Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 141-149 recibido 1-05-2009; aceptado 5-06-2009

On the Relevance of Search Space Reduction in Automatic Plagiarism Detection *

Sobre la importancia de la reducción del espacio de búsqueda en la detección automática de plagio

Alberto Barrón-Cedeño and Paolo Rosso
Natural Language Engineering Lab. - ELiRF
Dpto. Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
DSIC, edificio 1F Campus de Vera
Camino de Vera s/n, 46022 Valencia, Spain
[lbarron | prosso]@dsic.upv.es

Resumen: En la detección automática de plagio con referencia, los fragmentos de texto de un documento sospechoso son buscados de manera exhaustiva en un conjunto de documentos originales (de referencia) con el objetivo de determinar si han sido plagiados o no. Uno de los factores más importantes para el éxito de este tipo de aplicaciones es el tamaño del corpus de referencia el cual, al mismo tiempo, puede representar un problema al considerar el desempeño y la precisión. En este artículo, abordamos la detección automática de plagio con referencia analizando el impacto de una etapa previa de reducción del espacio de búsqueda (conformado por los documentos originales en el corpus de referencia). Nuestros experimentos sobre el corpus METER muestran una mejora en la Precisión y Cobertura de los resultados obtenidos cuando la reducción del espacio de búsqueda es realizada al principio del proceso de detección de plagio.
Palabras clave: detección de plagio, reducción del espacio de búsqueda, similitud de texto

Abstract: In automatic plagiarism detection with reference, the text fragments in a suspicious document are exhaustively searched in a set of original (reference) documents in order to determine whether they have been plagiarised or not. One of the most important factors for the success of this kind of applications is the size of the reference corpus that, at the same time, may represent a problem when we consider performance and precision. In this paper, we approach automatic plagiarism detection analysing the impact of a preliminary search space reduction (composed of the original documents in the reference corpus). Our experiments over the METER corpus show that the Precision and Recall of the obtained results are improved when a search space reduction is applied at the beginning of a plagiarism detection process.
Keywords: plagiarism detection, search space reduction, text similarity

1 Introduction

Plagiarism is a current problem in many scientific and cultural fields.[1] It is not uncommon to find documents that have not been originally written (be partially or completely) by their claimed authors. These cases of plagiarism arise from the facility of electronically accessing texts written by other people as well as retrieving documents directly from Internet Web pages. The amount of information reachable through a browser may be of great benefit, but when it is simply copied without any "rational" processing, plagiarism is committed. Fortunately, the problem that the information technology has caused can be approached by itself.

Automatic Plagiarism Detection, a closed

* We would like to thank Paul Clough for providing us the METER corpus and Germán Rigau for sharing with us the Les Luthiers's plagiarism sketch. This work was partially funded by the MCyT TIN2006-15265-C06-04 research project and the CONACyT-MEXICO 192021/302009 grant.

[1] See for instance http://www.youtube.com/watch?v=bYLWTxNgY6o

ISSN: 1135-5948

task to Authorship Attribution (Stein, Koppel, and Stamatatos, 2007), can be classified into two main approaches: *plagiarism detection with reference* and *intrinsic plagiarism analysis*. In the former approach a suspicious document is compared to a set of potential source documents in order to discover a possible case of plagiarism. In the latter approach stylometric and other text inherent features are considered in order to detect fragments of the suspicious document that could be plagiarised. This research work is focused on the former approach.

A key factor in plagiarism detection with reference is the size of the set of potential source documents considered (hereinafter reference corpus). The larger the reference corpus, the higher the probability of including the source of a plagiarised text fragment. The size of the reference corpus represents also a lack for any text comparison process due to the fact that it often composes a large search space. In this work we emphasise the relevance of reducing the search space before carrying out any text comparison strategy. The search space reduction is based on the Kullback-Leibler distance, which has been previously applied to documents clustering (Bigi, 2003) among other tasks. This research is mainly oriented to improve the quality of the obtained results (in terms of Precision and Recall) of automatic plagiarism detection.

