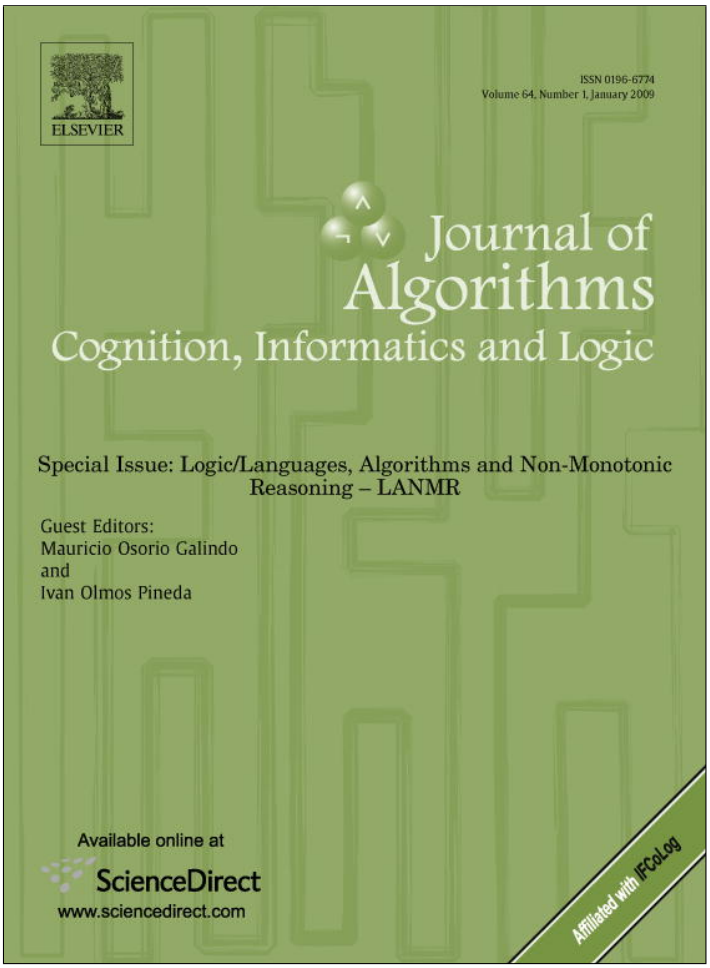

J. Algorithms 64 (2009) 51–60

Contents lists available at ScienceDirect

Journal of Algorithms
Cognition, Informatics and Logic

www.elsevier.com/locate/jalgor

A statistical approach to crosslingual natural language tasks

David Pinto [a,b,*], Jorge Civera [b], Alberto Barrón-Cedeño [b], Alfons Juan [b], Paolo Rosso [b]

[a] Facultad de Ciencias de la Computación, Benemérita Universidad Autónoma de Puebla, Mexico
[b] Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, Spain

ARTICLE INFO

Article history:
Received 13 February 2009
Available online 20 February 2009

Keywords:
Natural language processing
IBM translation models
Crosslingual data
Text classification
Information retrieval
Plagiarism analysis

ABSTRACT

The existence of huge volumes of documents written in multiple languages on Internet leads to investigate novel algorithmic approaches to deal with information of this kind. However, most crosslingual natural language processing (NLP) tasks consider a decoupled approach in which monolingual NLP techniques are applied along with an independent translation process. This two-step approach is too sensitive to translation errors, and in general to the accumulative effect of errors. To solve this problem, we propose to use a direct probabilistic crosslingual NLP system which integrates both steps, translation and the specific NLP task, into a single one. In order to perform this integrated approach to crosslingual tasks, we propose to use the statistical IBM 1 word alignment model (M1). The M1 model may show a non-monotonic behaviour when aligning words from a sentence in a source language to words from another sentence in a different, target language. This is the case of languages with different word order. In English, for instance, adjectives appear before nouns, whereas in Spanish it is exactly the opposite. The successful experimental results reported in three different tasks – text classification, information retrieval and plagiarism analysis – highlight the benefits of the statistical integrated approach proposed in this work.

