Cross Language Text Classification by Model Translation and Semi-Supervised Learning

Lei Shi
Yahoo! Global R&D
Beijing, China
lshi@yahoo-inc.com

Rada Mihalcea
University of North Texas
Denton, TX, U.S.A.
rada@cs.unt.edu

Mingjun Tian
Yahoo! Global R&D
Beijing, China
mingjun@yahoo-inc.com

Abstract

In this paper, we introduce a method that automatically builds text classifiers in a new language by training on already labeled data in another language. Our method transfers the classification knowledge across languages by translating the model features and by using an Expectation Maximization (EM) algorithm that naturally takes into account the ambiguity associated with the translation of a word. We further exploit the readily available unlabeled data in the target language via semi-supervised learning, and adapt the translated model to better fit the data distribution of the target language.

1 Introduction

Given the accelerated growth of the number of multilingual documents on the Web and elsewhere, the need for effective multilingual and cross-lingual text processing techniques is becoming increasingly important. There is a growing number of methods that use data available in one language to build text processing tools for another language, for diverse tasks such as word sense disambiguation (Ng et al., 2003), syntactic parsing (Hwa et al., 2005), information retrieval (Monz and Dorr, 2005), subjectivity analysis (Mihalcea et al., 2007), and others.

In this paper, we address the task of *cross-lingual text classification* (CLTC), which builds text classifiers for multiple languages by using training data in one language, thereby avoiding the costly and time-consuming process of labeling training data for each individual language. The main idea underlying our approach to CLTC is that although content can be expressed in different forms in different languages, there is a significant amount of knowledge that is shared for similar topics that can be effectively used to port topic classifiers across languages.

Previous methods for CLTC relied mainly on machine translation, by translating the training data into the language of the test data or vice versa, so that both training and test data belong to the same language. Monolingual text classification algorithms can then be applied on these translated data. Although intuitive, these methods suffer from two major drawbacks.

First, most off-the-shelf machine translation systems typically generate only their best translation for a given text. Since machine translation is known to be a notoriously hard problem, applying monolingual text classification algorithms directly on the erroneous translation of training or test data may severely deteriorate the classification accuracy.

Second, similar to domain adaptation in statistical machine learning, due to the discrepancy of data distribution between the training domain and test domain, data distribution across languages may vary because of the difference of culture, people's interests, linguistic expression in different language regions. So even if the translation of training or test data is perfectly correct, the cross language classifier may not perform as well as the monolingual one trained and tested on the data from the same language.

In this paper, we propose a new approach to CLTC, which trains a classification model in the source language and ports the model to the target language, with the translation knowledge learned using the EM algorithm. Unlike previous methods based on machine translation (Fortuna and Shawe-Taylor, 2005), our method takes into account dif-

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 1057–1067,
MIT, Massachusetts, USA, 9-11 October 2010.

ferent possible translations for model features. The translated model serves as an initial classifier for a semi-supervised process, by which the model is further adjusted to fit the distribution of the target language. Our method does not require any labeled data in the target language, nor a machine translation system. Instead, the only requirement is a reasonable amount of unlabeled data in the target language, which is often easy to obtain.

In the following sections, we first review related work. In section 3, we introduce our method that translates the classification model with the translation knowledge learned using the EM algorithm. Section 4 describes model adaptation by training the translated model with unlabeled documents in the target language. Experiments and evaluations are presented in section 5 and finally we conclude the paper in section 6.

2 Related Work

Text classification has rightfully received a lot of attention from both the academic and industry communities, being one of the areas in natural language processing that has a very large number of practical applications. Text classification techniques have been applied to many diverse problems, ranging from topic classification (Joachims, 1997), to genre detection (Argamon et al., 1998), opinion identification (Pang and Lee, 2004), spam detection (Sahami et al., 1998), gender and age classification (Schler et al., 2006).

Text classification is typically formulated as a learning task, where a classifier learns how to distinguish between categories in a given set, using features automatically extracted from a collection of documents. In addition to the learning methodology itself, the accuracy of the text classifier also depends to a large extent upon the amount of training data available at hand. For instance, distinguishing between two categories for which thousands of manually annotated examples are already available is expected to perform better than trying to separate categories that have only a handful of labeled documents.

Some of the most successful approaches to date for text classification involve the use of machine learning methods, which assume that enough annotated data is available such that a classification model can be automatically learned. These include algorithms such as Naive Bayes (Joachims, 1997; McCallum and Nigam, 1998), Rocchio classifiers (Joachims, 1997; Moschitti, 2003), Maximum Entropy (Nigam et al., 1999) or Support Vector Machines (Vapnik, 1995; Joachims, 1998). If only a small amount of annotated data is available, the alternative is to use semi-supervised bootstrapping methods such as co-training or self-training, which can also integrate raw unlabeled data into the learning model (Blum and Mitchell, 1998; Nigam and Ghani, 2000).

