On Cross-lingual Plagiarism Analysis using a Statistical Model

Alberto Barrón-Cedeño[1] and **Paolo Rosso**[1] and **David Pinto**[1,2] and **Alfons Juan**[3]

Abstract. The automatic detection of plagiarism is a task that has acquired relevance in the Information Retrieval area and it becomes more complex when the plagiarism is made in a multilingual panorama, where the original and suspicious texts are written in different languages. From a cross-lingual perspective, a text fragment in one language is considered a plagiarism of a text in another language if their contents are considered semantically similar no matter they are written in different languages and the corresponding citation or credit is not included.

Our current experiments on cross-lingual plagiarism analysis are based on the exploitation of a statistical bilingual dictionary. This dictionary is created on the basis of a parallel corpus which contains original fragments written in one language and plagiarised versions of these fragments written in another language.

The process for the automatic cross-lingual plagiarism analysis based on the statistical bilingual dictionary has shown good results and we consider that it could be useful also for the cross-lingual near-duplicate detection task.

1 INTRODUCTION

Nowadays people enjoy an easy access to a wide range of information in multiple languages via the World Wide Web. Unfortunately, this "free access" to the information has caused a big temptation: the plagiarism, also from one language to another one. In some way, cross-lingual plagiarism analysis is related to cross-lingual information retrieval [6, 4]. In fact, the aim is to retrieve those fragments that have been plagiarised in a language with respect to the one originally employed.

In this paper we present an approach to the task of cross-lingual plagiarism analysis based on the exploitation of a statistical bilingual dictionary, commonly used in the automatic machine translation and cross-language infomation retrieval tasks [1, 4, 6].

The rest of the paper is organised as follows. Section 2 describes some of the current work in the task of cross-lingual plagiarism analysis. Section 3 introduces the definition and estimation process of the model used in order to create the statistical bilingual dictionary. Section 4 gives a description of the preliminary experiments carried out. Finally, Section 5 includes discussion and future work.

[1] Natural Language Engineering Lab., Dpto. Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, Spain, email: {lbarron, prosso, dpinto}@dsic.upv.es

[2] Faculty of Computer Science, B. Autonomous University of Puebla, Mexico, email: dpinto@cs.buap.mx

[3] Pattern Recognition and Human Language Technology Group, Dpto. Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, Spain, email: ajuan@dsic.upv.es

2 PRELIMINARY APPROACH(ES) IN CROSS-LINGUAL PLAGIARISM ANALYSIS

Some efforts have been made in other research directions that could be useful for this task. There have been developed, for example, some methods for the automatic acquisition of translated web pages [9], based on the search of hyperlinks containing strings of the kind "Spanish version" in order to download all the language versions of a given page. Although these cases cannot be considered plagiarism, the method could be useful in order to retrieve some instances for the training phase when dealing with cross language plagiarism analysis.

In [8] it has been proposed a method based on a thesaurus. In order to search document translations they have used the Eurovoc Thesaurus[4] to decide whether a document is near to another one in a different language or not. As the authors point out, this approach could be useful in the plagiarism analysis, of course if a good thesaurus is available.

The automatic plagiarism analysis may be classified into two main approaches: one with a reference corpus [10, 3] and one without it, which is also known as intrinsic plagiarism analysis [5, 11]. In the first case, the idea is to compare fragments (x_i) of a suspicious document (D_s) with fragments y_j of documents in a reference corpus (C) which is composed by original documents, in order to find those similar fragments that could be considered plagiarised. In the other case the objective is the same, but the idea is to look for variations through the text of the suspicious document (D_s) -like syntax, grammatical categories use and content complexity- and do not exploit any reference corpus.

The state of the art in automatic plagiarism analysis allows to detect word by word plagiarism, even if fragments have been modified. However, to our knowledge, *Cross-Lingual Plagiarism Analysis* (CLiPA) nearly has been explored in the litherature.

The authors of [7] propose a method based on three main steps. Given a suspicious document d and a reference corpus C in a different language, the first step consists in retrieving a subset of candidate documents from C which could be sources of the plagiarised fragments of the document d. Then a semantic analysis is done among the sections of d and each $c_i \in C$. Finally, the similar sections are analysed in order to filter those cases where a proper citation has been made. Authors are currently working on the improvement of the analysis step.