1. Introduction

The fast growth of the Internet and the increasing presence of multilinguality on the web pose 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)[1] and the Text Analysis Conference (TAC).[2]

It is easy to find examples of NLP tasks in which more than one language is involved. In this paper, we focus on three specific multilingual tasks:

Multilingual text classification The proliferation and classification of multilingual documentation have become a common phenomenon in many official institutions and private companies. A good example is the EU parliament and commission, in which most official documents are translated into more than 20 languages, and classified according to the Eurovoc thesaurus [1]. Clearly, we could take advantage of redundancy and word correlation across languages

* Corresponding author at: Facultad de Ciencias de la Computación, Benemérita Universidad Autónoma de Puebla, Mexico.
E-mail addresses: dpinto@cs.buap.mx (D. Pinto), jcivera@dsic.upv.es (J. Civera), lbarron@dsic.upv.es (A. Barrón-Cedeño), ajuan@dsic.upv.es (A. Juan), prosso@dsic.upv.es (P. Rosso).

[1] http://www.clef-campaign.org/.
[2] http://www.nist.gov/tac/.

0196-6774/$ – see front matter
doi:10.1016/j.jalgor.2009.02.005

to improve the accuracy of text classifiers. Here we will focus on the particular case of *bilingual* text classification. The general multilingual case is left as future work.

Crosslingual information retrieval Conventional, monolingual information retrieval is limited to retrieve documents written in the language of the user query. However, in crosslingual information retrieval, the query can *cross* different languages, i.e., the user can retrieve documents written in languages other than the one of her/his query. This can be useful to simultaneously search relevant documents in many languages from a single, monolingual query. Also, it might be the case of a user who has some comprehension ability for a given language, but he is not sufficiently proficient to confidently specify a search query in that language. A search engine that may deal with this crosslingual problem would be of high benefit. In this work, we only consider the basic case, that is, the user query is expressed in a given *source* language and all the documents are written in a single, but different, *target* language.

Crosslingual plagiarism analysis Plagiarism (of text) is the use of original texts without providing appropriate citations. As in information retrieval, crosslingual plagiarism analysis can be seen as an extension of the monolingual problem in which the search is allowed to *cross* languages. More precisely, given a suspicious text in a certain (source) language, we are interested in checking if it has been obtained by plagiarism from a collection of original texts written in another (target) language; and, if so, we would also like to retrieve the particular plagiarised text. This problem is of high interest in the academia and journalism, and also in the Internet, where it is even easier to copy and translate original text data without adequate references.

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 tasks, we propose to use the IBM alignment model 1 (M1), which was firstly introduced for statistical Machine Translation (MT) [2]. One of the advantages of the statistical approach to MT is the possibility of applying these methods to any pair of languages whatever their alphabets are. The most remarkable examples of this fact are the NIST Open MT evaluations[3] in which Chinese and Arabic documents, among others, are translated into English.

The M1 model, the first of the IBM models, is basically defined as a statistical bilingual dictionary that captures word correlations across languages. In statistical MT, the M1 model has traditionally been an important component part in tasks 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 the lexical phrase scoring in phrase-based systems [8]. In MT as well as in applications such as bilingual text classification, crosslingual information retrieval and plagiarism analysis, the alignment of bilingual sentences may show a non-monotonic behaviour [9,10]. This is not the case, for instance, of speech recognition where there is a strict monotonic behaviour between the sequence of acoustic vectors and the sequence of recognised words or phonemes. Therefore, in our case the non-monotonic behaviour of the word alignment process makes the three crosslingual tasks enumerated above even harder.

On the other hand, we have also found that the M1 model can be directly applied to NLP tasks other than statistical MT and, in particular, to the three specific multilingual tasks described above. Indeed, we have already obtained comparatively good results in each of them separately [11–13]. Thus, it becomes clear that the M1 model is not just a statistical MT model, but a more general formalism by which many crosslingual NLP tasks can be approached.

In this work, we first discuss the M1 model as such general formalism in Section 2. Then, each of the three crosslingual NLP tasks previously discussed is reviewed in a separate subsection of Section 3. In each of these subsections, we first review related work and then we show how to apply the M1 model. In Section 4, new extended empirical results are reported to assess the usefulness of the proposed formalism, and to confirm the comparatively good results already published. Finally, conclusions and further work are discussed in Section 5.

2. The M1 model

2.1. The model

Let $x = x_1^J \equiv x_1 \dots x_j \dots x_J$ be a sentence, of known length J, in a certain source language over a given vocabulary $\mathcal{X}$. Similarly, let $y = y_1^I \equiv y_1 \dots y_i \dots y_I$ be a sentence, of length I, which is a translation of x into a target language over another vocabulary $\mathcal{Y}$.

[3] http://www.nist.gov/speech/tests/mt.