Despite the attention that monolingual text classification has received from the research community, there is only very little work that was done on cross-lingual text classification. The work that is most closely related to ours is (Gliozzo and Strapparava, 2006), where a multilingual domain kernel is learned from comparable corpora, and subsequently used for the cross-lingual classification of texts. In experiments run on Italian and English, Gliozzo and Strapparava showed that the multilingual domain kernel exceeds by a large margin a bag-of-words approach. Moreover, they demonstrated that the use of a bilingual dictionary can drastically improve the performance of the models learned from corpora.

(Fortuna and Shawe-Taylor, 2005; Olsson et al., 2005) studied the use of machine translation tools for the purpose of cross language text classification and mining. These approaches typically translate the training data or test data into the same language, followed by the application of a monolingual classifier. The performance of such classifiers very much depends on the quality of the machine translation tools. Unfortunately, the development of statistical machine translation systems (Brown et al., 1993) is hindered by the lack of availability of parallel corpora and the quality of their output is often erroneous. Several methods were proposed (Shi et al., 2006; Nie et al., 1999) to automatically acquire a large quantity of parallel sentences from the web, but such web data is however predominantly confined to a limited number of domains and language pairs.

(Dai et al., 2007) experimented with the use of transfer learning for text classification. Although in this method the transfer learning is performed across

different domains in the same language, the underlying principle is similar to CLTC in the sense that different domains or languages may share a significant amount of knowledge in similar classification tasks. (Blum and Mitchell, 1998) employed semi-supervised learning for training text classifiers. This method bootstraps text classifiers with only unlabeled data or a small amount of labeled training data, which is close to our setting that tries to leverage labeled data and unlabeled data in different languages to build text classifiers.

Finally, also closely related is the work carried out in the field of sentiment and subjectivity analysis for cross-lingual classification of opinions. For instance, (Mihalcea et al., 2007) use an English corpus annotated for subjectivity along with parallel text to build a subjectivity classifier for Romanian. Similarly, (Banea et al., 2008) propose a method based on machine translation to generate parallel texts, followed by a cross-lingual projection of subjectivity labels, which are used to train subjectivity annotation tools for Romanian and Spanish. A related, yet more sophisticated technique is proposed in (Wan, 2009), where a co-training approach is used to leverage resources from both a source and a target language. The technique is tested on the automatic sentiment classification of product reviews in Chinese, and showed to successfully make use of both cross-language and within-language knowledge.

3 Cross Language Model Translation

To make the classifier applicable to documents in a foreign language, we introduce a method where model features that are learned from the training data are translated from the source language into the target language. Using this translation process, a feature associated with a word in the source language is transferred to a word in the target language so that the feature is triggered when the word occurs in the target language test document.

In a typical translation process, the features would be translated by making use of a bilingual dictionary. However, this translation method has a major drawback, due to the ambiguity usually associated with the entries in a bilingual dictionary: a word in one language can have multiple translations in another language, with possibly disparate meanings. If an incorrect translation is selected, it can distort the classification accuracy, by introducing erroneous features into the learning model. Therefore, our goal is to minimize the distortion during the model translation process, in order to maximize the classification accuracy in the target language.

In this paper, we introduce a method that employs the EM algorithm to automatically learn feature translation probabilities from labeled text in the source language and unlabeled text in the target language. Using the feature translation probabilities, we can derive a classification model for the target language from a mixture model with feature translations.

3.1 Learning Feature Translation Probabilities with EM Algorithm

Given a document d from the document collection D in the target language, the probability of generating the document $P(d)$ is the mixture of generating d with different classes $c \in C$:

$$P(d) = \sum_c P(d|c)P(c)$$

In our cross-lingual setting, we view the generation of d given a class c as a two step process. In the first step, a pseudo-document d' is generated in the source language, followed by a second step, where d' is translated into the observed document d in the target language. In this generative model, d' is a latent variable that cannot be directly observed. Since d could have multiple translations d' in the source language, the probability of generating d can then be reformulated as a mixture of probabilities as in the following equation.

$$P(d) = \sum_c P(c) \sum_{d'} P(d|d', c)P(d'|c)$$

According to the bag-of-words assumption, the document translation probability $P(d|d', c)$ is the product of the word translation probabilities $P(w_i|w'_i, c)$, where w'_i in d' is the source language word that w_i is translated from. $P(d'|c)$ is the product of $P(w'_i|c)$. The formula is rewritten as:

$$P(d) = \sum_c P(c) \sum_{d'} \prod_{i=1}^{l} P(w_i|w'_i, c)P(w'_i|c)$$

where w_i is the i^{th} word of the document d with l words. The prior probability $P(c)$ and the probability of the source language word w' given class c are estimated using the labeled training data in the source language, so we use them as known parameters. $P(w_i|w'_i, c)$ is the probability of translating the word w'_i in the source language to the word w_i in the target language given class c, and these are the parameters we want to learn from the corpus in the target language.

