On Automatic Plagiarism Detection Based on n-Grams Comparison

Alberto Barrón-Cedeño and Paolo Rosso

Natural Language Engineering Lab.
Dpto. Sistemas Informáticos y Computación,
Universidad Politécnica de Valencia, Spain
{lbarron,prosso}@dsic.upv.es
http://www.dsic.upv.es/grupos/nle/

Abstract. When automatic plagiarism detection is carried out considering a reference corpus, a suspicious text is compared to a set of original documents in order to relate the plagiarised text fragments to their potential source. One of the biggest difficulties in this task is to locate plagiarised fragments that have been modified (by rewording, insertion or deletion, for example) from the source text.

The definition of proper text chunks as comparison units of the suspicious and original texts is crucial for the success of this kind of applications. Our experiments with the METER corpus show that the best results are obtained when considering low level word n-grams comparisons ($n = \{2, 3\}$).

Keywords: Plagiarism detection, reference corpus, n-grams, information extraction, text reuse.

1 Introduction

Automatic plagiarism detection is mainly focused, but not limited to, academic environments. Plagiarise means including another persons text in the own work without the proper citation (the easy access to the information via electronic resources, such as the Web, represent a high temptation to commit it). Plagiarism based on verbatim copy is the easiest to detect. However, when a plagiarism case implies rewording (changing words by synonyms or changing the order of part of the text), the task becomes significantly harder.

In *plagiarism detection with reference*, the suspicious text fragments are compared to a reference corpus in order to find the possible source of the plagiarism cases. We have carried out experiments based on the exhaustive comparison of reference and suspicious word-level n-grams. The obtained results show that low values of n, except $n = 1$ (unigrams), are the best option to approach this task.

2 Method Description

2.1 Related Work

Some methods have been developed in order to find original-plagiarised text pairs on the basis of flexible search strategies (able to detect plagiarised fragments

M. Boughanem et al. (Eds.): ECIR 2009, LNCS 5478, pp. 696–700, 2009.

even if they are modified from their source). If two (original and suspicious) text fragments are close enough, it can be assumed that they are a potential plagiarism case that needs to be investigated deeper. A simple option is to carry out a comparison of text chunks based on word-level n-grams. In *Ferret* [4], the reference and suspicious texts are split into trigrams, composing two sets that are after compared. The amount of common trigrams is considered in order to detect potential plagiarism cases. Another option is to split the documents into sentences. *PPChecker* [2] detects potentially plagiarised sentences on the basis of the intersection and complement of the reference and suspicious sentences vocabulary. Considering complement avoids detecting casual common text substrings as plagiarism cases. In this work, the suspicious sentence vocabulary is expanded based on Wordnet relations.

Our approach is mainly based on a combination of the main principles of PPChecker and Ferret. However, as we describe in the following section, the word-level n-grams comparison is not carried out considering sentences or entire documents, but in an asymmetric way (i.e., suspicious sentence versus reference document.

2.2 Proposed Method

Given a suspicious document s and a reference corpus D, our objective is to answer the question "*Is a sentence $s_i \in s$ plagiarised from a document $d \in D$?*". We must consider that plagiarised text fragments use to appear mixed and modified. The n-gram based comparison attempts to tackle this problem. We consider n-grams due to the fact that independent texts have a small amount of common n-grams. For instance, Table 1 shows how likely is that different documents include a common n-gram (note that the analysed documents were written by the same author and on the same topic). It is evident that the probability of finding common n-grams in different documents decreases as n increases.

Table 1. Common n-grams in different documents (avg. words per document: 3,700)

Documents	1-grams	2-grams	3-grams	4-grams
2	0.1692	0.1125	0.0574	0.0312
3	0.0720	0.0302	0.0093	0.0027
4	0.0739	0.0166	0.0031	0.0004

Additionally, due to the fact that a plagiarised sentence could be made of fragments from multiple parts of an original document, the reference documents should not be split into sentences, but simply into n-grams. Our method is based on the next four considerations:

1. The suspicious document s is split into sentences (s_i);
2. s_i is split into word n-grams. The set of n-grams represents the sentence;
3. a document d is not split into sentences, but simply into word n-grams; and
4. each sentence $s_i \in s$ is searched singleton over the reference documents.

