

# A statistical approach to crosslingual natural language tasks

David Pinto<sup>1,2</sup>, Jorge Civera<sup>2</sup>, Alfons Juan<sup>2</sup>,  
Paolo Rosso<sup>2</sup>, and Alberto Barrón-Cedeño<sup>2</sup>

<sup>1</sup> Facultad de Ciencias de la Computación,  
Benemérita Universidad Autónoma de Puebla, Mexico  
<sup>2</sup> Departamento de Sistemas Informáticos y Computación  
Universidad Politécnica de Valencia, Spain

dpinto@cs.buap.mx, {jcivera,ajuan,proso,lbarron}@dsic.upv.es

**Abstract.** The existence of huge volumes of documents written in multiple languages in Internet lead to investigate novel approaches to deal with information of this kind. We propose to use a statistical approach in order to tackle the problem of dealing with crosslingual natural language tasks. In particular, we apply the IBM alignment model 1 with the aim of obtaining a statistical bilingual dictionary which may further be used in order to approximate the relatedness probability of two given documents (written in different languages). The experimental results successfully obtained in three different tasks –text classification, information retrieval and plagiarism analysis– highlight the benefit of using the presented statistical approach.

## 1 Introduction

The fast growth of the Internet and the increasing presence of multilinguality on the web poses new challenges for Natural Language Processing (NLP) technology. This fact leads us to the necessity of developing novel techniques to manage multilingual data. Indeed, the growing demand of NLP systems that deal with multilingual information induces the development and evaluation of multilingual systems in international events such as the Cross Language Evaluation Forum (CLEF)<sup>3</sup> and the Text Analysis Conference (TAC)<sup>4</sup>.

It is easy to find examples of NLP applications in which more than one language is involved. In this paper, we focus on three specific multilingual applications: bilingual Text Classification (TC), crosslingual Information Retrieval (IR) and crosslingual plagiarism. The proliferation and categorisation of multilingual documentation has become a common phenomenon in many official institutions and private companies. The most significant case is the EU parliament and commission, in which most official documents are translated into more than 20 languages, and categorised according to the Eurovoc thesaurus [1].

---

<sup>3</sup> <http://www.clef-campaign.org/>

<sup>4</sup> <http://www.nist.gov/tac/>

In this scenario, we could take advantage of redundancy and word correlation across languages to improve the accuracy of text classifiers. Moreover, users may be interested in information which is in a language other than their own native one. A common language scenario is where a user has some comprehension ability for a given language but the user is not sufficiently proficient to confidently specify a search query in that language. Thus, a search engine that may deal with this crosslingual problem should be of a high benefit. Finally, another multilingual application would be crosslingual plagiarism. In particular, this latter application is a real problem which occurs very frequently, for instance, in academic environments. It consists of the detection of text fragments which have been translated or partially rewritten from one original language without the adequate reference to the original text. The crosslingual component added to traditional NLP tasks incorporates a higher level of complexity which must be studied adequately.

Most of the current approaches to crosslingual NLP use conventional monolingual NLP techniques that usually incorporate a decoupled translation process as a preprocessing step to bridge the crosslingual gap. However, this two-step approach is too sensitive to translation errors, and in general to the accumulative effect of errors. In fact, even if we have a highly accurate NLP system, translation errors may prevent us from obtaining the desired performance. To overcome this drawback, we propose to bring together source and target documents written in two different languages as input to a direct probabilistic crosslingual NLP system which integrates both steps, translation and the specific NLP task, into a single one. In order to carry out this integrated approach to crosslingual applications, we propose to employ the IBM alignment model 1 (M1), which was firstly introduced for statistical Machine Translation (MT) [2].

The M1 model, the first of the IBM models, is basically defined as a statistical bilingual dictionary that captures word correlation across languages. In statistical MT, the M1 model has traditionally been an important component part in applications such as the alignment of bilingual sentences [3], the alignment of syntactic tree fragments [4], the segmentation of bilingual long sentences for improved word alignment [5], the extraction of parallel sentences from comparable corpora [6], the estimation of word-level confidence measures [7] and, it has been an inspiration for lexicalised phrase scoring in phrase-based systems [8]. In contrast to this statistical MT applications, the M1 model has been recently applied to other NLP areas such as bilingual TC, crosslingual IR and plagiarism with promising results [9–11]. To our knowledge, the employment of the M1 model outside statistical MT tasks has barely been investigated.

The rest of this paper is structured as follows. Section 2 presents the M1 model together with the maximum likelihood estimation of its parameters using the Expectation-Maximisation (EM) algorithm. In Section 3 we introduce the three crosslingual applications in which the M1 model has been applied and we describe how the M1 model has been integrated. Section 4 shows the experimental results obtained on actual tasks related to the proposed crosslingual applications. Finally, conclusions and further work are discussed in Section 5.

