Using Information from the Target Language to Improve Crosslingual Text Classification

Gabriela Ramírez-de-la-Rosa[1], Manuel Montes-y-Gómez[1],
Luis Villaseñor-Pineda[1], David Pinto-Avendaño[2], and Thamar Solorio[3]

[1] Laboratory of Language Technologies,
National Institute for Astrophysics, Optics and Electronics
{gabrielarr,mmontesg,villasen}@inaoep.mx
[2] Faculty of Computer Science, Autonomous University of Puebla
dpinto@cs.buap.mx
[3] Department of Computer and Information Sciences,
University of Alabama at Birmingham
solorio@uab.edu

Abstract. Crosslingual text classification consists of exploiting labeled documents in a source language to classify documents in a different target language. In addition to the evident translation problem, this task also faces some difficulties caused by the cultural discrepancies manifested in both languages by means of different topic distributions. Such discrepancies make the classifier unreliable for the categorization task. In order to tackle this problem we propose to improve the classification performance by using information embedded in the own target dataset. The central idea of the proposed approach is that similar documents must belong to the same category. Therefore, it classifies the documents by considering not only their own content but also information about the assigned category to other similar documents from the same target dataset. Experimental results using three different languages evidence the appropriateness of the proposed approach.

Keywords: Crosslingual text classification, prototype-based method, unlabeled documents, text classification.

1 Introduction

Text classification is the task of assigning documents into a set of predefined classes or topics [1]. The leading approach for this task considers the application of machine learning techniques such as Support Vector Machines and Naïve Bayes, which require large labeled data sets to construct accurate classifiers. Unfortunately, due to the high costs associated with data tagging, for many applications in several languages these datasets are extremely small or, what is worst, they are not available.

Several approaches have recently proposed to alleviate the problem of lacking labeled data; one example is the *crosslingual text classification* (CLTC), which

H. Loftsson, E. Rögnvaldsson, S. Helgadóttir (Eds.): IceTAL 2010, LNAI 6233, pp. 305–313, 2010.

consists in exploiting labeled documents in a source language to classify documents in a different target language. Because of the inherent language-barrier problem of this approach, most current CLTC methods have mainly addressed different translation issues. In particular, they have explored the translation from one language to another by means of machine translation approaches as well as by multilingual lexical resources such as dictionaries and ontologies [2,3].

Although the language barrier is an important problem in CLTC, it is not the only one. It is clear that, in spite of a perfect translation, there are also some *cultural discrepancies* manifested in both languages that will affect the classification performance. That is, given that a language is the way of expression of a cultural and socially homogeneous community, 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 USA (in English); while the first will include more documents about soccer, rugby and cricket, the latter will mainly consider notes about baseball, basketball and American football. In order to tackle this problem, recent CLTC methods have proposed to enhance the classification model by iteratively incorporating information from the target language into the training phase [4,5,6]; their purpose is to obtain a classification model that is as close as possible to the target topic distribution.

The method proposed in this paper is a simple and inexpensive alternative for facing the problems caused the cultural discrepancies between both languages. Different to previous iterative approaches, it does not consider the modification or enrichment of the original classifier; instead, it attempts to improve the document classification by using more information to support the decision process. Mainly, it is based on the idea that similar documents must belong to the same category and, therefore, it classifies the documents by considering their own information (as usual) as well as the information about the assigned category to other similar documents from the same target dataset.

In the following section we describe the proposed method for CLTC. This method is based on the prototype-based classification approach [7], but modifies the traditional class-assignment strategy in order to incorporate information from the set of similar documents. Then, in Section 3 we define the experimental configuration and show results in six different pairs of languages that demonstrate the usefulness of the proposed approach for CLTC. Finally, in Section 4 we present our conclusions and some ideas for future work.

2 Prototype-Based CLTC Method

Given that prototype-based classification is very simple and has demonstrated to consistently outperform other algorithms such as Naïve Bayes, K-Nearest Neighbors and C4.5 in text classification tasks [7], we decided to implement the proposed approach using this classification algorithm. In general, our *prototype-based CLTC method* chooses a category for each document (from the target language) by determining the class which prototype (calculated from the source-language training set) is more similar to it and to its nearest neighbors (from the same target-language dataset).

