Cross-Lingual Text Categorization

Nuria Bel[1], Cornelis H.A. Koster[2], and Marta Villegas[1]

[1] Grup d'Investigació en Lingüística Computacional Universitat de Barcelona, 08028 - Barcelona, Spain. {nuria,tona}@gilc.ub.es
[2] Computer Science Dept., University of Nijmegen, Toernooiveld 1, 6525ED Nijmegen, The Netherlands. kees@cs.kun.nl

Abstract. This article deals with the problem of Cross-Lingual Text Categorization (CLTC), which arises when documents in different languages must be classified according to the same classification tree. We describe practical and cost-effective solutions for automatic Cross-Lingual Text Categorization, both in case a sufficient number of training examples is available for each new language and in the case that for some language no training examples are available.
Experimental results of the bi-lingual classification of the ILO corpus (with documents in English and Spanish) are obtained using bi-lingual training, terminology translation and profile-based translation.

1 Introduction

Text Categorization is an important but usually rather inconspicuous part of Document Management and (more generally) Knowledge Management. It is used in many information-providing institutions, either in the form of a hierarchical mono-classification ("where does this document belong in our topic hierarchy") or as a multi-classification, assigning zero or more keywords to the document, with the purpose of enhancing and simplifying retrieval.

Automatic Text Categorization techniques based on manually constructed class profiles have shown that a high accuracy can be achieved, but the cost of manual profile construction and maintenance is quite high. Automatic Text Categorization systems based on supervised learning [16] can reach a similar accuracy, so that the (semi)automatic classification of monolingual documents is becoming standard practice. Now the question arises how to deal efficiently with collections of documents in more than one language, that are to be classified according to the same Classification Tree.

This article describes the cross-lingual classification techniques developed in the PEKING project [1] and presents the results achieved in classifying the ILO corpus using the LCS classification engine.

In the following two sections we relate our research to previous research in Cross-Language Information Retrieval, describe the ILO corpus and our experimental approach. In section 4 we establish a baseline for mono-lingual classification of the ILO corpus, using different classification algorithms (Winnow

[1] http://www.cs.kun.nl/peking

and Rocchio). In sections 5 and 6 we propose three different solutions for cross-language classification, implying increasingly smaller (and therefore less costly) translation tasks. Then we describe our main experiments in multi-lingual classification, and compare the results to the baseline.

2 Previous research

When we embarked on this line of research, we did not find any publications addressing the area of Cross-Lingual Text Categorization as such. On the other hand, there is a rich literature addressing the related problem of Cross-Lingual Information Retrieval (CLIR).

Both CLIR and CLTC are based on some computation of the similarity between texts, comparing documents with queries or class profiles. The most important difference between them is the fact that CLIR is based on queries, consisting of a few words only, whereas in CLTC each class is defined by an extensive profile (which may be seen as a weighted collection of documents).

In developing techniques for CLTC, we want to keep in mind the lessons learned in CLIR.

2.1 Cross-Lingual Information Retrieval

CLIR is concerned with the problem of a user formulating a query in one language in order to retrieve documents in several (other) languages. Two approaches can be distinguished:

1. translation-based systems either translate queries into the document language or languages, or they translate documents into the query language
2. Intermediate representation systems transfer both queries and documents into some language-independent representation, be it a thesaurus, some ontological representation or a language-independent vector space model.

It is important to notice that all current approaches have inherent problems. Translation of documents into a given language for a large number of documents is a rather expensive approach, especially in terms of time demands. Using thesauri or ontological taxonomies requires the availability of parallel or comparable corpora, and the same is required by interlingual vector space techniques.

To collect and process this material is time consuming, but more crucially, these techniques based on statistical approaches reduce accuracy when not enough material is available.

The less expensive approach is to translate the queries. The most widely used techniques for translating queries proceed by first identifying content words (simple or multi-word units such as compounds) and then supplying all possible translations. These translations can be used in normal search engines, reducing the development costs.

In [12], the effect of the *quality* of the translation resource is investigated. Furthermore it compares the effects of pre- and post-expansion: A query consisting of a number of words is expanded, either before or after translation (or both), with related words from the lexicon or from some corpus.