## 2 The M1 model

### 2.1 The model

Let  $x = x_1 \dots x_j \dots x_{|x|}$  be a sentence in a certain source language of known length  $|x|$  and  $y = y_1 \dots y_i \dots y_{|y|}$  is its corresponding translation in a different target language of known length  $|y|$ . Let  $\mathcal{X}$  and  $\mathcal{Y}$  denote the source and target vocabularies, respectively.

To derive the M1 model we start from the target-conditional probability distribution  $p(x|y)$ , for which we define the alignment hidden variable  $a = a_1 \dots a_j \dots a_{|x|}$ . The alignment variable connects each source word to exactly one target word  $a_j = \{0, \dots, i, \dots, |y|\}$ , being 0 the position of the NULL<sup>5</sup> word

$$p(x|y) = \sum_{a \in \mathcal{A}(x,y)} p(x, a|y) \quad (1)$$

where  $\mathcal{A}(x, y)$  denotes the set of all possible alignments from  $x$  to  $y$ . Now, we can factorise the term  $p(x, a|y)$  at the word-level from left to right

$$\begin{aligned} p(x, a|y) &= \prod_{j=1}^{|x|} p(x_j, a_j | x_1^{j-1}, a_1^{j-1}, y) \\ &= \prod_{j=1}^{|x|} p(a_j | x_1^{j-1}, a_1^{j-1}, y) p(x_j | x_1^{j-1}, a_1^j, y) \end{aligned} \quad (2)$$

where  $p(a_j | x_1^{j-1}, a_1^{j-1}, y)$  is an alignment probability function (p.f.) and  $p(x_j | x_1^{j-1}, a_1^j, y)$  is a lexical p.f. or statistical dictionary.

The well-known M1 model is defined by making the following two assumptions. First, we assume that the probability of aligning a source position to a target position is uniform

$$p(a_j | x_1^{j-1}, a_1^{j-1}, y) := \frac{1}{|y| + 1}. \quad (3)$$

Then, we also assume that the probability of translating a source word does only depend on the target word to which is aligned

$$p(x_j | x_1^{j-1}, a_1^j, y) := p(x_j | y_{a_j}) \quad (4)$$

where  $p(x_j | y_{a_j})$  is a statistical bilingual dictionary. Thus, we can rewrite Eq. (2) under assumptions in Eqs. (3) and (4) as

$$p(x, a|y; \Theta) = \prod_{j=1}^{|x|} \frac{1}{|y| + 1} p(x_j | y_{a_j}) \quad (5)$$

---

<sup>5</sup> The NULL word represents the target word to which those source words with no direct translation are connected.

where the parameter vector

$$\Theta = \{p(u|v) \quad u \in \mathcal{X}, v \in \mathcal{Y}\} \quad (6)$$

is a statistical bilingual dictionary.

**The model using indicator vectors** Now we change the nature of the original alignment variable  $a_j \in \{0, \dots, |y|\}$  from an integer value into an indicator vector

$$\mathbf{a}_j = (a_{j0}, a_{j1}, \dots, a_{j|y|})^t. \quad (7)$$

in order to ease the presentation of the parameter estimation of the M1 model. The vector  $\mathbf{a}_j$  values one in the  $i$ th position and zeros elsewhere, if the source position  $j$  is aligned to the target position  $i$ . Equivalently to Eq. (5), we have

$$p(x, a | y; \Theta) = \prod_{j=1}^{|x|} \prod_{i=0}^{|y|} \left[ \frac{1}{|y|+1} p(x_j | y_i) \right]^{a_{ji}}. \quad (8)$$

According to this notation, the initial model in Eq. (1) can be rewritten as

$$p(x | y; \Theta) = \prod_{j=1}^{|x|} \sum_{i=0}^{|y|} \frac{1}{|y|+1} p(x_j | y_i). \quad (9)$$

Eq. (9) is the usual form of the M1 model. The M1 model makes the naive assumption that source words are conditionally independent given  $y$

$$p(x | y; \Theta) = \prod_{j=1}^{|x|} p(x_j | y) \quad (10)$$

where

$$p(x_j | y) = \sum_{i=0}^{|y|} \frac{1}{|y|+1} p(x_j | y_i) \quad (11)$$

is the average probability of  $x_j$  to be translated into a target word in  $y$ .

## 2.2 Parameter estimation

In this section we present the maximum likelihood estimation of the parameter vector  $\Theta$  for the M1 model with respect to a set of  $N$  independent bilingual samples  $(X, Y) = ((x_1, y_1), \dots, (x_n, y_n), \dots, (x_N, y_N))^t$ , where the sequence of source and target words of the  $n$ th sample is  $x_n = (x_{n1}, \dots, x_{nj}, \dots, x_{n|x|})$  and  $y_n = (y_{n1}, \dots, y_{ni}, \dots, y_{n|y|})$ , respectively.