Figure 1 shows the general schema of the proposed method. It consists of four main processes. The first one carries out the translation of the training documents from the source language (S) to the target language (T). The second process focuses on the construction of the class prototypes using the well-known normalized sum technique [8]. The third process involves the identification of the nearest neighbors for each document from the target language dataset (D_T). Finally, the fourth process computes the classification for each document $d \in D_T$ considering information from their own and their neighbors. Bellow we present a brief description of each one of these processes.

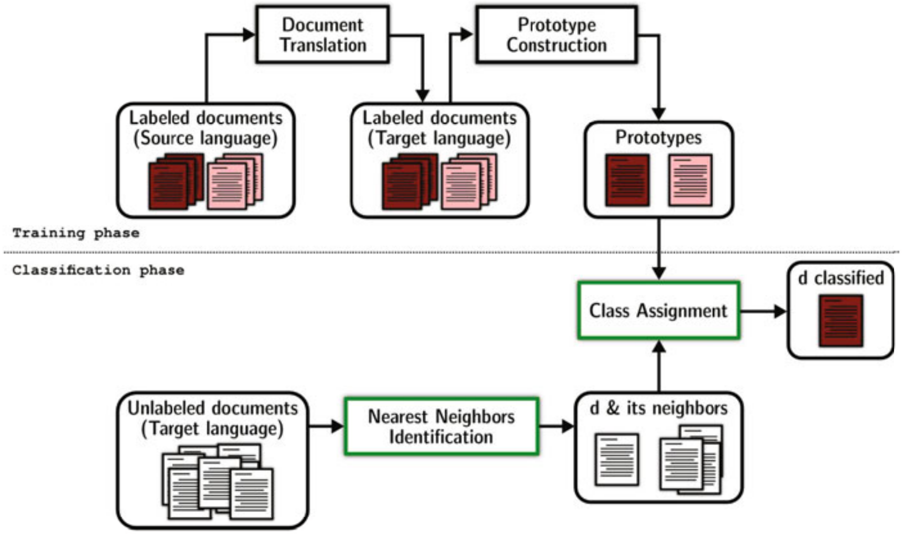

Document Translation
Prototype Construction
Labeled documents (Source language)
Labeled documents (Target language)
Prototypes
Training phase
Classification phase
Class Assignment
d classified
Unlabeled documents (Target language)
Nearest Neighbors Identification
d & its neighbors

Fig. 1. General scheme of the proposed text classification method

Document Translation. Two basic architectures have been explored for CLTC, one based on the translation of the target dataset to the source language, and another one based on the translation of the training set to the target language. We decided to adopt the latter option because training sets are commonly smaller than test sets and, therefore, their translation tend to be less expensive. In particular, the translation was achieved using the Worldlingo online translation machine[1].

Prototype Construction. This process carries out the construction of the class prototypes based on information from the –translated– training set; thus, the resulting prototypes are represented in the target language. In particular, given a set $D = \{d_1, d_2, ...\}$ of vectors of labeled documents (from the training

[1] http://www.worldlingo.com/es/products_services/worldlingo_translator.html

set) organized in a predefined set of classes C and represented in their own term space, it computes the prototype vector for each class $c_i \in C$ using Formula 1.

$$P_i = \frac{1}{||\sum_{d \in c_i} d||} \sum_{d \in c_i} d \tag{1}$$

Nearest Neighbors Identification. This process focuses on the identification of the k nearest neighbors for each document d_i from the target dataset D_T (refer to Formula 2). In order to do that we compute the similarity between two documents (d_i and all other d in D_T) using the cosine formula (refer to Formula 4).

$$N_k^{d_i} = argmax_{S_j \in \mathbb{S}_k} \left[\sum_{d \in S_j} sim(d, d_i) \right] \tag{2}$$

where $\mathbb{S}_k$ and $sim()$ are defined as follows:

$$\mathbb{S}_k = \{S | S \subseteq D_T \wedge |S| = k\} \tag{3}$$

$$sim(d_i, d_j) = \frac{d_i \cdot d_j}{||d_i|| \times ||d_j||} \tag{4}$$

Class Assignment. In prototype-based classification, the class of a document d from the target dataset is traditionally determined by Formula 5. Our proposal extends this class-assignment strategy by considering not only information from the document itself but also information about the assigned category to other similar documents from the same target dataset. In particular, given a document from the target dataset ($d \in D_T$) in conjunction with its k nearest neighbors (N_k^d), we assign a class to d using Formula 6.

$$class(d) = argmax_i \left(sim(d, P_i) \right) \tag{5}$$

$$class(d) = argmax_i \left(\lambda \, sim(d, P_i) + (1 - \lambda) \frac{1}{k} \sum_{n_j \in N_k^d} [inf(d, n_j) \times sim(n_j, P_i)] \right) \tag{6}$$

where,

- $sim(v_i, v_j)$ is the cosine similarity function defined in Formula 4.
- N_k^d is the set of k neighbors considered to provide information about document d (refer Formula 2).
- λ is a constant used to determine the relative importance of both, the information from the document (d) and the information from its neighbors. The smaller the value of λ is, the greater the contribution of the neighbors, and vice versa.

