An EM based training algorithm for Cross-Language Text Categorization

Leonardo Rigutini and Marco Maggini
Dipartimento di Ingegneria dell'Informazione
Università di Siena
Via Roma 56 I-53100 Siena, Italy
{rigutini,maggini}@dii.unisi.it

Bing Liu
Department of Computer Science
University of Illinois at Chicago
Chicago, Illinois - USA
liub@cs.uic.edu

Abstract

Due to the globalization on the Web, many companies and institutions need to efficiently organize and search repositories containing multilingual documents. The management of these heterogeneous text collections increases the costs significantly because experts of different languages are required to organize these collections. Cross-Language Text Categorization can provide techniques to extend existing automatic classification systems in one language to new languages without requiring additional intervention of human experts. In this paper we propose a learning algorithm based on the EM scheme which can be used to train text classifiers in a multilingual environment. In particular, in the proposed approach, we assume that a predefined category set and a collection of labeled training data is available for a given language L_1. A classifier for a different language L_2 is trained by translating the available labeled training set for L_1 to L_2 and by using an additional set of unlabeled documents from L_2. This technique allows us to extract correct statistical properties of the language L_2 which are not completely available in automatically translated examples, because of the different characteristics of language L_1 and of the approximation of the translation process. Our experimental results show that the performance of the proposed method is very promising when applied on a test document set extracted from newsgroups in English and Italian.

1 Introduction

Automatic Text Categorization is a well studied task with many effective techniques. Nowadays, the most popular and successful algorithms for text classification are based on machine learning techniques. Learning systems have the advantage of flexibility since the only required human effort is to provide a consistent set of labeled examples. However, labeling a dataset can be costly because different expertises may be required. In fact, in recent years also because of the growth in the popularity of the Web, many companies and organizations were required to manage documents in different languages. Multi-Language Text Classification became an important task. This task can be approached as a set of Mono-Language problems, but a different labeled dataset would be needed to train a classifier for each language. This labeling process can be quite costly since it needs an expert for each different language.

Cross-Language Text Categorization (CLTC) is a new area in text categorization. The CLTC task can be stated as follows: suppose we have a good classifier for a set of categories in a language L_1 and a large amount of unlabeled data in a different language L_2; how can we categorize this corpus according to the same categories defined for language L_1 without having to manually label any data in L_2? When using the machine learning paradigm, this problem can be reformulated as: how can we train a text classifier for language L_2 using the examples labeled for language L_1? An algorithm that is able to effectively perform this task would reduce the costs of building multi-language classification systems, since the human effort would be reduced to provide a training set in just one language.

The approach that we propose is based on two steps: first the training set available in the language L_1 is translated into the target language L_2 using an automatic translation system. This procedure can introduce noise in the translated documents since automatic translation is often approximate, especially if we use simple translation systems. Moreover, the translated documents can have different characteristics compared to the set of documents written in the target language L_2. In the second step, a text classifier for the target language L_2 is trained using the EM algorithm to take advantage both of the labeled examples obtained from the original language L_1 in the first step and of the set of unlabeled data in language L_2. In this way, the properties of the target language can be extracted from the unlabeled examples. The algorithm also requires a proper feature selection technique to avoid to converge to trivial solutions.

The paper is organized as follows. In the next section we discuss possible approaches to Cross-Language Text Classification and review the existing techniques proposed in the literature. Then, Section 3 describes the classical text categorization techniques that are used in the proposed system. In Section 4 the novel EM-based learning algorithm is described. Section 5 contains the experimental results obtained on a test set containing messages extracted from newsgroups in two different languages (English and Italian). Finally, in Section 6 the conclusions are drawn.

2 Cross-Language Text Classification

For reason of simplicity, we reduce the multi-lingual case with k languages to $k-1$ bi-lingual problems selecting one language as the principal one; thus studying the bi-lingual case is not restrictive with respect to the multi-lingual problem. Before describing some aspects of the Cross-Lingual Text Categorization task, we introduce some notations used in the following sections. We denote the two languages with L_1 and L_2 and with $L_{2\to1}$ the language L_1 generated by the translation from L_2. Moreover, we denote with TR_1 and TS_1 the training set and the test set in language L_1, and with $TR_{2\to1}$ and $TS_{2\to1}$ the training set TR_2 and the test set TS_2 translated into the language L_1.