The rest of the paper is structured as follows. Section 2 gives a brief overview of plagiarism detection and includes a description of some of the state of the art methods. Section 3 is intended to outline the corpus used in the research work. Section 4 describes the proposed reduction method as well as an exhaustive text comparison method. In Section 5 the results obtained by the different experiments carried out are discussed. Finally, Section 6 draws some conclusions and future work.

2 Plagiarism Detection Overview

Among the most frequent plagiarism cases, which are feasible to be automatically detected are: direct copy of text and text rewording (changing words by synonyms or changing the order of the text).[2] Maurer, Kappe, and Zaka (2006) have organised the automatic detection methods into three main categories including those which:

- Try to detect suspicious text fragments on the basis of stylometric analysis;
- Make an exhaustive comparison of suspicious versus reference texts in order to find the source of a potentially plagiarised text; and
- Select one characteristic text fragment of the suspicious document in order to search for it on the Web.

This simple classification entails to discriminate the plagiarism detection methods into the two main approaches aforementioned. Intrinsic plagiarism analysis is mainly inspired by the fact that humans can detect plagiarism cases by simply reading a text (without comparing it to any other document). The idea is to capture the writing style across the suspicious text in order to find fragments which are candidates of being plagiarised.

The main parameter considered by Meyer zu Eißen and Stein (2006) is known as Averaged Word Frequency Class (AWFC), where each word w is assigned to a class $c(w)$. Given a suspicious document s, the frequency of each word $w \in s$ ($f_s(w)$) is calculated. After this, the words are sorted by their frequency in decreasing order. The word with the maximum frequency is named w^* and is assigned to the class c_0. Then, any other word $w \in s$ is assigned to the class given by $\lfloor log_2(f_s(w^*)/f_s(w)) \rfloor$. The classes distribution reflects style and complexity as well as the vocabulary size. When analysing a document written by a single author (an original text), AWFC shows a small variance no matter the size of the analysed text. AWFC as well as other stylometric features are calculated for the entire document as well as for each paragraph in order to look for unexpected variations in the obtained values. This approach avoids the high cost of a text comparison process as well as the difficulty of compiling a good reference corpus. However, due to the method philosophy, no hint is obtained about the potential source of a presumably plagiarised text fragment.

Plagiarism detection with reference is based on the comparison of suspicious and reference texts. Once a suspicious text is found to be similar enough to an original

[2] Due to the unclear text dependency between the plagiarism and its source, plagiarism of citations and ideas are not considered automatically detectable.

one, it is considered a plagiarism candidate. Consider that one of the main difficulties in this approach (after having compiled a good enough reference corpus, which is by itself a hard task), is how to compare the text fragments considering the big size that a reference corpus should have. Following, we give some different approaches to this task.

The *CHECK* system approaches plagiarism detection on the basis of the documents structure (Si, Leong, and Lau, 1997). In CHECK the reference as well as the suspicious documents must be LaTeX files. The documents, whose structures are represented in a tree format, are compared with a depth first search strategy. In those cases where a leaf (composed of a pair of paragraphs) is reached, the similarity between suspicious and reference paragraphs is calculated on the basis of the *dot plot* technique. The weakness of this approach is that it is only able to process LaTeX files.

The *PPChecker* system carries out an exhaustive text comparison at the sentence level (Kang, Gelbukh, and Han, 2006). In this approach the sentences vocabularies are expanded considering *Wordnet* synonymic relationships. The method considers the vocabulary intersection and complement between suspicious and reference sentences in order to discriminate real plagiarism cases from casual vocabulary intersections. It is able to differentiate among exact copy of sentences, word insertion, word removal and rewording instead of simply determining whether there is a case of plagiarism or not.

Text splitting (considering sentences, paragraphs or any other text chunk) is not always necessary. The *Ferret* system (Lyon, Barrett, and Malcolm, 2004) bases the plagiarism analysis process on a word-level trigrams comparison. In this case, the trigrams in the suspicious text are obtained in order to after compare them to a set of reference texts (also codified as trigrams). The method defines whether or not the suspicious and reference documents become a potential case of plagiarism depending on the amount of common trigrams between both texts.