The expansion technique in the paper is a form of *pseudo relevance feedback*: using either the original query or its translated version and retrieving documents in the same language, the top 25 retrieved documents were taken as positive examples. From those, a set of 60 weighted query terms was composed including the original terms. This amounts to a combination of query expansion and term re-weighting.

The effect of degrading the quality of the linguistic resources turned out to be gradual. Therefore, it is to be expected that the effect of upgrading the resources should be gradual too. Weakness of the translation resources can be compensated for by query expansion.

In another recent paper[10] the use of Language Modeling in IR ([2, 7]) is extended to bi-lingual CLIR. For each query Q a *relevance model* is estimated consisting of a set of probabilities $P(w|R_Q)$ (the probability that a word sampled at random from a relevant document would be the word w).

In monolingual IR this relevance model is estimated by taking a set of documents relevant to the query. In CLIR, we need a relevance model for both the source language and the target language. The second can be obtained using either a parallel corpus or a bi-lingual lexicon giving translation probabilities.

The paper provides strong support for the Language Modeling approach in IR, in spite of the simplicity of the language models used (unigram). In using more informative representations (linguistically motivated terms) the effect should possibly be even larger.

2.2 Cross-lingual Text Categorization

Cross-lingual Text Categorization (CLTC) or Cross-lingual classification is a new research subject, about which no previous literature appears to be available. Still, it concerns a practical problem, which is increasingly felt in e.g. the documentation departments of multinationals and international organizations as they come to rely on automatic document classification. It is also manifest in many Search Engines on the web, which rely on a hierarchical classification of web pages to reduce search complexity and to raise accuracy: how should they combine this hierarchy with a classification on languages?

We shall distinguish two practical cases of CLTC:

- **poly-lingual training**:
  One classifier is trained on labeled documents written in different languages (or possible using different languages within one document).
- **cross-lingual training**: Labeled training documents are available only in one language, and we want to classify documents written in another language.

Most practical situations will be between these two extremes. Our experiments will show that the following is a feasible scenario:

> An organization, which already has an automatic classification system installed, wishes to extend this system to classify also documents in other languages. In order to ease the transition, some documents in those other languages are provided, either in untranslated form but manually supplied with a class label, or in translated form and without such a label. With limited manual intervention, a bootstrap of the system can be performed, so that documents in all those languages can be classified automatically in their original form by a single poly-lingual classifier.

By means of a number of experiments, we shall test the following hypotheses:

- **poly-lingual training**: simultaneous training on labeled documents in languages A and B will allow us to classify both A and B documents *with the same classifier*
- **cross-lingual training**: a monolingually trained classifier for language A plus a translation of the most important terms from language B to A allows to classify documents written in B.

2.3 Lessons from CLIR for CLTC?

In CLTC, for performing translations we shall have to use similar linguistic resources as in CLIR. Since our resources are less than ideal, should we compensate by implementing pre- and post-expansion? In CLTC, the role of the queries (with which test documents are compared) is played by the class profiles, which are composed from many documents; this may well have the same effect as explicit expansion of the documents or the profiles with morphological variants and synonyms. In fact, a class profile can be seen as an approximative (unigram) Language Model for the documents in that particular class.

3 The experimental procedure

All experiments were performed with Version 2.0 of the Linguistic Classification System LCS developed in the PEKING project[2], which implements the Winnow and Rocchio algorithms. It makes sense to compare those two algorithms, because we expect them to show qualitative differences in behaviour for some tasks. In Rocchio a class profile is essentially computed as a centroid, a weighted sum of the train documents, whereas Winnow[5, 6] by heuristic techniques computes (just like SVM)an optimal linear separator in the term space between positive and negative examples.

In the experiments we have used either a 25/75 or a 50/50 split of the data for training and testing, as stated in the text, with 12-fold or 16-fold cross-validation. Our goal is to compare the effect of different representations of the

[2] `www.cs.kun.nl/peking`

data rather than to reach the highest accuracy, and keeping the train sets small is good for performance (the cross-validation experiments are computationally very heavy). As a measure of Accuracy we have used the *micro-averaged F1 value*.

