Plagiarism Detection across Distant Language Pairs

Alberto Barrón-Cedeño **Paolo Rosso**
Natural Language Engineering Lab. - ELiRF
Universidad Politécnica de Valencia
{lbarron, prosso}@dsic.upv.es

Eneko Agirre **Gorka Labaka**
IXA NLP Group
Basque Country University
{e.agirre, gorka.labaka}@ehu.es

Abstract

Plagiarism, the unacknowledged reuse of text, does not end at language boundaries. Cross-language plagiarism occurs if a text is translated from a fragment written in a different language and no proper citation is provided. Regardless of the change of language, the contents and, in particular, the ideas remain the same. Whereas different methods for the detection of monolingual plagiarism have been developed, less attention has been paid to the cross-language case.

In this paper we compare two recently proposed cross-language plagiarism detection methods (CL-CNG, based on character n-grams and CL-ASA, based on statistical translation), to a novel approach to this problem, based on machine translation and monolingual similarity analysis (T+MA). We explore the effectiveness of the three approaches for less related languages. CL-CNG shows not be appropriate for this kind of language pairs, whereas T+MA performs better than the previously proposed models.

1 Introduction

Plagiarism is a problem in many scientific and cultural fields. Text plagiarism may imply different operations: from a simple cut-and-paste, to the insertion, deletion and substitution of words, up to an entire process of paraphrasing. Different models approach the detection of monolingual plagiarism (Shivakumar and García-Molina, 1995; Hoad and Zobel, 2003; Maurer et al., 2006). Each of these models is appropriate only in those cases where all the implied documents are written in the same language.

Nevertheless, the problem does not end at language boundaries. Plagiarism is also committed if the reused text is translated from a fragment written in a different language and no citation is provided. When plagiarism is generated by a translation process, it is known as cross-language plagiarism (CLP).

Less attention has been paid to the detection of this kind of plagiarism due to its enhanced difficulty (Ceska et al., 2008; Barrón-Cedeño et al., 2008; Potthast et al., 2010). In fact, in the recently held 1st International Competition on Plagiarism Detection (Potthast et al., 2009), no participants tried to approach it.

In order to describe the prototypical process of automatic plagiarism detection, we establish the following notation. Let d_q be a plagiarism suspect document. Let D be a representative collection of reference documents. D presumably includes the source of the potentially plagiarised fragments in d_q. Stein et al., (2007) divide the process into three stages[1]:

1. *heuristic retrieval of potential source documents*: given d_q, retrieving an appropriate number of its potential source documents $D^* \in D$ such that $|D^*| \lll |D|$;
2. *exhaustive comparison of texts*: comparing the text from d_q and $d \in D^*$ in order to identify reused fragments and their potential

[1]This schema was formerly proposed for monolingual plagiarism detection. Nevertheless, it can be applied without further modifications to the cross-language case.

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 37–45,
Beijing, August 2010

sources; and

3. *knowledge-based post-processing*: those detected fragments with proper citation are discarded as they are not plagiarised.

The result is offered to the human expert to take the final decision. In the case of cross-language plagiarism detection (CLPD), the texts are written in different languages: $d_q \in L$ and $d' \in L'$.

In this research we focus on step 2: *cross-language exhaustive comparison of texts*, approaching it as an Information Retrieval problem of cross-language text similarity. Step 1, *heuristic retrieval*, may be approached by different CLIR techniques, such as those proposed by Dumais et al. (1997) and Pouliquen et al. (2003).

Cross-language similarity between texts, $\varphi(d_q, d')$, has been previously estimated on the basis of different models: multilingual thesauri (Steinberger et al., 2002; Ceska et al., 2008), comparable corpora —CL-Explicit Semantic Analysis CL-ESA— (Potthast et al., 2008), machine translation techniques —CL-Alignment-based Similarity Analysis CL-ASA— (Barrón-Cedeño et al., 2008; Pinto et al., 2009) and n-grams comparison —CL-Character n-Grams CL-CNG— (Mcnamee and Mayfield, 2004).

A comparison of CL-ASA, CL-ESA, and CL-CNG was carried out recently by Potthast et al. (2010). The authors report that in general, despite its simplicity, CL-CNG outperformed the other two models. Additionally, CL-ESA showed good results in the cross-language retrieval of topic-related texts, whereas CL-ASA obtained better results in exact (human) translations.

