Bilingual Machine-Aided Indexing

Jorge Civera and Alfons Juan

Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València
{jcivera,ajuan}@dsic.upv.es

Abstract

The proliferation of multilingual documentation in our Information Society has become a common phenomenon. This documentation is usually categorised by hand, entailing a time-consuming and arduous burden. This is particularly true in the case of *keyword assignment*, in which a list of keywords (descriptors) from a controlled vocabulary (thesaurus) is assigned to a document. A possible solution to alleviate this problem comes from the hand of the so-called Machine-Aided Indexing (MAI) systems. These systems work in cooperation with professional indexer by providing a initial list of descriptors from which those most appropiated will be selected. This way of proceeding increases the productivity and eases the task of indexers. In this paper, we propose a statistical text classification framework for bilingual documentation, from which we derive two novel bilingual classifiers based on the naive combination of monolingual classifiers. We report preliminary results on the multilingual corpus Acquis Communautaire (AC) that demonstrate the suitability of the proposed classifiers as the backend of a fully-working MAI system.

1. Introduction

The proliferation of multilingual documentation in our Information Society has become a common phenomenon in many official institutions (EU parliament, the Canadian Parliament, UN sessions, Catalan and Basque Parliaments in Spain, etc.) and private companies (user's manuals, newspapers, books, etc.). In many cases, this textual information needs to be categorised by hand, entailing a time-consuming and arduous burden.

This fact is particularly true in *keyword assignment*, in which a list of keywords (descriptors) from a thesaurus is assigned to a document, without requiring the keywords to be explicitly present in the document. This task can be efficiently done using MAI tools (Hodge, 1998; Pouliquen and others, 2003). MAI tools assign to a document a list of keywords (descriptors) from a controlled vocabulary (thesaurus) for indexing purposes. This list of descriptors suggested by the system is reviewed by expert indexers to add and select those descriptors that are the most suitable.

The interest behind the development of indexing systems is not only the document classification capabilities *per se*, but also the cross-lingual information access possibilities (Pouliquen and others, 2003) through multilingual thesaurus, as EuroVoc (EC, 1995), AgroVoc (FAO, 1998), etc. However, current MAI systems do not take full advantage of multilinguality since they are based on monolingual text classifiers, both rule-based systems (Hlava and Hainebach, 1996; Loukachevitch and Dobrov, 2002) and statistical methods (Lin and Hovy, 2000; Pouliquen and others, 2003). A more sophisticated approach is to develop new classification models that make profit of multilingual information in order to boost the performance of MAI systems. In this paper, we will focus on bilingual text classification, even though the extension of our classification model to the multilingual case is straightforward.

The structure of this paper is as follows. Section 2. introduces the basic probabilistic framework for bilingual classification, together with two possible instantiations of bilingual classifiers. Section 3. is devoted to the EM parameter estimation of one of the models proposed. In Section 4., some preliminary monolingual and bilingual results obtained on the multilingual Acquis Communautaire (AC) corpus are presented. Finally, some conclusions and thoughts for future work are stated in Section 5..

Work supported by the Agència Valenciana de Ciència i Tecnologia under grant GRUPS03/031, the Spanish project TIC2003-08682-C02-02 and the Ministerio de Educación y Ciencia.

2. Bilingual text classification

Given a bilingual document (x, y), in which x and y are documents in different languages and translations of each other, we assign (x, y) to that class (descriptor):

$$\begin{aligned} c(x,y) &= \operatorname*{argmax}_c \; p(c)\, p(x,y \mid c) \\ &= \operatorname*{argmax}_c \; p(c)\, p(y \mid c)\, p(x \mid y,c) \end{aligned} \qquad (1)$$

where $p(c)$ is the *a priori* probability of class c and $p(x, y \mid c)$ is the probability of observing the bilingual document (x,y) in class c. This last term can be better understood when decomposed as a class-dependent language model $p(y \mid c)$, and a class-dependent translation model $p(x \mid y, c)$. Language models express the idea of how likely is a given sequence of words, while translation models represent the degree of correlation between sequence of words across languages. Note that the classification rule proposed in Eq. 1 can be easily extended to the multiclass case by considering the C most probable classes.