3 THE STATISTICAL MODEL

In this section we describe the statistical model (Sub-section 3.1) and the Expectation Maximisation method for the estimation of the prob-

[4] http://europa.eu/eurovoc/

abilities of the bilingual dictionary (Sub-section 3.2). This bilingual dictionary is the core of the CLiPA system used in this research work.

3.1 Model Definition

Let $x_1, x_2, \cdots x_V$ be fragments conforming a suspicious text V in a certain language, and let $y_1, y_2, \cdots, y_W$ be a collection of W original fragments in a different language (the reference corpus). Let $\mathcal{X}$ and $\mathcal{Y}$ be their associated vocabularies, respectively.

Given the suspicious fragment $x_j \in V$, our objective is to find the most similar original fragment $y_k \in W$. The obtained relations, could be the original and plagiarised pairs. In order to do this, we have followed a probabilistic approach in which the most similar original fragment is computed as the most probable given x, i.e.,

$$y_i^*(x) = \underset{y=y_1,\cdots,y_W}{\operatorname{argmax}} \; p(y \,|\, x) \tag{1}$$

In this work, $p(y \,|\, x)$ is modelled by using the well-known IBM alignment model 1 (IBM-1) for statistical machine translation [1, 2]. This model assumes that each word in the reference segment y_k is *connected to exactly one word* in the suspicious fragment x_j. Also, it is assumed that y_k has an initial "null" word to link it to those words in x_j with no direct connexion.

Formally, a hidden variable $a = a_1 a_2 \cdots a_{|y|}$ is introduced in order to reveal, for each position i in y_j, the suspicious fragment word position $a_i \in \{0, 1, \ldots, |x|\}$ to which it is connected. Thus,

$$p(y \,|\, x) = \sum_{a \in \mathcal{A}(x,y)} p(y, a \,|\, x) \tag{2}$$

where $\mathcal{A}(x, y)$ denotes the set of all possible alignments between x and y. The *alignment-completed* probability $p(y, a \,|\, x)$ can be decomposed in terms of individual, y_k position-dependent probabilities as:

$$p(y, a \,|\, x) = \prod_{i=1}^{|y|} p(y_i, a_i \,|\, a_1^{i-1}, y_1^{i-1}, x) \tag{3}$$

$$= \prod_{i=1}^{|y|} p(a_i \,|\, a_1^{i-1}, y_1^{i-1}, x) p(y_i \,|\, a_1^{i}, y_1^{i-1}, x) \tag{4}$$

In the case of the IBM-1 model, it is assumed that a_i is uniformly distributed

$$p(a_i \,|\, a_1^{i-1}, y_1^{i-1}, x) = \frac{1}{|x| + 1} \tag{5}$$

and that y_i only depends on the query word to which it is connected

$$p(y_i \,|\, a_1^{i}, y_1^{i-1}, x) = p(y_i \,|\, x_{a_i}) \tag{6}$$

By substitution of (5) and (6) in (4); and thereafter (4) in (2), we may write the IBM-1 model as follows by some straightforward manipulations:

$$p(y \,|\, x) = \sum_{a \in \mathcal{A}(x,y)} \prod_{i=1}^{|y|} \frac{1}{(|x| + 1)} p(y_i \,|\, x_{a_i}) \tag{7}$$

$$= \frac{1}{(|x| + 1)^{|y|}} \prod_{i=1}^{|y|} \sum_{j=0}^{|x|} p(y_i \,|\, x_j) \tag{8}$$

Note that this model is governed only by a *statistical dictionary* $\mathbf{\Theta}$=$\{p(w|v)$, for all $v \in \mathcal{X}$ and $w \in \mathcal{Y}\}$. The model assumes that the order of the words in the suspicious fragment is not important. Therefore, each position in a original fragment is equally likely to be connected to each position in the suspicious one. Although this assumption is unrealistic in machine translation, we do *not* actually perform any translation and we consider that the IBM-1 model is well-suited for approaching cross-lingual plagiarism analysis.