Using the Maximum Likelihood Estimation (MLE) framework, we learn the model parameters θ – the translation probability $P(w_i|w'_i, c)$ – by maximizing the log likelihood of a collection of documents in the target language:

$$\begin{aligned}\widehat{\theta} &= argmax_\theta \sum_{j=1}^{m} log(P(d_j, \theta)) \\ &= argmax_\theta \sum_{j=1}^{m} log(\sum_c P(c) \sum_{d'} \\ &\quad \prod_{i=1}^{l_j} P(w_i|w'_i, c)P(w'_i|c))\end{aligned}$$

where m is the number of documents in the corpus in the target language and l_j is the number of words in the document d_j.

In order to estimate the optimal values of the parameters, we use the EM algorithm (Dempster et al., 1977). At each iteration of EM we determine those values by maximizing the expectation using the parameters from the previous iteration and this iterative process stops when the change in the parameters is smaller than a given threshold. We can repeat the following two steps for the purpose above.

- E-step

$$\begin{aligned}P(w'c|w) &\leftarrow \frac{P(cw'w)}{P(w)} \\ &= \frac{P(w|w'c)P(w'c)}{\sum_c \sum_{w'} P(w|w'c)P(w'c)} \quad (1)\end{aligned}$$

- M-step

$$P(w|w'c) \leftarrow \frac{f(w)P(w'c|w)}{\sum_{w \in K} f(w)P(w'c|w)} \quad (2)$$

Algorithm 1 EM algorithm for learning translation probabilities

$D_l \leftarrow$ labeled data in the source language
$D_u \leftarrow$ unlabeled data in the target language
$L \leftarrow$ bilingual lexicon

1: Initialize $P_0(w|w'c) = \frac{1}{n_{w'}}$, where $(w, w') \in L$, otherwise $P_0(w|w'c) = 0$;
2: Compute $P(w'c)$ with D_l according to equation 3
3: **repeat**
4: Calculate $P_t(w'c|w)$ with D_u based on $P_{t-1}(w|w'c)$ according to equation 1
5: Calculate $P_t(w|w'c)$ based on $P_{t-1}(w'c|w)$ according to equation 2
6: **until** change of $P(w|w'c)$ is smaller than the threshold
7: **return** $P(w|w'c)$

Here $f(w)$ is the occurrence frequency of the word w in the corpus. K is the set of translation candidates in the target language for the source language word w' according to the bilingual lexicon. P(w'c) is the probability of occurrence of the source language word w' under the class c. It can be estimated from the labeled source language training data available as follows and it is regarded as a known parameter of the model.

$$P(w'c) = \frac{f(w'c)}{\sum_{w' \in V} f(w'c)} \quad (3)$$

where V is the vocabulary of the source language. Algorithm 1 illustrates the EM learning process, where $n_{w'}$ denotes the number of translation candidates for w' according to the bilingual lexicon.

Our method requires no labeled training data in the target language. Many statistical machine translation systems such as IBM models (Brown et al., 1993) learn word translation probabilities from millions of parallel sentences which are mutual translations. However, large scale parallel corpora rarely exist for most language pairs. (Koehn and Knight, 2000) proposed to use the EM algorithm to learn word translation probabilities from non-parallel monolingual corpora. However, this method estimates only class independent translation probabilities $P(w_i|w'_i)$, while our approach is able to learn class specific translation probabilities

$P(w_i|w'_i, c)$ by leveraging available labeled training data in the source language. For example, the probability of translating "bush" as "树丛" (small trees) is higher than translating as "布什" (U.S. president) when the category of the text is "botany."