However, most of the language pairs used in the reported experiments (English-{German, Spanish, French, Dutch, Polish}) are related, whether because they have common predecessors or because a large proportion of their vocabularies share common roots. In fact, the lower syntactical relation between the English-Polish pair caused a performance degradation for CL-CNG, and for CL-ASA to a lesser extent. In order to confirm whether the closeness among languages is an important factor, this paper works with more distant language pairs: English-Basque and Spanish-Basque.

The rest of the paper is structured as follows. Section 2 describes the motivation for working on this research topic, stressing the situation of cross-language plagiarism among writers in less resourced languages. A brief overview of the few works on CLPD is included. The three similarity estimation models compared in this research work are presented in Section 3. The experimental framework and the obtained results are included in Section 4. Finally, Section 5 draws conclusions and discusses further work.

2 Motivation

Cases of CLP are common nowadays because information in multiple languages is available on the Web, but people still write in their own language. This special kind of plagiarism occurs more often when the target language is a less resourced one[2], as is the case of Basque.

Basque is a pre-indoeuropean language with less than a million speakers in the world and no known relatives in the language families (Wikipedia, 2010a). Still, Basque shares a portion of its vocabulary with its contact languages (Spanish and French). Therefore, we decided to work with two language pairs: Basque with Spanish, its contact language, and with English, perhaps the language with major influence over the rest of languages in the world. Although the considered pairs share most of their alphabet, the vocabulary and language typologies are very different. For instance Basque is an agglutinative language.

In order to illustrate the relations among these languages, Fig. 1 includes extracts from the English (*en*), Spanish (*es*) and Basque (*eu*) versions of the same Wikipedia article. The fragments are a sample of the lexical and syntactic distance between Basque and the other two languages. In fact, these sentences are completely co-derived and the corresponding entire articles are a sample of the typical imbalance in text available in the different languages (around $2,000$, $1,300$, and only

[2]Less resourced language is that with a low degree of representation on the Web (Alegria et al., 2009). Whereas the available text for German, French or Spanish is less than for English, the difference is more dramatic with other languages such as Basque.

The Party of European Socialists (PES) is a European political party comprising thirty-two socialist, social democratic and labour parties from each European Union member state and Norway.

El Partido Socialista Europeo (PSE) es un partido político pan-europeo cuyos miembros son de partidos socialdemócratas, socialistas y laboristas de estados miembros de la Unión Europea, así como de Noruega.

Europako Alderdi Sozialista Europar Batasuneko herrialdeetako eta Norvegiako hogeita hamahiru alderdi sozialista, sozialdemokrata eta laborista biltzen dituen alderdia da.

Figure 1: First sentences from the Wikipedia articles "Party of European Socialists" (*en*),"Partido Socialista Europeo" (*es*), and "Europako Alderdi Sozialista" (*eu*) (Wikipedia, 2010b).

100 words are contained in the *en*, *es* and *eu* articles, respectively).

Of high relevance is that the two corpora used in this work were manually constructed by translating English and Spanish text into Basque. In the experiments carried out by Potthast et al. (2010), which inspired our work, texts from the JCR-Acquis corpus (Steinberger et al., 2006) and Wikipedia were used. The first one is a multilingual corpus with no clear definition of source and target languages, whereas in Wikipedia no specific relationship exists between the different languages in which a topic may be broached. In some cases (cf. Fig. 1) they are clearly co-derived, but in others they are completely independent.

CLPD has been investigated just recently, mainly by adapting models formerly proposed for cross-language information retrieval. This is the case of cross-language explicit semantic analysis (CL-ESA), proposed by Potthast et al. (2008). In this case the comparison between texts is not carried out directly. Instead, a comparable corpus $C_{L,L'}$ is required, containing documents on multiple topics in the two implied languages. One of the biggest corpora of this nature is Wikipedia. The similarity between $d_q \in L$ and every document $c \in C_L$ is computed based on the cosine measure. The same process is made for L'. This step generates two vectors $[cos(d_q, c_1), \ldots, cos(d_q, c_{|C_L|})]$ and $[cos(d', c'_1), \ldots, cos(d', c'_{|C_{L'}|})]$, where each dimension is comparable between the two vectors. Therefore, the cosine between such vectors can be estimated in order to —indirectly— estimate how similar d_q and d' are. The authors suggest that this model can be used for CLPD.