Although the ILO corpus is mono-classified (precisely one class per document) we allowed the classifiers to give 0-3 classes per document (which gives an indication of the Accuracy in multi-classification). Multi-classification gives more room for errors, and therefore has a somewhat lower Accuracy than Mono-classification.

For each representation, we first determined the optimal tuning and term selection parameters on the train set. The optimal parameter values depend on the corpus and on the document representation; their tuning is known to have an important effect on the Accuracy (see e.g. [8]), and without it the results from different experiments are hard to compare.

3.1 The ILO corpus

The ILO corpus is a collection of full-text documents, each labeled with one classname (mono-classification) which we have downloaded from the ILOLEX website of the International Labour Organisation[3]. 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".

The languages concerned are English, Spanish and French. From ILOLEX we extracted a bi-lingual corpus (only English and Spanish) of documents labeled for classification.

Although in the actual database every document has a translation, in constructing our corpus the documents were selected according to rough balance, avoiding total symmetry of documents in terms of language, that is, we have included some documents both in English and Spanish, and some in only one language. Some statistics of the ILO corpus:

1. the English version consists of 2165 documents. It comprises (after the removal of HTML tags) 4.2 million words, totalling 27 Mbytes. The average length of a document is 1942 words, and the document length varies widely, between 39 and 38646 words.
2. the Spanish version consists of 1590 documents. It comprises (after the removal of HTML tags) 4.7 million words, 30 Mbytes. The document length ranges from 117 to 7500 words. Most of the documents are around 2000 words.

The corpus is mono-classified into 12 categories, with a rather varying number of documents per category:

[3] `http://ilolex.ilo.ch:1567/Spanish/index.htm`

class name	# docs English	# docs Spanish	class description
02	123	74	Human rights
03	397	86	Conditions of employment
04	299	71	Conditions of work
05	22	23	Economic and social development
06	414	448	Employment
07	279	278	Labour Relations
08	85	81	Labour Administration
09	98	86	Health and Labour
10	156	148	Social Security
11	81	20	Training
12	131	154	Special prov. by category of persons
13	108	121	Special prov. by Sector of Econ. Act.
Total:	2165	1590	

4 The mono-lingual baseline

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

We also compared the traditional keyword representation with one in which multi-word terms were contracted into a single term (normalized keywords).

4.1 Monolingual keywords

The original documents were minimally preprocessed: de-capitalization, segmentation into words and elimination of certain special characters. In particular, no lemmatization was performed. The results (25/75 shuffle, 12-fold cross-validation) are as follows:

algorithm	representation	language	Accuracy	
			Multi 0:3	Mono 1:1
Winnow	keywords	English	.840±.013	.865±.007
Rocchio	keywords	English	.823±.010	.800±.010
Winnow	keywords	Spanish	.768±.014	.790±.015
Rocchio	keywords	Spanish	.755±.007	.764±.013

The Accuracy on the Spanish documents is significantly lower than on the English documents (according to Steiner's theorem for a Pierson-III distribution with bounds zero and one, $a \pm b$ and $c \pm d$ are different with risc $< 3\%$ when $\mid a - c \mid > \sqrt{b^2 + d^2}$, see page 929 of [1]), which is due not only to language characteristics but also to the fact that fewer train documents are available. In mono-classifying the English documents, Winnow is significantly more accurate than Rocchio.