When plagiarism detection with reference is carried out in "real life", a big drawback is the big size that a good reference corpus should reach. Precisely, this size (the number of documents in the reference corpus) represents the search space for every suspicious document to be analysed. Some efforts, such as fingerprinting techniques (Stein, 2007), have been made in the direction of improving the search speed. Instead of the original text fragments, a numerical value is assigned to each text chunk. The numerical values (fingerprints) of the reference and suspicious documents are compared in order to determine a possible case of plagiarism.

These approaches to plagiarism detection with reference share a common idea: a suspicious document must be exhaustively compared to all the documents contained in a reference corpus. As it will be shown across the following sections, assuming this behaviour can affect the quality of the obtained results. In this research work we describe our proposed solution to this problem: a previous selection of good source candidates of a suspicious text.

3 The METER Corpus

In order to experiment we have opted for considering a standard corpus. In this way, we offer results that could be after compared to those obtained by other methods. The *METER corpus* (Clough, Gaizauskas, and Piao, 2002) was created inside of the METER (MEasuring TExt Reuse) Project[3]. The objective of this project was working on "detecting and measuring text reuse".

This corpus, which is not a real plagiarism corpus, can be divided into two subcorpora. The first one is composed of a set of news reported by the Press Association (PA), the major UK news agency. The news reported by the PA are distributed to nine British newspapers (The Times, The Guardian, The Independent, The Telegraph, etc.) for their publication. This is precisely the second part of the corpus, which contains notes about the same news written in any of the different newspapers.

The notes in the newspaper subcorpus have been tagged by an expert journalist considering three main levels of derivation from the corresponding PA note: *wholly derived*, *partially derived* and *non derived*. These tags mean that the PA version of the note was the only source, one of the sources, or none of the sources of the newspaper note, respectively (Clough, Gaizauskas, and Piao, 2002). Additionally, a good part of the notes text

[3] http://www.dcs.shef.ac.uk/nlp/meter/

Feature	Entire	PA	Papers
$\|tokens\|$	526,427	226,427	299,767
$\|tokens\|$ (avg)	306.12	293.68	318.56
$\|types\|$	40,336	25,728	30,173
$\|types_s\|$	28,990	18,643	22,204

Table 1: METER Corpus statistics. Tokens in the entire subcorpora and per document; types before and after stemming.

fragments have been identified as:

verbatim if the text fragment is an exact copy of the PA version,

rewrite if it has been modified from the PA version, or

new if it has nothing to do with the PA version.

The entire METER corpus is composed of around 1,700 texts from July 1999 to June 2000. Table 1 shows some statistics about it including the number of tokens and types in the entire corpus as well as in the PA and newspapers notes.

For our purposes, the entire set of PA notes, 771 documents, are considered as source documents (i.e., the reference corpus). Those newspaper notes which text fragments are tagged (as verbatim, rewrite or new), 444 documents, compose the corpus of suspicious samples.

In order to illustrate the corpus samples, Figure 1 shows a story fragment written by the *PA* followed by the annotated version of the same story as published in *The Telegraph.* For an easier identification, the PA version contains those fragments that The Telegraph copied verbatim and rewrote highlighted (in **bold** and *italic* respectively). It can be noted from this example that the sentences in the "potentially plagiarised" documents are composed of a mix of verbatim, rewritten and new text fragments.

Press Association version

Celebrity **chef Marco Pierre White** *today* **won** *the battle of* **the Titanic and Atlantic restaurants**. **Oliver Peyton, owner of the Atlantic** Bar and Grill, **had tried to** *sink Marco's* **new Titanic restaurant housed in the same West End hotel in London by seeking damages against landlords Forte Hotels and** *an* **injunction in the High Court**. But today the Atlantic announced in court it had reached a confidential agreement with the landlords and was discontinuing the whole action.