The log-likelihood function of  $\Theta$ , which we would like to maximise, is

$$L(\Theta; X, Y) = \sum_{n=1}^N \sum_{j=1}^{|x_n|} \log \sum_{i=0}^{|y_n|} \frac{1}{|y_n|+1} p(x_{nj} | y_{ni}). \quad (12)$$

Now, let  $A$  be the set of alignment indicator vectors associated with the bilingual pairs  $(X, Y)$  with

$$A = (a_1, \dots, a_n, \dots, a_N)^t. \quad (13)$$

The variable  $A$  is the alignment missing data in the M1 model, since this information is not present in the bilingual samples  $(X, Y)$ . Indeed, if the alignment information were available, the estimation of the parameter  $p(u | v)$  would be as easy as counting how many times the source word  $u$  is aligned to the target word  $v$  in  $(X, Y)$  and normalise adequately. However, we do not know how the bilingual samples are aligned, and the maximisation of Eq. (12) in order to estimate  $\Theta$  is troublesome.

For this reason, we need to revert to the well-known EM algorithm that performs the maximum likelihood estimation of statistical models with missing data. The idea behind the EM algorithm is to estimate the parameter vector  $\Theta$  in two iterative steps. First, the so-called E-step computes the expected value of the missing data, in our case, an estimation of the actual value of the alignment data. Then, in the so-called M-step, given that we have an estimation of the missing data we can compute  $\Theta$ , in the case of M1 model, an estimation of the bilingual dictionary. This two-step process is repeated to refine the estimation of the missing data, and then improve the estimation of the parameter vector.

Formally, the E step computes the expected value of the logarithm of the term  $p(X, A | Y)$ , given the (incomplete) data samples  $(X, Y)$  and a current estimate of  $\Theta$  at iteration  $k$ ,  $\Theta^{(k)}$ . Given that the alignment variables in  $A$  are independent from each other, we can compute the E step,

$$Q(\Theta | \Theta^{(k)}) = \sum_{n=1}^N \sum_{j=1}^{|x_n|} \sum_{i=0}^{|y_n|} a_{nji}^{(k)} \left[ \log \frac{1}{|y_n| + 1} + \log p(x_{nj} | y_{ni}) \right] \quad (14)$$

with

$$a_{nji}^{(k)} = \frac{p(x_{nj} | y_{ni})^{(k)}}{\sum_{i'=0}^{|y_n|} p(x_{nj} | y_{ni'})^{(k)}}. \quad (15)$$

That is, the expectation of word  $x_{nj}$  to be aligned to  $y_{ni}$  is our current estimation of the probability of  $x_{nj}$  to be translated into  $y_{ni}$ , instead of any other word in  $y_n$  (including the NULL word).

In the M step, we maximise Eq. (14), in order to obtain the standard update formula for the M1 model,

$$p(u | v)^{(k+1)} = \frac{N(u, v)}{\sum_{u' \in \mathcal{X}} N(u', v)} \quad \forall u \in \mathcal{X}, v \in \mathcal{Y} \quad (16)$$

where

$$N(u, v) = \sum_{n=1}^N \sum_{j=1}^{|x_n|} \sum_{i=0}^{|y_n|} \delta(x_{nj} = u) \delta(y_{ni} = v) a_{nji}^{(k)}. \quad (17)$$

The estimation of  $p(u | v)$  can be seen as a normalised partial count of how many times the source word  $u$  is aligned to the target word  $v$ .

### 3 Crosslingual applications based on the M1 model

In this section we introduce three different natural language tasks which could take benefit from the application of the previously presented M1 probabilistic model. The three tasks –text classification, information retrieval and plagiarism analysis– are considered to be in the crosslingual scenario, i.e., some texts are written in one language, whereas other texts of the same collection are written in another one. The following sections explain into detail the manner we have used the M1 model in each of these tasks.

#### 3.1 Bilingual text classification

The purpose of TC is to convert an unstructured repository of documents into a structured one by automatically assigning documents to a predefined number of groups, in the case of text clustering, or to a set of predefined categories, in the case of text categorisation. Doing so, the task of storing, searching and browsing documents in these repositories is significantly simplified [12].

Among the diverse approaches to TC, the well-known naive Bayes classifier [13, 14] is one of the most popular. Being so, there have been several instantiations and generalisations of this classifier, from Bernoulli mixtures [15] to multinomial mixtures [16, 17]. Both generalisations seek to relax the naive Bayes feature independence assumption made when using a single Bernoulli or multinomial distribution per category.

The unrealistic assumption of the naive Bayes classifier is one of the main reasons explaining its comparatively poor results in contrast to other techniques such as *boosting-based classifier committees* (boosting) [18] and *support vector machines* (SVM) [19]. However, the performance of the naive Bayes classifier is significantly improved by using the generalisations mentioned above. Moreover, there are other recent generalisations (and corrections) that also overcome the weaknesses of the naive Bayes classifier and achieve competitive results [20–23].

