Using Nearest Neighbor Information to Improve Cross-Language Text Classification

Adelina Escobar-Acevedo, Manuel Montes-y-Gómez and Luis Villaseñor-Pineda

Laboratory of Language Technologies, Department of Computational Sciences,
National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico
{aescobar,mmontesg,villasen}@inaoep.mx

Abstract. Cross-language text classification (CLTC) aims to take advantage of existing training data from one language to construct a classifier for another language. In addition to the expected translation issues, CLTC is also complicated by the cultural distance between both languages, which causes that documents belonging to the same category concern very different topics. This paper proposes a re-classification method which purpose is to reduce the errors caused by this phenomenon by considering information from the own target language documents. Experimental results in a news corpus considering three pairs of languages and four categories demonstrated the appropriateness of the proposed method, which could improve the initial classification accuracy by up to 11%.

1 Introduction

Nowadays, there is a lot of digital information available from the Web. This situation has produced a growing need for tools that help people to find, filter and analyze all these resources. In particular, *text classification* (Sebastiani, 2002), the assignment of free text documents to one or more predefined categories based on their content, has emerged as a very important component in many information management tasks.

The state-of-the-art approach for text classification considers the application of a number of statistical and machine learning techniques, including Bayesian classifiers, support vector machines, nearest neighbor classifiers, and neuronal networks to mention some (Aas and Eikvil, 1999; Sebastiani, 2002). In spite of their great success, a major difficulty of this kind of supervised methods is that they require high-quality training data in order to construct an accurate classifier. Unfortunately, due to the high costs associated with data tagging, in many real world applications training data are extremely small or, what is even worst, they are not available.

In order to tackle this problem, three different classification approaches have recently proposed, each of them concerning a distinct circumstance. The first approach allow building a classifier by considering a small set of tagged documents along with a great number of unlabeled texts (Nigam et al., 2000; Krithara, et al., 2008; Guzmán-Cabrera et al., 2009). The second focuses on the construction of classifiers by reusing training sets from related domains (Aue and Gamon, 2005; Dai et al., 2007). Whereas, the third takes advantage of available training data from one language in order

to construct a classifier that will be applied in a different language. In particular, this paper focuses on this last approach, commonly referred to as *cross-language text classification* (CLTC).

As expected, one of the main problems that faces CLTC is the language barrier. In consequence, most current methods have mainly addressed different translation issues. For instance, some methods have proposed achieving the translation by means of multilingual ontologies (Olsson et al., 2005; Gliozzo and Strapparava, 2006; De Melo and Siersdorfer, 2007; Amine and Mimoun, 2007), while the majority tend to apply an automatic translation system. There are also methods that have explored the translation of complete documents as well as the translation of isolated keywords (Bel et al., 2003). In addition, there have been defined two main architectures for CLTC, based on the translation of the training and test corpus respectively (Rigutini et al., 2005; Jalam, 2003).

Although the language barrier is an important problem for CLTC, it is not the only one. It is clear that, in spite of a perfect translation, there is also a *cultural distance* between both languages, which will inevitably affect the classification performance. In other words, given that language is the way of expression of a cultural and socially homogeneous group, documents from the same category but different languages (i.e., different cultures) may concern very different topics. As an example, consider the case of news about sports from France (in French) and from US (in English); while the first will include more documents about soccer, rugby and cricket, the later will mainly consider notes about baseball, basketball and american football.

In this paper we propose a *post-processing method* for CLTC, which main purpose is to reduce the classification errors caused by the cultural distance between the source (training) and target (test) languages. This method takes advantage from the synergy between similar documents from the target corpus in order to achieve their *re-classification*. Mainly, it relies on the idea that similar documents from the target corpus are about the same topic, and, therefore, that they must belong to the same category.

The rest of the paper is organized as follows. Section 2 describes the proposed re-classification method. Section 3 details the experimental setup and shows the achieved results. Finally, Section 4 presents our conclusions and discusses some future work ideas.

2 The Re-Classification Method

As we previously mentioned, our proposal for CLTC consists in applying a two stage process (refer to Figure 1). The function of the first stage is to generate an *initial classification* of the target documents by applying any traditional CLTC approach. On the other hand, the purpose of the second is to *rectify the initial classification* of each document by using information from their neighbors. Following we describe the main steps of the proposed CLTC method.