However, it is common the case that documents labeled with the same class tend not only to devise about different topics, but also may consist of different kinds of sublanguages (legal texts, communications, questions, etc.) (Steinberger, 2001). For these reasons, it is appealing to consider the so-called mixture model, in which a class may contain documents from several unknown topics:

$$c(x,y) = \operatorname*{argmax}_c \sum_{t=1}^{T} p(c)\, p(t|c)\, p(y \mid t,c)\, p(x \mid y,t,c) \qquad (2)$$

where $p(y \mid t, c)$ and $p(x \mid y, t, c)$ are class and topic dependent language and translation models, respectively.

In the present work, we postpone the usage of translation models in bilingual text classification by considering x and y to be independent. Therefore, our simplified classification rules can be expressed as follows:

$$c(x,y) \approx \underset{c}{\operatorname{argmax}}\ p(c)\,p(y \mid c)\,p(x \mid c) \quad (3)$$

$$\approx \underset{c}{\operatorname{argmax}}\ \sum_{t=1}^{T} p(c)\,p(t|c)\,p(y \mid t,c)\,p(x \mid t,c) \quad (4)$$

In this work, Eq. 3 will be instantiated as a language-independent smooth n-gram model. Smooth n-gram models has been successfully applied in many different areas related to natural language processing. The parameter estimation of the smooth n-gram models is performed according to the maximum likelihood estimation paradigm along with powerful and well-founded smoothing techniques. These models were trained with the well-known and publicly available SRILM toolkit (Stolcke, 2002).
Conversely, Eq. 4 will be represented by a language-independent multinomial (unigram) mixture model. Mixture modelling is a standard pattern classification technique. In text classification, the use of multinomial mixtures (Novovicová and Malík, 2003) can be seen as a generalisation of the Naive Bayes text classifier by relaxing its feature independence assumption. Maximum likelihood estimation of mixture parameters can be reliably accomplished by the well-known *Expectation-Maximisation (EM)* algorithm (Dempster et al., 1977).
Monolingual classifiers can be easily derived from bilingual models in Eqs. 3 and 4 by ignoring one of the terms associated to one of the languages.

3. EM mixture parameter estimation

This section is devoted to the presentation of an instance of the EM algorithm that estimates the parameters of the mixture model presented in Eq. 4 for a given class c.
Let $(X,Y) = \{(x_1,y_1),\ldots,(x_N,y_N)\}$ be a set of samples to learn the parameters in Eq. 4. The only information retained is two vectors of word counts $x = (x_1,\ldots,x_D)$ and $y = (y_1,\ldots,y_E)$, where x_d and y_e are the number of occurrences of word d and e in the input and output sentences, respectively. D and E are the size of the input and output vocabularies.
The vector of unknown parameters is:

$$\Theta = (p(t); p(d \mid t); p(e \mid t)) \quad (5)$$

for all $t = 1,\ldots,T$, $d = 1,\ldots,D$ and $e = 1,\ldots,E$.
We are excluding the number of components from the estimation problem, as it is a crucial parameter to control model complexity and it is discussed in Section 4..
Following the maximum likelihood principle, the best parameter values maximise the log-likelihood function,

$$\mathcal{L}(\Theta|X,Y) = \sum_{n=1}^{N} \log \sum_{t=1}^{T} p(t)\,p(x_n|t)\,p(y_n|t) \quad (6)$$

In order to find these optimal values, it is useful to think of each sample pair (x_n,y_n) as an *incomplete* component-labelled sample, which can be completed by an indicator vector $z_n = (z_{n1},\ldots,z_{nT})$ with 1 in the position corresponding to the component generating (x_n,y_n) and zeros elsewhere. In doing so, a complete version of the log-likelihood function (6) can be stated as

$$\mathcal{L}_C(\Theta|X,Y,Z) = \sum_{n=1}^{N}\sum_{t=1}^{T} z_{nt} \log\big(p(t)p(x_n|t)p(y_n|t)\big) \quad (7)$$

where $Z = \{z_1,\ldots,z_N\}$ is the so-called *missing* data.
The form of the log-likelihood function given in (7) is generally preferred because it makes available the *EM* optimisation algorithm (for finite mixtures).
This algorithm proceeds iteratively in two steps. The E(xpectation) step computes the expected value of the missing data given the incomplete data and the current parameters. The M(aximisation) step finds the parameter values which maximise (7), on the basis of the missing data estimated in the E step. In our case, the E step replaces each z_{nt} by the posterior probability of (x_n,y_n) being actually generated by the t-th component,