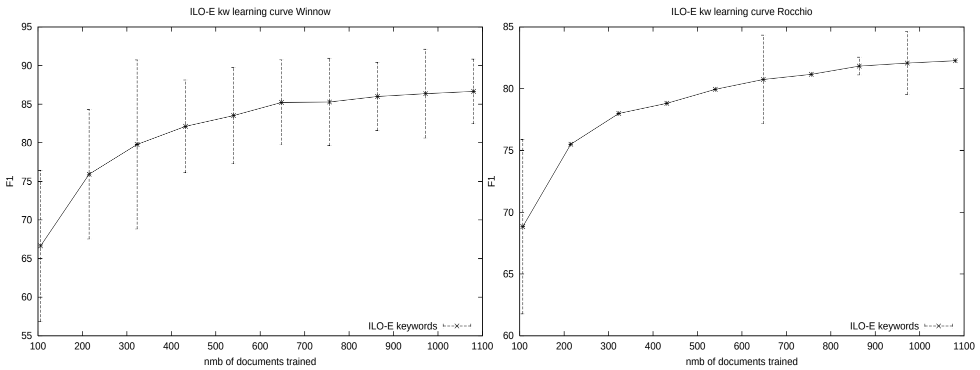

Fig. 1. Learning curves for English, Winnow and Rocchio

Figure 1 shows *learning curves* for the English documents, one for Winnow and one for Rocchio. (The learning curves for the Spanish documents are not given here, because they look quite similar.)

Using a 50/50 split of the English corpus, a classifier was trained in 10 epochs (= stepwise increasing subsets of the train set) and tested with the test set of 50% of the documents. This process was repeated for 16 different shuffles of the documents and the results averaged (16-fold cross-evaluation). The graphs show the Accuracy as a function of the number of documents trained, with error bars.

Notice that Winnow is on the whole more accurate than Rocchio, but that the variance is much larger for Winnow than for Rocchio.

4.2 Lemmatized keywords

Using the same pre-processing but in addition lemmatizing the noun and verb forms in the documents, the results are as follows:

algorithm	representation	language	Accuracy	
			Multi 0:3	Mono 1:1
Winnow	lemmatized keywords	English	.845±.008	.863±.006
Rocchio	lemmatized keywords	English	.797±.012	.817±.012
Winnow	lemmatized keywords	Spanish	.768±.012	.788±.015
Rocchio	lemmatized keywords	Spanish	.759±.010	.758±.017

In distinction to the situation in query-based Retrieval, in Text Categorization the lemmatization of terms does not seem to improve the Accuracy: although lemmatization enhances the Recall of terms, it may well hurt Precision more (see also [15]). In Text Categorization the positive effect of the conflation of morphological variants of a word is small: If two forms of a word are both important terms for a class, then they will both obtain an appropriate positive weight for that class provided they occur often enough, and if they don't occur often enough, their contribution is not important anyway.

4.3 Linguistically motivated terms

The use of n-grams instead of single words (unigrams) as terms has been advocated for Automatic Text Classification. Experiments like those of [11, 4], where only statistically relevant n-grams were used, did not show better results than the use of single keywords. For our experiment in CLTC the extraction of multi-word terms was required in order to be able to find proper translation equivalents, i.e. “trade union” vs. “sindicato” in Spanish.

In addition, for the monolingual experiments, we wanted to test to what extent linguistically motivated multi-word terms (for a survey on methods for automatic extraction of technical terms see [3]), rather than just statistically motivated ones, could make any improvement.

For Spanish, we extracted these Linguistically Motivated Terms (LMT) using both quantitative and linguistic strategies:

– A first list of candidates was extracted using Mutual Information and Likelihood Ratio measures over the available corpus
– The list of candidates was filtered by checking it against the list of well formed Noun Phrases that followed the patterns N+N, N+ADJ and N+prep+N.

This process ensured that all Spanish multi-words were both linguistically and statistically motivated and resulted in 303 bigrams (N+ADJ), and 288 trigrams (N+de+N mainly).

For want of a better term, we shall use the term *normalized* for a text in which important multi-word expressions have been contracted into one term (e.g. 'software_engineering' or 'Trabajadores_migrantes').

The list of English multi-word expressions was built from the multi-words present in the bilingual database (see section 6.1), that is, those resulting from the translation of Spanish terms and LMT's.

Training and testing on normalized documents gave the following results:

algorithm	representation	language	Accuracy	
			Multi 0:3	Mono 1:1
Winnow	normalized keywords	English	.840±.013	.867±.011
Rocchio	normalized keywords	English	.824±.010	.829±.011
Winnow	normalized keywords	Spanish	.762±.013	.800±.013
Rocchio	normalized keywords	Spanish	.769±.010	.779±.010

For English, the normalization has no effect, for Rocchio on Spanish there is a barely significant improvement. For Winnow there is no effect.