3.2 Maximum Likelihood Estimation

It is not difficult to derive an Expectation-Maximisation (EM) algorithm to perform maximum likelihood estimation of the statistical dictionary with respect to a collection of training samples $(X, Y) = \{(x_1, y_1), \ldots, (x_N, y_N)\}$. The *(incomplete)* log-likelihood function is:

$$L(\mathbf{\Theta}) = \sum_{n=1}^{N} \log \sum_{a_n} p(y_n, a_n | x_n) \tag{9}$$

with

$$p(y_n, a_n | x_n) = \frac{1}{(|x_n| + 1)^{|y_n|}} \prod_{i=1}^{|y_n|} \prod_{j=0}^{|x_n|} p(y_{ni} \,|\, x_{nj})^{a_{nij}} \tag{10}$$

where, for convenience, the alignment variable, $a_{ni} \in \{0, 1, \ldots, |x_n|\}$, has been rewritten as an indicator vector, $a_{ni} = (a_{ni0}, \ldots, a_{ni|x_n|})$, with 1 in the suspicious fragment position to which it is connected, and zeros elsewhere.

The so-called *complete* version of the log-likelihood function (9) assumes that the hidden (missing) alignments $a_1, \ldots, a_N$ are also known:

$$\mathcal{L}(\mathbf{\Theta}) = \sum_{n=1}^{N} \log p(y_n, a_n | x_n) \tag{11}$$

An initial estimate for $\mathbf{\Theta}$, $\mathbf{\Theta}^{(0)}$, is required for the EM algorithm to start. This can be done by assuming that the translation probabilities are uniformly distributed; i.e.,

$$p(w \,|\, v)^{(0)} = \frac{1}{|\mathcal{Y}|} \; \forall v \in \mathcal{X}, w \in \mathcal{Y} \tag{12}$$

After this initialisation, the EM algorithm maximises (9) iteratively, through the application of two basic steps in each iteration: the E(xpectation) step and the M(aximisation) step. At iteration k, the E step computes the expected value of (11) given the observed (incomplete) data, (X, Y), and a current estimation of the parameters, $\mathbf{\Theta}^{(k)}$. This reduces to the computation of the expected value of a_{nij}:

$$a_{nij}^{(k)} = \frac{p(y_{ni} \,|\, x_{nj})^{(k)}}{\sum_{j'} p(y_{ni} \,|\, x_{nj'})^{(k)}} \tag{13}$$

Then, the M step finds a new estimate of $\mathbf{\Theta}$, $\mathbf{\Theta}^{(k+1)}$, by maximising (11), using (13) instead of the missing a_{nji}. This results in:

$$P(w|v)^{(k+1)} = \frac{\sum_n \sum_{i=1}^{|y_n|} \sum_{j=0}^{|x_n|} a_{nij}^{(k)} \, \delta(y_{ni}, w) \, \delta(x_{nj}, v)}{\sum_{w'} \sum_n \sum_{i=1}^{|y_n|} \sum_{j=0}^{|x_n|} a_{nij}^{(k)} \, \delta(y_{ni}, w') \, \delta(x_{nj}, v)} \tag{14}$$

$$= \frac{\sum_n \frac{p(w \mid v)^{(k)}}{\sum_{j'} p(w \mid x_{nj'})^{(k)}} \sum_{i=1}^{|y_n|} \sum_{j=0}^{|x_n|} \delta(y_{ni}, w)\, \delta(x_{nj}, v)}{\sum_{w'} \left[\sum_n \frac{p(w' \mid v)^{(k)}}{\sum_{j'} p(w' \mid x_{nj'})^{(k)}} \sum_{i=1}^{|y_n|} \sum_{j=0}^{|x_n|} \delta(y_{ni}, w')\, \delta(x_{nj}, v) \right]} \quad (15)$$

for all $v \in \mathcal{X}$ and $w \in \mathcal{Y}$; where $\delta(a, b)$ is the Kronecker delta function; i.e., $\delta(a, b) = 1$ if $a = b$; 0 otherwise.

