A Refinement Framework for Cross Language Text Categorization

Ke Wu and Bao-Liang Lu*

Department of Computer Science and Engineering, Shanghai Jiao Tong University
800 Dong Chuan Road, Shanghai 200240, China
{wuke,bllu}@sjtu.edu.cn

Abstract. Cross language text categorization is the task of exploiting labelled documents in a source language (e.g. English) to classify documents in a target language (e.g. Chinese). In this paper, we focus on investigating the use of a bilingual lexicon for cross language text categorization. To this end, we propose a novel refinement framework for cross language text categorization. The framework consists of two stages. In the first stage, a cross language model transfer is proposed to generate initial labels of documents in target language. In the second stage, expectation maximization algorithm based on naive Bayes model is introduced to yield resulting labels of documents. Preliminary experimental results on collected corpora show that the proposed framework is effective.

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, thus creating the need of automatic organization of these multilingual documents. In addition, with the globalization of business environments, for many international companies and organizations, huge volume of documents in different languages need to be archived into common categories. On the other hand, in order to build a reliable model for automated text categorization, we typically need a large amount of manually labelled documents, which cost much human labor. Consequently, in multilingual scenario, how to employ the existing labelled documents written in a source language (e.g. English) to classify the unlabelled documents other than the language has become an important task, as it can be leveraged to alleviate cost of labelling. We refer to the mentioned-above task as cross language text categorization (CLTC).

Cross language information retrieval is highly related to CLTC. Also, the use of bilingual lexicon has been extensively studied in cross language information

* Corresponding author. This work was supported in part 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.

H. Li et al. (Eds.): AIRS 2008, LNCS 4993, pp. 401–411, 2008.

retrieval [1,2,3]. However, to our knowledge, there is little research on the direction for CLTC. This paper will focus on investigating the use of a bilingual lexicon. Accordingly, we propose a novel refinement framework for CLTC.

The basic idea is that we assume that initial and inaccurate labels from the transferred model can be refined in the original documents into better resulting labels. Specifically, the framework consists of two stages. In the first stage, a cross language model transfer is proposed to generate preliminary labels of documents in target language. In the second stage, expectation maximization algorithm (EM) [4] based on naive Bayes model is introduced to generate resulting labels of documents. Preliminary experimental results on collected corpora show that in the case of sufficient test data, with a small number of training documents, the proposed refinement framework can achieve better performance than monolingual text categorization and with a large number of training documents, it can also obtain promising results close to that of monolingual text categorization.

The remainder of this paper is organized as follows. Section 2 introduces related work. Section 3 presents the refinement framework. Section 4 performs evaluation over our proposed framework. Section 5 is conclusions and future work.

2 Related Work

Cross language text categorization is divided into two cases, which are poly-lingual training and cross-lingual training [5]. The term poly-lingual training indicates the case that enough training documents available for every language. However, such scenarios are not particularly interesting as they can be handled with separate monolingual solutions. The term cross-lingual training indicates that another case that enough training documents available for a language but no training documents for other languages. Currently, researchers focus their effort on the latter case. In this paper, we also focus on this case.

Typically, some external lexical resources are used for CLTC. Li and Shawe-Taylor [6] applied kernel canonical correlation analysis (KCCA) and latent semantic analysis (LSA) to parallel corpora and induced the semantic space for CLTC. Olsson et al. [7] used the probabilistic bilingual lexicon induced by parallel corpora to ensure that test data is translated into the language of training data. However, a good semantic space or accurate translation probabilities depend on the amount of parallel corpora. Unfortunately, large-scale parallel corpora are not easily obtained. To alleviate the difficulty, Gliozzo and Strapparava [8] exploit comparable corpora to induce a semantic space by LSA. Nevertheless, this method is applicable only for language pairs, which have common words for the same concepts. Furthermore, Fortuna and Shawe-Taylor [9] applied machine translation system to generate pseudo domain-specific parallel corpus. Rigutini et al. [10] used a machine translation system to bridge the gap between different languages. However, there are not machine translation systems for many language pairs and there is still wide gap of statistical characteristics between translated documents and original documents.