Bilingual TC is a novel application strongly characterised by word correlation across languages. This word correlation comes from the fact that the bilingual texts to be classified are mutual parallel translations. Given the latter scenario, we propose two main approaches to tackle bilingual TC. First, we may naively consider that bilingual texts were generated independently and therefore, there is not exist any crosslingual relation between words found in mutual translations. Alternatively, we may realistically assume that an underlying crosslingual word mapping exists and can be exploited to boost the performance of a bilingual classifier. Undoubtedly, the latter approach is significantly more complex than the former, however the crosslingual structure apprehended by the latter is a valuable information that cannot be neglected.

**The M1 model in bilingual text classification** Formally, our goal is to classify a bilingual parallel text  $(x, y)$  into one of the  $C$  supervised categories, so that we minimise the classification error. According to the optimal Bayes

decision (classification) rule [24], this can be achieved classifying the bilingual document  $(x, y)$  in the class with maximum posterior probability

$$\hat{c}(x, y) = \arg \max_{c=1, \dots, C} p(c | x, y) = \arg \max_{c=1, \dots, C} p(c) p(x, y | c) \quad (18)$$

where

$$p(x, y | c) = p(y | c) p(x | y, c) \quad (19)$$

can be factorised into a language p.f.,  $p(y | c)$ , and a translation p.f.,  $p(x | y, c)$ .

Given the bilingual classification rule stated in Eqs. (18) and (19), we can derive three different classification rules depending on the assumptions we make:

1. The *monolingual* rule only considers the contribution of one of the two languages

$$p(x, y | c) \approx p(x | c) \quad (20)$$

being  $p(x | c)$  modelled as a unigram model

$$p(x | c) := \prod_{j=1}^{|x|} p(x_j | c). \quad (21)$$

2. The bilingual *naive* factorisation rule unrealistically assumes that the bilingual parallel texts are independent from each other

$$p(x, y | c) \approx p(x | c) p(y | c) \quad (22)$$

where  $p(x | c)$  and  $p(y | c)$  are modelled as source and target unigram models, respectively. This rule incorporates a second source of information into the classifier that leads to believe in its superiority compared to the monolingual rule.

3. The *general* rule, as presented in Eqs. (18) and (19), models the language p.f.,  $p(y | c)$ , as a target unigram model and the translation p.f.,  $p(x | y, c)$ , as an M1 model. The integration of the M1 model allows to capture word correlation across languages enriching the structure of the bilingual text classifier, being theoretically superior to the monolingual and naive rules.

The maximum likelihood estimation of the source and target models is trivially computed by relative word frequency. The estimation of the M1 model involved in the general rule was already introduced in Section 2.

### 3.2 Crosslingual information retrieval

In CrossLingual Information Retrieval (CLIR), the usual approach consists of firstly translating the query into the target language and then retrieving documents in this language by using a conventional mono-lingual information retrieval system. The translation system might be of any type, rule-based, statistical or hybrid. In [25, 26], a statistical MT system is used, but it had to be

previously trained with parallel texts. See [27, 28] for a survey on CLIR. As previously mentioned, the above two-step approach is too sensitive to translation errors and, therefore, even if one information retrieval system performs well in a mono-lingual environment, its performance may be highly degraded in a multi-lingual scenario.

Probabilistic approaches which use parallel corpora in order to translate the input queries by means of a statistical dictionary in CLIR have been used in previous work [26]. However, our aim is *not* to translate queries but to obtain a set of associated words for a given query. In Figure 1 we may see the components of this novel approach that has just recently been explored in the literature [10]. The training phase is done by applying the M1 model to a set of pairs of query vs. relevant webpages. The obtained statistical dictionary is used in conjunction with the set of target webpages in order to show the most relevant ones given a query which is written in a different language of that of the webpages.

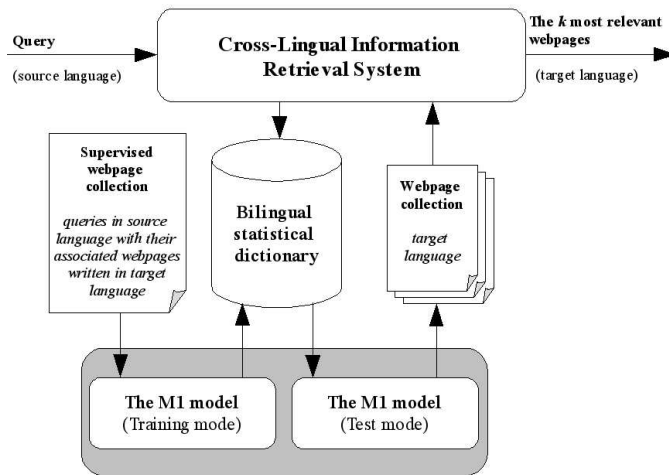


Fig. 1. The proposed crosslingual information retrieval approach.