4 PRELIMINARY EXPERIMENTS

We have carried out some preliminary experiments by selecting five document fragments from one author of the information retrieval area. The aim of this experiment was to obtain a personalised bilingual statistical dictionary which may be used to perform an author-focused CLiPA. The five original fragments $y_{\{1\cdots5\}}$ are the following:

- y_1 *Plagiarism analysis is a collective term for computed-based methods to identify a plagiarism offence. In connection with text documents we distinguish between corpus-based and intrinsic analysis: the former compares suspicious documents against a set of potential original documents, the latter identifies potentially plagiarised passages by analysing the suspicious document with respect to changes in writing style.*
- y_2 *Plagiarism is the practice of claiming, or implying, original authorship of someone elses written or creative work, in whole or in part, into ones own without adequate acknowledgement.*
- y_3 *A cluster algorithm takes a set D of objects as input and operationalizes a strategy to generate a clustering C. Informally stated, the overall objective of a cluster algorithm is to maximise the innercluster similarity and to minimise the intra-cluster similarity.*
- y_4 *Near-duplicate detection is mainly a problem of the World Wide Web: duplicate Web pages increase the index storage space of search engines, slow down result serving, and decrease the retrieval precision*
- y_5 *Intrinsic plagiarism analysis deals with the detection of plagiarised sections within a document d, without comparing d to extraneous sources*

For each original text fragment, we have constructed plagiarised cases based on two approaches: Machine Translation (MT) and Human Simulated (HS). In the former approach, we have used five popular on-line translators[5], whereas for the latter nine different people have "plagiarised" each original fragment written in English to fragments in Spanish.

In order to show the similarity between the plagiarised fragments based on the human process and on automatic machine translation, we show the Jaccard distance of the MT and HS pairs of plagiarised fragments in Tables 1 and 2, corresponding to the original fragments y_1 and y_3, respectively.

In both tables we can see that there is an important difference between MT and HS plagiarisms. Additionally, considering only one row, t_1 from Table 1 for example, we can see that there are significant differences between the HS plagiarisms simply considering the Jaccard distance with respect to t_1 or any of the other MT plagiarised fragments. The same behaviour fact can be appreciated for the MT plagiarisms fixing one HS column.

[5] Freetranslation ($www.freetranslation.com$)
Google ($www.google.com/language\_tools$)
Worldlingo ($www.worldlingo.com$)
Systran ($www.systransoft.com$)
Reverso ($www.reverso.net$)

Table 1. Jaccard distance (J_δ) for human plagiarised (h_i) and machine translation (t_j) fragments for y_1

	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9
t_1	0.50	0.58	0.58	0.50	0.52	0.74	0.58	0.44	0.37
t_2	0.54	0.59	0.48	0.45	0.52	0.73	0.52	0.41	0.41
t_3	0.51	0.60	0.60	0.54	0.51	0.75	0.61	0.45	0.43
t_4	0.56	0.64	0.60	0.56	0.54	0.76	0.63	0.48	0.43
t_5	0.67	0.74	0.73	0.70	0.66	0.78	0.67	0.65	0.63

Table 2. Jaccard distance (J_δ) for human plagiarised (h_i) and machine translation (t_j) fragments for y_3

	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9
t_1	0.47	0.70	0.59	0.55	0.48	0.69	0.56	0.49	0.53
t_2	0.46	0.64	0.56	0.51	0.53	0.68	0.57	0.48	0.55
t_3	0.49	0.72	0.58	0.54	0.56	0.71	0.60	0.54	0.58
t_4	0.49	0.69	0.56	0.51	0.56	0.71	0.60	0.51	0.55
t_5	0.64	0.81	0.67	0.66	0.62	0.84	0.70	0.66	0.61

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 Nine human simulated plagiarisms for each original fragment
iii Five automatic machine translations for each original fragment
iv Fourty six unplagiarised (independent) fragments about the plagiarism topic originally written in Spanish language

We have splitted the complete corpus into two datasets: training and test. The training dataset, which is used to construct the statistical bilingual dictionary, is made up of 50 pairs composed of original fragments and their corresponding plagiarised versions. The plagiarised versions were those obtained by 3 MT and 7 HS plagiarisms. In the test dataset, we employed 46 Unplagiarised Text Fragments (UTF) distributed as follows: 20 text fragments obtained by rewriting the same original concept, but mostly with other words (UTF_1), and 26 text fragments without any relation with the one of the original text fragments (UTF_2)[6].

In order to verify the similarity among the text fragments of the test dataset, we have represented each text by using the vector space model and, thereafter, we calculated the cosine of the similarity among them. We were particularly interested in observing the similarity arithmetic mean obtained only among the Plagiarised Text Fragments (PTF), of the same original fragment. This in order to confirm that those texts are similar enough and, at the same time, they are different enough to be considered as a challenge. Moreover, we also calculated the arithmetic mean of the similarity between the plagiarised vs. unplagiarised text fragments. The obtained results are shown in Table 3. We may observe that the plagiarised documents obtained an average of 0.44 which we consider to be good for the purposes of this preliminary investigation. The unplagiarised documents obtained instead, very low average of similarity.

