4 occurrences; for English corpus, we used the stop words list from SMART system (Buckley, 1985) to eliminate common words. Finally, We randomly split both the English and Chinese document collection into 75% for training and 25% for testing.

we compiled a large general-purpose English-Chinese lexicon, which contains 276,889 translation pairs, including 53,111 English entries and 38,517 Chinese entries. Actually we used a subset of the lexicon including 20,754 English entries and 13,471 Chinese entries , which occur in our corpus.

Table 2: Distribution of documents over categories

Categories	English	Chinese
Sports	1797	2375
Business	951	1212
Science	843	1157
Education	546	692
Entertainment	1325	575
Total	5462	6011

4.2 Baseline Algorithms

To investigate the effectiveness of our algorithms on cross-language text categorization, three baseline methods are used for comparison. They are denoted by ML, MT and LSI respectively.

ML (Monolingual). We conducted text categorization by training and testing the text categoriza-

[1] http://www.rssreader.com/

[2] http://projects.ldc.upenn.edu/Chinese/LDC_ch.htm

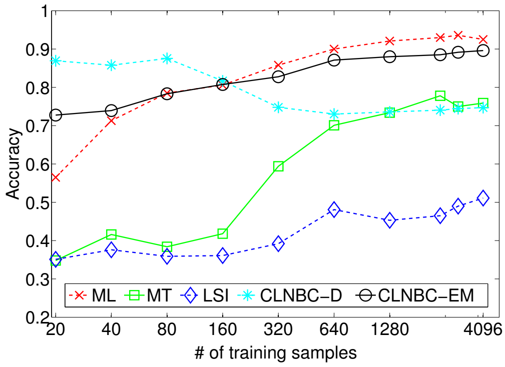

Figure 1: Comparison of the best performance of different methods with various sizes of training set and the entire test set. Training is conducted over Chinese corpus and testing is conducted over English corpus in the cross language case, while both training and testing are performed over English corpus in the monolingual case.

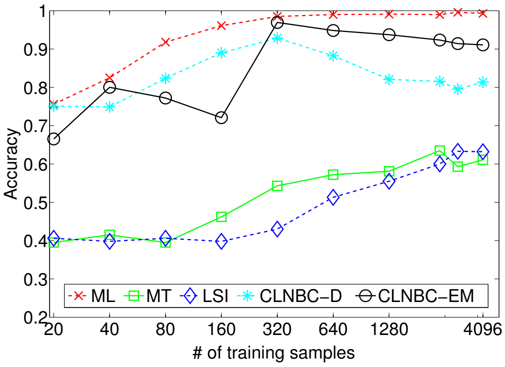

Figure 2: Comparison of the best performance of different methods with various sizes of training set and the entire test set. Training is conducted over English corpus and testing is conducted over Chinese corpus in the cross language case, while both training and testing are performed over Chinese corpus in the monolingual case.

tion system on document collection in the same language.

MT (Machine Translation). We used Systran premium 5.0 to translate training data into the language of test data, since the machine translation system is one of the best machine translation systems. Then use the translated data to learn a model for classifying the test data.

LSI (Latent Semantic Indexing). We can use the LSI or SVD technique to deduce language-independent representations through a bilingual parallel corpus. In this paper, we use SVDS command in MATLAB to acquire the eigenvectors with the first K largest eigenvalues. We take K as 400 in our experiments, where best performance is achieved.

In this paper, we use SVMs as the classifier of our baselines, since SVMs has a solid theoretic foundation based on structure risk minimization and thus high generalization ability. The commonly used one-vs-all framework is used for the multi-class case. SVMs uses the SVM^{light} software package(Joachims, 1998). In all experiments, the tradeoff parameter C is set to 1.

4.3 Results

In the experiments, all results are averaged on 5 runs. Results are measured by accuracy, which is defined as the ratio of the number of labelled correctly documents to the number of all documents. When investigating how different training data have effect on performance, we randomly select the corresponding number of training samples from the training set 5 times. The results are shown in Figure 1 and Figure 2. From the two figures, we can draw the following conclusions. First, CLNBC-EM has a stable and good performance in almost all cases. Also, it can achieve the best performance among cross language methods. In addition, we notice that CLNBC-D works surprisingly better than CLNBC-EM, when there are enough test data and few training data. This may be because the quality of the probabilistic bilingual lexicon derived from CLNBC-EM method is poor, since this bilingual lexicon is trained from insufficient training data and thus may provide biased translation probabilities.

To further investigate the effect of varying the amount of test data, we randomly select the corresponding number of test samples from test set 5 times. The results are shown in Figure 3 and Figure 4, we can draw the following conclusions . First, with the increasing test data, performance of our two approaches is improved. Second, CLNBC-EM statistically significantly outperforms CLNBC-D.

From figures 1 through 4, we also notice that MT and LSI always achieve some poor results. For MT,

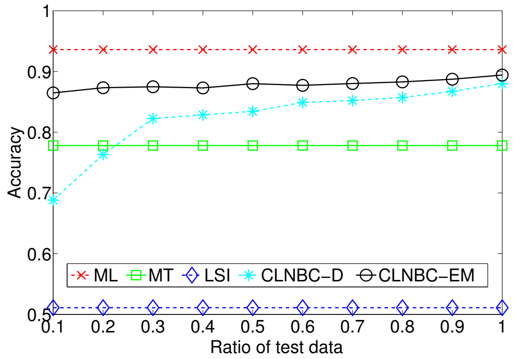

Figure 3: Comparison of the best performance of different methods with the entire training set and various sizes of test set. Training is conducted over Chinese corpus and testing is conducted over English corpus in the cross language case, while both training and testing are performed over English corpus in the monolingual case.