$$z_{nt} = \frac{p(t)\,p(x_n \mid t)\,p(y_n \mid t)}{\sum_{t'=1}^{T} p(t')\,p(x_n \mid t')\,p(y_n \mid t')} \quad (8)$$

for all $n = 1,\ldots,N$ and $t = 1,\ldots,T$, while the M step finds the maximum likelihood estimates for the priors,

$$p(t) = \frac{1}{N}\sum_{n=1}^{N} z_{nt} \qquad (t = 1,\ldots,T) \quad (9)$$

and the component prototypes,

$$p(d \mid t) = \frac{1}{\sum_{n=1}^{N} z_{nt} \sum_{d'=1}^{D} x_{nd'}} \sum_{n=1}^{N} z_{nt}\,x_{nd} \quad (10)$$

$$p(e \mid t) = \frac{1}{\sum_{n=1}^{N} z_{nt} \sum_{e'=1}^{E} y_{ne'}} \sum_{n=1}^{N} z_{nt}\,y_{ne} \quad (11)$$

for all $t = 1,\ldots,T$, $d = 1,\ldots,D$ and $e = 1,\ldots,E$.

4. Experimental results

4.1. Dataset

Experiments were carried out on the Acquis Communautaire (AC) corpus (Steinberger et al., 2006). This large text collection contains documents selected from the European Union (EU) legislation in all the EU languages. Most of these documents have been manually classified according to the EuroVoc thesaurus. Each document is assigned a set of EuroVoc descriptors out of 6645 possible, even though only those 990 descriptors ocurring at least 5 times were considered in this work for evaluation purposes. Before training our text classifier, the AC corpus underwent a basic preprocessing consisting in downcasing, isolation of punctuation marks and replacement of numbers by a generic label. Some statistics of the preprocessed French-English partition of this corpus are shown in Table 1.
However, we prefered not to apply any language-dependent preprocessing, such as lemmatisation, multi-word mark-up or stopword lists, since our models are intended to deal with multilingual text. This linguistic preprocessing would improve the accuracy of classifiers and we plan to consider it in future work (Pouliquen and others, 2003).

	Fre	Eng
documents	5108	
average length	1819	1564
vocabulary	36.6K	32.5K
singletons	10.6K	10.5K
running words	9.3M	8.0M

Table 1: Basic statistics of the preprocessed French-English partition of the AC corpus.

4.2. Experimental results

We evaluated the bilingual classifier discussed above, and also its monolingual counterpart, on random 80%-20% train-test splits of the French-English AC partition. Classifiers were assessed in terms of precision and recall on a per-document basis, since this measure is closer to user needs. Also, it should be considered that the number of EuroVoc descriptors varies from one document to another, therefore a strategy to select the right number of descriptors for each document is required. As a first, preliminary approach, we have simply extract five descriptors per document, which is the average number of descriptors in the whole corpus.

The preliminary results obtained for the smooth n-gram (straight lines) and mixture (curves) multinomial classifiers are shown in Figure 1, both for the best monolingual (English-only) and the bilingual classifier. In the case of mixture-based classifiers, an analysis of the evolution of the precision and recall values as a function of the number of mixture components per class was performed. Each plotted point along mixture curves is an average over values obtained from 6 randomised trials. In the case of smooth n-gram models, a single experiment for each parameter setting was considered.

From the results in Figure 1, we can observe that the trigram (3g) classifier performs the best on its monolingual and bilingual versions, followed by the bigram (2g) classifier, the mixture multinomial (mix 1g) classifier and the unigram (1g) classifier. This perfomance directly correlates with the increasing complexity of the models that support these classifiers. Additional experiments demonstrated that smooth n-gram models beyond trigrams provides no accuracy improvement at all.

When analysing the behaviour of multinomial mixture on monolingual and bilingual classifiers clearly outstands the advantages of multiple component over single component modelling. Indeed, we could consider that the optimal number of unkown topics (components) in our mixture model is about 10, thereafter the precision and recall values seem to follow a steady trend.