**The M1 model in crosslingual information retrieval** Let  $x$  be a query text in a certain (source) language, and let  $y_1, y_2, \dots, y_n$  be a collection of  $n$  web pages in a different (target) language. Given a number  $k < n$ , we are interested in finding the  $k$  most relevant web pages with respect to the source query  $x$ . For this purpose, we have employed a probabilistic approach in which the  $k$  most relevant web pages are computed as those most probable ones given  $x$ , i.e.,

$$\hat{S}_k(x) = \arg \max_{\substack{S \subset \{y_1, \dots, y_n\} \\ |S|=k}} \arg \min_{y \in S} p(y|x) \quad (23)$$



In this work,  $p(y|x)$  is modelled by using the M1 model. The M1 model assumes that the order of the words in the query is not important and, therefore, each position in a document is equally likely to be connected to each position in the query. Although this assumption is unrealistic in MT, we consider the M1 model to be particularly well-suited for CLIR.

### 3.3 Crosslingual plagiarism

Plagiarism is the practice of rewriting someone else’s creative work, in whole or in part, without the adequate credit of the original authorship. Plagiarism may be carried out in the same language or across different languages (i.e., crosslingual plagiarism). In some way, crosslingual plagiarism analysis is related to the crosslingual information retrieval field [10, 29]. In fact, the aim is to retrieve those fragments that have been plagiarised from a source text originally written in another language.

Whereas some research works have been carried out for the automatic plagiarism analysis [30, 31], to our knowledge, *Cross-Lingual Plagiarism Analysis* (CLiPA) is a NLP task that nearly has been studied in the literature. In [32], it is proposed an automatic method to assign descriptors (keywords) drawn from the multilingual Eurovoc thesaurus to documents that can be found in different languages. Given the multilingual nature of these descriptors (but with a unique descriptor id) the authors suggest the possibility of automatically identifying document translations on the basis of common descriptors. This approach could be useful in the plagiarism analysis but it has not been investigated any further. In [33], the authors propose a preliminar method based on semantic analysis in order to identify documents that may be plagiarised in a different language.

**The M1 model in crosslingual plagiarism** Let  $x$  be a text fragment drawn from a suspicious document in a given (source) language, and let  $y_1, y_2, \dots, y_n$  be a set of original text fragments that may be the source of plagiarised texts in a different (target) language (the reference corpus).

Given the suspicious fragment  $x$ , the aim is to find the most probable original text fragment  $\hat{y}$  obtained as a plagiarised translation of  $x$

$$\hat{y}(x) = \arg \max_{y \in \{y_1, \dots, y_n\}} p(y|x) \quad (24)$$

Again,  $p(y|x)$  is modelled using the M1 model described in Section 2.

## 4 Experimental results

In this section we present the results obtained by applying the statistical approach, based on the M1 model, to three different crosslingual natural language tasks. Bilingual text classification is presented first, crosslingual information retrieval follows and, finally, the experiments with crosslingual plagiarism are shown.

#### 4.1 Bilingual text classification

The three bilingual text classifiers introduced in Section 3.1 were assessed in terms of classification error rate on two categorised parallel corpora. First, we describe these two corpora and then, we present the experimental setting employed to evaluate the proposed monolingual and bilingual text classifiers.

The INTERSECT corpus is a collection of sentence-aligned parallel texts in English, French and German drawn from different subjects. The English-French partition contains extracts coming from the Bible, the Canadian Hansard, fiction books, user manuals, news, scientific-technical reports and official documents from international organisations. These seven subjects constitute the categories in which bilingual parallel sentences are classified. The statistics of this corpus can be found in Table 1.

OPUS [34] is a growing sentence-aligned multilingual corpus of translated open source documents freely available on the Internet<sup>6</sup>. The collections extracted from OPUS for experimental purposes were:

- OpenOffice.org documentation<sup>7</sup>.
- KDE manuals including KDE system messages<sup>8</sup>.
- PHP manuals<sup>9</sup>.
- European constitution.

These four collections were considered as independent categories in which bilingual parallel sentences had to be classified. Their corresponding joint statistics are presented in Table 1.

**Table 1.** Statistics for INTERSECT and OPUS corpora ( $K = \times 10^3$ ,  $M = \times 10^6$ ).

	INTERSECT		OPUS	
	English	French	English	French
sentence pairs (K)	60		129	
average length	25	28	13	15
vocabulary (K)	35	47	20	26
running words (M)	1.5	1.7	1.7	1.9
singletons (K)	13	18	7	9

For experimental purposes, these corpora were partitioned into three sets, devoting 80% for training, 5% for development and 15% test sets. This partitioning process was randomly carried out 30 times to compute confidence intervals on test error. The parameters of the statistical models proposed were automatically learnt on the training set, additional smoothing parameters were manually

<sup>6</sup> <http://urd.let.rug.nl/tiedeman/OPUS/>

<sup>7</sup> OpenOffice.org is an open source office suite.