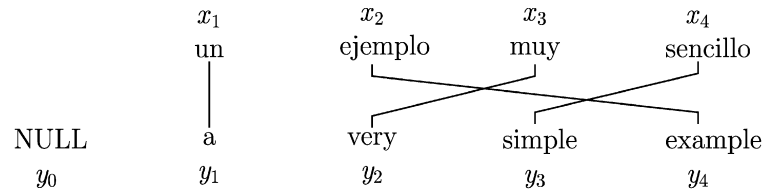

Fig. 1. Example of alignment between a source sentence of length $J = 4$, $x =$"`un ejemplo muy sencillo`", and a target sentence of equal length, $I = 4$, $y =$"`a very simple example`", in which the artificial NULL word at position 0 has been explicitly represented. The alignment variable in this example is $a = a_1 a_2 a_3 a_4$ with $a_1 = 1$, $a_2 = 4$, $a_3 = 2$ and $a_4 = 3$.

As discussed in [2], each IBM model and, in particular, the M1 model, gives a prescription for computing the probability of x to be translated into a *given* y; that is, the conditional probability $p(x|y)$. Following [2], to derive the M1 model, we first introduce the idea of *alignment* between x and y, as a J-dimensional (vector) variable $a = a_1^J \equiv a_1 \cdots a_j \cdots a_J$ in which, for each source position j, a_j indicates the target position to which it is connected. Thus, the alignment variable connects each source word to exactly one target word $a_j \in \{0, 1, \ldots, i, \ldots, I\}$, including the NULL[4] word at position 0. An example of alignment is shown in Fig. 1.

Let $\mathcal{A}(x, y)$ be the set of all possible alignments from x to y. In order to compute $p(x|y)$, we consider each possible alignment $a \in \mathcal{A}(x, y)$ as follows

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

where $p(x, a|y)$ can be understood as the probability of x to be translated into a *given* y, *in accordance with* the alignment a. In practice, the alignment information a is not provided together with x and y and, therefore, we revert to sum over each possible alignment in order to compute $p(x|y)$ as a marginalisation of $p(x, a|y)$.

Now, we proceed to factorise the term $p(x, a|y)$ at the word-level from left to right

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

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

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

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

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

where $p(x_j | y_{a_j})$ is a statistical bilingual dictionary. Thus, we can rewrite Eq. (2) under the assumptions presented in Eqs. (3) and (4) as[5]

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

where the parameter vector

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

is a statistical bilingual dictionary.

[4] The NULL word represents the target word to which those source words with no direct translation are connected.

[5] To show the dependence of $p(x, a|y)$ on $\boldsymbol{\Theta}$ (see Eq. (6)) explicitly, we write $p(x, a|y)$ as $p(x, a|y; \boldsymbol{\Theta})$. Some authors prefer to write $p(x, a|y, \boldsymbol{\Theta})$ but, strictly speaking, the notation $p(x, a|y, \boldsymbol{\Theta})$ would imply that $\boldsymbol{\Theta}$ is a random variable and this is not the case.

2.1.1. The model using indicator vectors

Now we change the nature of the original alignment variable $a_j \in \{0, \ldots, I\}$ from an integer value into an indicator vector

$$\boldsymbol{a}_j = (a_{j0}, a_{j1}, \ldots, a_{jI})^t \tag{7}$$

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

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

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

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

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

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

where

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

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 $\boldsymbol{\Theta}$ for the M1 model with respect to a set of N independent bilingual samples $(X, Y) = ((x_1, y_1), \ldots, (x_n, y_n), \ldots, (x_N, y_N))^t$. We denote the sequence of source and target words of the nth sample as $x_n = (x_{n1}, \ldots, x_{nj}, \ldots, x_{nJ_n})$ and $y_n = (y_{n1}, \ldots, y_{ni}, \ldots, y_{nI_n})$, respectively.

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

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

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

$$A = (a_1, \ldots, a_n, \ldots, a_N)^t. \tag{13}$$

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 $\boldsymbol{\Theta}$ 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 $\boldsymbol{\Theta}$ in two iterative steps. First, the so-called E-step computes the expected value of the missing data which, in our case, is 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 $\boldsymbol{\Theta}$; that is, an estimation of the bilingual dictionary in the case of the M1 model. This two-step process is repeated to refine the estimation of the missing data, improving 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 $\boldsymbol{\Theta}$ at iteration k, $\boldsymbol{\Theta}^{(k)}$. Given that the alignment variables in A are independent from each other, we can compute the E step,

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

with

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

That is, the expectation of word x_{nj} to be aligned with 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 updated formula for the M1 model,

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

where

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

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

3. Crosslingual tasks based on the M1 model

As discussed in the introduction, this section includes a separate subsection for each of the three crosslingual NLP tasks considered in this work. In each subsection, we first review related work and then we show how to apply the M1 model to the specific task considered.