- $inf()$ is an influence function used to weight the contribution of each neighbor n_j to the classification of d. The purpose of this function is to give more relevance to the closer neighbors. In particular, we define this influence in direct proportion to the similarity between each neighbor and d calculated using the cosine formula (refer to Formula 4).

3 Evaluation

3.1 Datasets

For the experiments we considered a subset of the Reuters RCV-1 Corpus [9]. This subset considers three languages (English, French and Spanish), and the news reports corresponding to four classes (Crime, Disasters, Politics, and Sports). For each language we used 320 documents; 80 per each class[2].

3.2 Evaluation Measure

The evaluation of the performance of the proposed method was carried out by means of the F-measure. This measure is a linear combination of the precision and recall values from all class $c_i \in C$. It is defined as follows:

$$F - Measure = \frac{1}{|C|} \sum_{i=1}^{|C|} \left[\frac{2 \times Recall(c_i) \times Precision(c_i)}{Recall(c_i) + Precision(c_i)} \right] \tag{7}$$

$$Recall(c_i) = \frac{number\ of\ correct\ predictions\ of\ c_i}{number\ of\ examples\ of\ c_i} \tag{8}$$

$$Precision(c_i) = \frac{number\ of\ correct\ predictions\ of\ c_i}{number\ of\ predictions\ as\ c_i} \tag{9}$$

3.3 Baseline Experiments

The goal of these experiments was to evaluate the performance of a traditional CLTC approach, where documents from a source language are used to classify documents from a different target language. For these experiments we applied the following standard procedure: first, we translated the training documents from the source language to the target language (using Worldlingo); then, we constructed a classifier (in the target language) using the translated training set; finally, we used the built classifier to determine the class of each document from the target-language dataset. For the construction of the classifier we considered three of the most used methods for text classification, namely, Naïve Bayes (NB), Support Vector Machines (SVM)[3], and a prototype-based method (PBC)

[2] This corpus can be downloaded from http://ccc.inaoep.mx/∼mmontesg/resources/CLTC/RCV-Subset.txt

[3] For NB and SVM we used the implementation and default configuration of WEKA [10].

Table 1. F-measure results for six crosslingual experiments using a traditional CLTC approach

Source language	Target language	Experiment	PBC	NB	SVM
English	French	$E_F - F$	0.616	0.753	0.764
English	Spanish	$E_S - S$	0.814	0.791	0.625
French	English	$F_E - E$	0.956	0.931	0.616
French	Spanish	$F_S - S$	0.879	0.882	0.658
Spanish	English	$S_E - E$	0.851	0.891	0.486
Spanish	French	$S_F - F$	0.790	0.802	0.723

using the class-assignment function described in Formula 5. Table 1 shows the F-measure results obtained by these methods in six crosslingual experiments, which correspond to all possible pair-combinations of the three selected languages. From these results, those by PBC are of special interest since our method is an extension of this approach.