Another recent model is *MLPlag*, proposed by Ceska et al. (2008). It exploits the *EuroWordNet Thesaurus*[3], that includes sets of synonyms in multiple European languages, with common identifiers across languages. The authors report experiments over a subset of documents of the English and Czech sections of the JRC-Acquis corpus as well as a corpus of simplified vocabulary[4]. The main difficulty they faced was the amount of words in the documents not included in the thesaurus (approximately 50% of the vocabulary).

This is a very similar approach to that proposed by Pouliquen et al. (2003) for the identification of document translations. In fact, both approaches have something in common: translations are searched at document level. It is assumed that an entire document has been reused (translated). Nevertheless, a writer is free to plagiarise text fragments from different sources, and compose a mixture of original and reused text.

A third model is the cross-language alignment-based similarity analysis (CL-ASA), proposed by Barrón-Cedeño et al. (2008), which is based on statistical machine translation technology. This model was proposed to detect plagiarised text fragments (similar models have been proposed for extraction of parallel sentences from comparable corpora (Munteanu et al., 2004)). The authors report experiments over a short set of texts from which simulated plagiarism was created from English to Spanish. Human as well as automatic machine translations were included in the collection. Further descriptions of this model are included in Section 3, as it is one of those being assessed in this research work.

To the best of our knowledge, no work (including the three previously mentioned) has been done considering less resourced languages. In this research work we approach the not uncommon problem of CLPD in Basque, with source texts written in Spanish (the co-official language of the

[3] `http://www.illc.uva.nl/EuroWordNet/`

[4] The authors do not mention the origin of the documents.

	low	tok	pd	bd	sd	lem
T+MA	■	■				■
CL-ASA	■	■				■
CL-CNG	■		■	■	■	

Table 1: Text preprocessing operations required for the different models. low=lowercasing, tok=tokenization, pd=punctuation marks deletion, bd=blank space deletion, sd=symbols deletion, lem=lematization.

Basque Country) and English (the language with most available texts in the world).

We compare three cross-language similarity analysis methods: T+MA (translation followed by monolingual analysis), a novel method based on machine translation followed by a monolingual similarity estimation; CL-CNG, a character n-gram based comparison model; and CL-ASA a model that combines translation and similarity estimation in a single step. Neither MLPlag nor CL-ESA are included in the comparison. On the one hand, we are interested in plagiarism at sentence level, and MLPlag is designed to compare entire documents. On the other hand, in previous experiments over exact translations, CL-ASA has shown to outperform it on language pairs whose alphabet or syntax are unrelated (Potthast et al., 2010). This is precisely the case of *en-eu* and *es-eu* language pairs. Additionally, the amount of Wikipedia articles in Basque available for the construction of the required comparable corpus is insufficient for the CL-ESA data requirements.

3 Definition of Models

In this section, we describe the three cross-language similarity models we compare. For experimental purposes (cf. Section 4) we consider d_q to be a suspicious sentence written in L and D' to be a collection of potential source sentences written in L' ($L \neq L'$). The text pre-processing required by the different models is summarised in Table 1. Examples illustrating how the models work are included in Section 4.3.

3.1 Translation + Monolingual Analysis

$d_q \in L$ is translated into L' on the basis of the Giza++ (Och and Ney, 2003), Moses (Koehn et al., 2007) and SRILM (Stolcke, 2002) tools, generating d'_q. The translation system uses a log-linear combination of state-of-the-art features, such as translation probabilities and lexical translation models on both directions and a target language model. After translation, d'_q and d' are lexically related, making possible a monolingual comparison.

Multiple translations from d_q into d'_q are possible. Therefore, performing a monolingual similarity analysis based on "traditional" techniques, such as those based on word n-grams comparison (Broder, 1997) or hash collisions (Schleimer et al., 2003), is not an option. Instead, we take the approach of the bag-of-words, which has shown good results in the estimation of monolingual text similarity (Barrón-Cedeño et al., 2009). Words in d'_q and d' are weighted by the standard tf-idf, and the similarity between them is estimated by the cosine similarity measure.

