Cross-effective cross-lingual document classification

Núria Bel*, Cornelis H.A. Koster, Marta Villegas*****

*IULA, Universitat Pompeu Fabra,
nuria.bel@upf.edu
** Computer Science Dept., University of Nijmegen,
kees@cs.kun.nl
***Grup d'Investigació en Lingüística Computacional, Universitat de Barcelona,
tona@gilc.ub.es

Abstract

This article addresses the question of how to deal with text categorization when the set of documents to be classified belong to different languages. The figures we provide demonstrate that cross-lingual classification where a classifier is trained using one language and tested against another is possible and feasible provided we translate a small number of words: the most relevant terms for class profiling. The experiments we report, demonstrate that the translation of these most relevant words proves to be a cost-effective approach to cross-lingual classification.

1. Introduction

Automatic text classification is an area of document management technologies that has developed for the last years with success. One of the challenging areas now is working in multilingual documentation scenarios. This paper addresses the point of finding methods for text categorization with only one classifier although the documents to be classified are written in different languages. In the present paper, we report on the experiments done during the Peking[1] project and concerning "cross-lingual text categorization"[2]. We concentrate on describing the experiments where a classifier was trained with documents in one particular language for being used on a set of documents written in another language. The method followed was combining term selection and term translation as the main objective was looking for a cost-effective method of doing cross-lingual classification. The figures we report demonstrate that translating only the most relevant words –as indicated by the classification system-- proves to be a cost-effective approach to cross-lingual classification.

All experiments reported in this paper were performed with Version 2.0 of the Linguistic Classification System, LCS, developed in the Peking project[3], which implements both the Winnow and Rocchio algorithms for multi and mono-classification. Spanish and English documents for the experiments were downloaded from ILOLEX, a database of classified documents compiled by the International Labour Organization. ILOLEX describes itself as "a trilingual database containing ILO Conventions and Recommendations, ratification information, comments of the Committee of Experts and the Committee on Freedom of Association, representations, complaints, interpretations, General Surveys, and numerous related documents." Although the actual database contains original documents and its translation into the two other languages (English, Spanish and French), in constructing our (just bilingual, English-Spanish) corpus, the documents were selected to simulate an occasional corpus, representative of the type of corpus we could find in a real scenario. Thus, we avoided parallel documents as much as possible, that is, when enough non-parallel text was available. The final experimental corpus included some documents both in English and Spanish, but most of them in only one language. Detailed information of the ILOLEX-PEKING corpus is supplied below.

- English consists of 2165 documents. 4.2 million words. Document length: between 39 and 38,646 words.
- Spanish consists of 1590 documents. 4.7 million words. Document length: between 117 and 7500 words.

The corpus documents were mono-classified into 12 categories, each of these with a rather varying number of documents. Table 1 describes the content of the ILOLEX-PEKING corpus per category.

#English docs.	#Spanish docs.	Class Name
123	74	Human rights
397	86	Conditions of employment
299	71	Conditions of work
22	23	Economic and social development
414	448	Employment
279	278	Labour Relations
85	81	Labour Administration
98	86	Health and Labour
156	148	Social Security
81	20	Training
131	154	Special provisions by category of persons
108	121	Special provisions by Sector of Econ. Activity
2165	1590	

Table 1: ILOLEX-PEKING corpus

2. The experiments

In order to establish a baseline with which to compare the results of cross-lingual classification, we first measured the accuracy achieved by LCS in mono-lingual classification for the Spanish and English documents.

[1] PEKING-People and Knowledge Cross-Lingual Information Gathering – EU, 5FP IST-2533.
[2] For more information on text categorization in the Peking project (Bel, Koster, Villegas 2003).
[3] www.cs.kun.nl/peking

Experimentation with documents in different languages was concerned with measuring accuracy first in bi-lingual classification and later in cross-lingual classification. By "bi-lingual classification" we mean building/training a classifier with a set of documents in two languages and testing its performance when classifying either in one or another language. We reserve the term "cross-lingual classification" for the case where training is made in one language, in our case English, and the classifier is asked to classify documents in another language, in our case Spanish. Thus, we mirror the use of 'cross-lingual' made in Cross-lingual Information retrieval. For Cross-Lingual Text Categorization, three translation strategies may be distinguished. The first two are taken from Cross-Lingual Information Retrieval (CLIR):

- Document translation. Although feasible, it is considered to be too expensive and for document classification worthless as a rather reduced number of elements of the text are actually considered by the classifier, as we will see below.
- Terminology translation. Using bilingual, or multilingual glossaries for the domain that can be used for translating the terms. It is expected that these include all or most of the terms which will be relevant for classification
- Profile-based translation. As we will see, this is the final approach we followed for a cost-effective solution. It consists on finding the equivalents in another language of the linguistic terms, the words, that appear in the profile or list of terms that the classifier is to work with.