We compute the probability $p(y|x)$ of each original text fragment given a suspicious one of the test subcorpus, on the basis of the statistical bilingual dictionary that we have obtained during the training phase (Section 3). The entire process of the experiment is illustrated in Figure 1.

We have conducted experiments in order to define the number of necessary iterations of the EM algorithm. We have calculated bilingual dictionaries with $k = \{10, 20, \cdots, 100\}$ where k is the number

[6] Our CLiPA corpus is freely available at www.dsic.upv.es/grupos/nle/downloads.html

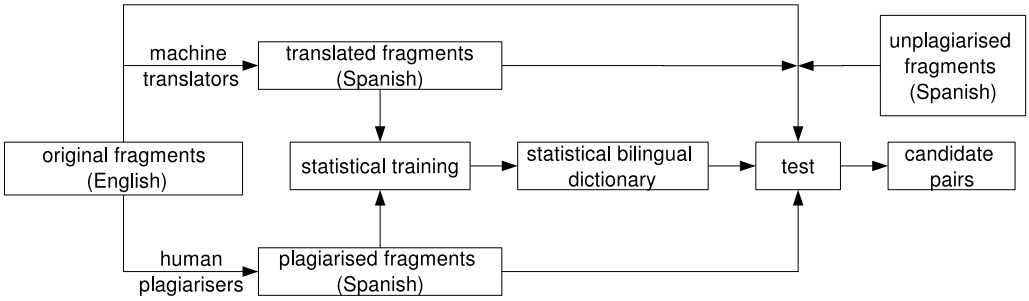

Figure 1. Experiment description (including training and test)

Table 3. Analysis of similarities for the test dataset

Fragment-Fragment	Similarity average	Minimum similarity	Maximum similarity
PTF-PTF	0.447	0.153	0.929
PTF-UTF_1	0.089	0.002	0.344
PTF-UTF_2	0.028	0.002	0.133

of iterations in the EM algorithm. We are only interested in defining a good number of iterations for the training phase. We consider the training phase good enough when the association probability between the original fragment y_i and its plagiarised one x_i is greater than the probability obtained with any other text fragment x_j, i. e., when it is fulfilled the following condition:

$$p(y_i|x_i) > p(y_j|x_i) \ \forall j \neq i \tag{16}$$

We have used a variation of the Precision measure: *Precision at 1* ($P@1$)[7]. Figure 2 shows the behaviour of $P@1$ for different EM iterations. In agreement with [6], we have considered a maximum number of 100 iterations in order to avoid over-training.

In Figure 2 we can observe that the $P@1$ value reaches a certain stability after 60 iterations. The preliminary results are interesting and they encourage to further continue in this research direction. However, the results need to be validated in the future on a bigger corpus.

In order to obtain a discrimination of the good candidates, we have tested different threshold values. Figure 3 shows the behaviour of the F-measure based on different thresholds. The curve in this figure must be analysed from right to left. In the highest values of the threshold, the F-measure is low due to the fact that Recall is near to zero. Meanwhile the threshold descends, more actual plagiarised fragments are considered and the F-measure is incremented. The best value is obtained when $Threshold = 4.31 * 10^{-9}$ where a good part of the real plagiarised fragments have a probability of being detected as plagiarised that is little higher than the threshold value. After this peak, both precision and F-measure decrease. This is the reason why we opted for using this value as threshold for which $P = 1$ and $R = 0.8$ (F-measure=0.88).

In order to clarify the obtained results, we consider the following three Spanish text fragments. The first two are examples of plagiarised fragments in Spanish of y_5 and y_1, respectively. The third one is an example of unrelated text.

[7] In P@1 only the best ranked item in the output is considered for the precision calculation.