In order to determine if s_i is a candidate of being plagiarised from $d \in D$, we compare the corresponding sets of n-grams. Due to the difference in size of these sets, an asymmetric comparison is carried out on the basis of the *containment* measure [3]:

$$C(s_i \mid d) = \frac{|N(s_i) \cap N(d)|}{|N(s_i)|} \quad , \tag{1}$$

where $N(\cdot)$ is the set of n-grams in $(\cdot)$. If the maximum $C(s_i \mid d)$, after considering every $d \in D$, is greater than a given threshold, s_i becomes a candidate of being plagiarised from d.

3 Experimental Results

The aim of our experiments is to define the best n-gram level to detect plagiarism cases. We have proved n-gram levels in the range $[1, \cdots, 10]$. Subsection 3.1 describes the used corpus. The obtained results are discussed in Subsection 3.2.

3.1 The Corpus

In our experiments, we have used the XML version of the *METER corpus* [1]. This corpus is composed of news written by the Press Association (PA) as well as notes about the same news written by nine British newspapers. The newspapers are allowed to use the PA notes as a source for their own publications.

Around 750 PA notes compose our reference corpus. 444 from the 942 newspaper notes compose the suspicious documents set. We selected them because the fragments in their sentences are identified as *verbatim*, *rewrite* or *new*, for exact copy of the PA note, rewritten from the PA note or nothing to do with the PA note, respectively. A sentence s_i is considered plagiarised if a high percentage of its words belong to verbatim or rewritten fragments; in particular, if it fulfils the inequality $|s_{i_V} \cup s_{i_R}| > 0.4|s_i|$, where s_{i_V} and s_{i_R} are the words in verbatim and rewritten fragments in s_i, respectively. This estimation avoids considering sentences with incidental common fragments (such as named entities) as plagiarised. The distribution of verbatim, rewritten and new fragments in all the suspicious sentences is {43, 17, 39}%, respectively. When considering only the plagiarised sentences, it is {65, 26, 7}%.

The average number of words in the reference documents is 293 (330 for the suspicious ones). The reference corpus has a vocabulary of 18,643 words (14,796 for the suspicious one). The global vocabulary length is of 24,574 words. The pre-processing consists of words and punctuation marks splitting (for instance, "*cases, respectively.*" becomes "*cases , respectively .*") and stemming [5].

3.2 Obtained Results

In the experiments we carried out a 5-fold cross validation process. We have varied the containment threshold in order to decide whether a suspicious sentence is plagiarised or not. Precision, Recall and F-measure were estimated by considering 4 sets of suspicious documents. The threshold with the best F-measure t^* was after applied to the fifth set (unseen during the estimation). Fig. 1 shows the obtained results by considering n-grams with n in the range $[1, 5]$ (higher n values obtain worst results). Note that the results obtained by considering t^* over the test sets were exactly the same ones than those obtained during the estimation.