In CLTC, we can imagine three different scenarios:

- **Poly-lingual training**: a labeled training set is available for each language and one classifier is trained using training examples from all the different languages. The feature space used by the classifier corresponds to an unique and enormous language obtained by the union of all the vocabularies. The classifier parameters are estimated by merging the training sets of the different languages.
- **Cross-lingual training**: the labeled training set is available for only one language and we have to use that to classify documents in other languages. This approach is the more interesting one and is what we are interested in this paper. Clearly, this problem is more difficult to solve. In this setting, we can individuate a sub-categorization depending on the approach used to tackle it:
  - **Training-set Translation**: This approach requires the translation of the labeled set into the target language which then is used to train a classifier for this language. We can take advantage of the availability of unlabeled documents in the second language.
  - **Test-set Translation**: In this approach a classifier is trained by using documents in the first language TR_1 and it is tested on the $TS_{2\to1}$ sets.
  - ***"Esperanto"* language**: This approach uses an universal reference language which all documents are translated to. The *esperanto* language (from the legendary language spoken in all the world) should contain all properties of the languages of interest and be organized in a semantic way: words indicating the same concepts in the languages should be translated to the same terms in the *esperanto* language. Some linguistic tools exist to create an universal language. We were not able to find systems in the literature using this approach.

There are very few systems for *CLTC* described in the literature. An interesting work is presented in [7] where some methods are tested on a bi-lingual task for the ILO (International Labor Organization) corpus. The considered languages were English and Spanish and two different classifiers were evaluated (the Rocchio classifier and the Winnow classifier). As baseline they used the results for the *mono-lingual* categorization which yielded an accuracy of about 86%. Using the poly-lingual approach, in which English and Spanish labeled documents were used together to train an unique classifier, they obtained an accuracy of about 81%. Finally, following the *test-set translation* approach they proposed two different methods: *Terminology Translation* and *Profile-Based Translation*. In the first method they simply trained a classifier using the TR_E set and tested it using the $TS_{S\to E}$ (the Spanish test set translated into English). In the second strategy (*Profile-Based Translation*), a reduced vocabulary was created by selecting the 150 most important terms of each class from a preliminary categorization of the English data set and the translation process was performed using this vocabulary. When using this second technique the average accuracy was around 73%, which is still very far from the mono-lingual case of 86%.

3 Text Categorization

Different machine learning models have been applied to Automatic Text Categorization, but the most successful are the *Naive-Bayes (NB)* [5] and *Support Vector Machines (SVM)* [4].

The *Naive-Bayes* is a probabilistic classifier which estimates the probability of the document to be in the modeled class. More formally, a document d_i belongs to the class C_j such that:

$$C_j = \arg\max_{C_r} P(C_r|d_i)\ . \tag{1}$$

Using the Bayes rule, the probability $P(C_j|d_i)$ can be expressed as:

$$P(C_j|d_i) = \frac{P(C_j) \times P(d_i|C_j)}{P(d_i)}\ . \tag{2}$$

Since $P(d_i)$ is a common factor in the models for each class (it does not depend on the class), it can be neglected. The class a-priori probability $P(C_j)$ can be easily estimated from the document distribution in the training set or otherwise can be considered as a constant. Finally, the most important assumption, the *naive* assumption, is that the presence of each word in the documents is an independent event. Even if this hypothesis is not true in practice, it allows to write:

$$P(d_i|C_j) = \prod_{w_t \in d_i} P(w_t|C_j)^{N(w_t,d_i)} , \qquad (3)$$

where $N(w_t, d_i)$ is the number of occurrences of the word w_t in the document d_i. Assuming that each document is drawn from a multinomial distribution of words, the probability of w_t in class C_j can be estimated as:

$$P(w_t|C_j) = \frac{\sum_{i=1}^{|D|} N(w_t, d_i) P(C_j|d_i)}{\sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N(w_v, d_i) P(C_j|d_i)} \qquad (4)$$

where $P(C_j|d_i) \in \{0, 1\}$ and $P(w_t|C_j) = 1$ if the training document d_i is labeled with the class C_j.