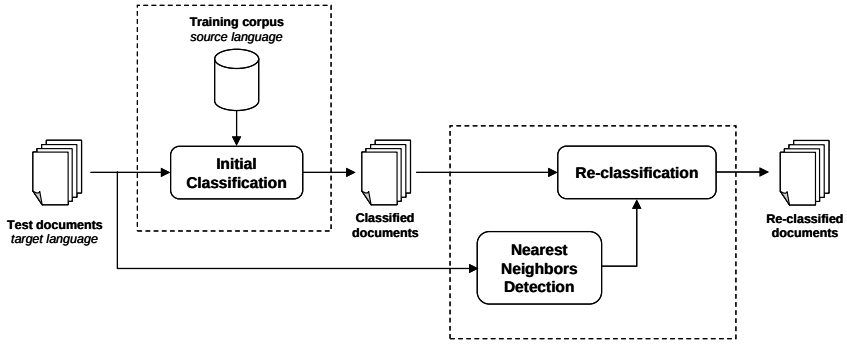

Figure 1. Proposed two-step method for cross-language text classification

1. Build a classifier (C_l) using a specified learning method (l) and a given training set (S) in the source language.
   Depending on the used CLTC approach, the construction of this classifier may or may not consider the translation of the training corpus to the target language.
2. Classify each document (d_i) from the test set (T), in the target language, using the built classifier (C_l). The result of this step is the initial classification of the test documents. We represent the initial class of $d_i \in T$ as $c^0(d_i)$.
   Similar to the previous step, depending on the used CLTC approach, the documents from the test set may or may not be translated to the source language.
3. Determine the set of k nearest neighbors for each document $d_i \in T$, which is represented by NN_i.
   In our experiments we represented documents as set of words, and measured their similarity using the Dice coefficient. That is, the similarity between documents d_i and d_j is computed as indicated below; where $|d_x|$ indicates the number of words in document d_x, and $|d_i \cap d_j|$ their common vocabulary.

$$sim(d_i, d_j) = \frac{2 \times |d_i \cap d_j|}{|d_i| + |d_j|}$$

4. Modify the current class of each test document d_i (represented by $c^n(d_i)$), by considering information from their neighbors. We contemplate two different situations:
   a. If all neighbors of d_i belong to the same class, then:

$$c^{n+1}(d_i) = c^n(d_j) : d_j \in NN_i$$

b. In the case that the neighbors of d_i do not belong to the same class, maintain the current classification of d_i:

$$c^{n+1}(d_i) = c^n(d_i)$$

5. Iterate σ times over step 4 (being σ a user specified threshold), or repeat until no document changes their category. That is, iterate until:

$$\forall (d_i \in T) : c^n(d_i) \equiv c^{n-1}(d_i)$$

3 Experiments

For the experiments we used a subset of the *Reuters Corpus RCV-1* (Lewis et al., 2004). We considered three languages: English, Spanish and French; and the news corresponding to four different classes: crime (GCRIM), disasters (GDIS), politics (GPOL) and sports (GSPO). For each language, we employed 200 news reports for training and 120 for test, which correspond to 50 and 30 news per class respectively.

The used evaluation measure was the classification *accuracy*, which indicates the percentage of test documents that were correctly categorized by the classifier.

Following we describe the performed experiments. In particular, Section 3.1 presents the results from two traditional approaches for CLTC, which correspond to our initial classification (refer to Figure 1); whereas, Section 3.2 presents the results achieved by the proposed re-classification method.

Table 1. Results from traditional approaches for cross-language text classification

Source language (training set)	**Target language** (test set)	**Vocabulary** (training set)	**Vocabulary** (test set)	**Vocabulary intersection**	**Percentage intersection** (w.r.t test set)	**Accuracy**
Translating training set to target language						
French	English	11338	7658	3700	48%	0.858
Spanish	English	9012	7658	3351	44%	0.817
French	Spanish	14684	8051	3920	49%	0.833
English	Spanish	13453	8051	3640	45%	0.717
Spanish	French	10666	9258	3793	41%	0.808
English	French	12426	9258	4131	45%	0.758
Translating test set to source language						
English	French	10892	7731	3697	48%	0.767
English	Spanish	10892	6314	3295	52%	0.750
Spanish	French	12295	9398	3925	42%	0.792
Spanish	English	12295	9190	3826	42%	0.850
French	Spanish	14071	7049	3749	53%	0.800
French	English	14071	8428	4194	50%	0.867

3.1 Traditional Cross-Language Classification

There are two traditional architectures for CLTC, one based on the translation of the training corpus to the target language, and the other based on the translation of the test corpus to the source language. Table 1 shows the results from these two approaches. In both cases, we used the Wordlingo free translator and performed the classification by means of a Naïve Bayes classifier based on word features with a Boolean weighting.