The rest of the paper refers to the experiments and results for each of the last two scenarios just described.

3. Monolingual classification

Monolingual classification experiment goal was to fix a baseline for comparison of results. In addition, for each language and classifying algorithm, we performed three different rounds of experiments using the same set of ILOLEX documents, but after different degrees of pre-processing, as this pre-processing was also a requirement of the cross-lingual classification task. In the first case, pre-processing merely included de-capitalization and elimination of certain special characters. In the second case, pre-processing consisted of noun and verb lemmatisation. In the third case, documents were pre-processed in such a way that multiword terms were taken as single units (e.g. 'software_engineering' or 'trabajadores_migrantes'). This 'multiword chunking' was motivated by the cross-lingual experiment mainly because some translational equivalences could not be done on a word-to-word basis. The list of multiwords to be identified and chunked was, however, compiled in different ways depending on the language.

For Spanish texts, multiword terms were extracted using both quantitative and linguistic strategies. A first list of candidates was extracted using Mutual Information and Likelihood Ratio measures over the available corpus. This list of candidates was filtered by checking it against another list made of well formed Noun Phrases (basically N+N, N+ADJ and N+prep+N). This process guaranteed that all Spanish multi-words were both linguistically and statistically motivated. The final list consisted of 303 bi-grams (N+ADJ), and 288 tri-grams (N+prep+N). As Table 2 data shows, and in accordance with the results by (Riloff 1995) and (Larkey, 1999) for English, any pre-processing has little effect on accuracy as far as monolingual classification is concerned also for Spanish texts.

The English multi-word expression list was build as required by the cross-lingual experiment. As we will explain in detail in section 5, the cross-lingual experiment implied the translation of the list of most 'important terms' for the classifier. Thus, for the monolingual experiment we used this list of multiword terms in English present in the bilingual database, that is, those resulting from the translation of Spanish terms.

alg.	repr.	lang.	accuracy	
			Multi0:3	Mono1:1
W.	keywords	En	.840$±.013	.865$±.007
R.	keywords	En	.823$±.010	.800$±.010
W.	keywords	Sp	.768$±.014	.790$±.015
R.	keywords	Sp	.755$±.007	.764$±.013
W.	lemmat. k.	En	.845$±.008	.863$±.006
R.	lemmat. k.	En	.797$±.012	.817$±.012
W.	lemmat. k.	Sp	.768$±.012	.788$±.015
R.	lemmat. k.	Sp	.759$±.010	.758$±.017
W.	chunked k.	En	.840$±.013	.867$±.011
R.	chunked k.	En	.824$±.010	.829$±.011
W.	chunked k.	Sp	.762$±.013	.800$±.013
R.	chunked k.	Sp	.769$±.010	.779$±.010

Table 2: Monolingual classification accuracy for Winnow and Rocchio algorithms.

4. Bilingual classification

This is the case where a classifier was trained and tested using a bilingual corpus, i.e. documents in English and documents in Spanish. In this experiment no translation or pre-processing was performed. The 2167 English and 1590 Spanish ILO documents, labelled with the same class-labels, were combined at random into one working corpus. Then, this corpus was randomly split into 4 training sets, each containing 15% (563) of the documents. For each, we fixed a test set of 40% of the documents, and the rest as training set. We got the results shown in Table 3.

alg.	repr.	lang.	accuracy	
			Multi0:3	Mono1:1
W.	keywords	En/Sp	.785$±.013	.811$±.014
R.	keywords	En/Sp	.739$±.009	.758$±.014

Table 3: Bilingual classification accuracy for Winnow and Rocchio algorithms

As we can see, the results proved to be more than acceptable, as are comparable to those obtained for monolingual classification with no processing.

5. Cross-lingual classification

Finally, in the cross-lingual classification experiment, the classifier was trained using only labeled English documents. And the classifier was tested against Spanish

documents, which were previously pre-processed according to two different lines.