3.2 Model Translation

In order to classify documents in the target language, a straightforward approach to transferring the classification model learned from the labeled source language training data is to translate each feature from the bag-of-words model according to the bilingual lexicon. However, because of the translation ambiguity of each word, a model in the source language could be potentially translated into many different models in the target language. Thus, we think of the probability of the class of a target language document as the mixture of the probabilities by each translated model from the source language model, weighed by their translation probabilities.

$$P(c|d, m_t) \approx \sum_{m'_t} P(m'_t|m_s, c)P(c|d, m'_t)$$

where m_t is the target language classification model and m'_t is a candidate model translated from the model m_s trained on the labeled training data in the source language. This is a very generic representation for model translation and the model m could be any type of text classification. Specifically in this paper, we take the Maximum Entropy (ME) model(Berger et al., 1996) as an example for the model translation across languages, since the ME model is one of the most widely used text classification models. The maximum entropy classifier takes the form

$$P(c|d) = \frac{1}{Z(d)} \prod_{w \in V} e^{\lambda_w f(w,c)}$$

where: V is the vocabulary of the language; $f(w, c)$ is the feature function associated with the word w and class c and its value is set to 1 when w occurs in d and the class is c or otherwise 0. λ_w is the feature weight for $f(w_i, c)$ indicating the importance of the feature in the model. During model translation, the feature weight for $f(w_i, c)$ is transferred to $f(w'_i, c)$ in the target language model, where w'_i is the translation of w_i. $Z(d)$ is the normalization factor which is invariant to c and hence we can omit it for classification since our objective is to find the best c. According to the formulation of the Maximum Entropy model, the document can be classified as follows.

$$\widehat{c} = argmax_{c \in C} \sum_{m'_t} P(m'_t|m_s, c) \prod_{i=1}^{v} e^{\lambda_{w_s^i} f(w_t^i, c)}$$

The model translation probability $P(m'_t|m_s, c)$ can be modeled as the product of the translation probabilities of each of its individual bag-of-words features $P(m'_t|m_s, c) \approx \prod_{i=1}^{l} P(w_t^i|w_s^i, c)$ and the classification model can be further written as

$$\widehat{c} = argmax_{c \in C} \sum_{m'_t} \prod_{i=1}^{v} P(w_t^i|w_s^i, c) e^{\lambda_{w_s^i} f(w_t^i, c)}$$

where feature translation probabilities $P(w_t^i|w_s^i, c)$ are estimated with the EM algorithm described in the previous section. Note that if the average number of translations for a word w is n and v is the number of words in the vocabulary there are n^v possible models m'_t translated from m_s. However, we can do the following mathematical transformation on the equation which leads to a polynomial time complexity algorithm. The idea is that instead of enumerating the exponential number of different translations of the entire model, we will instead handle one feature at a time.

$$\sum_{m'_t} \prod_{i=1}^{v} P(w_t^i|w_s^i, c) e^{\lambda_{w_s^i} f(w_t^i, c)} =$$
$$\sum_{j=1}^{n_1} P(w_t^{1j}|w_s^1, c) e^{\lambda_1 f(w_t^{1j}, c)} \sum_{m_t^{2,v}} \prod_{i=2}^{v} P(w_t^i|w_s^i, c) e^{\lambda_i f(w_t^i, c)}$$

Here w_1 is the first word in the vocabulary of the source language and w_{1j} is a translation of w_1 in the target language with n denoting the number its translations according to the bilingual lexicon. $\sum_{m_t^{2,v}}$ are all the target language models translated from the model consisting of the rest of the words w_2 ... w_v in the source language. This process is recursive until the last word w_s^v of the vocabulary and this transforms the equation into a polynomial form as

follows.

$$\sum_{m'_t} \prod_{i=1}^{v} P(w_t^i|w_s^i, c) e^{\lambda_{w_s^i} f(w_t^i, c)}$$
$$= \prod_{i=1}^{v} \sum_{j=1}^{n_i} P(w_t^{ij}|w_s^i, c) e^{\lambda_{w_s^i} f(w_t^{ij}, c)}$$

Based on the above transformation, the class $\widehat{c}$ for the target language document d is then calculated with the following equation.

$$\widehat{c} = argmax_{c \in C} \prod_{i=1}^{v} \sum_{j=1}^{n_i} P(w_t^{ij}|w_s^i, c) e^{\lambda_{w_s^i} f(w_t^{ij}, c)}$$

The time complexity of computing the above equation is $n \times v$.

4 Model Adaptation with Semi-Supervised Learning

In addition to translation ambiguity, another challenge in building a classifier using training data in a foreign language is the discrepancy of data distribution in different languages. Direct application of a classifier translated from a foreign model may not fit well the distribution of the current language. For example, a text about "sports" in (American) English may talk about "American football," "baseball," and "basketball," whereas Chinese tend to discuss about "soccer" or "table tennis."