The Telegraph version

```
<R> THE </R>
<V> chef Marco Pierre White </V>
<R> yesterday </R>
<V> won </V>
<R> a dispute over </R>
<V> the Titanic and Atlantic restaurants.
    </V>
<V> Oliver Peyton, owner of the Atlantic,
    had tried to </V>
<R> close White's </R>
<V> new Titanic restaurant, housed in the
    same West End hotel in London, by
    seeking damages against </V>
<R> the </R>
<V> landlords, Forte Hotels, and </V>
<R> a </R>
<V> High Court injunction.</V>
<R> He </R>
<V> claimed that the Titanic was a
    replica of the Atlantic and should
    not be allowed to trade in competition
    at the Regent Palace Hotel. </V>
```

Figure 1: *PA* and *The Telegraph* notes of the news item "Titanic restaurant case discontinued". <R> = rewrite, <V> = verbatim

The experiments carried out in this research work are based on the search of potentially plagiarised sentences. A sentence is considered plagiarised if a high percentage of its words have been copied verbatim or rewritten from the original PA source note. In order to avoid considering incidental common fragments (such as named entities) as plagiarism cases, a sentence must fulfil the following inequality to be considered as plagiarised:

$$|w_V \cup w_R| > 0.4|w| \ , \tag{1}$$

where $|w|$ is the number of words in the entire sentence, whereas $|w_V \cup w_R|$ is the number of words in verbatim and rewritten fragments in the sentence.

The corpus preprocessing includes punctuation marks and words splitting as well as lowercasing and stemming. For stemming the Porter stemmer has been applied (Porter, 1980).[4] Note that splitting punctuations marks and words implies that given the text fragment "*here we are, born*" the resulting bigrams are {here we}, {we are}, {are ,} and {, born}.

[4] The Vivake Gupta implementation of the Porter stemmer has been used in this research. It is available at http://tartarus.org/∼martin/PorterStemmer/.

4 Method Definition

Due to the nature of the corpus, composed of short texts (around 300 words per document in average), we consider that the plagiarism detection task is solved if given a plagiarised sentence $s_i \in s$, its source document d is accurately retrieved from the reference corpus.

The common schema of automatic plagiarism detection methods with reference is an exhaustive comparison of sentences, paragraphs or any other text chunk $s_i \in s$ to the text in $d \in D$. The processing cost of making all the necessary comparisons is $O(n \cdot m)$, being n and m the length of s and D in fragments, respectively. If D contains several hundreds of documents with thousands of words, the output could be affected as well as the time invested in order to obtain it. Due to this reason D must be short enough to obtain the wished results in a reasonable time.

Given s and D, our propose is carrying out a preliminary selection of those documents $d \in D$ with the highest probabilities of being the source of the potentially plagiarised fragments $s_i \in s$. The selected documents compose the set $D' \subset D$, such that $|D'| \ll |D|$. After D' has been obtained, an exhaustive comparison method can be applied in order to relate a fragment $s_i \in s$ to its potential source document $d \in D'$.

The proposed method for the selection of reference documents, which implies the search space reduction, is based on the Kullback-Leibler distance (Kullback and Leibler, 1951) (Subsection 4.1). For the exhaustive text comparison we have opted for the containment measure (Lyon, Malcolm, and Dickerson, 2001), over word-level bigrams, which have previously shown good results in this task (Barrón-Cedeño and Rosso, 2009) (Subsection 4.2).

4.1 Search Space Reduction

Different methods to relate a potentially plagiarised text to its source have shown good results (Si, Leong, and Lau, 1997; Kang, Gelbukh, and Han, 2006; Lyon, Malcolm, and Dickerson, 2001). However, a difficult which has not been considered is the size that a reference corpus can reach. As it can be seen in the following section, the large size of the reference corpus (the search space) can affect the quality of the obtained results in this kind of analysis. In order to solve this problem, we propose a method for reducing the search space of the plagiarism detection task.

The method is based on the *Kullback-Leibler distance* (KL_δ) between two probability distributions which characterise the reference as well as the suspicious documents. KL_δ is a symmetric version of the Kullback-Leibler divergence (Kullback and Leibler, 1951). It measures how different (or equal) two probability distributions P and Q, over a feature vector $\mathcal{X}$, are. In agreement with Bigi (2003) we define KL_δ as:

$$KL_\delta(P||Q) = \sum_{x \in \mathcal{X}} (P(x) - Q(x)) log \frac{P(x)}{Q(x)} \ . \tag{2}$$