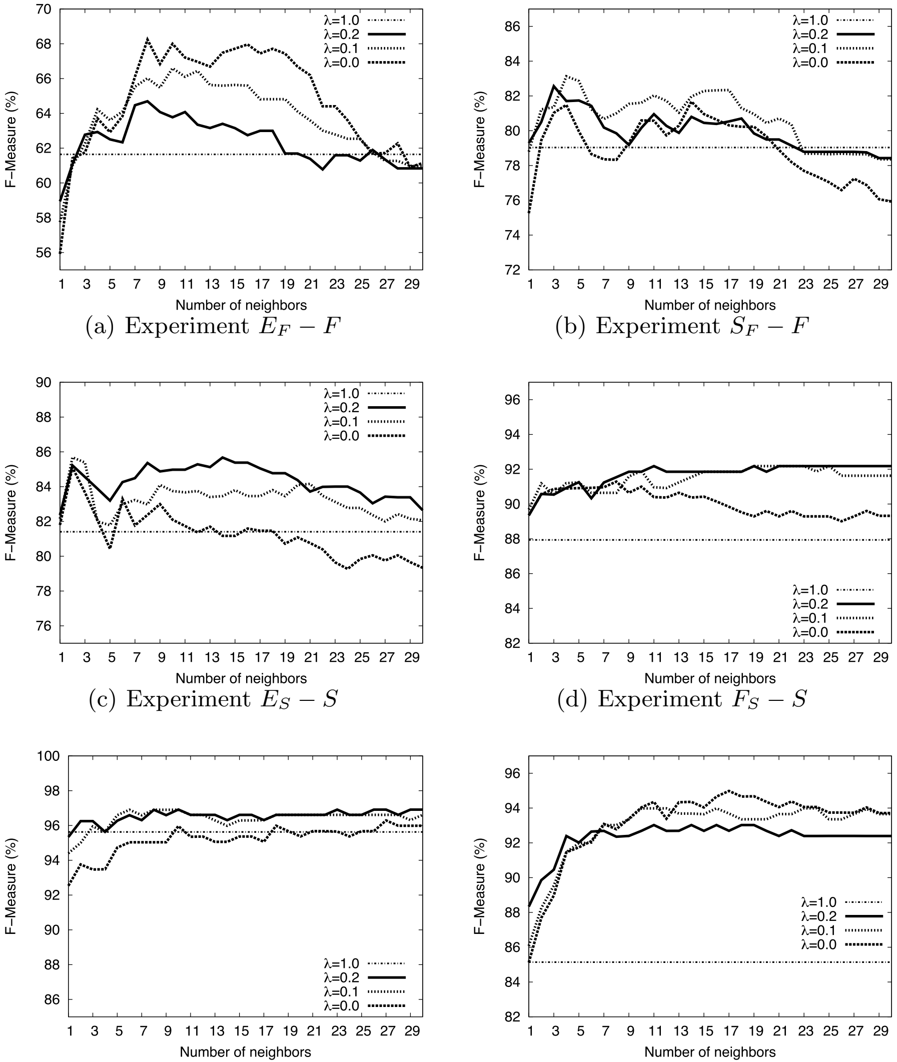
3.4 Results from the Proposed Method

As described in Section 2, the main idea of the proposed method is to classify the documents by considering not only their own content but also information from other similar documents from the same target dataset. Particularly, we adapted the traditional prototype-based approach (PBC) to capture this information (refer to Formula 6), being λ a constant that determines the relative importance of both components.

Considering the proposed method, we designed some experiments in such a way that we could evaluate the impact on the classification results caused by the selection of different values of λ, as well as the impact caused by the usage of different number of neighbor documents into the class assignment process. In particular, we used $\lambda = 0, 0.1, 0.2, ..., 1$, and $k = 1, ..., 30$.

Experiments showed that the best results were achieved when using small values of λ, indicating that information from the neighbor documents is of great relevance. On the other hand, they could not indicate a clear conclusion about the appropriate number of neighbors, since several different values allowed to obtain similar classification improvements. Figure 2 shows some results of the proposed method in the six crosslingual experiments. These results correspond to three different values of λ: 0, 0.1 and 0.2. This figure also shows the results from the traditional prototype-based approach, which correspond to our method results using $\lambda = 1$. The achieved results indicate that the proposed method clearly outperforms the traditional prototype-based approach.

In order to summarize the results from the experimental evaluation, Table 2 presents the best results achieved by the proposed method. Comparing these results against those from Table 1, it is possible to notice that our method outperformed all used classification algorithms in all except one of the crosslingual



(a) Experiment $E_F - F$

(b) Experiment $S_F - F$

(c) Experiment $E_S - S$

(d) Experiment $F_S - S$

(e) Experiment $F_E - E$

(f) Experiment $S_E - E$

Fig. 2. F-measure results of the proposed method in the six crosslingual experiments, using different values of λ and numbers of neighbors (k). The straight line corresponds to the PBC baseline result ($\lambda = 1$).

experiments, demonstrating the usefulness of considering information from the target dataset in crosslingual text classification.