The method is very simple and fast and it is the most used in text categorization since, despite the strong simplificative hypotheses, it yields good performances in most cases. However, the estimation for those words that never appear in the training set can yield wrong results. In fact, if a word doesn't appear in the training set, it will have a zero probability of appearing in that class. Thus when it is found in a test document the probability computed in (2) will be null too, even if the document contained many other on-topic terms. This problem is known as the *zero-estimation problem* and can be solved using the smoothing techniques [5]. In particular, we applied the *Good-Turing smoothing* [3], which was showed to have better performance and corresponds to set

$$P(w(0)) = \frac{\#w(1) \in C_j}{\#w \in C_j} , \qquad (5)$$

where $w(n)$ is the event of a word appearing exactly n times. The *Good-Turing smoothing* states that the probability of finding a never-seen word is equal to the probability of having a once-seen word. The idea is that the once-seen words are rare words that have been observed in the training set only by accident, while the never-seen words are rare words not appearing in the training set by chance. Thus the probability of observing a rare word is estimated by counting the frequency of once-seen words.

Term Filtering: Information Gain

In text classification, the goal is to assign labels to the documents according their contents. A very important assumption of most automatic text classifiers is that many words and structures used in natural language are not significant in representing the content of the text document: adverbs, pronouns, generic terms and so on can be safely removed without loosing information. Moreover, many experiments show that they often represent noisy features, which can harm the classifier. Two different approaches to removing non informative words can be pursued [8]:

- editing a list of not relevant words (stop-words) and removing all of them from the vocabulary (and from the data);
- defining an informativeness function that assigns scores to the words, thus removing the words below a given threshold.

Usually both techniques are used in cascade. In categorization tasks, the most common measure of informativeness is the *Information Gain (IG)*. This function measures the gain in information given by the presence and absence of a class C_k by knowing the presence or absence of a word w_i [1]:

$$IG(w_i, C_k) = \sum_{c \in (C_k, \overline{C}_k)} \sum_{w \in (w_i, \overline{w}_i)} P(w, c) \log_2 P(w, c), \qquad (6)$$

where $P(\overline{w}_i, C_k)$ is the probability that, for a random document x belonging to the class C_k, the word w_i doesn't appear in x. This equation refers to a specific category: in order to obtain the score of w_i in a global and category-independent sense, a globalization technique is applied. Usually the sum over the classes is used:

$$IG(w_i) = \sum_{k=1}^{|C|} IG(w_i, C_k) . \qquad (7)$$

If the *IG* value is high, that feature is considered important and informative for a topic and it should not be removed, otherwise it does not provide useful information to determine the topic and it can be removed. Usually using the *IG* measure we can remove a very high number of features, reducing the vocabulary dimension to $100 - 1000$ words and improving the results of the system.

4 An EM based training algorithm for Cross-Lingual classifiers

We designed a Cross-Language Text Classification system taking into account the following requirements. We assume that two different sets of data are available: a set of manually labeled documents in the known language L_1 and a large amount of unlabeled documents in the unknown language L_2. In this scenario, we would like to use the labeled set on L_1 (and eventually the information hidden in the unlabeled data for L_2) to train a classifier for organizing the

collection of documents in the language L_2. Since the unlabeled dataset can contain a large number of documents, an approach based on the test-set translation can be very time consuming as translation is a slow process. Thus, the best solution seems to be the adoption of the training-set translation which requires only the translation of the training set available in language L_1. The so trained classifier $C_{1\rightarrow 2}$ can then be applied on the documents directly in their original language L_2.