Compared with the above lexical resources, bilingual lexicon is a kind of cheap resource, which is readily available. However, there is little research on the use of a bilingual lexicon alone for CLTC. In this paper, we wish to concentrate on the direction.

3 Refinement Framework

Figure 1 shows the overall architecture of our refinement framework. L_1 denotes the source language (i.e. the language in which documents are manually labelled); L_2 denotes the target language (i.e. the language in which documents are to be classified according to the categories from language L_1). The framework consists of two stages. The preliminary labels are generated in the first stage and a refinement with the preliminary labels is performed in the second stage. In the following two sections, we shall explain the two stages in details.

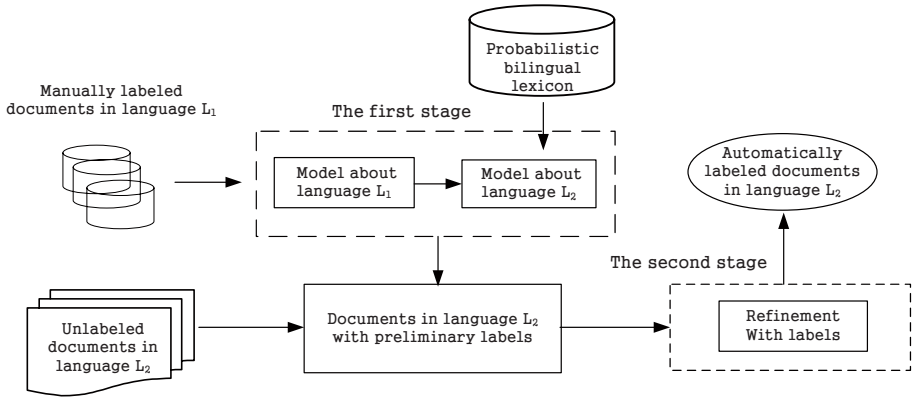

Fig. 1. Refinement framework for cross language text categorization

3.1 The First Stage

In this stage, preliminary labels about documents in L_2 are generated. Accordingly, a learning model about L_2 needs to be generated for label assignment of the documents in L_2. However, a learning model about L_2 can not be derived directly. As a result, we propose an approach which transfers the trained model in L_1 into the new model in L_2 via a bilingual lexicon. This approach is called as `cross language model transfer` (CLMT). In this paper, we choose to investigate the transfer of the naive Bayes model from L_1 to L_2, since naive Bayes model is efficient and effective for multi-class case. The details of naive Bayes model can be referred to [11]. For naive Bayes model in language L_2, we need estimate two parameters, $P(w_f|c)$ and $P(c)$, where $P(w_f|c)$ denotes the probability that word w_f in L_2 occurs, given class c and $P(c)$ denotes the probability that class c occurs in language L_2.

In cross language information retrieval, a unigram language model in one language is combined with a probabilistic bilingual lexicon to yield a unigram language model in another language. Our approach is inspired by this well-known technique. We extend this by using a class-conditional bilingual lexicon. It can be formalized as follows:

$$P(w_f|c) = \sum_{w_e \in \mathcal{V}^{L_1}} P(w_e|c)P(w_f|w_e, c) \tag{1}$$

where $\mathcal{V}^{L_1}$ denotes the vocabulary in L_1; $P(w_e|c)$ denotes the probability that word w_e in L_1 occurs given class c; and $P(w_f|w_e, c)$ denotes the probability that word w_e is translated into word w_f given class c.

$P(w_e|c)$ is derived from the parameter estimation of model about L_1. There are two solutions for $P(w_f|w_e, c)$. First, a naive and direct method is that we simply assume for each class c, a word w_e is translated into w_f with the same probability, which is a uniform distribution on a word's translations. If a word w_e has n translations in our bilingual lexicon $\mathcal{L}$, each of them will be assigned equal probability, i.e. $P(w_f|w_e, c) = \frac{1}{n}$, where w_f is a translation of w_e in $\mathcal{L}$; otherwise $P(w_f|w_e, c) = 0$.