To alleviate this problem, we employ semi-supervised learning in order to adapt the model to the target language. Specifically, we first start by using the translated classifier from English as an initial classifier to label a set of Chinese documents. The initial classifier is able to correctly classify a number of unlabeled Chinese documents with the knowledge transferred from English training data. For instance, words like "game(比赛)," "score(比分)," "athlete(运动员)," learned from English can still effectively classify Chinese documents. We then pick a set of labeled Chinese documents with high confidence to train a new Chinese classifier. The new classifier can then learn new knowledge from these Chinese documents. E.g. it can discover that words like "soccer(足球)" or "badminton(羽毛球)" occur frequently in the Chinese "sports" documents, while words that are frequently occurring in English documents such as "superbowl(超级碗)" and "NHL(全美冰球联盟)" do not occur as often. Re-training the classifier with the Chinese documents can adjust the feature weights for these words so that the model fits better the data distribution of Chinese documents, and thus it improves the classification accuracy. The new classifier then re-labels the Chinese documents and the process is repeated for several iterations. Algorithm 2 illustrates this semi-supervised learning process.

Algorithm 2 Semi-supervised learning for cross-lingual text classification

$L_s \leftarrow$ labeled data in the source language
$U_t \leftarrow$ unlabeled data in the target language

1: $C_s = train(L_s)$
2: $C_t = translate(C_s)$
3: **repeat**
4: $\quad Label(U, C_t)$
5: $\quad L \leftarrow select(confidence(U, C_t))$
6: $\quad C_t \leftarrow train(L)$
7: **until** stopping criterion is met
8: **return** C_t

The confidence score associated with the documents is calculated based on the probabilities of the class. For a binary classifier the confidence of classifying the document d is calculated as:

$$confidence(d) = \left| log(\frac{P(c|d)}{P(\overline{c}|d)}) \right|$$

An unlabeled document is selected as training data for a new classifier when its confidence score is above a threshold.

5 Experiments and Evaluation

To evaluate the effectiveness of our method, we carry out several experiments. First, we compare the performance of our method on five different categories, from five different domains, in order to see its generality and applicability on different domains. We also run experiments with two different language pairs - English-Chinese and English-French - to see if the distance between language families influences the effectiveness of our method.

To determine the performance of the method with respect to other approaches, we compare the classification accuracy with that of a machine translation

approach that translates the training (test) data from the source language to the target language, as well as with a classifier trained on monolingual training data in the target language.

Finally, we evaluate the performance of each of the two steps of our proposed method. First, we evaluate the model translated with the parameters learned with EM, and then the model after the semi-supervised learning for data distribution adaptation with different parameters, including the number of iterations and different amounts of unlabeled data.

5.1 Data Set

Since a standard evaluation benchmark for cross-lingual text classification is not available, we built our own data set from Yahoo! RSS news feeds. The news feed contains news articles from October 1st 2009 to December 31st 2009. We collected a total of 615731 news articles, categorized by their editors into topics such as "sports" or "business". We selected five categories for our experiments, namely "sports", "health", "business", "entertainment", "education". The Yahoo! RSS news feed includes news in many languages, including English, Chinese, French, Spanish, and others.

We experimented on two language pairs, English-Chinese and English-French, selected for their diversity: English and Chinese are disparate languages with very little common vocabulary and syntax, whereas English and French are regarded as more similar. We expect to evaluate the impact of the distance of languages on the effectiveness of our method. In both cases, English is regarded as the source language, where training data are available, and Chinese and French are the target languages for which we want to build text classifiers. Note that regardless of the language, the documents are assigned with one of the five category labels mentioned above. Table 1 shows the distribution of documents across categories and across languages.

Category	English	Chinese	French
sports	23764	14674	18398
health	15627	11769	12745
business	34619	23692	28740
entertainment	26876	21470	23756
education	16488	14353	15753

Table 1: number of documents in each class

Before building the classification model, several preprocessing steps are applied an all the documents. First, the HTML tags are removed, and advertisements and navigational information are also eliminated. For the Chinese corpus, all the Chinese characters with BIG5 encoding are converted into GB2312 and the Chinese texts are segmented into words. For the translation, we use the LDC bilingual dictionary[1] for Chinese English and "stardict" [2] for Spanish English.

