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.