Second, we propose to apply EM algorithm to deduce the conditional translation probabilities given class c, via the bilingual lexicon $\mathcal{L}$ and the training document collection at hand. This idea is inspired by the work of word translation disambiguation [12]. We can assume that given class c, each word w_e in language L_1 is independently generated by a finite mixture model according to $P(w_e|c) = \sum_{w_f} P(w_f|c)P(w_e|w_f, c)$.

Therefore we can use EM algorithm to estimate the parameters of the model. Specifically, $p(w_f|w_e, c)$ is initialized through the first solution and then the following two steps are iterated until $p(w_f|w_e, c)$ remains unchanged.

– E-step:

$$P(w_f|w_e, c) = \frac{P(w_f|c)P(w_e|w_f, c)}{\sum_{w_f \in \mathcal{V}^{L_2}} P(w_f|c)P(w_e|w_f, c)} \tag{2}$$

– M-step:

$$P(w_e|w_f, c) = \frac{N(w_e, c)P(w_f|w_e, c)}{\sum_{w_e \in \mathcal{V}^{L_1}} N(w_e, c)\, P(w_f|w_e, c)} \tag{3}$$

$$P(w_f|c) = \frac{\sum_{w_e \in \mathcal{V}^{L_1}} N(w_e, c)P(w_f|w_e, c)}{\sum_{w_e \in \mathcal{V}^{L_1}} N(w_e, c)} \tag{4}$$

where $N(w_e, c)$ denotes the times of co-occurrence of w_e and c.

For $P(c)$, there are two solutions, too. A simple solution is that we use estimation from the labelled documents in langauge L_1, since we assume that documents from different languages have the same class distribution. Another solution is that we can assume that the class distribution for documents in L_2 conforms to the uniform distribution, i.e. $P(c) = \frac{1}{|\mathcal{C}|}$. The true class priors for

documents in language L_2 may be different from those for documents in language L_1. We do not simply estimate $P(c)$ in language L_2 from the documents in language L_1. That is, we have no idea of any information about the class distribution for documents in language L_2. According to principle of maximum entropy, we can assume that the class distribution for documents in L_2 conforms to the uniform distribution.

3.2 The Second Stage

In this stage, preliminary labels of documents in language L_2 from the first stage are used as input and an EM algorithm is introduced to obtain the final labels of document in language L_2. The iterations of EM are a hill-climbing algorithm in parameter space that locally maximizes the entire log likelihood of documents in the collection. In this paper, we use naive Bayes model for the EM, similar to [11]. The algorithm we use is an unsupervised clustering whereas [11] is a semi-supervised learning. The basic idea is that EM is initialized to onto a right hill and then hill-climb the top. Specifically, $P(c|d)$ is initialized based on the preliminary labels of documents in language L_2 and then the following two steps are iterated until $P(c|d)$ stays unchanged, where d denotes a document.

– E-step:

$$P(c|d) = \frac{P(c)P(d|c)}{\sum_c P(c)P(d|c)} \tag{5}$$

– M-step:

$$P(w_f|c) = \frac{1 + \sum_d N(w_f, d)P(c|d)}{|\mathcal{V}^{L_2}| + \sum_c \sum_d N(w_f, d)P(c|d)} \tag{6}$$

$$P(c) = \frac{1 + \sum_d P(c|d)}{|\mathcal{C}| + |\mathcal{D}|} \tag{7}$$

where the calculation of $P(d|c)$ is referred to [11]. The resulting labels of documents in language L_2 are assigned according to the following equation:

$$c = \arg\max_c P(c|d) \tag{8}$$

Notice that in this stage only original documents in language L_2 are involved.