3.1. Bilingual text classification

The purpose of text classification 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 [14].

Among the diverse approaches to text classification, the well-known naive Bayes classifier [15,16] is one of the most popular. Being so, there have been several instantiations and generalisations of this classifier, from Bernoulli mixtures [17] to multinomial mixtures [18,19]. 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) [20] and *support vector machines* (SVM) [21]. 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 [22–25].

Bilingual text classification is a novel task 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 text classification. 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.

3.1.1. 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 [26], this can be achieved by classifying the bilingual document (x, y) in the class with maximum posterior probability

$$c(x,y) = \arg\max_{c=1,\ldots,C} p(c|x,y) = \arg\max_{c=1,\ldots,C} p(c)p(x,y|c) \tag{18}$$

where

$$p(x,y|c) = p(y|c)p(x|y,c) \tag{19}$$

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

$$p(x,y|c) \approx p(x|c) \tag{20}$$

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

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

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

$$p(x, y|c) \approx p(x|c)p(y|c) \tag{22}$$

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, the usual approach consists in first translating the query into the target language and then retrieving documents in this language by using a conventional (monolingual) system. The translation system may be of any type, rule-based, statistical or hybrid. In [27,28], a statistical MT system is used, but it had to be previously trained with parallel texts. See [29,30] for a survey on crosslingual information retrieval.

As said in the introduction, this two-step approach is too sensitive to translation errors. That is, even if one system performs well in a monolingual environment, it might not be so good in the multilingual case due to translation errors in the query. To circumvent this difficulty, our basic idea is to integrate translation and retrieval into a single model by which documents in the target language are directly associated with queries in a different, source language. This basic idea can be easily instantiated by using the M1 model, as shown below.

3.2.1. The M1 model in crosslingual information retrieval

Lex x be a query text in the source language, and let $y_1, y_2, \ldots, y_n$ be a collection of n web pages in the target language. Given a number $k < n$, we are interested in finding a set of k most relevant web pages to x,

$$S_k(x) = \underset{\substack{S\subset\{y_1,\ldots,y_n\}\\ |S|=k}}{\arg\max} \; \underset{y\in S}{\arg\min}\, p(y|x) \tag{23}$$

where $p(y|x)$ denotes a probabilistic model of relevance of y to x.

In this work, $p(y|x)$ is approximated using the M1 model. It is worth noting that, as widely accepted in information retrieval, this model explicitly assumes that all word orderings in the query are equally probable. This makes the M1 model a better choice for information retrieval than other (superior) IBM models which, on the other hand, are far more complex and difficult to estimate.

3.3. Crosslingual plagiarism analysis

Whereas some research works have been carried out for conventional (monolingual) plagiarism analysis [31,32], to our knowledge, *crosslingual* plagiarism analysis is a NLP task that nearly has been studied in the literature. In [33], an automatic method is proposed 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 identifier) 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 [34], the authors propose a preliminary method based on semantic analysis in order to identify documents that may be plagiarised in a different language.

3.3.1. The M1 model in crosslingual plagiarism analysis

Let x be a suspicious text in a certain (source) language. As discussed in the introduction, we are interested in checking if it has been obtained by plagiarism from a collection $y_1, y_2, \ldots, y_n$ of original texts written in another (target) language; and, if so, we would also like to retrieve the particular plagiarised text. This can be formally stated by first defining a

probabilistic measure for the original text y to be plagiarised by x, $p(y|x)$. Deciding between a plagiarism case or not can be expressed in terms of the following two-class decision rule:

$$c(x) = \begin{cases} \text{plagiarism} & \text{if } p(y(x)|x) > \lambda, \\ \text{non-plagiarism} & \text{otherwise} \end{cases} \tag{24}$$

where $y(x)$ is the original text most probably plagiarised by x,

$$y(x) = \underset{y \in \{y_1, \ldots, y_n\}}{\arg\max}\ p(y|x) \tag{25}$$

and λ is a decision threshold that has to be empirically adjusted.