3.2 CL-Alignment-based Similarity Analysis

In this model an estimation of how likely is that d' is a translation of d_q is performed. It is based on the adaptation of the Bayes rule for MT:

$$p(d' \mid d_q) = \frac{p(d')\ p(d_q \mid d')}{p(d_q)}. \quad (1)$$

As $p(d_q)$ does not depend on d', it is neglected. From an MT point of view, the conditional probability $p(d_q \mid d')$ is known as *translation model probability* and is computed on the basis of a statistical bilingual dictionary. $p(d')$ is known as *language model probability*; it describes the target language L' in order to obtain grammatically acceptable translations (Brown et al., 1993).

Translating d_q into L' is not the concern of this method, rather it focuses on retrieving texts written in L' which are potential translations of d_q. Therefore, Barrón-Cedeño et al. (2008) proposed replacing the language model (the one used in T+MA) by that known as *length model*. This model depends on text's character lengths instead of language structures.

Multiple translations from d into L' are possible, and it is uncommon to find a pair of translated texts d and d' such that $|d| = |d'|$. Nevertheless, the length of such translations is closely related to a translation length factor. In accordance with Pouliquen et al. (2003), the length model is defined as:

$$\varrho(d') = e^{-0.5\left(\frac{\frac{|d'|}{|d_q|}-\mu}{\sigma}\right)^2}, \quad (2)$$

where μ and σ are the mean and the standard deviation of the character lengths between translations of texts from L into L'. If the length of d' is not the expected given d_q, it receives a low qualification.

The translation model probability is defined as:

$$p(d \mid d') = \prod_{x \in d} \sum_{y \in d'} p(x, y), \quad (3)$$

where $p(x, y)$, a statistical bilingual dictionary, represents the likelihood that x is a valid translation of y. After estimating $p(x, y)$ from a parallel corpus, on the basis of the IBM statistical translation models (Brown et al., 1993), we consider, for each word x, only the k best translations y (those with the highest probabilities) up to a minimum probability mass of 0.4. This threshold was empirically selected as it eliminated noisy entries without discarding an important amount of relevant pairs.

The similarity estimation based on CL-ASA is finally computed as:

$$\varphi(d_q, d') = \varrho(d')\ p(d_q \mid d'). \quad (4)$$

3.3 CL-Character n-Gram Analysis

This model, the simplest of those compared in this research, has been used in (monolingual) Authorship Attribution (Keselj et al., 2003) as well as cross-language Information Retrieval (Mcnamee and Mayfield, 2004). The simplified alphabet considered is $\Sigma = \{a, \ldots, z, 0, \ldots, 9\}$; any other symbol is discarded (cf. Table 1). The resulting text strings are codified into character 3-grams, which are weighted by the standard tf-idf (considering this n has previously shown to produce the best results). The similarity between such representations of d_q and d' is estimated by the cosine similarity measure.

4 Experiments

The objective of our experiments is to compare the performance of the three similarity estimation models. Section 4.1 introduces the corpora we have exploited. The experimental framework is described in Section 4.2. Section 4.3 illustrates how the models work, and the obtained results are presented and discussed in Section 4.4.

4.1 Corpora

In other Information Retrieval tasks a plethora of corpora is available for experimental and comparison purposes. However, plagiarism implies an ethical infringement and, to the best of our knowledge, there is no corpora of actual cases available, other than some seminal efforts on creating corpora of text reuse (Clough et al., 2002), artificial plagiarism (Potthast et al., 2009), and simulated plagiarism (Clough and Stevenson, 2010). The problem is worse for cross-language plagiarism.

Therefore, in our experiments we use two parallel corpora: *Software*, an *en-eu* translation memory of software manuals generously supplied by Elhuyar Fundazioa[5]; and *Consumer*, a corpus extracted from a consumer oriented magazine that includes articles written in Spanish along with their Basque, Catalan, and Galician translations[6] (Alcázar, 2006). *Software* includes $288,000$ parallel sentences; 8.66 (6.83) words per sentence in the English (Basque) section. *Consumer* contains $58,202$ sentences; 19.77 (15.20) words per sentence in Spanish (Basque). These corpora also reflect the imbalance of text available in the different languages.