4 Evaluation

4.1 Setting

We chose English and Chinese as our experimental languages, for we can easily setup our experiments and they are quite different languages. Standard evaluation benchmark is not available and thus we developed a test data from the Internet, containing Chinese Web pages and English Web pages. We applied

Table 1. Source of Chinese Web pages

Chinese Web sites	Number of Web pages
people.com.cn	464
sina.com.cn	4814
tom.com	94
xinhuanet.com	408
chinanews.com.cn	41
jfdaily.com	32
voanews.com	18
takungpao.com	140
Total	6011

Table 2. Source of English Web pages

English Web sites	Number of Web pages
abcnews.go.com	232
allafrica.com	110
english.people.com.cn	416
football.guardian.co.uk	191
gradschool.about.com	19
news.bbc.co.uk	794
news.xinhuanet.com	142
soccernet.espn.go.com	237
yahoo.com	963
cbc.ca	81
cnn.com	335
nba.com	353
news.gov.hk	56
nytimes.com	740
soccerway.com	164
sportnetwork.net	246
uefa.com	290
voanews.com	93
Total	5462

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. The data consists of news during December 2005. There are total 5462 English Web pages which are from 18 news Web sites and 6011 Chinese Web pages which are from 8 news Web sites. The details of the sources of Web pages are shown in Table 1 and Table 2. Data distribution over categories is shown in Table 3.

Some preprocessing steps are applied to those 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

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

Table 3. 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

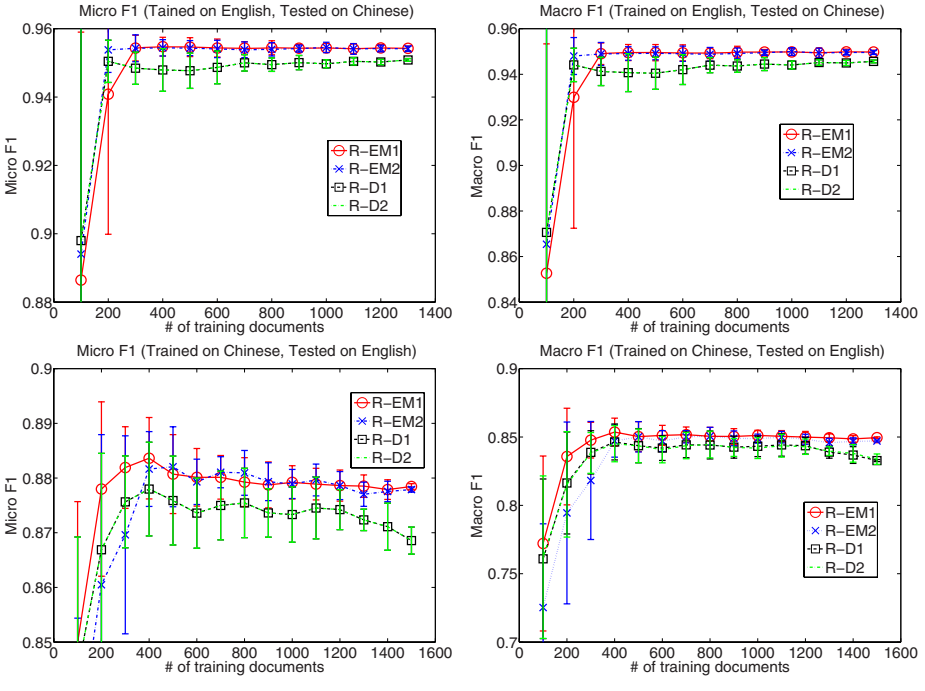

Fig. 2. Performance comparison of our refinement framework with different parameter estimations. The entire test data is used. Each point represents the mean performance for 10 arbitrary runs. The error bars show standard deviations for the estimated performance.

converted into ones with GB2312 encoding, applied a Chinese segmenter tool[2] by Zhibiao Wu from linguistic data consortium (LDC) to our Chinese corpus and removed words with one character and less than 4 occurrences; for English corpus, we used a stop list from SMART system [13] to eliminate common words. Finally, We randomly split both the English and Chinese documents into 2 sets, 25% for training and 75% for test.