<sup>8</sup> The K Desktop Environment (KDE) is a free graphical desktop environment.

<sup>9</sup> Hypertext Preprocessor (PHP) is a widely-used general purpose scripting language.

tuned on the development set and the accuracy of the different text classifiers was assessed on the test set.

Table 2 presents the results for the monolingual, naive and general classifiers on the test sets of the INTERSECT and OPUS corpora. As observed in both corpora, the monolingual classifier is outperformed by the naive and general classifiers that incorporate additional information obtained from the second language. However, remarkably, the general classifier that capture word correlation across language using the M1 model is superior to the naive classifier in which each language is modelled independently. This proves the benefits of the M1 model to improve the accuracy of bilingual text classifiers. These results are consistent with those presented in [9].

**Table 2.** Classification Error Rates (CER) for the monolingual, naive and general classifiers on the INTERSECT and OPUS corpora.

CER(%)	Monolingual	Naive	General
INTERSECT	$8.0 \pm 0.7$	$6.4 \pm 0.6$	$5.5 \pm 0.4$
OPUS	$11.2 \pm 0.4$	$9.6 \pm 0.3$	$6.8 \pm 0.4$

## 4.2 Crosslingual information retrieval

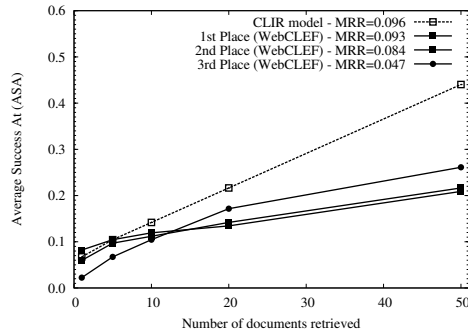
The presented approach (*CLIR Model*) was 10-fold cross-validated on the EuroGOV corpus [35]. In the training process we used its 134 supervised English queries. The obtained results were compared against the three best results reported at the bilingual “English to Spanish” subtrack of WebCLEF 2005<sup>10</sup>. A complete explanation of the systems/runs evaluated at WebCLEF 2005 may be found in [36]. The performance of each system is evaluated by using the Mean Reciprocal Rank (MRR). The reciprocal rank of a query response is the multiplicative inverse of the rank of the correct answer. The MRR is the average of the reciprocal ranks of the results for a sample of queries [37].

In Figure 2 it is presented the name of each run together with its MRR. The Average Success At (ASA) different number of documents retrieved (1, 5, 10, 20 and 50) is also shown in this figure. It is quite obvious the improvement that may be obtained when using the presented M1 model instead of traditional ones such as those which represent documents by using the vector space model. The main contribution of the M1 model in CLIR consists of its direct approach (translation and indexing/searching) over crosslingual data.

## 4.3 Crosslingual plagiarism

We have carried out some preliminary experiments by selecting five document fragments,  $y_1, \dots, y_5$ , from one author of the information retrieval area (e.g.

<sup>10</sup> <http://www.clef-campaign.org/>



**Fig. 2.** Comparison results over 134 English topics

*y*<sub>5</sub>: *Intrinsic plagiarism analysis deals with the detection of plagiarised sections within a document  $d$ , without comparing  $d$  to extraneous sources*). The aim of this experiment was to obtain an author-based bilingual statistical dictionary which can be used to perform an author-focused CLiPA.

For each original text fragment, we have constructed plagiarised cases by using both, machine and human translators. In the former approach, we have used five popular online translators<sup>11</sup>, whereas for the latter five different people have “plagiarised” each original fragment written in English to fragments in Italian. In general, the complete corpus is made up of the following text fragments:

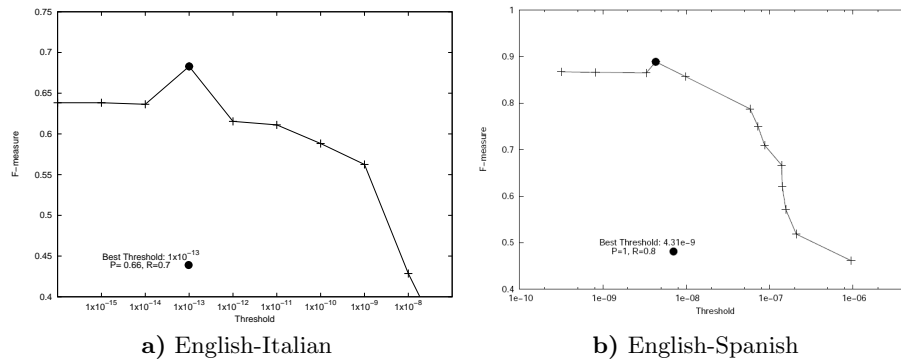
- i* Five original fragments written in English by a unique author
- ii* Five human simulated plagiarisms for each original fragment (in Italian)
- iii* Five automatic machine translations for each original fragment (in Italian)
- iv* Five unplagiarised versions for each original fragment (in Italian) obtained by rewriting the same original concept but mostly with other words.
- iv* Twenty unplagiarised (independent) fragments about the plagiarism topic originally written in Italian language

We have splitted the complete corpus into two datasets: training (60%) and test (40%). The training dataset, which is used to construct the statistical bilingual dictionary, is made up of 40 pairs composed of original fragments and their corresponding plagiarised versions. In the test dataset, we employed 10 plagiarised text fragments plus 45 unplagiarised text fragments.