Nevertheless, these results surprisingly reflect that there is little difference between the performance of the monolingual and bilingual classifiers. Even though, this is not the rule but the exception, as revealed in previous work (Juan and Civera, 2005; Civera and Juan, 2005).

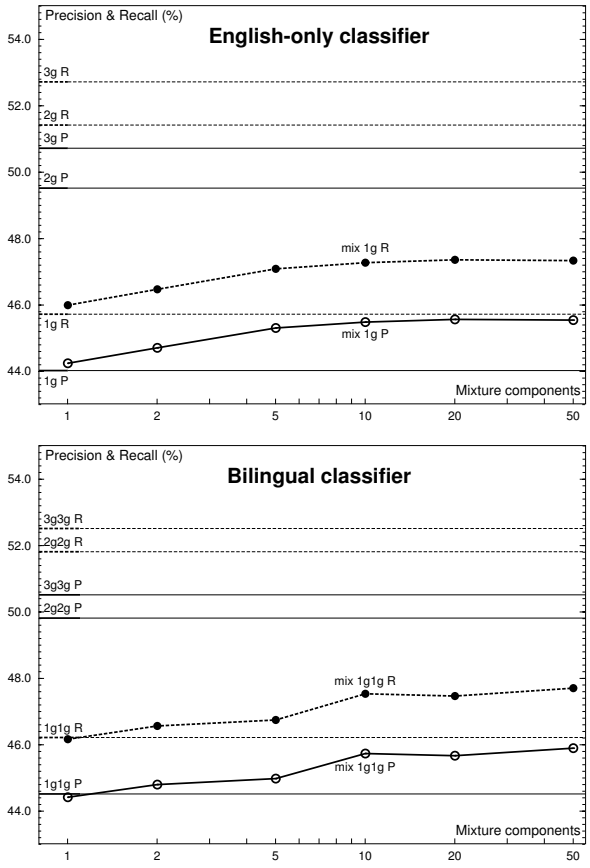

Figure 1: Precision (P) and recall (R) curves as a function of the number of mixture components for the English-only (top) and bilingual (bottom) multinomial (mix) classifiers. Precision and recall straight lines are plotted for the English-only and bilingual single component n-gram (ng) classifier.

4.3. Discussion

The excellence of these results should be assessed bearing in mind the complexity of this task and how MAI systems work. On the one hand, professional indexers do not completely agree on the most suitable descriptors for a given document. Indeed, previous studies (Pouliquen and others, 2003) on annotator agreement mantain that keyword overlapping among indexers is about 70% to 80%.

On the other hand, MAI systems work by providing a lengthy list of descriptors from which an indexer would select those ones considered most appropriated. For evaluation purposes, we decided that our MAI system should provide only 5 descriptors for each document, seeking a balance between precision and recall. However, in a MAI scenario, we would be more interested in recall, since we would like that our system provides a longer list of descriptors, from which a indexer would filter out those unsuitable descriptors.

Taking this into account, we conducted some additional experiments to evaluate the recall values that we would obtain if we considered a longer list of descriptors. These experiments revealed that our MAI system would be offering up to 68.9% of the correct descriptors for a list of 10 descriptors, and up to 78.7% for a list of 20 descriptors. These figures clearly convey the possibility of a MAI system which suggests most of the desired descriptors.

5. Conclusions and future work

In the current work, we have presented two bilingual text classifiers and their corresponding monolingual counterparts based on multinomial mixture and smooth n-gram models. The performance of these classifiers was assessed on the recently released preliminary version of the multilingual AC corpus.

Three outstanding conclusions can be stated from the results presented. First, multinomial (unigram) mixture-based classifiers surpass single component unigram classifiers. In fact, we have taken advantage of the flexibility of the mixture modelisation over the "single component" approach to further improve the precision and recall values achieved. Second, smooth n-gram models clearly outperform multinomial mixture models. This is so, because smooth n-gram models go beyond the bag-of-words representation and make profit of the context information (Peng and others, 2003; Scheffer and Wrobel, 2002). Third, bilingual classifiers show similar performance to their monolingual counterparts, although previous work on simpler datasets exhibit different behaviour (Juan and Civera, 2005; Civera and Juan, 2005). As said above, we think that this may be due to the relatively high complexity of the AC task. Nonetheless, the accuracy of our smooth n-gram classifier is good enough to support a MAI system, that would be providing on average about 80% of the correct descriptors associated with a document.