First, we tried the translation line and all Spanish documents were 'term-translated' into English. It was considered that there was no need for performing text translation as in fact the only interesting elements for the classification system were words as features. Thus, in this 'term-translation' process, only nouns, verbs and adjectives with a certain frequency threshold in Spanish documents (>30) were translated into English. The resulting list consisted of 4462 wordforms (out of 4,619,681 tokens) for Spanish and 5258 (out of 4,609,670 tokens) for English.

Under the assumption that for a corpus as ours, which included documents in different languages but classified under the same classes, most frequent wordforms in one language should have a translational equivalent in the other language, we checked the list of candidates against a bilingual dictionary selecting only those forms that were in the list of candidates of the target language. Additionally, translations for all chunked multi-words as described in section 2.1 were added.

The resulting bilingual database was used to translate the Spanish documents into pseudo-English documents made of the English translational equivalents of the Spanish words present in the database. When a Spanish word had several translations, all target candidates were included in the pseudo-English documents. In order to avoid noise and to guarantee that chunked multiwords were correctly translated, the process favoured the translation of longer matches, avoiding the translation of parts of a given multiword when found. The problem of finding the good translation was tackled by using all possible translational equivalences for the Spanish words although among those that belong to the domain.

The classifier was tested using the resulting 'pseudo English' documents. The results showed that while Rocchio classification was worst than the Spanish monolingual baseline, for Winnow the results were slightly better for cross-lingual classification than for Spanish monolingual classification as shown in Table 4.

alg.	repr.	lang.	accuracy	
			Multi0:3	Mono1:1
W.	keywords	En/pseudoEn	.696$±.051	.792$±.012
R.	keywords	En/pseudoEn	.592$±.025	.709$±.012

Table 4: Cross-lingual classification accuracy for Winnow and Rocchio algorithms with pseudo-English documents

However, compiling bilingual glossaries, as the one we just described, is a time and resource consuming task. Thus, this solution, even although having revealed as effective was considered not cost-effective.

We followed a second line of development. After having seen the encouraging results of the first experiment, we investigated the results of combining 'term selection' and 'term translation'. The key point was to see the performance when translating only those words which were selected as relevant terms by the classification engine.

5.1.1. Term selection

Most document classification techniques are crucially based in automatic feature/term selection. The reason is because one of the difficulties these techniques face is the high dimensionality of the search space: terms/features are the words that occur in documents. Depending on the corpus, terms can be hundreds of thousands. Usually, quantitative methods are applied for selecting those terms which will be used for classification computations. (Yang & Pedersen 1997) report from their experiments comparing different methods, that the Text Classification methods that performed better are those that favored the most common terms in their final selected lists. Only Support Vector Machine based system can avoid term selection. But still, and as (Joachims 2001) has pointed out, there is a connection between the statistical properties of the text, i.e. word frequency, with the generalization performance of the learner. The classifier needs mostly high discriminative terms in the high-frequency range. We called these the Most Important Terms.

The LCS system belongs to the class of classifiers that extract a list of terms to work with in classification. As explained in Peters & Koster (2001), using a suitable term selection algorithm, acceptable accuracy can be achieved although using a very small number of terms per class (between 40 and 150). And once we have this profile, the list of terms the classifier is to work with, we know what are the terms that the classifier will be looking for in any document. Thus, we foresaw that cross-lingual classification could work by finding the equivalents in Spanish to the words in English present in the selected list of terms for classification of English documents. In translating from Spanish into English only these Spanish words, the classifier could work in the same way as it was originally an English text.

5.1.2. Term translation

As explained above, the second cross-lingual classification experiment was to investigate the effect of translating *only* towards the words occurring in the class profiles, or selected terms. The experiment proceed as follows:

The Winnow algorithm was trained on the English documents, and for each class the 150 terms with the highest weight (Most Important Terms) were taken from the class profile for each class and English documents. The result was a vocabulary of 923 words. The actual list of terms selected in this fashion contained not only obviously relevant words, but also some non-words and stop words, which no human would consider as 'important concepts' for classification, but which were no doubt chosen for good statistical reasons.

For each English word in this vocabulary, all possible Spanish words present in the bilingual database we described above that may be translated into it were identified.