In order to use the information available in the unlabeled collection TS_2 in language L_2, the proposed system exploits the EM algorithm [2] by considering the class labels for the documents in TS_2 as unknown variables, that can be estimated in the E step. A similar approach has been proposed for text classification for learning from both labeled and unlabeled data [6]. The idea is that, even if the labels are not available, useful statistical properties can be extracted by looking at the distribution of terms in unlabeled texts. This additional source of knowledge can be particularly useful for Cross-Language learning, since the labeled documents are written in a different language than the target one and their translated versions usually do not have the same statistical properties of the target language (also because of the approximation and errors introduced by the automatic translation process). In fact, using the classifier trained also with documents (although not labeled) written in the same language as the test-set, the results should be better than without using them (both the test-set translation and the training-set translation approaches do not use documents originally written in the target language).

The basic algorithm

The algorithm consists of an initialization step (step 0) and an iteration step (step t) that repeats the E and M phases until we reach convergence. Thus, the algorithm can be sketched as follows:

1. *Initialization*. Train a cross-language classifier $C_{1\rightarrow 2}$ using the labeled set available for language L_1 after it is automatically translated to language L_2 (i.e. the set $TR_{1\rightarrow 2}$). Set $t = 0$ and $C^0 = C_{1\rightarrow 2}$.

2. *E step*. Guess the unknown labels on the collection for the language L_2 (i.e. the set TS_2) using the classifier C^t obtained after t iterations. Define a labeled set $TR_2^t = \{(d, l)|d \in TS_2 \wedge l = C^t(d)\}$.

3. If $TR_2^t = TR_2^{(t-1)}$ for $t > 2$, stop.

4. *M step*. Use the labeled set TR_2^t obtained in the E step to estimate the parameters of the new classifier $C^{(t+1)}$ using the appropriate supervised learning algorithm.

5. Set t=t+1 and go to step 2.

However, this basic algorithm may fail to find useful solutions, as it confirmed by the experimental results presented in the next section. This is due to the fact that the EM algorithm needs to apply some regularization technique to avoid trivial solutions. For example, we can obtain a good explanation of the unlabeled set by assigning all the examples to the same class, which is an easy task to learn. In the experiments we found a very frequent behavior which caused one of the classes to end up with an empty set of examples. This effect is due to many noisy terms in the documents which overwhelm the real informative terms causing the progressive shift of all the examples in one class to another one.

The improved algorithm using IG term filtering

In order to introduce a regularization effect in the EM iteration, a term filtering technique was exploited to keep only the most informative features. We decided to use the Information Gain (IG) scores to identify the most informative words with respect to the classification task. In the modified algorithm a feature selection step is applied before training the classifier at each iteration. In particular the IG filter is applied at the initialization step, selecting the k_1 top score words from the set $TR_{1\rightarrow 2}$. Then a similar feature selection is performed at each iteration of the EM procedure. In particular, for each iteration t, the IG values for the terms are computed using the labeled set TR_2^t and only the first k_2 most informative words are kept in the document representation. In the experiments we set $k_1 < k_2$. In fact, from the experimental evidence, we found that it is preferable to select few informative terms from the translated documents to avoid the inclusion of many noisy words. On the contrary, to capture a good model in the target language a larger dictionary is needed, since many informative terms are extracted from the unlabeled data.

The scheme of the modified algorithm is shown in figure 1.

5 Experimental results

We performed a set of experiments in the bi-lingual case; the extension to the multi-lingual case can be easily done by replicating the system for each language. In the following we describe the experimental setup and the results obtained by different approaches. All the experiments where performed using a *Naive Bayes* classifier with *Good-Turing* smoothing.

The Multi-Lingual dataset

We collected a bi-lingual dataset by downloading messages from newsgroups in English (language L_E) and Ital-

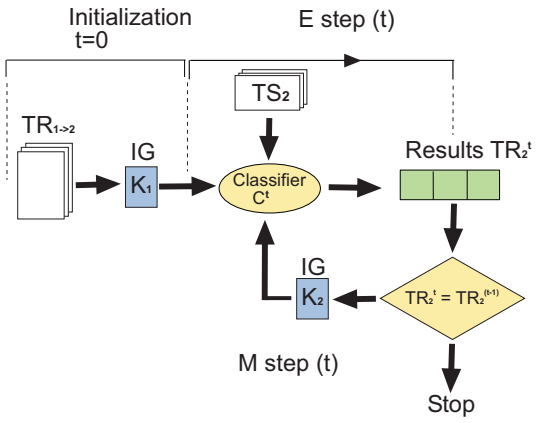

Figure 1. The scheme of algorithm with IG feature selection.