P_d and Q_s are the probability distributions characterising the documents d and s, respectively. The feature selection technique considered in order to compose P_d is the well known *tfidf* (Salton, Fox, and Wu, 1983). The *tfidf* value of a given term x in a document d is defined as:

$$tfidf_{x,d} = tf_{x,d} \cdot idf_x \ , \tag{3}$$

where

$$tf_{x,d} = \frac{n_{x,d}}{\sum_k n_{k,d}} \ , \tag{4}$$

($n_{\cdot,d}$ is the frequency of $\cdot$ in d), and

$$idf_x = log \frac{|D|}{|\{d \mid x \in d\}|} \ . \tag{5}$$

Equation 3 is used to define the terms composing the vector $\mathcal{X}$ of P_d (i.e. the probability distribution corresponding to each reference document). Once $\mathcal{X}$ is defined, the probability of each considered term $x \in \mathcal{X}$ is calculated as in Equation 4, i.e., $P(x \mid d) = tf_{x,d}$. Considering distributions composed only of the top 20% of the terms in d, ranked by their *tfidf* value, gives a good result comparable to the one obtained by considering the entire set of terms (Barrón-Cedeño, Rosso, and Benedí, 2009).

The probability distributions P_d have to be re-calculated only when a new document is added to the reference corpus. This process can be carried out *off-line* before any suspicious document is analysed. After obtaining the distributions P_d for each document $d \in D$, the set $D' \subset D$ related to a suspicious document s can be obtained.

Given a document s, a preliminary distribution Q' is obtained on the basis of Equation 4, i.e., $Q'(x \mid s) = tf_{x,s}$. However, when comparing it to each probability distribution P_d, Q'_s must be adapted to it. The reason is that when analysing a document s with respect to each document $d \in D$, the vocabulary composing the corresponding distributions will be different in most cases. Calculating the distance between two distributions composed of different terms obtains $KL_\delta(P_d||Q'_s) = \infty$ when $\exists x \in d \wedge x \notin s$ and $KL_\delta(P_d||Q'_s) = 0$ vice versa. The distribution Q_s depends on the distribution P_d and it must be composed of the same terms (i.e. the same vocabulary). If $x \in (P_d \cap Q'_s)$, $Q(x, s)$ is smoothed from $Q'(x, s)$, otherwise $Q(x, s) = \epsilon$. This is simply a back-off smoothing of Q_s and, in agreement with Bigi (2003), the final probability $Q(x, s)$ is calculated as:

$$Q(x,s) = \begin{cases} \gamma \cdot tf_{x,s} & \text{if } x \in d \cap s \\ \epsilon & \text{it } x \in d \setminus s \end{cases} , \quad (6)$$

where γ is a normalisation coefficient estimated by:

$$\gamma = 1 - \sum_{x \in d \setminus s} \epsilon \ , \quad (7)$$

respecting the condition:

$$\sum_{x \in s} \gamma \cdot Q(x,s) + \sum_{x \in d \setminus s} \epsilon = 1 \ . \quad (8)$$

As it is expected, ϵ is smaller than the minimum probability of a term x in a document d. The search space reduction process is resumed in Figure 2.

Algorithm 1: Search space reduction. Given s:

```
$Q'_s = \{[x, tf_{x,s}] \forall x \in s\}$
for all $d \in D$
    $Q_s = (Q'_s \mid P_d)$
    Calculate $KL_\delta(P_d||Q_s)$
$LD$ = Ranked list of $d \in D$ based on $KL_\delta$
$D'$ = [top 10 documents in $LD$]
```

Figure 2: Search space reduction process.

Calculating $KL_\delta(P_d||Q_s) \, \forall d \in D$ the subset of possible source documents of the plagiarised fragments in s is obtained. The length of the reference corpus subset $D' \subset D$ is only $|D'| = 10$. After obtaining D', an exhaustive text comparison strategy can be applied in order to detect plagiarism candidates. The exhaustive text comparison method we have opted for is based on word-level n-grams.