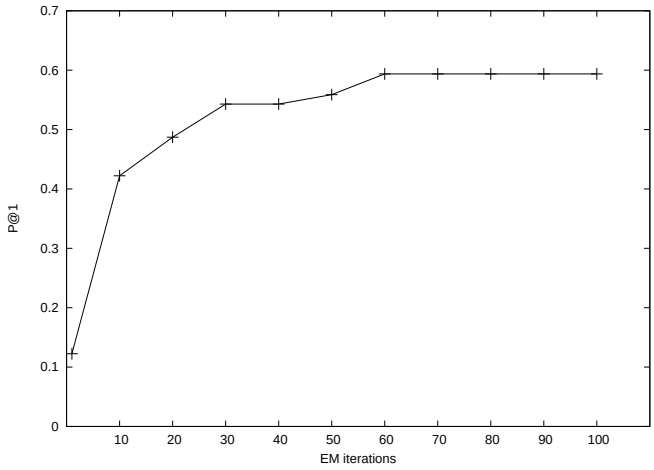

Figure 2. P@1 with different EM iterations

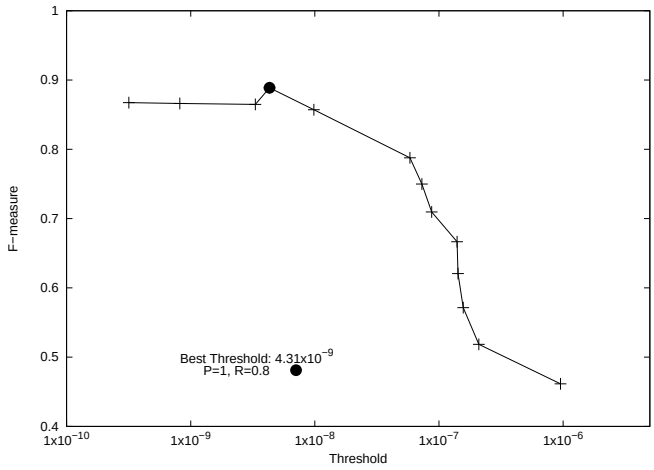

Figure 3. F-measure for different threshold values

x_1 *El análisis del plagio intrinseco tiene que ver con la detección de secciones plagiadas de un documento d, sin comparar d con fuentes externas*

x_2 *El analisis del plagio es un término colectivo para que los metodos computar-basados identifiquen una ofensa del plagio. Con respecto a documentos del texto distinguimos entre el análisis recopilación-basado e intrínseco: el anterior compara documentos sospechosos contra un sistema de documentos originales potenciales, el último identifica pasos potencialmente plagiados analizando el documento sospechoso con respecto a cambios en estilo de escritura.*

z_1 *Hipótesis La perplejidad de un fragmento perteneciente a un escritor con respecto a otro, será mayor que la de dos documentos escritos por el mismo autor. Aquellos párrafos que tengan mayor perplejidad sera los mejores candidatos a ser fragmentos plagiados.*

x_1 is one case of a HS plagiarism. One translation of this fragment could be *"Intrinsic plagiarism analysis has to do with the detection of plagiarised sections from a document d without comparing d to external sources*" and, obviously, is a plagiarism of y_5. In this case $p(y_5|x_1) = 33.1 \cdot 10^{-5}$ which exceeds the previously defined threshold and, therefore, x_1 is considered a plagiarism of y_5. x_2 has been generated from y_1 by using an on-line machine translator. In this case $p(y_1|x_2) = 10.28 \cdot 10^{-9}$ and, therefore, x_2 is considered a plagiarism of y_1.

With respect to z_1, $p(y_i|z_1) \approx 0$ and, therefore, z_1 is not considered to be a plagiarism of any original text fragment of the reference corpus. For instance, the following words *hipótesis*, *párrafos*, *perplejidad* and *mejores* (hypothesis, paragraphs, perplexity and best) do not have any relation with the English words in the reference corpus and, therefore, the association probability between them is close to zero.

5 CONCLUSIONS

In this paper we have approached the cross-lingual plagiarism analysis with a probabilistic method which calculates a bilingual statistical dictionary on the basis of the IBM-1 model. In order to generate the bilingual dictionary, we have used a set of original documents written in English and Spanish plagiarised examples. Our proposal calculates the probabilistic association between two terms in two different languages. The main contribution of this paper is that the probabilistic model is trained with a data set made of pairs of fragments of text from a particular author. The aim of our approach is to investigate the cross-lingual plagiarism with respect to a specific author.