5.2 Model Translation

To transfer a model learned in one language to another, we can translate all the bag-of-word features according to a bilingual lexicon. Due to the translation ambiguity of each feature word, we compare three different ways of model translation. One method is to equally assign probabilities to all the translations for a given source language word, and to translate a word we randomly pick a translation from all of its translation candidates. We denote this as "EQUAL" and it is our baseline method. Another way is to calculate the translation probability based on the frequencies of the translation words in the target language itself. For instance, the English word "bush" can be translated into "布什" , "树丛" or "套管" . We can obtain the following unigram counts of these translation words in our Yahoo! RSS news corpus.

count	translation	sense
582	布什	Goerge W. Bush
43	树丛	small trees
2	套管	canula

We can estimate that $P(布什|bush) = 582/(582 + 43+2) = 92.8\%$ and so forth. This method often allows us to estimate reasonable translation probabilities and we use "UNIGRAM" to denote this method. And finally the third model translation approach is to use the translation probability learned with the EM algorithm proposed in this paper. The initial parameters of the EM algorithm are set to the probabilities calculated with the "UNIGRAM" method and we use 4000 unlabeled documents in Chinese

[1] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002L27

[2] http://stardict.sourceforge.net/Dictionaries.php

to learn translation probabilities with EM. We first train an English classification model for the topic of "sport" and then translate the model into Chinese using translation probabilities estimated by the above three different methods. The three translated models are applied to Chinese test data and we measure the precision, recall and F-score as shown in Table 2.

Method	P	R	F
$EQUAL$	71.1	70.6	70.8
$UNIGRAM$	79.5	77.8	78.6
EM	83.1	84.7	83.9

Table 2: Comparison of different methods for model translation

From this table we can see that the baseline method has lowest classification accuracy due to the fact that it is unable to handle translation ambiguity since picking any one of the translation word is equally likely. "UNIGRAM" shows significant improvement over "EQUAL" as the occurrence count of the translation words in the target language can help disambiguate the translations. However occurrence count in a monolingual corpus may not always be the true translation probability. For instance, the English word "work" can be translated into "工作(labor)" and "工厂(factory)" in Chinese. However, in our Chinese monolingual news corpus, the count for "工厂(factory)" is more than that of "工作(labor)" even though "工作(labor)" should be a more likely translation for "work". The "EM" algorithm has the best performance as it is able to learn translation probabilities by looking at documents in both source language and target language instead of just a single language corpus.

5.3 Cross Language Text Classification

To evaluate the effectiveness of our method on cross language text classification, we implement several methods for comparison. In each experiment, we run a separate classification for each class, using a one-versus-all binary classification.

ML (Monolingual). We build a monolingual text classifier by training and testing the text classification system on documents in the same language. This method plays the role of an upper-bound, since the best classification results are expected when monolingual training data is available.

MT (Machine Translation). We use the Systran 5.0 machine translation system to translate the documents from one language into the other in two directions. The first direction translates the training data from the source language into the target language, and then trains a model in the target language. This direction is denoted as **MTS**. The second direction trains a classifier in the source language and translates the test data into the source language. This direction is denoted as **MTT**. In our experiments, Systran generates the single best translation of the text as most off-the-shelf machine translation tools do.

EM (Model Translation with EM). This is the first step of our proposed method. We used 4,000 unlabeled documents to learn translation probabilities with the EM algorithm and the translation probabilities are leveraged to translate the model. The rest of the unlabeled documents are used for other experimental purpose.

SEMI (Adapted Model after Semi-Supervised Learning). This is our proposed method, after both model translation and semi-supervised learning. In the semi-supervised learning, we use 6,000 unlabeled target language documents with three training iterations.

In each experiment, the data consists of 4,000 labeled documents and 1,000 test documents (e.g., in the cross-lingual experiments, we use 4,000 English annotated documents and 1,000 Chinese or French test documents). For a given language, the same test data is used across all experiments.

Table 3 shows the performance of the various classification methods. The **ML** (Monolingual) classifier has the best performance, as it is trained on labeled data in the target language, so that there is no information loss and no distribution discrepancy due to a model translation. The **MT** (machine translation) based approach scores the lowest accuracy, probably because the machine translation software produces only its best translation, which is often error-prone, thus leading to poor classification accuracy. In addition, the direct application of a classification model from one language to an-