4.2 Experimental Framework

We consider D_q and D' to be two entire documents from which plagiarised sentences and their source are to be detected. We work at this level of granularity, and not entire documents, for two main reasons: (*i*) we are focused on the exhaustive comparison stage of the plagiarism detection process (cf. Section 1); and (*ii*) even a single sentence could be considered a case of plagiarism, as it transmits a complete idea. However, a plagiarised sentence is usually not enough to automatically negate the validity of an entire document. This decision is left to the human expert, which can examine the documents where several plagiarised sentences occur. Note that the task becomes computationally more expensive as, for every sentence, we are looking through thousands

[5] http://www.elhuyar.org

[6] http://revista.consumer.es

	es-eu		en-eu	
	μ	σ	μ	σ
f_1	1.1567	0.2346	1.0561	0.5497
f_2	1.1569	0.2349	1.0568	0.5510
f_3	1.1571	0.2349	1.0566	0.5433
f_4	1.1565	0.2363	1.0553	0.5352
f_5	1.1571	0.2348	1.0553	0.5467
$avg.$	1.1569	0.2351	1.0560	0.5452

Table 2: Length models estimated for each training partition $f_{1,\ldots,5}$. The values describe a normal distribution centred in $\mu \pm \sigma$, representing the expected length of the source text given the suspicious one.

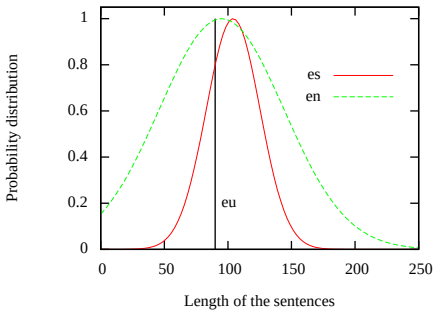


Figure 2: Example length factor for a sentence written in Basque (*eu*) d_q, such that $|d_q| = 90$. The normal distributions represent the expected lengths for the translation d', either in Spanish (*es*) or English (*en*).

of topically-related sentences that are potential sources of d_q, and not only those of a specific document.

CLPD is considered a ranking problem. Let $d_q \in D_q$ be a plagiarism suspicious sentence and $d' \in D'$ be its source sentence. We consider that the result of the process is correct if, given d_q, d' is properly retrieved. A 5-fold cross validation for both *en-eu* and *es-eu* was performed. Bilingual dictionaries, language and length models were estimated with the corresponding training partitions. The computed values for μ and σ are those included in Table 2. The values for the different partitions are very similar, showing the low variability in the translation lengths. On the basis of these estimated parameters, an example of length factor for a specific sentence is plotted in Fig. 2.

In the test partitions, for each suspicious sentence d_q, $11,640$ source candidate sentences exist for *es-eu* and $57,290$ for *en-eu*. This results in more than 135 million and 3 billion comparisons carried out for *es-eu* and *en-eu* respectively.

x_{eu}	y_{en}	$p(x,y)$	x_{eu}	y_{en}	$p(x,y)$
beste	another	0.288	**beste**	**other**	**0.348**
dokumentu	**document**	**0.681**	batzu	some	0.422
makro	**macro**	**0.558**	**ezin**	**not**	**0.179**
ezin	cannot	0.279	izan	is	0.241
izan	the	0.162	atzi	access	0.591
.	**.**	**0.981**			

Table 3: Entries in the bilingual dictionary for the words in d_q. Relevant entries for the example are in bold.

4.3 Illustration of Models

In order to clarify how the different models work, consider the following sentence pair, a suspicious sentence d_q written in Basque and its source d' written in English (sentences are short for illustrative purposes):

d_q *beste dokumentu batzuetako makroak ezin dira atzitu.*
d' *macros from other documents are not accessible.*

CL-CNG Example

In this case, symbols and spaces are discarded. Sentences become:

d_q *bestedokumentubatzuetakomakroakezindiraatzitu*
d' *macrosfromotherdocumentsarenotaccessible*

Only three 3-grams appear in both sentences (*ume*, *men*, *ent*). In order to keep the example simple, the 3-grams are weighted by tf only (in the actual experiments, tf-idf is used), resulting in a dot product of 3. The corresponding vectors magnitudes are $|\mathbf{d_q}| = 6.70$ and $|\mathbf{d'}| = 5.65$. Therefore, the estimated similarity is $\varphi(d_q, d') = 0.079$.