Results from Table 1 indicate that both architectures achieved similar results (around 80% of accuracy), being slightly better the one based on the translation of the test set to the source language. This table also evidences the enormous difference in the vocabularies from the training and test sets, which somehow reflects the relevance of the cultural distance problem.

Analyzing the results from cross-language text classification

Given that we had available training and test data for the three languages, we were able to perform the three monolingual classification experiments. Table 2 shows the results from these experiments. Somehow, these results represent an upper bound for CLTC methods.

Table 2. Results from the monolingual classification experiments

Source language (Training set)	**Target language** (test set)	**Vocabulary** (training set)	**Vocabulary** (test set)	**Vocabulary intersection**	**Percentage intersection** (w.r.t test set)	**Accuracy**
English	English	10892	7658	5452	71%	0.917
Spanish	Spanish	12295	8051	5182	64%	0.917
French	French	14072	9258	6000	65%	0.933

The comparison of results from Tables 1 and 2 evidences an important drop in accuracy for cross-language experiments with respect to the monolingual exercises. We presume that this effect is consequence of the small intersection between the source and target languages (30% less than for the monolingual exercises), which indicates that the training data do not contain all relevant information for the classification of the test documents. However, it is important to point out that it was not possible to establish a direct relation between the cardinality of this intersection and the classification accuracy.

With the aim of understanding the causes of the low intersection between the vocabularies of the training (source language) and test (target language) datasets, we carried out an experiment to evaluate the impact of the translation errors. In particular, we translated the training and test datasets from Spanish to English and French. Using these new datasets, we trained and evaluated two monolingual classifiers. The first classifier (with all data in English) achieved an accuracy of 91.66%, whereas, the second (with all data in French) achieved an accuracy of 90.83%. Comparing these results against the original accuracy from the Spanish monolingual exercise (91.66%), we may conclude that the lost of accuracy introduced by the translation process in practically insignificant, and, therefore, that the *cultural distance* arises as

that main problem of cross-language classification; at least for these kinds of corpora. In other words, these results evidence that news from different countries (in different languages), even though belonging to the same category, tend to report very different events, which generates a great lexical discrepancy in the vocabularies, and, therefore, a noticeable decrement in the classification performance.

3.2 Results from the Re-Classification Method

As we exposed in Section 2, the proposed method considers a first stage where an initial classification is performed by applying some CLTC approach, and a second stage where initial classifications are rectified by considering information from their neighbors.

In particular, the evaluation of the proposed re-classification method (second stage) considered the results from two traditional architectures for CLTC as the initial classifications (refer to Table 1), and employed information from 2 to 5 neighbors to modify or confirm these classifications. Table 3 shows the accuracy results from this experiment. In addition, it also indicates in parenthesis the number of iterations required at each case.

The results achieved by the proposed method are encouraging. In the majority of the cases they outperformed the initial accuracies, confirming the relevance of taking into account information from the own test documents (target language) for their classification. It was interesting to notice that in all cases our method obtained results better than the initial classification, and that the best results were achieved when the initial accuracy was very confident (higher than 0.80).

Results from Table 3 also indicate that the best accuracy results were achieved using only three neighbors. In this case, the average improvement was of 4.33% and the maximum was of 11.65%. For the cases where we used information from four and five neighbors the average improvement was 2.87% and the maximum improvements were 10.15% and 8.07% respectively.

Table 3. Accuracy results obtained after the re-classification process

Source language (Training set)	**Target language** (test set)	**Initial Accuracy**	**Number of Neighbors**		
			3	**4**	**5**
Translating training set to target language					
French	English	0.858	**0.958** (1)	0.925 (1)	0.925 (2)
Spanish	English	0.817	**0.900** (1)	**0.900** (2)	0.883 (3)
French	Spanish	0.833	**0.842** (1)	**0.842** (1)	**0.842** (1)
English	Spanish	0.717	0.725 (3)	**0.733** (4)	0.725 (1)
Spanish	French	0.808	**0.833** (1)	0.817 (1)	0.825 (1)
English	French	0.758	0.775 (1)	**0.767** (1)	**0.767** (1)
Translating test set to source language					
English	French	0.767	0.758 (2)	0.767 (1)	0.767 (1)
English	Spanish	0.750	0.750 (0)	0.750 (0)	0.750 (0)
Spanish	French	0.792	0.808 (1)	0.808 (1)	**0.817** (1)
Spanish	English	0.850	**0.908** (1)	0.892 (1)	0.892 (1)
French	Spanish	0.800	**0.817** (1)	0.808 (1)	**0.817** (1)
French	English	0.867	**0.925** (2)	0.892 (1)	0.892 (1)