Even when using linguistic phrases rather than statistical phrases, document normalization seems to make no significant improvement to automatic classification (see also [9, 8]).

4.4 Comparing the learning curves

In order to faciltiate their comparison, figure 2 shows, for each combination of language and classification algorithm, the learning curves (50/50 split, 10 epochs, 16-fold cross-validation) for each of the three document representations.

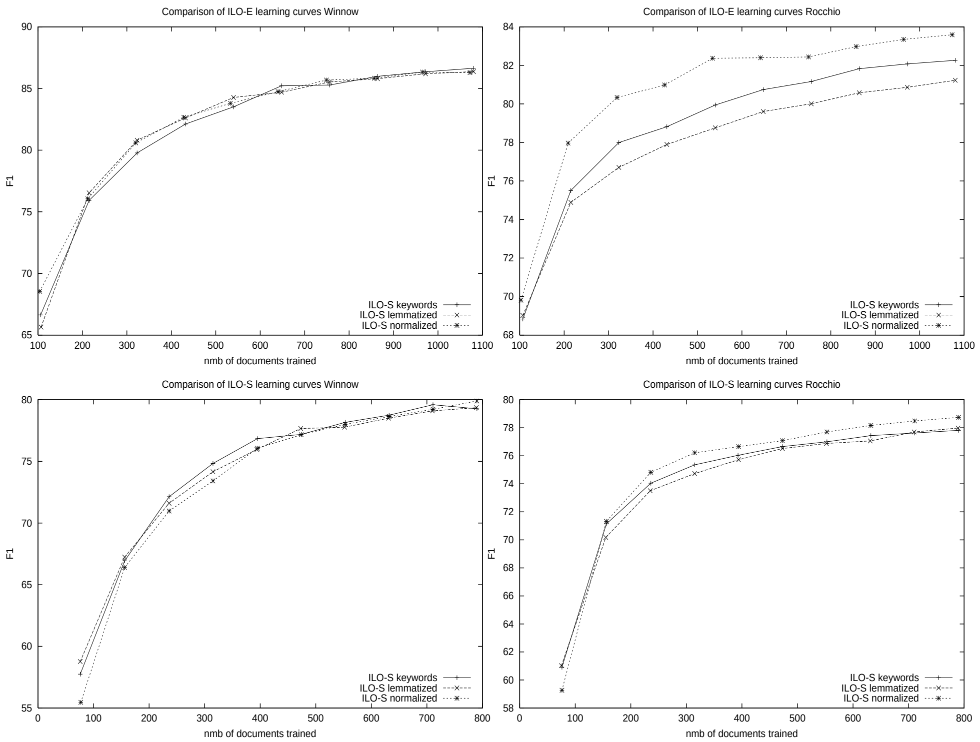

Fig. 2. Learning curves (English and Spanish, Winnow and Rocchio

For Winnow, the representation chosen makes no difference. Rocchio gains somewhat by normalisation, especially for English, whereas lemmatization has a small negative impact. Observe also that lemmatization and normalization do not improve the classification accuracy for small numbers of training documents, where it might be expected that term conflation would be more effective.

Since Winnow is the most accurate algorithm, we are more interested in its behaviour than in that of Rocchio, and therefore we may ignore the influence of lemmatization and normalization implied in the translation processes in the following sections.

5 Poly-lingual training and testing

In this section we shall investigate the effect of training on labeled documents written in a mix of languages. Since we have a bi-lingual corpus, we shall restrict ourselves (without loss of generality) to the bi-lingual case. The bi-lingual training approach amounts to building a single classifier from a set of labeled train documents in both languages, which will classify documents in any of the two trained languages, without translating anything and even without trying to find out what language the documents are in. We exploit the strong statistical properties of the classification algorithms, and use no linguistic resources.