4.2 Exhaustive Text Comparison

For the n-gram based text comparison, the entire documents in the reference corpus are codified as word-level n-grams. However, the document s is first split into sentences and thereafter each sentence $s_i \in s$ is codified as n-grams. In this way an asymmetric comparison (sentence versus document) is carried out. The length of the set of n-grams of each $s_i \in s$ and d are different; in general, $|N(s_i)| \ll |N(d)|$, where $N(\cdot)$ is the set of n-grams in $\cdot$. Due to this fact, we consider the *containment measure* (Lyon, Malcolm, and Dickerson, 2001), a value in the range of $[0, 1]$, to determine if a fragment s_i is plagiarised from d:

$$C(s_i, d) = \frac{|N(s_i) \cap N(d)|}{|N(s_i)|} \quad (9)$$

If the maximum $C(s_i, d)$, after considering every $d \in D'$, is greater than a given threshold, s_i is considered a candidate of plagiarism from d. The exhaustive text comparison process is resumed in Figure 3. Both, suspicious sentence and reference document, are represented as a bag of n-grams.

Algorithm 2: Exhaustive text comparison. Given s and D':

```
for each sentence $s_i \in s$
    $N(s_i)$ = [$n$-grams in $s_i$]
    for each $d \in D'$
        $N(d)$ = [$n$-grams in $d$]
        Calculate $C(N(s_i)\, N(d))$
if $(M = \max_{d \in D}(C(N(s_i), N(d)))) \geq thresh$
    $s_i$ is plagiarised from $\arg\max_{d \in D}(M)$
```

Figure 3: Exhaustive text comparison process.

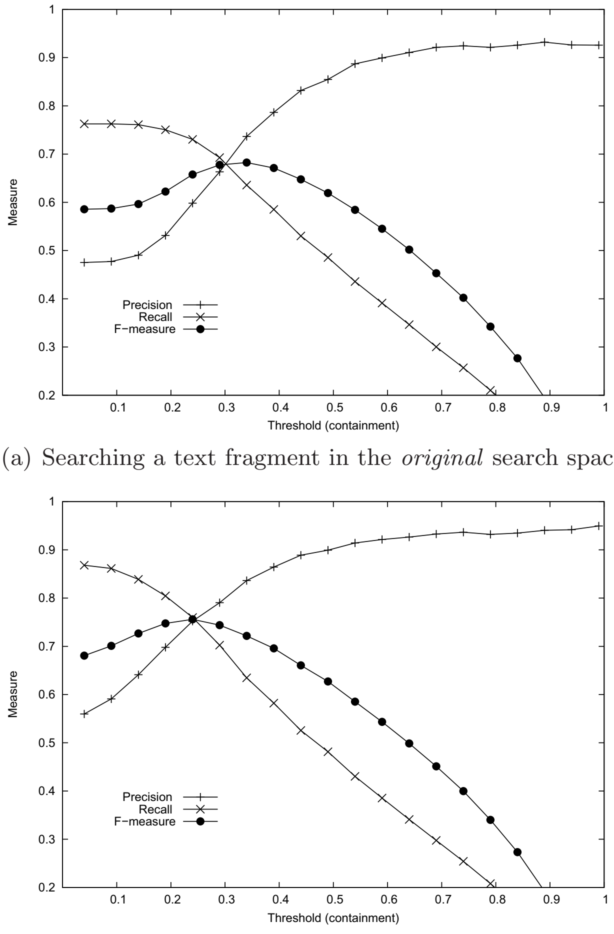
5 *Experiments Description*

The aim of our experiments is to investigate the impact of applying a preliminary search space reduction to a plagiarism detection method. In order to analyse this impact, given s and D, we have carried out two experiments considering different stages: (a) exhaustive text comparison; and (b) search space reduction + exhaustive text comparison. The exhaustive text comparison is based on the containment measure, while the search space reduction is based on the Kullback-Leibler distance.

The main purpose of these experiments is to compare how the quality of the obtained output is influenced by the reduction stage. As it has been pointed out (Section 3), the reference corpus is composed of 771 PA notes while the suspicious corpus contains 444 newspaper notes.

Different probability distributions have been considered in order to characterise the suspicious document s as well as the reference documents $d \in D$. We have considered multi-word terms with a length of n in the range $[1, \ldots, 5]$. As it is expected, the best result for the search space reduction has been obtained by considering $n = 1$. Higher n-gram levels produce distributions where a good part of the terms have a probability close to 1. These distributions (where almost all the terms have the same probability), do not allow KL_δ to effectively determine how close two documents are.