The application of a statistical machine translation technique (without any translation), has demonstrated to be a valuable resource for the CLiPA task. Due to the fact that we determine the similarity between suspicious and original text fragments based on a dictionary, the order of the words in the fragment is not relevant and we are able to find good candidates even when the plagiarised text has been modified.

We believe that this approach is not only useful for the cross-lingual plagiarism analysis, but for the near-duplicate analysis too. As further work, we would like to validate the results we obtained in this preliminary experiment on a bigger corpus. Unfortunately, the construction of a cross-lingual corpus with the required characteristics, in size and quality, seems to be by itself a sufficiently difficult task which makes cross-language plagiarism analysis even more challenging.

References of the Original Text Fragments

y_1 Preface of the Proc. of the International Workshop on Plagiarism Analysis, Authorship Identification , and Near-Duplicate Detection (PAN 2007).

y_2 Introduction of the lecture on "Technology for Plagiarism Analysis" given by Benno Stein at the UPV in March of 2008.

y_3 B. Stein and M. Busch. 'Density-based Cluster Algorithms in Low-dimensional and High-dimensional Applications'. Stein and Meyer zu Eissen (eds.), 2nd Int. Workshop on Text-Based Information Retrieval (TIR 05), *Germany*, 45-56, (2005).

y_4 See reference of y_1.

y_5 B. Stein and S. Meyer zu Eissen. 'Intrinsic Plagiarism Analysis with Meta Learning'. SIGIR Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 07), *Netherlands*, 45-50, (2007).

ACKNOWLEDGEMENTS

We would like to thank the MCyT TIN2006-15265-C06-04 research project for partially funding this work as well as to the CONACyT-MEXICO 192021 grant and the BUAP-701 PROMEP/103.5/05/1536 grant.

REFERENCES

[1] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vicent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin, 'A statistical approach to machine translation', *Computational Linguistics*, **16**(2), 79–85, (1990).

[2] Jorge Civera and Alfons Juan, 'Mixtures of ibm model 2', *Proc. of the EAMT Conference*, 159–167, (2006).

[3] Parvati Iyer and Abhipsita Singh, 'Document similarity analysis for a plagiarism detection system', *2nd Indian Int. Conf. on Artificial Intelligence (IICAI-2005)*, 2534–2544, (2005).

[4] Wessel Kraaij, Jian-Yun Nie, and Michel Simard, 'Embedding web-based statistical translation models in cross-language information retrieval', *Computational Linguistics*, **29**(3), 381–419, (2003).

[5] Sven Meyer zu Eissen and Benno Stein, 'Intrinsic plagiarism detection', *Lalmas et. al. (Eds.): Advances in Information Retrieval Proc. of the 28th European Conf. on IR research, ECIR 2006, London*, 565–569, (2006).

[6] David Pinto, Alfons Juan, and Paolo Rosso, 'Using query-relevant documents pairs for cross-lingual information retrieval', *TSD 2007. Springer-Verlag*, **LNAI (4629)**, 630–637, (2007).

[7] Martin Potthast, Benno Stein, and Maik Anderka, 'A wikipedia-based multilingual retrieval model', *Macdonald, Ounis, Plachouras, Ruthven and White (eds.). 30th European Conf. on IR Research (ECIR 2008), Glasgow, Springer-Verlag*, **LNCS (4956)**, 522–530, (2008).

[8] Bruno Pouliquen, Ralf Steinberger, and Camelia Ignat, 'Automatic identification of document translations in large multilingual document collections', *Proc of the Int. Conf. Recent Advances in Natural Language Processing (RANLP-2003), Borovets, Bulgaria*, 401–408, (2003).

[9] Philip Resnik, 'Mining the web for bilingual text', *37th Annual Meeting of the Association for Computational Linguistics (ACL 99), Maryland*, (1999).

[10] Antonio Si, Hong Va Leong, and Rynson W. H. Lau, 'Check: a document plagiarism detection system', *Proc. of the 1997 ACM Symposium on Applied Computing, San Jose, California, United States*, 70–77, (1997).

[11] Benno Stein and Sven Meyer zu Eissen, 'Intrinsic plagiarism analysis with meta learning', *B. Stein, M. Koppel, and E. Stamatatos, Eds., SIGIR Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 07), Amsterdam, Netherlands*, 45–50, (2007).