ian (language L_I). We selected three different discussion groups available in the two languages: *'Hardware'*, *'Auto'* and *'Sports'*. We downloaded about 27.000 messages: about 8.000 messages from each Italian newsgroup and 1.000 messages from the English ones. The Italian dataset was split into a training set TR_I and a test set TS_I as detailed in table 1. For each experiment 10 different configurations of these sets were generated by choosing their elements at random in order to perform a ten-fold cross-validation scheme. Notice that the training set TR_I was only used for the baseline Mono-Lingual experiment. Instead, the English training set contained the same 3.000 documents for all the experiments. Finally, we used the

	TRAIN		TEST
	TR_I	TR_E	TS_I
Auto	1.000	1.000	6.988
Hw	1.000	1.000	6.991
Sport	1.000	1.000	6.984
total	3.000	3.000	20.963

Table 1. Composition of the bi-lingual training and test sets used in the experiments.

newsgroup topics as class labels.

The automatic translation of the documents from English to Italian was performed using the trial version of the *Office Translator Idiomax*[1], a plug-in for the *Office XP* suite. We used this tool without any pre-processing on the input text, obtaining the translated training set $TR_{E\rightarrow I}$.

[1] www.idiomax.com

Baseline experiment: mono-lingual classification

The baseline experiment was the mono-lingual categorization test as in [7]. We trained a *Naive-Bayes* classifier with *Good-Turing* smoothing by using the set TR_I and then we tested it on the set TS_I. We removed the most common words by using a stop-word list and we repeated the experiment in ten-fold cross-validation, evaluating precision and recall. The results are shown in table 2. We can treat these as the optimal results because labeled Italian training sets are used. Below, we want to see how close to these results we can achieve using the English training set, i.e., without using any labeled Italian documents.

TS_I test set		Recall	Precision
Auto	6988	$94.01\% \pm 1.03\%$	$93.76\% \pm 1.09\%$
Hw	6991	$96.21\% \pm 0.93\%$	$93.01\% \pm 0.45\%$
Sport	6984	$92.89\% \pm 1.12\%$	$96.74\% \pm 1.24\%$
Total	20963	$94.43\% \pm 0.90\%$	$94.43\% \pm 0.90\%$

Table 2. Recall and precision for the Italian mono-lingual classification. Results are averages on a ten-fold cross-validation.

Basic training-set translation

In the first cross-language experiment, we applied the basic training-set translation approach. We used the translated learning set $TR_{E\rightarrow I}$ to train the classifier and then tested it on the Italian test set TS_I. Notice that this is the configuration obtained by the initialization step of the proposed algorithm without feature selection. The results are obviously worse than in the mono-lingual case (table 3).

TS_I test set		Recall	Precision
Auto	6988	$69.56\% \pm 5.34\%$	$66.56\% \pm 4.76\%$
Hw	6991	$87.24\% \pm 2.02\%$	$63.35\% \pm 3.72\%$
Sport	6984	$50.95\% \pm 6.28\%$	$88.22\% \pm 4.36\%$
Total	20963	$69.26\% \pm 4.22\%$	$69.26\% \pm 4.22\%$

Table 3. Recall and precision obtained by the basic training-set translation approach. Results are averages on a ten-fold cross-validation.

From the results it is evident that the three classes have quite different behaviors: the class *Hw* has a quite high recall and a low precision; the class *Sport* instead has a very low recall and very high precision; finally, the class *Auto* has both low recall and precision. These results are due to a phenomenon existing in dealing with multi-language document corpora, which is analyzed in the next subsection.

In order to evaluate the effect of the feature selection using the IG scores, we performed the basic training-set

translation learning applying the IG filtering to the translated dataset. The results in table 4 show a significant improvement in performance, supporting the introduction of the IG-based feature selection. This approach is basically the Profile-based translation technique proposed in [16] (the only difference is the direction of the translation step).