Again, as in previous crosslingual NLP tasks, $p(y|x)$ is approximated using the M1 model.

4. Experimental results

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

- OpenOffice.org documentation.[7]
- KDE manuals including KDE system messages.[8]
- PHP manuals.[9]
- 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.

For experimental purposes, these corpora were partitioned into three sets, devoting 80% for training, 5% for development and 15% for 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 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 is superior to the naive classifier, and this can be only explained by the fact that it takes advantage of crosslingual word correlations captured by the M1 model. This proves the benefits of the M1 model to improve the accuracy of bilingual text classifiers. These results are consistent with those presented in [11].

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

[6] http://urd.let.rug.nl/tiedeman/OPUS/.
[7] OpenOffice.org is an open source office suite.
[8] The K Desktop Environment (KDE) is a free graphical desktop environment.
[9] Hypertext Preprocessor (PHP) is a widely-used general purpose scripting language.

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 ± 0.7	6.4 ± 0.6	$\mathbf{5.5} \pm 0.4$
OPUS	11.2 ± 0.4	9.6 ± 0.3	$\mathbf{6.8} \pm 0.4$

Table 3
Comparison results over 134 English topics.

Run name	MRR	Average Success At (ASA)		
		10	20	50
CLIR model	0.096	0.142	0.216	0.440
1st Place (WebCLEF)	0.093	0.119	0.134	0.209
2nd Place (WebCLEF)	0.084	0.112	0.142	0.216

Table 4
Overview of the corpus for crosslingual plagiarism analysis. There are 5 original texts from a single author of the information retrieval area. Source text counts denote: N_1) plagiarisms obtained manually; N_2) automatic machine translations; N_3) non-plagiarised versions; N_4) independent (non-plagiarised) texts.

Original texts in the target language			Source text counts			
n	l_n	y_n (abbreviated)	N_1	N_2	N_3	N_4
1	56	Plagiarism analysis is …in writing style.	5	5	5	–
2	28	Plagiarism is the …adequate acknowledgement.	5	5	5	–
3	41	A cluster algorithm …intra-cluster similarity.	5	5	5	–
4	31	Near-duplicate detection is …retrieval precision.	5	5	5	–
5	20	Intrinsic plagiarism …extraneous sources.	5	5	5	–
Totals			25	25	25	20

4.2. Crosslingual information retrieval

The task presented in Section 3.2.1 was 10-fold cross-validated on the EuroGOV corpus [36]. In the training process we used its 134 supervised English queries. The results obtained were compared against the three best results reported at the bilingual "English to Spanish" subtrack of WebCLEF 2005.[10] A complete explanation of the systems/runs evaluated at WebCLEF 2005 may be found in [37]. 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 [38].

In Table 3 it is presented the name of each run together with its MRR. The Average Success At (ASA) different number of documents retrieved (10, 20 and 50) is also shown in this Table. 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 analysis

As discussed in Section 3.3, crosslingual plagiarism analysis is a NLP task that has been nearly studied in the literature, and thus no standard, publicly available corpora exist for empirical assessment. This lack of data led us to build a small yet useful corpus of crosslingual plagiarism in which original texts are in English and plagiarised texts are in Italian. An overview of this corpus is shown in Table 4. There are 5, original texts from a single author of the information retrieval area. For each of them, 10 plagiarised versions were obtained: 5 manually and 5 automatically. The former were obtained by simulating a plagiarised case of each original text, while the latter were generated using 5 different on-line translators. Also, from each original text, 5 non-plagiarised versions were manually derived by writing some new text about the same topic. Finally, 20 independent (non-plagiarised) text fragments in Italian, about the plagiarism topic, were added to the corpus. Summarising, the Italian part of the corpus contains 95 texts: 50 plagiarised and 45 non-plagiarised. There is also a Spanish version of the corpus which has been obtained by following the same procedure described before for Italian.

The crosslingual plagiarism corpus so built was partitioned into training and test sets for evaluation purposes. Plagiarised texts were divided into 40 texts for training and 10 for testing, whereas the 45 non-plagiarised texts were always included

[10] http://www.clef-campaign.org/.

Table 5
Classification errors for plagiarism identification in Italian and Spanish.

	Italian	Spanish
plagiarism identification	13.4 ± 6.2	4.9 ± 3.9

in the test set. This partitioning process was randomly carried out 30 times to compute confidence intervals for classification errors in the test set.