For the exhaustive text comparison, it has been previously shown that considering $n = 2$ gives the best results in terms of Precision and Recall (Barrón-Cedeño and Rosso, 2009). Comparable results have been reported by considering $n = 3$ (Lyon, Malcolm, and Dickerson, 2001; Barrón-Cedeño and Rosso, 2009). However, due to the fact that bigrams comparisons are simpler, we have opted for considering this n-gram level. By using a bag of words ($n = 1$), relevant factors for plagiarism detection such as plagiarism detection and writing style are eliminated. Additionally, the probability of a single word of appearing in an entire document is too high. Considering $n > 3$ would make the search strategy too rigid and a plagiarised sentence with just small changes would be more difficult to be detected.



(a) Searching a text fragment in the *original* search space

(b) Searching a text fragment in the *reduced* search space

Figure 4: Evaluation of searching results combining different options

Exp.	thresh	P	R	F
s, D *	0.34	0.74	0.62	0.68
s, D	0.34	0.73	0.63	0.68
s, D'	**0.25**	**0.77**	**0.74**	**0.75**

Table 2: P, R and F-measure obtained by the different experiments (results considering the best F-measure). The first row corresponds to the same experiment of the second one but without carrying out a stemming process.

Figure 4 compares the final detection results with and without search space reduction. The evaluation is carried out on the basis of the standard measures *P*recision, *R*ecall and F-measure. The figure shows the evolution of the three measures while the *detection threshold* is varied. As it is evident, Precision and Recall show a normal behaviour as the threshold increases: Precision increases while Recall decreases (both in a soft way). It can be seen that in general the curves corresponding to those experiments including the search space reduction stage evolve across higher values of P, R and hence F-measure. The best combinations of values obtained for this metrics are resumed in Table 2.

In the first experiment (Figure 4(a), second row of Table 2), an exhaustive comparison of $s_i \in s$ is carried out by considering the entire set of documents in the reference corpus D. The containment-based comparison technique obtains good results by itself (F-*measure* $= 0.68$ with *threshold* $= 0.34$). However, considering too many reference documents which are unrelated to the suspicious one, produces noise in the output, affecting P as well as R.

In the second experiment (Figure 4(b), third row of Table 2) the best F-measure is higher than in the previous one, which is obtained by considering a lower threshold. This behaviour is due to the fact that when s is compared to the entire corpus D, each s_i is compared to too many documents that are not even related to the topic of s (and that cannot be possible sources of it). However, common n-grams are found in documents of D which are not related at all with s. By reducing the set of potential sources in the reference corpus, less noisy comparisons are carried out, improving the Precision. Additionally, as the actual source documents of the plagiarism sentences are correctly retrieved during the reduction, the Recall is not affected. On the contrary, it is improved due to the fact that lower thresholds can be considered.

An additional experiment without stemming nor search space reduction is also reported (first row of the Table 2). Comparing the results to those obtained after stemming is quite interesting. There is no significant difference between the obtained results. The stemming process, that frequently improves the output of other information retrieval tasks, does not affect the results in this case. However, the search space reduction causes an important increase in the quality of the plagiarism analysis output.

The results displayed in Figure 4 correspond to the estimation stage of the experiment (variation of the threshold). However in a further experiment, based on a 5-fold cross-validation, the results obtained with the corresponding test sets did not present any variation with respect to those obtained during the estimation. This fact reflects the stability of the method.

A secondary aspect to consider is the processing time. Carrying out a search space reduction before the exhaustive search process causes an important decrease in the processing time. This is due to three main reasons: (1) the probability distribution P_d is pre-calculated for every reference document; (2) the probability distribution Q'_s is calculated only once and simply adapted to define the probability distribution Q_s given each P_d; and (3) s is compared to the reduced set D', which only contains the 10 of the original set of documents D with the highest probabilities of being the source of the potentially plagiarised sentences in s. The average time needed to analyse a suspicious document including search space reduction and exhaustive search over the minimised reference corpus is 10 times lower than carrying out an exhaustive text comparison of a suspicious document in the entire reference corpus.