TS_I test set		Recall	Precision
Auto	6988	$76.08\% \pm 0.83\%$	$67.16\% \pm 0.54\%$
Hw	6991	$76.88\% \pm 1.01\%$	$81.38\% \pm 0.62\%$
Sport	6984	$67.03\% \pm 0.98\%$	$74.06\% \pm 0.92\%$
Total	20963	$73.36\% \pm 0.97\%$	$73.36\% \pm 0.97\%$

Table 4. Precision and recall for the basic training-set translation technique using the IG feature selection with $k_1 = 300$. Results are averages on a ten-fold cross-validation.

Named Entities

A particular phenomenon may happen in a multi-lingual corpora especially when the contents concern aspects of everyday life, as it is in newsgroups. Even if the general topic is the same, the actual contents are strongly dependent on the nation where the writer lives. For example, if we consider the messages in the group *Sport*, the subtopic distribution of the Italian messages (popular sub-topics are *soccer*, *F1*, *basketball*, etc.) are very different from those of American messages (they mostly deal with *baseball*, *football*, *basketball*, etc.). Beside the different distribution among the sub-topics for a broad category such as *Sport*, the terms used may differ depending on the countries. For instance, in the class *Auto* we can find different names of car models, brands and car accessories in different languages. In some other classes this effect has lower impact as it happens in the class *Hardware* where the terms referring to computer brands, device brands, softwares are basically the same for any language. These observations give a possible explanation for the results in table 3, where the recall for the class *Hw* is significantly higher than for the other two classes.

The terms used in messages on the same topic but from different languages have two other peculiarities which can cause problems to the design of cross-language systems:

- the use of terms related to proper names peculiar to a particular geographical area (e.g. teams, players, car models, brands).

- the presence in the L_1 language (in our case English) of words or groups of words which should not be translated, since they are used in the original form also in the L_2 language. An example is the word *windows* referring to the computer operating system which should not be translated because in Italian the same word is used. This problem affects the quality of the translated training set $TR_{E \to I}$ and it can be reduced by using more sophisticated automatic translation systems which take into account domain knowledge.

In both cases a solution would be to tag *Named Entities*, mapping them to cross-lingual tags. The recognition of named entities and the use of appropriate tags would prevent the automatic translator from producing funny translations and give a better representation of the text by using fewer but more significant features. We used a simple *NER* preprocessing step only to avoid the translation of a predefined set of words (a subset of them is reported in table 5). The results in table 6 show that this technique did not yield

Auto	**Hardware**	**Sport**
news	windows	white sox
jaguar	mouse	giants
civic	notebook	braves
honda	apple	hawks
caravan	hard disk	predators
taurus	getaway	eagles
maxima	cache	bears
viper	socket	bulls

Table 5. Some Named Entities which are not translated.

a significant improvement when using the basic training set translation method. In fact, many of these words refer to entities which are peculiar only to the English domain. Since NER in the general case is a quite complex task requiring considerable costs for providing appropriate rules, we decided not to use NER in the final system.

TS_I test set		Recall	Precision
Auto	6988	$72.32\% \pm 1.25\%$	$37.53\% \pm 1.05\%$
Hw	6991	$85.89\% \pm 0.86\%$	$74.28\% \pm 1.74\%$
Sport	6984	$52.85\% \pm 4.32\%$	$92.31\% \pm 1.83\%$
Total	20963	$70.46\% \pm 2.10\%$	$70.46\% \pm 2.10\%$

Table 6. Precision and recall for the basic training-set translation approach when using the *NER* preprocessing step. Results are averages on a ten-fold cross-validation.