In accordance with the formal approach described in Section 3.3.1, the experimental results are presented at two levels. First, we have to decide whether a suspicious text is plagiarised or not. The solution to this problem is given by the decision rule stated in Eq. (24). We refer to this problem as *plagiarism identification*. Secondly, once we have identified a suspicious text as a case of plagiarism, we would like to know which original text has been the source of plagiarism. This information is provided by Eq. (25). This second problem is referred to as *source identification*.

Table 5 shows the number of classification errors for the problem of plagiarism identification in Italian and Spanish. Most of the classification errors are due to the fact that there are non-plagiarised texts that were wrongly classified as being cases of plagiarism. However, for those suspicious texts correctly identified as plagiarism cases, there are no source identification errors. This optimal result about source identification would not surely apply to collections of plagiarised texts larger than that considered here.

5. Conclusions and future work

In this work, we have proposed the M1 model as a general formalism by which many crosslingual NLP tasks can be approached. In particular, after a brief review of the M1 model, we have shown how it can be applied to the task of bilingual text classification, crosslingual information retrieval and crosslingual plagiarism analysis tasks. New results has been reported on these tasks, from which the usefulness of the proposed formalism is clearly confirmed.

As a future work, we plan to apply the M1 model to more challenging tasks in bilingual text classification such as the JRC-Acquis corpus [39]. Moreover, the extension of the bilingual text classifier to the multilingual case is yet another appealing idea that we plan to study.

In the case of plagiarism analysis, we plan to validate the obtained results on a larger corpus. At present we are working on the compilation of a cross-lingual plagiarism corpus with the required characteristics.

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 EC (FEDER) and the Spanish government under the MIPRCV "Consolider Ingenio 2010" research programme (CSD2007-00018), the research projects iTransDoc (TIN2006-15694-CO2-01) and MiDEs (TIN2006-15265-C06-04), and the FPU fellowship AP2003-0342. It has been also supported by the BUAP-701 PROMEP/103.5/05/1536 and CONACyT 192021/302009 grants.

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, http://europa.eu.int/celex/eurovoc, 1995.
[2] P.F. Brown, et al., The mathematics of statistical machine translation: Parameter estimation, Comput. Linguist. 19 (2) (1993) 263–311.
[3] R. Moore, Fast and accurate sentence alignment of bilingual corpora, in: Proc. of AMTA'02, 2002, pp. 135–244.
[4] Y. Ding, D. Gildea, M. Palmer, An algorithm for word-level alignment of parallel dependency trees, in: Proc. of MT Summit IX, 2003, pp. 95–101.
[5] F. Nevado, F. Casacuberta, E. Vidal, Parallel corpora segmentation using anchor words, in: Proc. of EAMT/CLAW'03, 2003, pp. 33–40.
[6] D. Munteanu, A. Fraser, D. Marcu, Improved machine translation performance via parallel sentence extraction from comparable corpora, in: Proc. of HLT-NAACL'04, 2004, pp. 265–272.
[7] N. Ueffing, H. Ney, Word-level confidence estimation for machine translation, Comput. Linguist. 33 (1) (2007) 9–40.
[8] P. Koehn, F. Och, D. Marcu, Statistical phrase-based translation, in: Proc. of NAACL'03, 2003, pp. 48–54.
[9] C. Tillmann, H. Ney, Word reordering and a dynamic programming beam search algorithm for statistical machine translation, Comput. Linguist. 29 (1) (2003) 97–133.
[10] F.J. Och, H. Ney, A systematic comparison of various statistical alignment models, Comput. Linguist. 29 (1) (2003) 19–51.
[11] J. Civera, A. Juan, Unigram-IBM model 1 mixtures for bilingual text classification, in: Proc. of LREC'08, 2008.
[12] D. Pinto, A. Juan, P. Rosso, Using query-relevant documents pairs for cross-lingual information retrieval, in: Proc. of TSD'07, 2007, pp. 630–637.
[13] A. Barrón-Cedeño, P. Rosso, D. Pinto, A. Juan, On cross-lingual plagiarism analysis using a statistical model, in: Proc. of PAN'08, 2008, pp. 9–13.
[14] F. Sebastiani, Classification of text, automatic, in: K. Brown (Ed.), The Encyclopedia of Language and Linguistics, vol. 2, 2nd edition, Elsevier Science Publishers, Amsterdam, NL, 2006, pp. 457–463.
[15] D.D. Lewis, Naive Bayes at forty: The independence assumption in information retrieval, in: Proc. of ECML'98, 1998, pp. 4–15.
[16] A. McCallum, K. Nigam, A comparison of event models for naive Bayes text classification, in: Proc. of AAAI/ICML-98: Workshop on Learning for Text Categorization, 1998, pp. 41–48.
[17] A. Juan, E. Vidal, On the use of Bernoulli mixture models for text classification, Pattern Recognition 35 (12) (2002) 2705–2710.
[18] K. Nigam, et al., Text classification from labeled and unlabeled documents using EM, Mach. Learn. 39 (2/3) (2000) 103–134.
[19] J. Novovicová, A. Malík, Application of multinomial mixture model to text classification, in: Proc. of IbPRIA 2003, in: Lecture Notes in Comput. Sci., vol. 2652, Springer-Verlag, 2003, pp. 646–653.