CL-ASA Example

In this case, the text must be tokenised and lemmatised, resulting in the following string:

d_q *beste dokumentu batzu makro ezin izan atzi .*
d' *macro from other document be not accessible .*

The sentences' lengths are $|d_q| = 38$ and $|d'| = 39$. Therefore, on the basis of Eq. 2, the length factor between them is $\varrho(d_q, d') = 0.998$.

The relevant entries of the previously estimated dictionary are included in Table 3. Such entries are substituted in Eq. 3, and the overall process results in a similarity $\varphi(d_q, d') = 2.74$. Whereas not a stochastic value, this is a weight used when ranking all the potential source sentences in D'.

T+MA Example

In this case, the same pre-processing than in CL-ASA is performed. In T+MA d_q is translated into L', resulting in the new pair:

d'_q *other document macro cannot be access .*
d' *macro from other document be not accessible .*

Note that d'_q is a valid translation of d_q. Nevertheless, it has few syntactic relation to d'. Therefore, applying more sophisticated codifications than the cosine measure over bag-of-words is not an option. The example is again simplified by weighting the words based on tf. Five words appear in both sentences, resulting in a dot product of 5. The vectors magnitudes are $|\mathbf{d'_q}| = |\mathbf{d'}| = \sqrt{7}$. The estimation by T+MA is $\varphi(d_q, d') = 0.71$, a high similarity level.

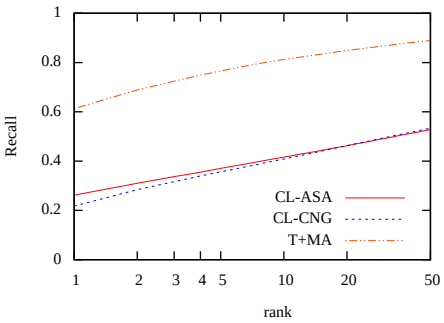
4.4 Results and Discussion

For evaluation we consider a standard measure: Recall. More specifically Recall after n texts have been retrieved ($n = [1 \ldots, 50]$). Figure 3 plots the average Recall value obtained in the 5-folds with respect to the rank position (n).

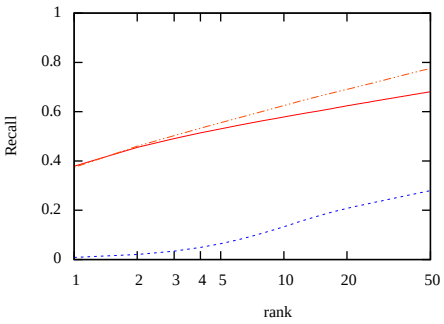
In both language pairs, CL-CNG obtained worse results than those reported for English-Polish by Potthast et al. (2010): $R@50 = 0.68$ vs. $R@50 = 0.53$ for *es-eu* and 0.28 for *en-eu*. This is due to the fact that neither the vocabulary nor its corresponding roots keep important relations. Therefore, when language pairs have a low syntactical relationship, CL-CNG is not an option. Still, CL-CNG performs better with *es-eu* than with *en-eu* because the first pair is composed of contact languages (cf. Section 1).

About CL-ASA, the results obtained with *es-eu* and *en-eu* are quite different: $R@50 = 0.68$ for *en-eu* and $R@50 = 0.53$ for *es-eu*. Whereas in the first case they are comparable to those of CL-CNG, in the second one CL-ASA completely outperforms it. The improvement of CL-ASA obtained for *en-eu* is due to the size of the training corpus available in this case (approximately five times the number of sentences available for *es-eu*). This shows the sensitivity of the model with respect to the size of the available resources.

Lastly, although T+MA is a simple approach that reduces the cross-language similarity estimation to a translation followed by a monolingual process, it obtained a good performance (R@50= 0.77 for *en-eu* and R@50=0.89 for es-eu). Moreover, this method proved to be less sensitive than CL-ASA to the lack of resources. This could be due to the fact that it considers both directions of the translation model (*e[n|s]-eu* and *eu-e[n|s]*). Additionally, the language model, applied in order to compose syntactically correct translations, reduces the amount of wrong translations and, indirectly, includes more syntactic information in the process. On the contrary, CL-ASA only considers one direction translation model *eu-e[n|s]* and completely disregards syntactical relations between the texts.