Category	English → Chinese														
	ML			MTS			MTT			EM			SEMI		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
sports	96.1	94.3	95.2	80.6	81.7	81.2	81.7	83.8	82.7	83.1	84.7	83.9	92.1	91.8	91.9
health	95.1	93.1	94.1	80.8	81.5	81.2	81.6	83.5	82.6	84.5	85.8	85.2	90.2	91.7	90.9
business	91.6	93.1	92.4	81.3	81.9	81.6	80.7	81.0	80.9	81.6	82.0	81.8	87.3	89.3	88.3
entertainment	88.1	88.3	88.2	76.1	78.8	77.5	75.3	78.9	77.1	76.8	79.7	78.2	83.2	83.8	83.5
education	79.1	82.2	80.6	70.2	72.5	71.8	71.1	72.0	71.6	71.2	73.7	72.5	76.2	79.8	78.0
	English → French														
sports	95.8	95.0	95.4	82.8	83.6	83.2	82.1	83.0	82.5	85.3	87.1	86.2	92.5	92.1	92.3
health	94.2	94.5	94.3	82.6	83.9	83.2	81.8	83.0	82.4	86.2	87.2	86.6	92.0	92.2	92.1
business	90.1	92.2	91.1	81.4	82.1	81.7	81.3	81.8	81.8	84.4	84.3	84.4	88.3	89.2	88.8
entertainment	87.4	87.2	87.3	76.6	79.1	77.8	76.0	78.8	77.4	78.9	81.0	80.0	84.3	85.5	84.9
education	78.8	81.8	80.3	72.1	74.8	73.5	72.3	72.7	72.5	73.8	76.2	75.0	76.3	80.1	78.2

Table 3: Comparison of different methods and different language pairs

other does not adapt to the distribution of the second language, even if the documents belong to the same domain. Comparing the two **MT** alternatives, we can see that translating the training data (**MTS**) has better performance than translating the test data (**MTT**). The reason is that when the model is trained on the translated training data, the model parameters are learned over an entire collection of translated documents, which is less sensitive to translation errors than translating a test document on which the classification is performed individually.

Our **EM** method for translating model features outperforms the machine translation approach, since it does not only rely on the best translation by the machine translation system, but instead takes into account all possible translations with knowledge learned specifically from the target language. Additionally, the **SEMI** (semi-supervised) learning is shown to further improve the classification accuracy. The semi-supervised learning is able to not only help adapt the translated model to fit the words distribution in the target language, but it also compensates the distortion or information loss during the model translation process as it can down-weigh the incorrectly translated features.

The improvement in performance for both the **EM** and the **SEMI** methods is consistent across the five different domains, which indicates that the methods are robust and they are insensitive to the domain of the data.

The performance of the two language pairs English-Chinese and English-French shows a difference as initially hypothesized. In both the **EM** and the **SEMI** models, the classification accuracy of English-French exceeds that of English-Chinese, which is probably explained by the fact that there is less translation ambiguity in similar languages, and they have more similar distributions. Note that the monolingual models in French and Chinese perform comparably, which means the difficulty of the test data is similar between the two target languages.

5.4 Model Adaptation with Semi-Supervised Learning

Finally, to gain further insights into our proposed adaptation method, we run several experiments with different parameters for the semi-supervised learning stage. As these experiments are very time consuming, we run them only on Chinese.

For each of the five categories, we train a classification model using the 4,000 training documents in English and then translate the model into Chinese with the translation parameters learned with **EM** on 20,000 unlabeled Chinese documents. Then we further train the translated model on a set of unlabeled Chinese documents using a different number of iterations and a different amount of unlabeled documents. Figures 1 and 2 show the results of these evaluations.

As the plots show, the use of unlabeled data in the target language can improve the cross-language classification by learning new knowledge in the target language. Larger amounts of unlabeled data in general help, although the marginal benefit drops with increasing amounts of data. Regarding the number of iterations, the best performance is

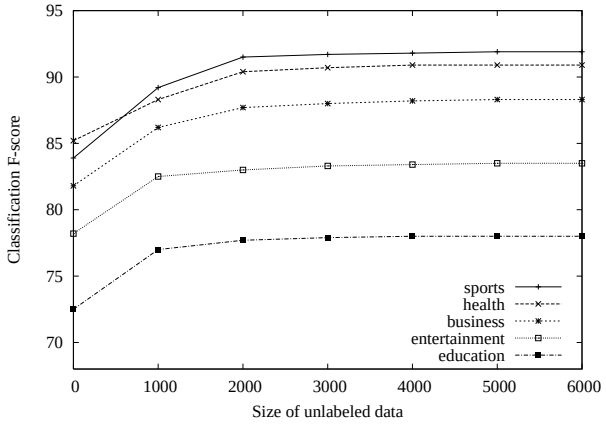

Figure 1: Change in classification F-score for an increasing amount of unlabeled data in the target language

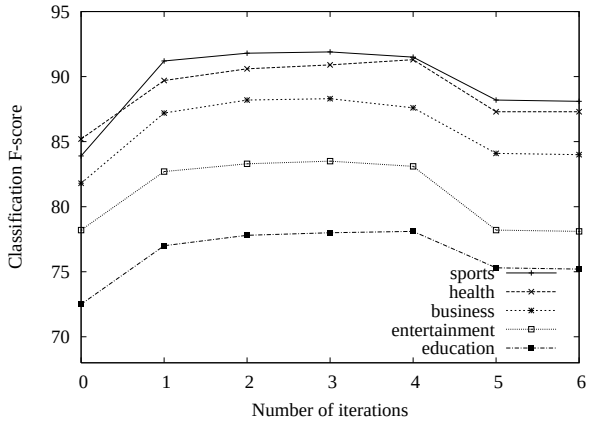

Figure 2: Change in classification F-score for a different number of iterations

achieved after 3-4 iterations.