[20] R.E. Schapire, Y. Singer, Boostexter: A boosting-based system for text categorization, Mach. Learn. 39 (2–3) (2000) 135–168.
[21] T. Joachims, Text categorization with support vector machines: Learning with many relevant features, in: Proc. of ECML'98, 1998, pp. 137–142.
[22] T. Scheffer, S. Wrobel, Text classification beyond the bag-of-words representation, in: Proc. of ICML'02: Workshop on Text Learning, 2002, pp. 28–35.
[23] J. Rennie, et al., Tackling the poor assumptions of naive Bayes text classifiers, in: Proc. of ICML'03, 2003, pp. 616–623.
[24] D. Pavlov, et al., Document preprocessing for naive Bayes classification and clustering with mixture of multinomials, in: Proc. of KDD'04, 2004, pp. 829–834.
[25] F. Peng, et al., Augmenting naive Bayes classifiers with statistical language models, Inf. Retr. 7 (3) (2004) 317–345.
[26] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, Wiley, 1973.
[27] M. Franz, J.S. McCarley, S. Roukos, Ad-hoc and multilingual information retrieval at IBM, in: Proc. of the TREC-7 Conference, 1998, pp. 157–168.
[28] W. Kraaij, J.Y. Nie, M. Simard, Embedding web-based statistical translation models in cross-language information retrieval, Comput. Linguist. 29 (3) (2003) 381–419.
[29] N. Fuhr, Probabilistic models in information retrieval, Comput. J. 35 (3) (1992) 243–255.
[30] C.J.V. Rijsbergen, Information Retrieval, 2nd edition, Dept. of Computer Science, University of Glasgow, 1979.
[31] A. Si, H.V. Leong, R.W.H. Lau, Check: A document plagiarism detection system, in: Proc. of the 1997 ACM Symposium on Applied Computing, ACM, 1997, pp. 70–77.
[32] B. Stein, S. Meyer zu Eissen, Intrinsic plagiarism analysis with meta learning, in: Proc. of PAN'07, 2007, pp. 45–50.
[33] B. Pouliquen, R. Steinberger, C. Ignat, Automatic annotation of multilingual text collections with a conceptual thesaurus, in: Proc. of EUROLAN'03, 2003.
[34] M. Potthast, B. Stein, M. Anderka, A wikipedia-based multilingual retrieval model, in: Proc. of ECIR'08, in: Lecture Notes in Comput. Sci., vol. 4956, Springer-Verlag, 2008, pp. 522–530.
[35] J. Tiedemann, L. Nygaard, The opus corpus – parallel & free, in: Proc. of LREC'04, Lisbon, Portugal, 2004, pp. 1183–1186.
[36] B. Sigurbjörnsson, J. Kamps, M. de Rijke, Eurogov: Engineering a multilingual web corpus, in: Proc. of WebCLEF'06, in: Lecture Notes in Comput. Sci., vol. 4022, Springer-Verlag, 2006, pp. 825–836.
[37] B. Sigurbjörnsson, J. Kamps, M. de Rijke, Overview of WebCLEF 2005, in: Proc. of WebCLEF'06, in: Lecture Notes in Comput. Sci., vol. 4022, Springer-Verlag, 2006, pp. 810–824.
[38] E. Voorhees, The TREC-8 question answering track report, in: Proc. of the 8th Text Retrieval Conference, 1999, pp. 77–82.
[39] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, A. Ceausu, D. Varga, The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages, in: Proc. of LREC'06, 2006.