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

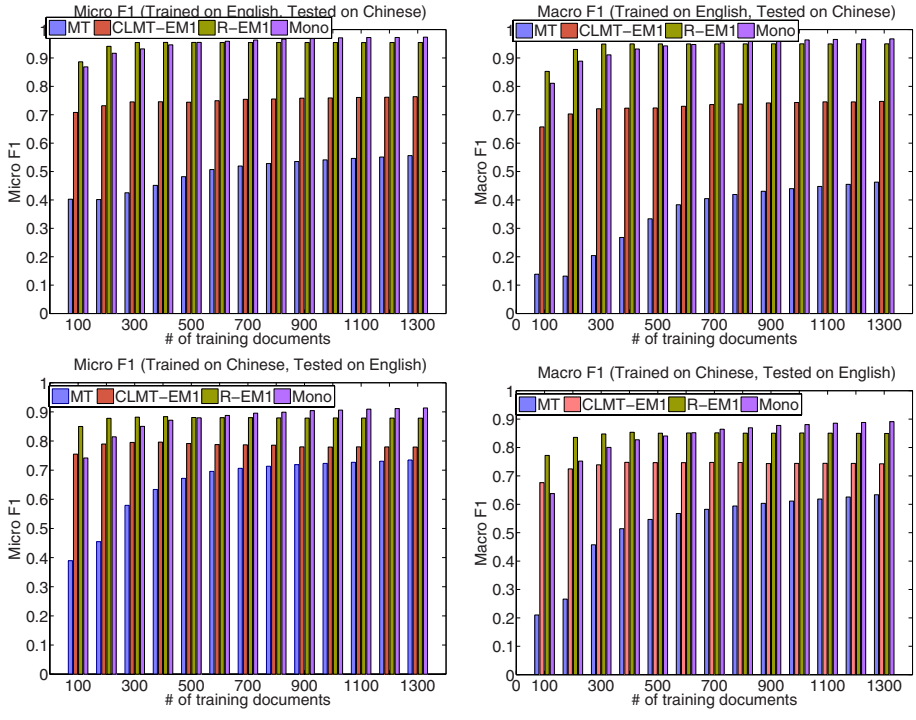

Fig. 3. Performance comparisons of different methods. The entire test data is used. Each value represents the mean performance for 10 arbitrary runs.

We compiled a 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.

4.2 Evaluation Measures

The performance of the proposed methods was evaluated in terms of conventional precision, recall and $F1$-measures. Furthermore, there are two conventional methods to evaluate overall performance averaged across categories, namely micro-averaging and macro-averaging [14]. Micro-averaging gives equal weight to each document while macro-averaging assigns equal weight to each category. In this paper, it is a multi-class case. Micro F1 and Macro F1 are short for micro-averaging F1 and macro-averaging F1.

4.3 Results

In our experiments, all results are averaged over 10 arbitrary runs. For the proposed CLMT approach for initial labels of documents in language L_2, four