EM Cross-Language learning

Table 7 reports the results obtained when applying the basic algorithm using the EM iteration. The bad results for the class *Sport* are due to the fact that the initial classifier trained on the translated messages is not sufficiently selective because of the different subtopic distributions in the Italian and in the English newsgroups. This fact implies that some examples are incorrectly assigned to the other classes

TS_I test set		Recall	Precision
Auto	6988	$71.32\% \pm 1.05\%$	$51.40\% \pm 1.00\%$
Hw	6991	$98.04\% \pm 1.01\%$	$61.55\% \pm 0.98\%$
Sport	6984	$0.73\% \pm 0.00\%$	$65.41\% \pm 0.05\%$
Total	20963	$56.32\% \pm 1.10\%$	$56.32\% \pm 1.10\%$

Table 7. Precision and recall for the basic EM-based Cross-Language learning algorithm. Results are averages on a ten-fold cross-validation.

and this is progressively amplified during the following EM iterations.

Finally, we tested the improved algorithm, which exploits the IG feature selection, by using different configurations for the values k_1 and k_2: $k_1 = 1000$ and $k_2 = 1000$, $k_1 = 300$ and $k_2 = 1000$, $k_1 = 300$ and $k_2 = 300$, $k_1 = 100$ and $k_2 = 1000$. We report only the results for the best configuration, which corresponds to $k_1 = 300$ and $k_2 = 1000$. In this case, we apply a selection on the words in the translated training set, while a larger value for k_2 defines a sufficiently large dictionary for the target language. Table 8 shows the results which are extremely good even compared to the baseline mono-lingual test in table 2. Moreover, comparing with the results in table 4, we can see that the EM iterations are very effective yielding a significant improvement with respect to the corresponding baseline experiment using feature selection. Thus, the proposed algorithm shows a very good accuracy improvement with respect to the other methods proposed so far.

TS_I test set		Recall	Precision
Auto	6988	$92.59\% \pm 1.05\%$	$87.07\% \pm 1.02\%$
Hw	6991	$87.88\% \pm 0.98\%$	$92.78\% \pm 0.88\%$
Sport	6984	$91.01\% \pm 1.03\%$	$92.28\% \pm 0.90\%$
Total	20963	$90.64\% \pm 0.96\%$	$90.64\% \pm 0.96\%$

Table 8. Precision and recall for the improved EM-based Cross-Language learning algorithm using IG filtering with $k_1 = 300$ and $k_2 = 1000$. Results are averages on a ten-fold cross-validation.

6 Conclusions

In this paper we presented a new technique to categorize text documents in a cross-language environment. It is motivated by the availability of documents written in different languages and also by the fact that companies need to build categorization systems for these multi-lingual documents. Manually labeling a large number of documents in each language is very labor intensive and time consuming. Although Cross-Language Text Categorization is an important problem, little research has been done so far. The proposed approach is based on the idea that we can use a known training set in one language to initialize the EM iterations on an unlabeled set of documents written in a different language. Once labeled, these documents are used to train a classifier iteratively without using the labeled documents in the first language. This solves the problem that the document sets in the two languages may be quite dissimilar, or with different distributions. We also show that feature selection is extremely important in this setting. Experimental results demonstrated that our approach is highly effective and reaches recall and precision values comparable with the monolingual case, where manually labeled documents are prepared for each language.

Acknowledgments

We would like to thank Amir Ashkenazi for giving us the problem. The work of Bing Liu is supported by the National Science Foundation (NSF) under the NSF grant IIS-0307239.

References

[1] F. Debole and F. Sebastiani. Supervised term weighting for automated text categorization. In *Proceedings of SAC-03, 18th ACM Symposium on Applied Computing*, 2003.

[2] A. Dempster, N. M. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistic Society*, 39:1–38, 1977.

[3] I. J. Good. The populations frequencies of species and estimation of population parameters. *Biometrika*, 1953.

[4] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In C. Nédellec and C. Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, pages 137–142, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.

[5] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *AAAI/ICML-98 Workshop on Learning for Text Categorization*, pages 41–48. AAAI Press, 1998.

[6] K. Nigam, A. McCallum, and S. T. ans T. Mitchell. Learning to classify text from labeled and unlabeled documents. In *AAAI-98*, 1998.

[7] Nuria Bel, Cornelis H.A. Koster, and Marta Villegas. Cross-lingual text categorization. *European Conference on Digital Libraries (ECDL)*, 18(11):613–620, 2003.

[8] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.