(a) es-eu



(b) en-eu

Figure 3: Evaluation of the cross-language ranking. Results plotted as rank versus Recall for the three evaluated models and the two language pairs ($R@[1, \ldots, 50]$).

Note that the better results come at the cost of higher computational demand. CL-CNG only requires easy to compute string comparisons. CL-ASA requires translation probabilities from aligned corpora, but once the probabilities are estimated, cross-language similarity can be computed very fast. T+MA requires the previous translation of all the texts, which can be very costly for large collections.

5 Conclusions and Further Work

In a society where information in multiple languages is available on the Web, cross-language

plagiarism is occurring every day with increasing frequency. Still, cross-language plagiarism detection has not been approached sufficiently due to its intrinsic complexity. Though few attempts have been made, even less work has been made to tackle this problem for less resourced languages, and to explore distant language pairs.

We investigated the case of Basque, a language where, due to the lack of resources, cross-language plagiarism is often committed from texts in Spanish and English. Basque has no known relatives in the language family. However, it shares some of its vocabulary with Spanish.

Two state-of-the-art methods based on translation probabilities and n-gram overlapping, and a novel technique based on statistical machine translation were evaluated. The novel technique obtains the best results in both language pairs, with the n-gram overlap technique performing worst. In this sense, our results complement those of Potthast et al. (2010), which includes closely related language pairs as well.

Our results also show that better results come at the cost of more expensive processing time. For the future, we would like to investigate such performance trade-offs in more demanding datasets.

For future work we consider that exploring semantic text features across languages could improve the results. It could be interesting to further analyse how the reordering of words through translations might be relevant for this task. Additionally, working with languages even more distant from each other, such as Arabic or Hindi, seems to be a challenging and interesting task.

Acknowledgements

The research work of the first two authors is partially funded by CONACYT-Mexico and the MICINN project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i). The research work of the last two authors is partially funded by the MICINN projects OPENMT-2 TIN2009-14675-C03-01 and KNOW2 TIN2009-14715-C04-01.

References

Alcázar, Asier. 2006. Towards Linguistically Searchable Text. In *Proceedings of the BIDE 2005*, Bilbao, Basque Country.

Alegria, Iñaki, Mikel L. Forcada, and Kepa Sarasola, editors. 2009. *Proceedings of the SEPLN 2009 Workshop on Information Retrieval and Information Extraction for Less Resourced Languages*, Donostia, Basque Country. University of the Basque Country.

Barrón-Cedeño, Alberto, Paolo Rosso, David Pinto, and Alfons Juan. 2008. On Cross-lingual Plagiarism Analysis Using a Statistical Model. In Stein, Stamatatos, and Koppel, editors, *ECAI 2008 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2008)*, pages 9–13, Patras, Greece. CEUR-WS.org.

Barrón-Cedeño, Alberto, Andreas Eiselt, and Paolo Rosso. 2009. Monolingual Text Similarity Measures: A Comparison of Models over Wikipedia Articles Revisions. In Sharma, Verma, and Sangal, editors, *ICON 2009*, pages 29–38, Hyderabad, India. Macmillan Publishers.

Broder, Andrei Z. 1997. On the Resemblance and Containment of Documents. In *Compression and Complexity of Sequences (SEQUENCES'97)*, pages 21–29. IEEE Computer Society.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Ceska, Zdenek, Michal Toman, and Karel Jezek. 2008. Multilingual Plagiarism Detection. In *Proceedings of the 13th International Conference on Artificial Intelligence*, pages 83–92. Springer Verlag Berlin Heidelberg.

Clough, Paul and Mark Stevenson. 2010. Developing a Corpus of Plagiarised Short Answers. *Language Resources and Evaluation: Special Issue on Plagiarism and Authorship Analysis*.

Clough, Paul, Robert Gaizauskas, and Scott Piao. 2002. Building and Annotating a Corpus for the Study of Journalistic Text Reuse. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, volume V, pages 1678–1691, Las Palmas, Spain.