Figure 3(a) shows the performance of the proposed statistical-based system identifying plagiarized documents at different thresholds. The maximum value obtained on the basis of the M1 model indicates the correct association between the original and the corresponding plagiarised text fragment. The obtained results in Italian-written plagiarisms are consistent with those reported in [11] for

<sup>11</sup> Freetranslation ([www.freetranslation.com](http://www.freetranslation.com)), Systran ([www.systransoft.com](http://www.systransoft.com)), Google ([www.google.com/language\\_tools](http://www.google.com/language_tools)), Worldlingo ([www.worldlingo.com](http://www.worldlingo.com)), and Reverso ([www.reverso.net](http://www.reverso.net))

documents written in English and plagiarized in Spanish language (see Figure 3(b)). Table 3 presents the CER results for the crosslingual classifiers on the test sets of the English-Italian and English-Spanish corpora.



**Fig. 3.** Performance of crosslingual plagiarism analysis with different thresholds

**Table 3.** Classification Error Rates for crosslingual plagiarism.

CER(%)	$1 \times 10^{-14}$	$1 \times 10^{-13}$	$1 \times 10^{-12}$	$1 \times 10^{-9}$	$1 \times 10^{-8}$	$1 \times 10^{-7}$
Italian	$0.25 \pm 0.26$	$0.28 \pm 0.27$	$0.34 \pm 0.29$	$0.38 \pm 0.30$	$0.40 \pm 0.30$	$0.46 \pm 0.30$
Spanish	$0.09 \pm 0.18$	$0.09 \pm 0.17$	$0.09 \pm 0.17$	$0.14 \pm 0.21$	$0.22 \pm 0.25$	$0.33 \pm 0.29$

## 5 Conclusions and future work

In this paper we have presented the application of the M1 statistical model to the bilingual TC and the crosslingual IR and plagiarism tasks. The M1 translation model has been widely employed in statistical machine translation but still unexplored in many other crosslingual NLP tasks.

The aim of the presented approach is to *directly* capture word correlation across languages, in contrast to current approaches that ignore or do not take full advantage of multilinguality. The experimental results obtained in different NLP tasks highlight the benefits of the M1 model and the usefulness of learning crosslingual information in multilingual applications.

As a future work, we plan to apply the M1 model for the bilingual TC task by using the challenging JRC-Acquis corpus. Moreover, the extension of the bilingual text classifier to the multilingual case is yet another appealing idea that we would like to study.

It would also be worth exploring superior IBM translation models, like the IBM model 2, that refine the M1 model by learning a crosslingual source-target position mapping. This refinement should be also analysed in other NLP tasks such as summarization, headlines generation and word sense disambiguation.

## Acknowledgments

The authors would like to thank Raphael Salkie of the University of Brighton for providing access to the INTERSECT corpus. This work has been partially supported by the MCyT TIN2006-15265-C06-04, TIN2006-15694-CO2-01 and CSD2007-00018 research projects, the BUAP-701 PROMEP/103.5/05/1536 grant and the FPU fellowship AP2003-0342.

## References

1. EC: 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 (1995) <http://europa.eu.int/celex/eurovoc>.
2. Brown, P.F., et al.: The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics* **19**(2) (1993) 263–311
3. Moore, R.: Fast and accurate sentence alignment of bilingual corpora. In: Proc. of AMTA'02. (2002) 135–244
4. Ding, Y., Gildea, D., Palmer, M.: An algorithm for word-level alignment of parallel dependency trees. In: Proc. of MT Summit IX. (2003) 95–101
5. Nevado, F., Casacuberta, F., Vidal, E.: Parallel corpora segmentation using anchor words. In: Proc. of EAMT/CLAW'03. (2003) 33–40
6. Munteanu, D., Fraser, A., Marcu, D.: Improved machine translation performance via parallel sentence extraction from comparable corpora. In: Proc. of HLT-NAACL'04. (2004) 265–272
7. Ueffing, N., Ney, H.: Word-level confidence estimation for machine translation. *Computational Linguistics* **33**(1) (2007) 9–40
8. Koehn, P., Och, F., Marcu, D.: Statistical phrase-based translation. In: Proc. of NAACL'03. (2003) 48–54
9. Civera, J., Juan, A.: Unigram-IBM Model 1 Mixtures for Bilingual Text Classification. In: Proc. of LREC'08. (2008)
10. Pinto, D., Juan, A., Rosso, P.: Using query-relevant documents pairs for cross-lingual information retrieval. In: Proc. of TSD'07. (2007) 630–637
11. Barrón-Cedeño, A., Rosso, P., Pinto, D., Juan, A.: On cross-lingual plagiarism analysis using a statistical model. In: Proc. of PAN-08. (2008) *in print*
12. Sebastiani, F.: Classification of text, automatic. In Brown, K., ed.: *The Encyclopedia of Language and Linguistics*. Volume 2. Second edn. Elsevier Science Publishers, Amsterdam, NL (2006) 457–463
13. Lewis, D.D.: Naive Bayes at Forty: The Independence Assumption in Information Retrieval. In: Proc. of ECML'98. (1998) 4–15
14. McCallum, A., Nigam, K.: A Comparison of Event Models for Naive Bayes Text Classification. In: Proc. of AAAI/ICML-98: Workshop on Learning for Text Categorization. (1998) 41–48