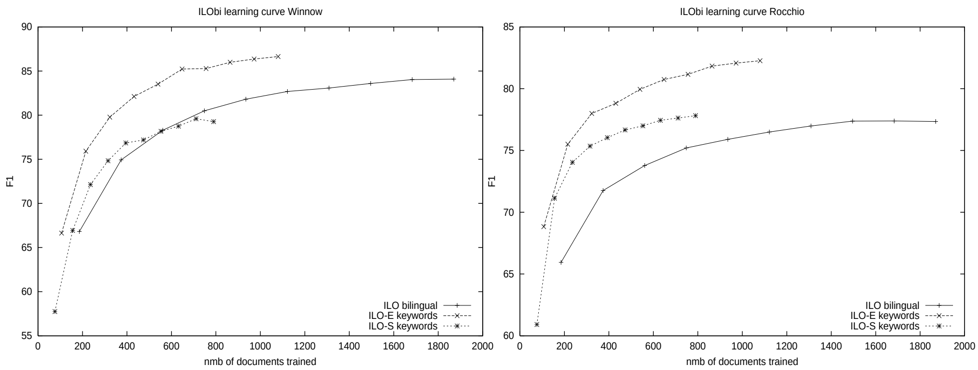

Fig. 3. Learning curves (English, bilingual and Spanish, Winnow and Rocchio)

The 2167 English and 1590 Spanish ILO documents (labeled with the same class-labels) were combined at random into one corpus. Then this corpus was randomly split into 4 train sets each containing 15% (563) of the documents and a fixed test set of 40% of the documents, a train set of size comparable to the above experiments, and tested with the remaining 40% as test set, with the following results:

algorithm	representation	language	Accuracy	
			Multi 0:3	Mono 1:1
Winnow	keywords	English and Spanish	.785±.013	.811±.014
Rocchio	keywords	English and Spanish	.739±.009	.758±.014

Using the Winnow classifier, the Accuracy achieved for the mixture of Spanish and English documents lies after 563 train documents above that for Spanish documents alone. But at this point only about 225 Spanish documents have been trained, so that it is quite surprising that the Accuracy is so high.

In graph 3 the learning curve for bi-lingual training (50/50 split, 16-fold cross-validation) is compared with those for the Spanish and English mono-lingual corpora.

Again, keeping in mind the number of documents trained in each language, the curve for bi-lingual classification with Winnow is nicely in the middle. Although the vocabularies of the two languages are very different, Winnow trains a classifier which is good at either. Rocchio on the other hand is quite negatively impacted. It attempts to construct a centroid out of all documents in a class, and is confused by a document set that has two very different centroids.

As an afterthought, we tested how well an English classifier understands Spanish, by training Winnow mono on 2164 English documents and testing on 1590 Spanish documents, without any translation whatsoever. We found an Accuracy of 10.75%! In spite of the difference in vocabulary, there are still some terms shared, probably mostly non-linguistic elements (proper names, abbreviations like ceacr, maybe even numbers) which are the same in both languages.

6 Cross-lingual training and testing

For Cross-Lingual Text Categorization, three translation strategies may be distinguished. The first two are familiar from Cross Language Information Retrieval (CLIR):

- **document translation:** Although translating the complete document is workable, it is not popular in CLIR, because automatic translations are not satisfactory and manual translations are too expensive
- **terminology translation:** constructing a terminology for each of the relevant domains (classes), and translating all domain terms. It is expected that these include all or most of the terms which are relevant for classification.
- **profile-based translation**: translate only the terms actually occurring in the class profiles (Most Important Terms or MIT's).

Translation of the complete document (either manually or automatically) has not been evaluated by us, since it costs much more effort than the other approaches possibility, without promising better results. Our experiments with the other techniques are described below.

6.1 The linguistic resources

We know from Cross-Lingual Information Retrieval applications that existing translation lexica are very limited. In order to enlarge their coverage it is also possible to extract translation equivalences from aligned corpora. but both approaches show some drawbacks [14]. While bi-lingual dictionaries and glossaries provide reliable information, they propose more than one translation per term without preference information. Aligned corpora for very innovative domains, such as technical ones, offer contextualized translations, but the errors introduced by statistical processing of texts in order to align them are considerable.

Our translation resources were built using a corpus-driven approach, following a frequency criterion to include nouns, adjectives and verbs with a frequency higher than 30 occurrences in the bilingual lexicon. 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.