The English vocabulary was also used to train a classifier on all English documents. And, finally, the classifier was tested on all Spanish documents that suffered a pre-processing as every Spanish term that could have a translation towards a word in the English vocabulary was indeed substituted

The resulting accuracy is .730 as mentioned in Table 5. Taking into account that the best accuracy achieved in the monolingual experiment of the Spanish documents

was .775, the results of the last cross-lingual experiment are encouraging: .724 for Winnow.

alg.	repr.	lang.	accuracy	
			Multi0:3	Mono1:1
W.	profile	En/Sp	.605$±.071	.724$±.035
R.	profile	En/Sp.	.681$±.048	.730$±.019

Table 5: Profile translation classification accuracy for Winnow and Rocchio algorithms

The results must also be considered on the light of the costs of working that way. First, we must take into account that no Spanish labelled (or pre-classified) documents were needed, which is the cost of the bilingual scenario described in 3. The classifier, in this case, was trained with only English labelled texts. And second, the translation effort required for achieving these results was pretty controlled and can be foreseen beforehand: in our experiment we worked with an average of 60 translated terms per class, as they were the linguistic terms actually used by the classifier.

6. Conclusions

We have presented the results of several experiments whose goal was to find cost-effective ways of performing cross-lingual text classification. Figure 1 shows graphically the results obtained with the two algorithms used and it compares for each of them with the baseline, i.e. Spanish monolingual classification.

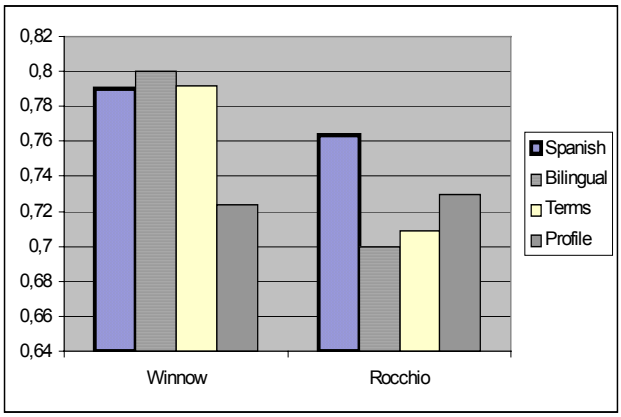

Figure 1: Accuracy for different translation approaches

The figures demonstrate that the best results were for bilingual training with Winnow. Comparison with the results for the bilingual experiment with Rocchio shows that not all algorithms perform the same when handling the disjointedness of the resulting class profile. However, enough (bilingual in this case o multilingual in a real scenario) documents classified under the same system and for all the languages handled are required for this method. This seems not to be the most current real scenario for document classification, and makes this method a non practical solution.

For Term Translation method, the second best rated, the main drawback is the requirement of having dictionaries for all languages involved into the one that was the basis for system training. As we have already mentioned, this is a costly exercise.

As for the third experiment, the results are still good enough if compared against the ones obtained with the first two methods. We consider that they can be even acceptable if we consider that in this case no Spanish labelled documents were needed and that the translation effort was really small. These figures demonstrate that cross-lingual classification where a classifier is trained using one language and tested against another is possible and feasible provided we translate a small number of words: the most relevant terms for class profiling. In our experiment, with a Spanish corpus of about 4.5 million words consisting of 22000 different lemmas we only need to provide a translation for 923. Accuracy is slightly lower, but the approach significantly reduces the investment required for adding documents in new languages.

References

Bel, N. C. Koster and M. Villegas (2003). Cross-lingual Text Categorization. Proceedings of ECDL 2003. Springer-Verlag LNCS, (pp. 126-139) Berlin.

Joachims, Thorsten (2001). A statistical learning model of text classification with support vector machines. In W. B. Croft, D. J. Harper, D.H. Kraft and J. Zobel editors, Proceedings of SIGIR'01, 24th ACM International Conference on Research and Development in Information Retrieval, (pp. 128-136), New-Orleans, US. ACM Press, New York.

Larkey, L.S. (1999), A patent search and classification system. Proceedings of DL-99, 4th ACM Conference on Digital Libraries (pp 179-187).

Peters, C and C.H.A. Koster, (2001). Uncertainty-based Noise Reduction and Term Selection in Text Categorization, Proceedings of the 24th BCS-IRSG European Colloquium on IR Research, Springer LNCS 2291, pp 248-267.

Riloff, E., (1995). Little words can make a big difference for text classification. In Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval,

Yang, Y. and J.O. Pedersen (1997), A comparative study on feature selection in text categorization. In Proceedings of ICML-97, 14th International Conference on Machine Learning.