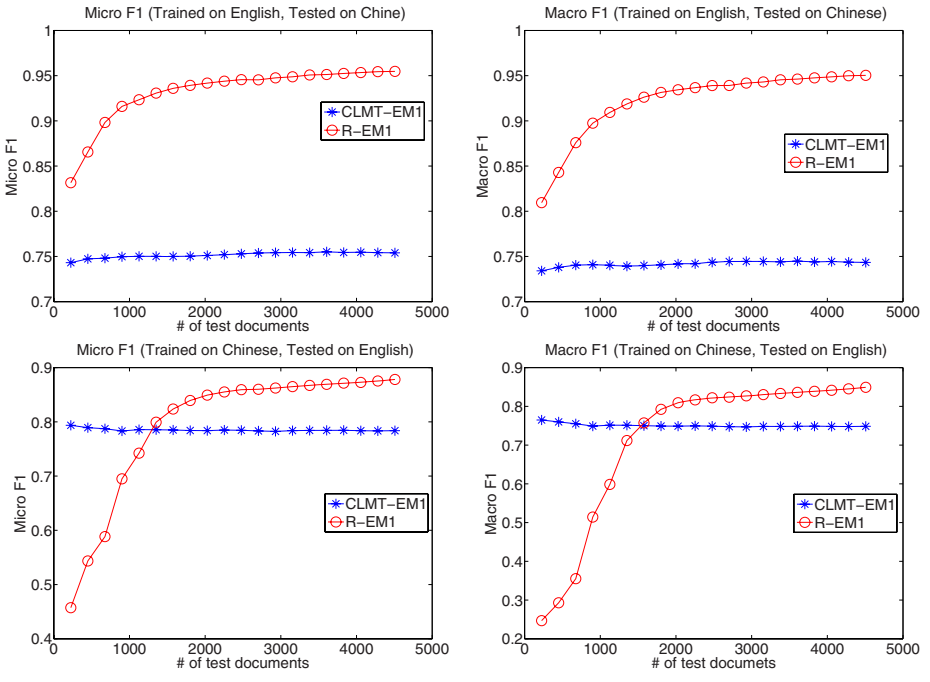

Fig. 4. Performance comparisons of CLMT-EM1 and R-EM1 varying the size of test data. The entire training data is used. Each value represents the mean performance for 10 arbitrary runs.

variants are naturally yielded as different parameter estimations may be used. As a result, we first investigate the impact on resulting performance, through varying different parameter estimations of CLMT. For ease of description, we call them as R-D1, R-D2, R-EM1 and R-EM2, where R indicates refinement framework, D indicates the first solution to estimate $P(w_f|w_e, c)$, EM indicates the second solution to estimate $P(w_f|w_e, c)$, digit 1 denotes the first solution to estimate $P(c)$ in language L_2, and digit 2 denotes the second solution to estimate $P(c)$ in language L_2. Their results on collected corpora are shown in Fig. 2. We can notice that R-EM1 and R-EM2 consistently work better than R-D1 and R-D2 over experiments trained on English documents and tested on Chinese documents or trained on Chinese documents and tested on English documents. In addition, R-EM1 performs slightly better than R-EM2.

For further evaluation of our framework, we compare our approach with the following three baselines. In our experiments, we use Naive Bayes as our classifier for fair comparison.

Mono (**Monolingual text categorization**)**.** Training and testing are performed on documents in the same language.

CLMT-EM1. It is used to generate preliminary labels for R-EM1. It sets a starting point of refinement for R-EM1, which is used as representative of our methods, since it perform better than other methods.

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 commercial machine translation systems. Then use the translated data to learn a statistical model for classifying the test data.

The results are shown in Fig. 3. We notice that with fewer training documents, R-EM1 works best among all methods and with more training documents, R-EM1 achieves a performance close to monolingual text categorization. In addition, we observe that MT obtains poor performance. This may be because statistical property of the translated documents is quite different from that of the original documents, although human can understand the translated documents produced by Systran premium 5.0.

To examine how the size of test data affects resulting performance, we compare CLMT-EM1 with R-EM1, varying the size of test data. The results are shown in Fig. 4. Experiments show that higher performance benefits from more test data. Meanwhile, we can also notice that when applied on a small portion of English test data set, EM based on naive Bayes model obtains results contrary to what we expect. The EM does not improve the performance of initial labels. On the contrary, it makes resulting performance worse than initial performance. It may be because there are too many parameters to be estimated but few data do not provide potential of accurate parameter estimation.