6.2 Terminology translation

In the approach based on terminology translation, these resources were used as follows:

1. training a classifier on all 2167 normalized English documents
2. using this classifier to classify the 1590 pseudo-English (Spanish) documents.

Our experiments (training on subsets of 25% of the English documents, testing on all pseudo-English documents, 12-fold cross-validation, and similarly for Spanish) gave the following results:

algorithm	representation	language	Accuracy	
			Multi 0:3	Mono 1:1
Winnow	keywords	English and pseudo-English	.696±.051	.792±.012
Rocchio	keywords	English and pseudo-English	.592±.025	.709±.012
Winnow	keywords	Spanish and pseudo-Spanish	.552±.062	.617±.062
Rocchio	keywords	Spanish and pseudo-Spanish	.538±.045	.589±.029

Winnow's mono-classification of pseudo-English documents after training on English documents is quite good (as good as when training and testing on Spanish keywords), but when translating English documents to pseudo-Spanish the result is not good (which is only partly explained by the lower number of train examples). Rocchio is in all cases much worse than for monolingual classification.

Both algorithms are much worse in multi-classification. A closer look at the classification process shows why: the test documents obtain very low and widely varying thresholds. Without forcing each document to obtain one class, 25% of the pseudo-English documents and nearly 50% of the pseudo-Spanish documents are not accepted by Winnow for any class. We have violated the fundamental assumption that train- and test-documents must be sampled from the same distribution, on which the threshold computation (and indeed the whole classification approach) is based. In the pseudo-English documents most English words from the train set are missing, and therefore the thresholds are too high. Furthermore, the many synonyms generated as a translation for a single term in the original distort their frequency distribution. This thresholding problem can be solved by filtering the words in the english train set and using a validation set to set the thresholds (not tried here).

In spite of the thresholding problem, terminology translation is a viable approach for cross-lingual mono-classification.

6.3 Profile-based translation

Why don't we ask the classifier what terms it would like to find? When using an English classifier on Spanish terms, we should need for each English term in the profile a list of only those terms in Spanish that can be translated to that English term – including morphological variation, spelling variation and synonymy. We need a translation for only the terms actually occurring in the profile, and not for any other term (because it would not contribute anything).

Our previous research [13] has shown that, using a suitable Term Selection algorithm, a surprisingly small number of terms per class (40-150) gives optimal Accuracy. All other terms can safely and even profitably be eliminated. Based on this observation, we have investigated the effect of translating *only* towards the words occurring in the class profiles, performing the following experiment:

1. we determined the best 150 terms in classifying all English documents with Winnow, and combined the results into a vocabulary of 923 different words (out of 22000)
2. a translation table from Spanish to English was constructed, comprising for each English word in the vocabulary those Spanish words that may be translated to it

3. a classifier was trained on all English documents but using only the words in the vocabulary
4. this classifier was tested on all Spanish documents, translating only the Spanish terms having a translation towards a word in the vocabulary.

The resulting accuracy when using profile-translation (training a classifier on English documents and classifying with it Spanish documents in which just the profile words have been translated) gave the following results:

algorithm	representation	language	Accuracy	
			Multi 0:3	Mono 1:1
Winnow	keywords	profile translation Eng/Spa	.605±.071	.724±.035
Rocchio	keywords	profile translation Eng/Spa	.681±.048	.730±.019

Taking into account that the best accuracy achieved in the mono-classification of Spanish documents was .775, that no labeled Spanish documents were needed and that the required translation effort is very small, an Accuracy of .724 in cross-lingual classification is not bad. On this corpus, Rocchio does as well as Winnow in mono-classification and even significantly better in multi-classification.

7 Conclusion

Cross-lingual Text Categorization is actually easier than Cross-lingual Information Retrieval, for the same reason that lemmatization and term normalization have much less effect in CLTC than in CLIR: the law of large numbers is with us. Given an abundance of training documents, our statistical classification algorithms will function well, even in the absence of term conflation, which is the CLTC equivalent of expansion in CLIR. We do not have to work hard to ensure that all linguistically related forms or synonyms of a word are conflated: If two equivalent forms of a word occur frequently enough to have an impact on classification, they will also do so as independent terms.