At this point it is important to clarify that several different configurations of our method (as shown in Figure 2) allowed obtaining competitive classification results. One example is the configuration defined by $\lambda = 0.1$ and $k = 11$, which

Table 2. Best F-measure results of the proposed method

Experiment	**Baselines**		**Best results**		**Configuration**
	PBC$\star$	Best†		$[k, \lambda]$	$[k = 11, \lambda = 0.1]$
$E_F - F$	0.616	-	0.682 $\star$	[8, 0.0]	0.661
$E_S - S$	0.814	0.814	0.857 $\star$†	[2, 0.1]	0.837
$F_E - E$	0.956	-	0.969	[10, 0.2]	0.966
$F_S - S$	0.879	0.882	0.922 $\star$†	[11, 0.2]	0.910
$S_E - E$	0.851	0.891	0.950 $\star$†	[17, 0.0]	0.940
$S_F - F$	0.790	-	0.831 $\star$	[4, 0.1]	0.820

also outperformed most baseline results as shown in the last column of Table 2. We evaluated the statistical significance of the best achieved results using the z-test with a confidence of 95%; a $\star$ indicates that the improvement over the PCB is statistically significant, whereas, a † indicates the same regarding the best baseline result.

4 Conclusions and Future Work

In addition to the evident translation problem, crosslingual text classification (CLTC) also faces some difficulties caused by the cultural discrepancies manifested in both languages by means of different topic distributions. In this paper we proposed a simple and inexpensive approach for facing this problem. This approach is based on the idea that similar documents must belong to the same category and, therefore, it classifies the documents by considering their own information (as usual) as well as the information about the assigned category to other similar documents.

In particular, we implemented the proposed approach using the prototype-based classification algorithm. In our implementation the decision about the category of each document (from the target language) is determined by the class whose prototype (calculated from the training set) is more similar to it and to its nearest neighbors (from the same target-language dataset). This way, the proposed method determines the category of documents taking advantage of information from the two languages.

As future work we plan to carry out an extensive analysis of several crosslingual experiments (using different languages and a larger number of documents) to establish a simple criterion for determining the appropriate values for parameters λ and k. Once defined this criterion, we also plan to use the proposed approach in conjunction with a semi-supervised method as the one described by Rigutini et al. [4]. Our goal is to enhance the selection of the documents that will be iteratively included in the training set, and, consequently, to obtain a classification model that is as close as possible to the target-language distribution.

Acknowledgments. This work was done under partial support of CONACyT-Mexico (project grants 83459, 82050, 106013 and 106625, and scholarship 239516).

References

1. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34, 1–47 (2002)
2. Bel, N., Koster, C.H.A., Villegas, M.: Cross-lingual text categorization. In: Koch, T., Sølvberg, I.T. (eds.) ECDL 2003. LNCS, vol. 2769, pp. 126–139. Springer, Heidelberg (2003)
3. de Melo, G., Siersdorfer, S.: Multilingual text classification using ontologies. In: Amati, G., Carpineto, C., Romano, G. (eds.) ECiR 2007. LNCS, vol. 4425, pp. 541–548. Springer, Heidelberg (2007)
4. Rigutini, L., Maggini, M., Liu, B.: An EM based training algorithm for cross-language text categorization. In: WI 2005: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, Washington, DC, USA, pp. 529–535. IEEE Computer Society, Los Alamitos (2005)
5. Ling, X., Xue, G.R., Dai, W., Jiang, Y., Yang, Q., Yu, Y.: Can Chinese web pages be classified with English data source? In: WWW 2008: Proceeding of the 17th International Conference on World Wide Web, pp. 969–978. ACM, New York (2008)
6. Wan, X.: Co-training for cross-lingual sentiment classification. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Suntec, Singapore, Association for Computational Linguistics, pp. 235–243 (2009)
7. Han, E.H., Karypis, G.: Centroid-based document classification: Analysis and experimental results. In: Zighed, D.A., Komorowski, J., Żytkow, J.M. (eds.) PKDD 2000. LNCS (LNAI), vol. 1910, pp. 424–431. Springer, Heidelberg (2000)
8. Cardoso-Cachopo, A., Oliveira, A.L.: Semi-supervised single-label text categorization using centroid-based classifiers. In: SAC 2007: Proceedings of the 2007 ACM Symposium on Applied Computing, pp. 844–851. ACM, New York (2007)
9. Lewis, D.D., Yang, Y., Rose, T.G., Li, F.: Rcv1: A new benchmark collection for text categorization research. J. Mach. Learn. Res. 5, 361–397 (2004)
10. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)