15. Juan, A., Vidal, E.: On the use of Bernoulli mixture models for text classification. *Pattern Recognition* **35**(12) (2002) 2705–2710
16. Nigam, K., et al.: Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning* **39**(2/3) (2000) 103–134
17. Novovicová, J., Malík, A.: Application of Multinomial Mixture Model to Text Classification. In: Proc. of IbPRIA 2003. Volume 2652 of Lecture Notes in Computer Science., Springer-Verlag (2003) 646–653
18. Schapire, R.E., Singer, Y.: Boostexter: A boosting-based system for text categorization. *Machine Learning* **39**(2-3) (2000) 135–168
19. Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: Proc. of ECML'98. (1998) 137–142
20. Scheffer, T., Wrobel, S.: Text Classification Beyond the Bag-of-Words Representation. In: Proc. of ICML'02: Workshop on Text Learning. (2002) 28–35
21. Rennie, J., et al.: Tackling the Poor Assumptions of Naive Bayes Text Classifiers. In: Proc. of ICML'03. (2003) 616–623
22. Pavlov, D., et al.: Document Preprocessing For Naive Bayes Classification and Clustering with Mixture of Multinomials. In: Proc. of KDD'04. (2004) 829–834
23. Peng, F., et al.: Augmenting Naive Bayes classifiers with statistical language models. *Information Retrieval* **7**(3) (2004) 317–345
24. Duda, R.O., Hart, P.E.: *Pattern Classification and Scene Analysis*. Wiley (1973)
25. Franz, M., McCarley, J.S., Roukos, S.: Ad-hoc and multilingual information retrieval at ibm. In: Proc. of the TREC-7 Conference. (1998) 157–168
26. Kraaij, W., Nie, J.Y., Simard, M.: Embedding web-based statistical translation models in cross-language information retrieval. *Computational Linguistics* **29**(3) (2003) 381–419
27. Fuhr, N.: Probabilistic models in information retrieval. *The Computer Journal* **35**(3) (1992) 243–255
28. Rijsbergen, C.J.V.: *Information Retrieval*, 2nd edition. Dept. of Computer Science, University of Glasgow (1979)
29. Kraaij, W., Nie, J.Y., Simard, M.: Embedding web-based statistical translation models in cross-language information retrieval. *Computational Linguistics* **29**(3) (2003) 381–419
30. Si, A., Leong, H.V., Lau, R.W.H.: Check: a document plagiarism detection system. In: Proc. of the 1997 ACM Symposium on Applied Computing, ACM (1997) 70–77
31. Stein, B., Meyer zu Eissen, S.: Intrinsic plagiarism analysis with meta learning. In: Proc. of PAN-07. (2007) 45–50
32. Poulliquen, B., Steinberger, R., Ignat, C.: Automatic annotation of multilingual text collections with a conceptual thesaurus. In: Proc. of EUROLAN'03. (2003)
33. Potthast, M., Stein, B., Anderka, M.: A wikipedia-based multilingual retrieval model. In: Proc. of ECIR'08. Volume 4956 of Lecture Notes in Computer Science., Springer-Verlag (2008) 522–530
34. Tiedemann, J., Nygaard, L.: The opus corpus - parallel & free. In: Proc. of LREC'04, Lisbon, Portugal (2004) 1183–1186
35. Sigurbjörnsson, B., Kamps, J., de Rijke, M.: Eurogov: Engineering a multilingual web corpus. In: Proc. of WebCLEF'06. Volume 4022 of Lecture Notes in Computer Science., Springer-Verlag (2006) 825–836
36. Sigurbjörnsson, B., Kamps, J., de Rijke, M.: Overview of WebCLEF 2005. In: Proc. of WebCLEF'06. Volume 4022 of Lecture Notes in Computer Science., Springer-Verlag (2006) 810–824
37. Voorhees, E.: The TREC-8 question answering track report. In: Proc. of the 8th Text Retrieval Conference. (1999) 77–82