We have found viable solutions for two extreme cases of Cross-Lingual Text Categorization, between which all practical cases can be situated. On the one hand we found that *poly-lingual training*, training one single classifier to classify documents in a number of languages, is the simplest approach to cross-lingual Text Categorization, provided that enough training examples are available in the respective languages (tens to hundreds), and the classification algorithm used is immune to the evident disjointedness of the resulting class profile (as is the case for Winnow but not for Rocchio).

At the other extreme, when for the new language no labeled train documents are available it is possible to use *terminology translation*: find, buy or construct a translation resource from the new language to the language in which the classifier has been trained, and translate just the typical terms of the documents. Finally, it is possible to translate only the terms in the class profile. Although the accuracy is somewhat lower, this *profile-based translation* provides a very cost-effective way to perform Cross-lingual Classification: in our experiment an average of 60 terms per class had to be translated.

In a practical classification system, the above techniques can be combined, by using terminology translation or profile-based translation to generate examples for poly-lingual training and then bootstrap the poly-lingual classifier (with some manual checking of uncertain classifications).

References

1. M. Abramowitz and Irena A. Stegun (1970), Handbook of Mathematical Functions, 9th edition.
2. A. Berger and J. Lafferty (1999), Information Retrieval as statistical translation, Proceedings ACM SIGIR'99, pp. 222-229.
3. Cabré, M.T., R. Estopà and J. Vivaldi (2001), Automatic Term Detection: A review of current systems, In: *Recent Advances in Computational Terminology*, John Benjamins, Amsterdam.
4. M.F. Caropreso, S. Matwin and F. Sebastiani (2000), A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization, A. G. Chin (Ed.), *Text Databases and Document Management: Theory and Practice*, Idea Group Publishing, Hershey, US, pp. 78–102.
5. I. Dagan, Y. Karov, D. Roth (1997), Mistake-Driven Learning in Text Categorization. *Proceedings of the Second Conference on Empirical Methods in NLP*, pp. 55-63.
6. A. Grove, N. Littlestone, and D. Schuurmans (2001), General convergence results for linear discriminant updates. Machine Learning 43(3), pp. 173-210.
7. Djoerd Hiemstra and F. de Jong (1999), Disambiguation strategies for cross-language Information Retrieval, Proceedings ECDL'99, Springer SLNC vol 1696 pp. 274-293.
8. Cornelis H.A. Koster and Marc Seutter (2002), Taming Wild Phrases, Proceedings 25th European Conference on IR Research (ECIR'03), Springer LNCS 2633, pp 161-176.
9. Leah S. Larkey (1999), A patent search and classification system, Proceedings of DL-99, 4th ACM Conference on Digital Libraries, pp. 179-187.
10. V. Lavrenko, M. Choquette and W. Bruce Croft (2002), Cross-Lingual Relevance Models, Proceedings ACM SIGIR'02, pp. 175-182.
11. Lewis, D.D. (1992), An evaluation of phrasal and clustered representations on a text categorization task. Proceedings ACM SIGIR'92.
12. Paul McNamee and James Mayfield (2002), Comparing Cross-Language Query Expansion Techniques by Degrading Translation Resources, Proceedings ACM SIGIR'02, pp. 159-166.
13. C. Peters and C.H.A. Koster, Uncertainty-based Noise Reduction and Term Selection in Text Categorization, Proceedings 24th BCS-IRSG European Colloquium on IR Research, Springer LNCS 2291, pp 248-267.
14. Philip Resnik, Douglas W. Oard, and Gina-Anne Levow (2001), Improved Cross-Language Retrieval using Backoff Translation, Human Language Technology Conference (HLT), San Diego, CA, March 2001.
15. E. Riloff (1995), Little Words Can Make a Big Difference for Text Classification, Proceedings ACM SIGIR'95, pp. 130-136.
16. F. Sebastiani (2002), Machine learning in automated text categorization. ACM Computing Surveys, Vol 34 no 1, 2002, pp. 1-47.