Dumais, Susan T., Todd A. Letsche, Michael L. Littman, and Thomas K. Landauer. 1997. Automatic Cross-Language Retrieval Using Latent Semantic Indexing. In *AAAI-97 Spring Symposium Series: Cross-Language Text and Speech Retrieval*, pages 24—26. Stanford University.

Hoad, Timothy C. and Justin Zobel. 2003. Methods for Identifying Versioned and Plagiarized Documents. *Journal of the American Society for Information Science and Technology*, 54(3):203–215.

Keselj, Vlado, Fuchun Peng, Nick Cercone, and Calvin Thomas. 2003. N-gram-based Author Profiles for Authorship Attribution. In *Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING'03*, pages 255–264, Halifax, Canada.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic.

Maurer, Hermann, Frank Kappe, and Bilal Zaka. 2006. Plagiarism - A Survey. *Journal of Universal Computer Science*, 12(8):1050–1084.

Mcnamee, Paul and James Mayfield. 2004. Character N-Gram Tokenization for European Language Text Retrieval. *Information Retrieval*, 7(1-2):73–97.

Munteanu, Dragos S., Alexander Fraser, and Daniel Marcu. 2004. Improved Machine Translation Performace via Parallel Sentence Extraction from Comparable Corpora. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL 2004)*, Boston, MA.

Och, Frank Josef and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51. See also http://www.fjoch.com/GIZA++.html.

Pinto, David, Jorge Civera, Alberto Barrón-Cedeño, Alfons Juan, and Paolo Rosso. 2009. A Statistical Approach to Crosslingual Natural Language Tasks. *Journal of Algorithms*, 64(1):51–60.

Potthast, Martin, Benno Stein, and Maik Anderka. 2008. A Wikipedia-Based Multilingual Retrieval Model. In Macdonald, Ounis, Plachouras, Ruthven, and White, editors, *30th European Conference on IR Research, ECIR 2008, Glasgow*, volume 4956 LNCS of *Lecture Notes in Computer Science*, pages 522–530, Berlin Heidelberg New York. Springer.

Potthast, Martin, Benno Stein, Andreas Eiselt, Alberto Barrón-Cedeño, and Paolo Rosso. 2009. Overview of the 1st International Competition on Plagiarism Detection. In Stein, Rosso, Stamatatos, Koppel, and Agirre, editors, *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)*, pages 1–9, San Sebastian, Spain. CEUS-WS.org.

Potthast, Martin, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. 2010. Cross-Language Plagiarism Detection. *Language Resources and Evaluation, Special Issue on Plagiarism and Authorship Analysis*.

Pouliquen, Bruno, Ralf Steinberger, and Camelia Ignat. 2003. Automatic Identification of Document Translations in Large Multilingual Document Collections. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2003)*, pages 401–408, Borovets, Bulgaria.

Schleimer, Saul, Daniel S. Wilkerson, and Alex Aiken. 2003. Winnowing: Local Algorithms for Document Fingerprinting. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, New York, NY. ACM.

Shivakumar, Narayanan and Hector García-Molina. 1995. SCAM: A Copy Detection Mechanism for Digital Documents. In *Proceedings of the 2nd Annual Conference on the Theory and Practice of Digital Libraries*.

Stein, Benno, Sven Meyer zu Eissen, and Martin Potthast. 2007. Strategies for Retrieving Plagiarized Documents. In Clarke, Fuhr, Kando, Kraaij, and de Vries, editors, *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 825–826, Amsterdam, The Netherlands. ACM.

Steinberger, Ralf, Bruno Pouliquen, and Johan Hagman. 2002. Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC. *Computational Linguistics and Intelligent Text Processing. Proceedings of the CICLing 2002*, 2276:415—424.

Steinberger, Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, volume 9, Genoa, Italy.

Stolcke, Andreas. 2002. SRILM - An Extensible Language Modeling toolkit. In *Intl. Conference on Spoken Language Processing*, Denver, Colorado.

Wikipedia. 2010a. Basque language. [Online; accessed 5-February-2010].

Wikipedia. 2010b. Party of European Socialists | Partido Socialista Europeo | Europako Alderdi Sozialista . [Online; accessed 10-February-2010].