5 Conclusions and Future Work

This paper proposes a novel refinement framework for cross language text categorization. Our preliminary experiments on the collected data show that our refinement framework is effective for CLTC. This work has the following three main contributions. First, we are apparently the first to investigate the use of a bilingual lexicon alone for cross language text categorization. Second, a refinement framework is proposed for the use of a bilingual lexicon on cross language text categorization. Third, a cross language model transfer approach is proposed for the transfer of naive Bayes models from different languages via a bilingual lexicon and an EM algorithm based on naive Bayes model is introduced for the refinement of initial labels yielded by the proposed cross language model transfer method.

In the future, we shall improve our work from the following three directions. First, our data set is limited and the predefined categories are coarse. we plan to collect larger data collection with finer categories and test our proposed refinement framework on it. Second, different monolingual text categorization algorithms will be explored with the framework and accordingly new cross language model transfer approaches need to be proposed. Third, the EM algorithm is easily trapped into local optima. Therefore, we plan to propose a new refinement approach to avoid this case. Finally, people have recently tried to automatically

collect bilingual corpora from web [15,16], and therefore we may benefit by using the translation probabilities trained from the bilingual corpora.

References

1. Gao, J., Xun, E., Zhou, M., Huang, C., Nie, J.Y., Zhang, J.: Improving query translation for cross-language information retrieval using statistical models. In: ACM SIGIR 2001, pp. 96–104 (2001)
2. Gao, J., Nie, J.Y.: A study of statistical models for query translation: finding a good unit of translation. In: SIGIR 2006, pp. 194–201. ACM Press, New York (2006)
3. Liu, Y., Jin, R., Chai, J.Y.: A maximum coherence model for dictionary-based cross-language information retrieval. In: SIGIR 2005, pp. 536–543 (2005)
4. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society,Series B 39, 1–38 (1977)
5. Bel, N., Koster, C.H.A., Villegas, M.: Cross-Lingual Text Categorization. In: Koch, T., Sølvberg, I.T. (eds.) ECDL 2003. LNCS, vol. 2769, pp. 126–139. Springer, Heidelberg (2003)
6. Li, Y., Shawe-Taylor, J.: Using KCCA for Japanese-English cross-language information retrieval and document classification. Journal of Intelligent Information Systems 27, 117–133 (2006)
7. Olsson, J.S., Oard, D.W., Hajič, J.: Cross-language text classification. In: Proceedings of SIGIR 2005, pp. 645–646. ACM Press, New York (2005)
8. Gliozzo, A.M., Strapparava, C.: Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In: Proceedings of ACL 2006, The Association for Computer Linguistics (2006)
9. Fortuna, B., Shawe-Taylor, J.: 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) (2005)
10. Rigutini, L., Maggini, M., Liu, B.: An EM based training algorithm for cross-language text categorization. In: Proceedings of WI 2005, Washington, pp. 529–535. IEEE Computer Society, Los Alamitos (2005)
11. Nigam, K., McCallum, A., Thrun, S., Mitchell, T.: Text classification from labeled and unlabeled documents using EM. Machine Learning 39, 103–134 (2000)
12. Li, C., Li, H.: Word translation disambiguation using bilingual bootstrapping. In: Proceedings of ACL 2002, pp. 343–351 (2002)
13. Buckley, C.: Implementation of the SMART information retrieval system. Technical report, Ithaca, NY, USA (1985)
14. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Fisher, D.H. (ed.) Proceedings of ICML 1997, 14th International Conference on Machine Learning, Nashville, US, pp. 412–420. Morgan Kaufmann Publishers, San Francisco (1997)
15. Zhang, Y., Wu, K., Gao, J., Vines, P.: Automatic Acquisition of Chinese–English Parallel Corpus from the Web. In: Lalmas, M., MacFarlane, A., Rüger, S.M., Tombros, A., Tsikrika, T., Yavlinsky, A. (eds.) ECIR 2006. LNCS, vol. 3936, pp. 420–431. Springer, Heidelberg (2006)
16. Resnik, P., Smith, N.A.: The web as a parallel corpus. Comput. Linguist. 29, 349–380 (2003)