6 Conclusions and Further Work

In this paper we have investigated the impact that the search space reduction may have on the task of automatic plagiarism detection with reference. The search space reduction method, based on the Kullback-Leibler symmetric distance, measures how close two probability distributions are. The probability distributions are composed of a set of unigram terms from the reference and suspicious documents.

In the experiments we carried out, a comparison of the obtained results was made (also in terms of time performance) employing a method that exhaustively searches for bigrams of the suspicious document in the reference corpus. When the search space reduction is applied, the entire reference corpus (composed of approximately 700 documents) is reduced to only 10 reference documents. In these optimised conditions, the quality of the obtained output in terms of Precision, Recall and therefore F-measure is importantly increased (particularly, F-measure increases from 0.68 to 0.75).

As future work, we would like to investigate the impact of the search space reduction in automatic plagiarism detection with larger documents and reference corpora including other registers such as scholar and literary.

References

Barrón-Cedeño, A. and P. Rosso. 2009. On Automatic Plagiarism Detection based on n-grams Comparison. In M. Boughanem, C. Berrut, J. Mothe, and C. Soule-Dupuy, editors, *Proceedings of the 31st Euro-*

pean Conference on IR Research, volume 5478 of *Lecture Notes in Computer Science*, pages 696–700, Toulouse, France. Springer.

Barrón-Cedeño, A., P. Rosso, and J.M. Benedí. 2009. Reducing the Plagiarism Detection Search Space on the Basis of the Kullback-Leibler Distance. In A. Gelbukh, editor, *Proceedings of the CICLing 2009*, volume 5449 of *Lecture Notes in Computer Science*, pages 523–534, Mexico city, Mexico. Springer.

Bigi, B. 2003. Using Kullback-Leibler Distance for Text Categorization. In *Proceedings of the 25th European Conference on IR Research*, volume 2633 of *Lecture Notes in Computer Science*, pages 305–319, Pisa, Italy. Springer.

Clough, P., R. Gaizauskas, and S. 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-02)*, volume V, pages 1678–1691, Las Palmas de Gran Canaria, Spain.

Kang, N., A. Gelbukh, and S. Han. 2006. PPChecker: Plagiarism Pattern Checker in Document Copy Detection. In *Proceedings of the TSD-2006: Text, Speech and Dialogue*, volume 4188 of *Lecture Notes in Artificial Intelligence*, pages 661–667, Brno, Czech Republic.

Kullback, S. and R. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

Lyon, C., R. Barrett, and J. Malcolm. 2004. A theoretical basis to the automated detection of copying between texts, and its practical implementation in the Ferret plagiarism and collusion detector. In *Proceedings of Plagiarism: Prevention, Practice and Policies Conference*, Newcastle, UK.

Lyon, C., J. Malcolm, and B. Dickerson. 2001. Detecting short passages of similar text in large document collections. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 118–125, Pennsylvania, USA.

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

Meyer zu Eißen, S. and B. Stein. 2006. Intrinsic Plagiarism Detection. In M. Lalmas, A. MacFarlane, S. Rüger, A. Tombros, T. Tsikrika, and A. Yavlinsky, editors, *Proceedings of the 28th European Conference on IR Research*, volume 3936 of *Lecture Notes in Computer Science*, pages 565–569, London, UK. Springer.

Porter, M.F. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Salton, G., E.A. Fox, and H. Wu. 1983. Extended Boolean Information Retrieval. *Communications of the ACM*, 26(11):1022–1036.

Si, A., H.V. Leong, and R.W.H. Lau. 1997. CHECK: a document plagiarism detection system. In *Proceedings of the 1997 ACM Symposium on Applied Computing*, pages 70–77, San Jose, CA.

Stein, B. 2007. Principles of Hash-based Text Retrieval. In *Proceedings of the 30th Annual International ACM SIGIR Conference*, pages 527–534, Amsterdam, Netherlands.

Stein, B., M. Koppel, and E. Stamatatos. 2007. Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN' 07). *SIGIR Forum*, 41(2):68–71.