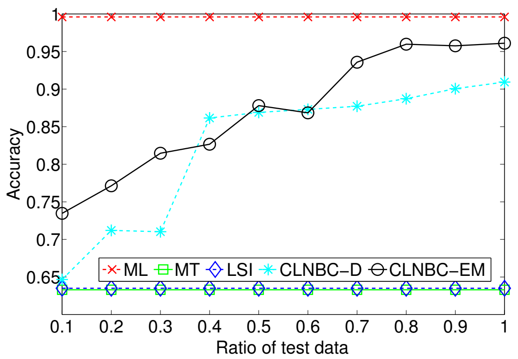

Figure 4: Comparison of the best performance of different methods with the entire training set and various sizes of test set. Training is conducted over English corpus and testing is conducted over Chinese corpus in the cross language case, while both training and testing are performed over Chinese corpus in the monolingual case.

maybe it is due to the large difference of word usage between original documents and the translated ones. For example, 骑士 (Qi Shi) has two common translations, which are *cavalier* and *knight*. In sports domain, it often means a basketball team of National Basketball Association (NBA) in U.S. and should be translated into *cavalier*. However, the translation *knight* is provided by Systran translation system we use in the experiment. In term of LSI method, one possible reason is that the parallel corpus is too limited. Another possible reason is that it is out-of-domain compared with the domain of the used document collections.

From Table 3, we can observe that our algorithm is more efficient than three baselines. The spent time are calculated on the machine, which has a 2.80GHz Dual Pentium CPU.

5 Conclusions and Future Work

In this paper, we addressed the issue of how to conduct cross language text categorization using a bilingual lexicon. To this end, we have developed a cross language naive Bayes classifier, which contains two main steps. In the first step, we deduce a probabilistic bilingual lexicon. In the second step, we adopt naive Bayes method combined with EM to conduct cross language text categorization. We have proposed two different methods, namely CLNBC-D and CLNBC-EM, for cross language text categorization. The preliminary experiments on collected data collections show the effectiveness of the proposed two methods and verify the feasibility of achieving performance near to monolingual text categorization using a bilingual lexicon alone.

As further work, we will collect larger comparable corpora to verify our algorithm. In addition, we will investigate whether the algorithm can be scaled to more fine-grained categories. Furthermore, we will investigate how the coverage of bilingual lexicon have effect on performance of our algorithm.

Table 3: Comparison of average spent time by different methods, which are used to conduct cross-language text categorization from English to Chinese.

Methods	Preparation	Computation
CLNBC-D	-	∼1 Min
CLNBC-EM	-	∼2 Min
ML	-	∼10 Min
MT	∼48 Hr[a]	∼14 Min
LSI	∼90 Min[b]	∼15 Min

[a] Machine Translation Cost

[b] SVD Decomposition Cost

Acknowledgements. The authors would like to thank three anonymous reviewers for their valuable suggestions. This work was partially supported by the National Natural Science Foundation of China under the grants NSFC 60375022 and NSFC 60473040, and the Microsoft Laboratory for Intelligent Computing and Intelligent Systems of Shanghai Jiao Tong University.

References

Nuria Bel, Cornelis H. A. Koster, and Marta Villegas. 2003. Cross-lingual text categorization. In *ECDL*, pages 126–139.

Chris Buckley. 1985. Implementation of the SMART information retrieval system. Technical report, Ithaca, NY, USA.

Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu. 2007. Transferring naive Bayes classifiers for text classification. In *Proceedings of Twenty-Second AAAI Conference on Artificial Intelligence (AAAI 2007)*, pages 540–545, July.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Pedro Domingos and Michael J. Pazzani. 1997. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3):103–130.

Blaž Fortuna and John Shawe-Taylor. 2005. The use of machine translation tools for cross-lingual text mining. In *Learning With Multiple Views, Workshop at the 22nd International Conference on Machine Learning (ICML)*.

Alfio Massimiliano Gliozzo and Carlo Strapparava. 2006. Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. The Association for Computer Linguistics, July.

Thorsten Joachims. 1998. Making large-scale support vector machine learning practical. In A. Smola B. Schölkopf, C. Burges, editor, *Advances in Kernel Methods: Support Vector Machines*. MIT Press, Cambridge, MA.

Cong Li and Hang Li. 2002. Word translation disambiguation using bilingual bootstrapping. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 343–351.

Yaoyong Li and John Shawe-Taylor. 2006. Using KCCA for Japanese-English cross-language information retrieval and document classification. *Journal of Intelligent Information Systems*, 27(2):117–133.

Bing Liu, Wee Sun Lee, Philip S. Yu, and Xiaoli Li. 2002. Partially supervised classification of text documents. In *ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning*, pages 387–394, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Andrew McCallum and Kamal Nigam. 1998. A comparison of event models for naive bayes text classification. In *Proceedings of AAAI-98, Workshop on Learning for Text Categorization*.

Kamal Nigam, Andrew McCallum, Sebastian Thrun, and Tom Mitchell. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134.

J. Scott Olsson, Douglas W. Oard, and Jan Hajič. 2005. Cross-language text classification. In *Proceedings of the 28th Annual international ACM SIGIR Conference on Research and Development in information Retrieval*, pages 645–646, New York, NY, August. ACM Press.

Leonardo Rigutini, Marco Maggini, and Bing Liu. 2005. An EM based training algorithm for cross-language text categorization. In *Proceedings of Web Intelligence Conference (WI-2005)*, pages 529–535, Compiègne, France, September. IEEE Computer Society.