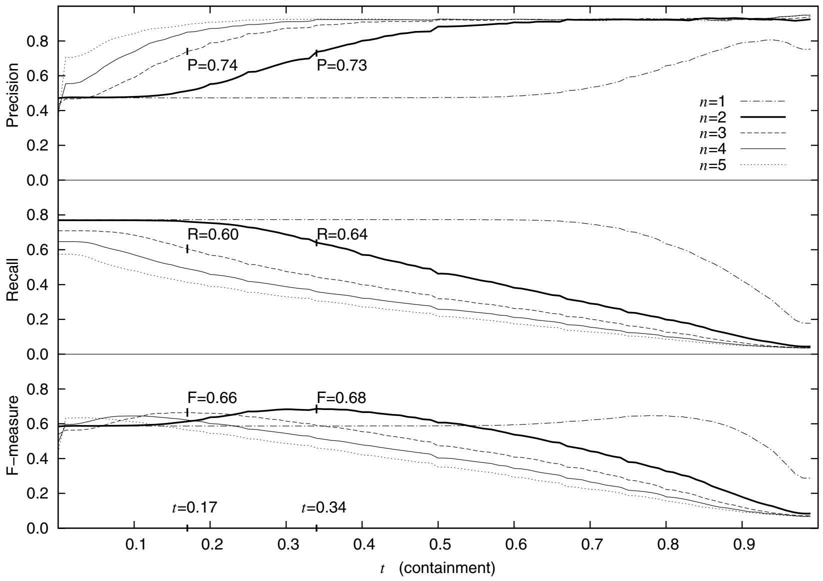

Fig. 1. Results considering different n-gram levels and $t$thresholds

Considering $n = 1$ (i.e., bag of words) a a good Recall is obtained (practically constant until $t = 0.7$). However, the probability of a document d of containing the entire vocabulary of a sentence s_i is too high. Due to this reason, the obtained Precision is the lowest among all the experiments. On the other side, considering $n = 4$ (and higher n) produces a rigid search strategy. Minor changes in a plagiarised sentence avoids its detection, resulting in the lowest Recall values.

The best results are obtained by considering $n = \{2, 3\}$ (best F-measures are 0.68 and 0.66, respectively). In both cases, the word n-grams are short enough to handle modifications in the plagiarised sentences and long enough to compose strings with a low probability of appearing in any (but the plagiarism source) text. Trigram based search is more rigid, resulting in a better Precision. Bigram based search is more flexible, allowing better Recall. The difference is reflected in the threshold where the best F-measure values are obtained for both cases: 0.34 for bigrams versus 0.17 for trigrams. Selecting bigrams or trigrams depends on the interest of catching as most as possible plagiarised fragments or leaving out some of them with the aim of after reviewing less candidates.

4 Conclusions

In this paper we have explored the search of plagiarism suspicious text over a reference corpus (commonly named plagiarism detection with reference). Our flexible search strategy is based on the asymmetric search of suspicious sentences across a set of reference documents (both codified as n-grams). Comparing sentences to entire documents becomes the search strategy even more flexible.

The experimental results show that bigrams and trigrams are the best comparison units for this task. Bigrams favour Recall while trigrams favour Precision, obtaining an F-measure of 0.68 and 0.66, respectively.

As future work, we would like to carry out some further experiments to extend the n-grams vocabulary in order to handle synonymic and other kinds of words substitutions. Additionally, results should be validated by considering another kind of documents (i.e., not necessarily journalistic notes), such as student papers.

Acknowledgements. We would like to thank Paul Clough for providing us the METER corpus. This work was partially funded by the MCyT TIN2006-15265-C06-04 research project as well as the CONACyT-MEXICO 192021/302009 grant have partially funded this work.

References

1. Clough, P., Gaizauskas, R., Piao, S.: Building and Annotating a Corpus for the Study of Journalistic Text Reuse. In: 3rd International Conference on Language Resources and Evaluation (LREC 2002), V, pp. 1678–1691. Las Palmas, Spain (2002)
2. Kang, N., Gelbukh, A.: PPChecker: Plagiarism Pattern Checker in Document Copy Detection. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2006. LNCS, vol. 4188, pp. 661–667. Springer, Heidelberg (2006)
3. Lyon, C., Malcolm, J., Dickerson, B.: Detecting Short Passages of Similar Text in Large Document Collections. In: Conference on Empirical Methods in Natural Language Processing, Pennsylvania, pp. 118–125 (2001)
4. Lyon, C., Barrett, R., Malcolm, J.: A Theoretical Basis to the Automated Detection of Copying Between Texts, and its Practical Implementation in the Ferret Plagiarism and Collusion Detector. In: Plagiarism: Prevention, Practice and Policies Conference, Newcastle, UK (2004)
5. Porter, M.F.: An Algorithm for Suffix Stripping. Program 14(3), 130–137 (1980)