As a future work, there are several research lines that would be interesting to explore. First of all, the accuracy of multinomial mixture classifier may be significantly boosted by incorporating some of the techniques proposed in (Rennie and others, 2003; Pavlov and others, 2004). Extensions of smooth n-gram models provide an interesting starting point for more versatile language models, as mixtures of smooth n-gram models (Iyer and Ostendorf, 1999) or smooth n-gram models that incorporate automatically learned word classes (Brown and others, 1992). Other appealing approaches consider the problem of text classification under the maximum entropy framework (Nigam et al., 1999) or the application of multi-label text classifiers (McCallum, 1999).

All in all, the two bilingual classifiers described in this work are relatively simple models for the statistical distribution of bilingual texts. More sophisticated models, such as IBM statistical translation models (Brown and others, 1990), may be better in describing the statistical distribution of bilingual, correlated texts.

6. References

P. F. Brown et al. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85.

P. F. Brown et al. 1992. Class-based n-gram models of natural language. *Comput. Linguistics*, 18(4):467–479.

J. Civera and A. Juan. 2005. Multinomial Mixture Modelling for Bilingual Text Classification. Technical report DSIC-II/10/05, UPV.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum Likelihood from Incomplete Data via the *EM* Algorithm. *Journal of the Royal Statistical Soc.*, 39(B):1–38.

EC. 1995. Thesaurus eurovoc - volume 2: Subject-oriented version. Annex to the index of the Official Journal of the EC, Office for Official Publications of the EC. http://europa.eu.int/celex/eurovoc.

FAO. 1998. Multilingual agricultural thesaurus. World Agricultural Information Center. http://www.fao.org/scripts/agrovoc/frame.htm.

M. Hlava and R. Hainebach. 1996. Multilingual Machine Indexing. In *Proc. of NIT'96*, pages 105–121.

G. Hodge. 1998. CENDI agency indexing system descriptors: A Baseline Report. Technical report, Information International Associates, Inc.

R. M. Iyer and M. Ostendorf. 1999. Modelling long distance dependence in language: Topic mixtures versus dynamic cache models. *IEEE Transactions on Speech & Audio Processing*, 7(1):30–39.

A. Juan and J. Civera. 2005. Parallel Multinomial Mixtures for Bilingual Text Classification. Technical report DSIC-II/09/05, DSIC, Polytechnical Univ. of Valencia.

C. Lin and E. Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proc. of CoLing'00*, pages 495–501.

N. Loukachevitch and B. Dobrov. 2002. Crosslingual IR based on Multilingual Thesaurus specifically created for Automatic Text Processing. In *Proc. of SIGIR'02*, pages 105–121.

A. McCallum. 1999. Multi-label text classification with a mixture model trained by EM. In *Proceedings of the AAAI'99 Workshop on Text Learning*.

K. Nigam, J. Lafferty, and A. McCallum. 1999. Using maximum entropy for text classification. In *Proc. of IJCAI-99*, pages 61–67.

J. Novovicová and A. Malík. 2003. Application of Multinomial Mixture Model to Text Classification. In *Proc. of IbPRIA 2003*, pages 646–653.

D. Pavlov et al. 2004. Document Preprocessing For Naive Bayes Classification and Clustering with Mixture of Multinomials. In *Proc. of KDD'04*, pages 829–834.

F. Peng et al. 2003. Augmenting Naive Bayes classifiers with statistical language models. *Information Retrieval*, 7(3):317–345.

B. Pouliquen et al. 2003. Automatic annotation of multilingual text collections with a conceptual thesaurus. In *Proc. of EUROLAN'03*.

J. Rennie et al. 2003. Tackling the Poor Assumptions of Naive Bayes Text Classifiers. In *Proc. of ICML'03*, pages 616–623.

T. Scheffer and S. Wrobel. 2002. Text Classification Beyond the Bag-of-Words Representation. In *Proc. of ICML'02*.

Ralf Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, A. Ceausu, and D. Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proc. of LREC'06*.

R. Steinberger. 2001. Cross–lingual keyword assignment. In *Proc. of SEPLN'01*, pages 273–280.

A. Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proc. of ICSLP'02*.