6 Conclusions

In this paper, we proposed a novel method for cross-lingual text classification. Our method ports a classification model trained in a source language to a target language, with the translation knowledge being learned using the EM algorithm. The model is further tuned to fit the distribution in the target language via semi-supervised learning. Experiments on different datasets covering different languages and different domains show significant improvement over previous methods that rely on machine translation. Moreover, the cross-lingual classification accuracy obtained with our method was found to be close to the one achieved using monolingual text classification.

Acknowledgments

The work of the second author has been partially supported by National Science Foundation awards #0917170 and #0747340. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

S. Argamon, M. Koppel, and G. Avneri. 1998. Style-based text categorization: What newspaper am i reading? In *AAAI-98 Workshop on Learning for Text Categorization*, Madison.

C. Banea, R. Mihalcea, J. Wiebe, and S. Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, Honolulu, Hawaii.

A. Berger, S. Della Pietra, and V. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

A. Blum and T. Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers*, June.

P. Brown, S. della Pietra, V. della Pietra, and R. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2).

W. Dai, G. Xue, Q. Yang, and Y. Yu. 2007. Transferring naive bayes classifiers for text classification. In *In Proceedings of the 22nd AAAI Conference on Artificial Intell igence*, pages 540–545.

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 (Methodological)*, 39(1).

B. Fortuna and J. 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)*.

A. Gliozzo and C. Strapparava. 2006. Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In *Proceedings of the Conference of the Association for Computational Linguistics*, Sydney, Australia.

R. Hwa, P. Resnik, and A. Weinberg. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*. Special issue on Parallel Texts, editors R. Mihalcea and M. Simard.

T. Joachims. 1997. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning*, Nashville, US.

T. Joachims. 1998. Text categorization with Support Vector Machines: learning with mny relevant features. In *Proceedings of the European Conference on Machine Learning*, pages 137–142.

P. Koehn and K. Knight. 2000. Estimating word translation probabilities from unrelated monolingua l corpora using the em algorithm. In *National Conference on Artificial Intelligence (AAAI 2000) Lang kilde*, pages 711–715.

A. McCallum and K. Nigam. 1998. A comparison of event models for Naive Bayes text classification. In *Proceedings of AAAI-98 Workshop on Learning for Text Categorization*.

R. Mihalcea, C. Banea, and J. Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the Association for Computational Linguistics*, Prague, Czech Republic.

C. Monz and B.J. Dorr. 2005. Iterative translation disambiguation for cross-language information retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil.

A. Moschitti. 2003. A study on optimal paramter tuning for Rocchio text classifier. In *Proceedings of the European Conference on Information Retrieval*, Italy.

H.T. Ng, B. Wang, and Y.S. Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, Sapporo, Japan, July.

J.-Y. Nie, M. Simard, P. Isabelle, and R. Durand. 1999. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*.

K. Nigam and R. Ghani. 2000. Analyzing the effectiveness and applicability of co-training. In *Proceedings of the Conference on Information and Knowledge Management (CIKM 2000)*, McLean, VA, November.

K. Nigam, J. Lafferty, and A. McCallum. 1999. Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*.

J.S. Olsson, D. W. Oard, and J. Hajic. 2005. Cross-language text classification. In *Proceedings of the 28th Annual international ACM SIGIR Conference on Research and Development in information Retrieval*.

B. Pang and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics*, Barcelona, Spain, July.

M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. 1998. A Bayesian approach to filtering junk e-mail. In *AAAI-98 Workshop on Learning for Text Categorization*, Madison.

J. Schler, M. Koppel, S. Argamon, and J. Pennebaker. 2006. Effects of age and gender on blogging. In *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, pages 199–204, Stanford.

L. Shi, C. Niu, M. Zhou, and J. Gao. 2006. A dom tree alignment model for mining parallel data from the web. In *Proceedings of the Annual Meeting of the Association for Computational Lingusitics (ACL 2006)*, Sydney, Australia.

V. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer, New York.

X. Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the Association of Computational Linguistics and the International Joint Conference on Natural Language Processing*, Singapore, August.