Regarding the convergence of the method, the numbers in the parenthesis in Table 3 help to confirm that, due to the strong condition imposed to perform the iterations (which considers that all neighbors must belong to the same category to generate a re-classification), our method requires just a few iterations to reach the final classification. These results also show, as was expected, that augmenting the number of neighbors, the number of iterations tend to decrease.

4 Conclusions

The analysis presented in this paper showed that the problematic of *cross-language text classification* (CLTC) goes beyond the translation issues. In particular, our experiments indicated that the *cultural distance* manifested in the source and target languages greatly affects the classification performance, since documents belonging to the same category tend to concern very different topics.

In order to reduce the classification errors caused by this phenomenon, we proposed a *re-classification method* that uses information from the target-language documents for improving their classification. The experimental results demonstrated the appropriateness of the proposed method, which could improve the initial classification accuracy by up to 11%.

The results also indicated that the proposed method is independent of the approach employed for generating the initial classification, given that it achieved satisfactory results when training documents were translated to the target language as well as when test documents were translated to the source language.

Finally, it was interesting to notice that relevant improvements were only achieved when initial classification accuracies were very confident (higher than 0.80). In relation to this point, as future work we plan to apply, in conjunction with the re-classification method, a semi-supervised classification approach that allows incorporating information from the target language into the construction of the classifier.

Acknowledgements: This work was done under partial support of CONACYT (project grant 83459 and scholarship 212424).

References

1. Aas K., and Eikvil L. Text Categorisation: A Survey. Technical Report. Norwegian Computing Center, 1999.
2. Amine B. M., and Mimoun M. WordNet based Multilingual Text Categorization. Journal of Computer Science, Vol. 6, Num. 4, 2007.
3. Aue A., and Gamon M. Customizing Sentiment Classifiers to New Domains: a Case Study. International Conference on Recent Advances in Natural Language Processing (RANLP-2005). Borovets, Bulgaria, 2005.
4. Bel N., Koster C., and Villegas M. Cross-Lingual Text Categorization. European Conference on Digital Libreries (ECDL-2003). Trondheim, Norway, August 2003.
5. Dai W., Xue G., Yang Q., and Yu Y. Co-clustering based classification for out-of-domain documents. Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD-07). San Jose, California, USA. August, 2007.

6. De Melo G., and Siersdorfer S. Multilingual Text Classification using Ontologies. 29th European Conference on IR Research (ECIR 2007). Rome, Italy, April 2007.
7. Gliozzo A., and Strapparava C. Exploiting Corporable Corpora and Biligual Dictionaries for Cross-Language Text Categorization. Proceedings of the 21st International Conference on Computational Linguistics (Coling-2006). Sydney, Australia, 2006.
8. Guzmán-Cabrera R., Montes-y-Gómez M., Rosso P., Villaseñor-Pineda L. Using the Web as Corpus for Self-training Text Categorization. Journal of Information Retrieval, Volume 12, Number 3, June, 2009.
9. Jalam, Radwan: Apprentissage automatique et catégorisation de textes multilingues. PhD Tesis, Université Lumiere Lyon 2, Lyon, France, 2003.
10. Krithara A., Amini M., Renders J.M., and Goutte C. Semi-Supervised Document Classification with a Mislabeling Error Model. 30th European Conference on Information Retrieval (ECIR 2008), Glasgow, Scotland, 2008.
11. Lewis D., Yang Y., Rose T. G., Dietterich G., Li F. RCV1: A New Benchmark Collection for Text Categorization Research. Journal of Machine Learning Research, Vol. 5, 2004.
12. Nigam K., Mccallum A. K., Thrun S., and Mitchell T., Text classification from labeled and unlabeled documents using EM, Machine Learning, 39(2/3):103–134, 2000.
13. Olsson J. S., Oard D., and Hajic J. Cross-Language Text Classification. Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2005) New York, NY, USA, 2005.
14. Rigutini L., Maggini M., and Liu B. An EM based training algorithm for Cross-Language Text Categorization. Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence. Compiegne, France. September 2005.
15. Sebastiani, F. Machine learning in automated text categorization. ACM Computing Surveys 34(1), 2002.