Optimized Fuzzy Text Alignment for Plagiarism Detection
Notebook for PAN at CLEF 2012

Fernando Sánchez-Vega, Manuel Montes-y-Gómez and Luis Villaseñor-Pineda

Laboratorio de Tecnologías del Lenguaje
Instituto Nacional de Astrofisica, Óptica y Electrónica (INAOE), México
{fer.callotl,mmontesg,villasen}@inaoep.mx

Abstract. This paper describes a method for plagiarism detection based on a fuzzy alignment between a given pair of documents. The proposed method assigns a weight to each word of the suspicious document according to the straightness of its alignment to the source document; this weight is used as a kind of plagiarism probability measure for each word of the suspicious document. The paper also presents a strategy to optimize the alignment of the two documents based on the evaluation of all possible matches in a limited context. Evaluation results on the test set of the PAN corpus show that the method is relatively fast and that it could detect 35% of the plagiarized words with accuracy greater than 50%.

1 Introduction

Copying another author's text and claiming its authorship is called plagiarism]1]. Current research on automatic plagiarism detection has mainly focused on two tasks: candidate document retrieval and detailed comparison. While the first task consists in retrieving –from the Web– a set of candidate source documents for a given suspicious document, the second task considers the identification of all plagiarized passages from the suspicious document and their corresponding passages from the source document.

In this paper we describe a method for the detailed comparison task carried out at the PAN 2012 competition. The proposed method is supported on the premise that a strong alignment between two –different– documents is an indicator of plagiarism, whereas the lack of a strong alignment on them points to a plagiarism free situation. The following section briefly describes the main parts of the proposed method.

2 Our plagiarism detection method

The alignment of the documents is carried out by a method that matches the words from the suspicious and source documents; the straighter the alignment, the greater the probability that plagiarism has occurred. The method has two main modules. The

first module assigns a weight to each word from the suspicious document that indicates its probability to belong to a plagiarized section from the source document. This weight is computed by a fuzzy alignment strategy optimized by the use of multiple exploratory particles. The second module determines the plagiarized sections of the documents by applying some threshold functions on the weights computed by the first module. Sections 2.1 and 2.2 describe in detail these two modules.

2.1 Fuzzy document alignment

The fuzzy alignment strategy matches the words from the suspicious and source documents without considering any word order restriction. It assigns a weight to each word from the suspicious document as indicated by the following formula:

$$P(w_i) = \frac{1}{1 + \left| pos_O^h(w_i) - pos_O^r(w_i) \right|}$$

where $P(w_i)$ indicates the probability that word w_i has been plagiarized from the source document, $pos_O^h(w_i)$ is the hypothetical position[1] that w_i should have in the source document to be part of a copy-&-paste sequence, and $pos_O^r(w_i)$ is the real position of w_i in the source document.

The "plagiarism probability" of each word, $P(w_i)$, is equivalent to the inverse of the distance between the current position of the word in the source document and its expected position caused by a copy-&-paste action. Intuitively, $P(w_i)$ measures the degree of change needed in the suspicious document to produce an exact copy of the source document. A special case is when w_i does not occur in the source document; it is assumed that $pos_O^r(w_i) \to \infty$, and therefore $P(w_i) = 0$.

The position $pos_O^h(w_i)$ is computed using the last word evaluated before w_i (i.e., w_{i-1}). Basically, $pos_O^h(w_i)$ is the next position after $pos_O^r(w_{i-1})$ if and only if $P(w_{i-1}) \neq 0$ as it is showed in the following formula.

$$pos_O^h(w_i) = \begin{cases} pos_O^r(w_{i-1}) + 1 & if\ P(w_{i-1}) \neq 0 \\ pos_O^h(w_{i-1}) & other\ case \end{cases}$$

On the other hand, the position $pos_O^r(w_i)$ has multiple possible values when w_i occurs several times in the suspicious document. In order to select one value from all possible values for $pos_O^r(w_i)$, the most straightforward strategy is to choose the $pos_O^r(w_i)$ that maximizes the value of $P(w_i)$. The problem with this kind of solution is that it provides a local maximum, which often is far from the global maximum. To tackle this problem we considered an optimization strategy based on the use of multiple exploratory particles.

[1] Positions are expressed in number of words from the beginning of the documents.

Optimization based on multiple exploratory particles

To get the best alignment of the two given documents it is necessary to select the $pos_O^r(w_i)$ values that maximize all $P(w_i)$ values for the whole suspicious document. Unfortunately, the achievement of this goal is extremely expensive, and, therefore, we use an optimization strategy that considers the best choice of $pos_O^r(w_i)$ after examining all options for the k next words using multiple particles to perform the exploration. This strategy allows determining the best $pos_O^r(w_i)$ value in the context $w_i,\ldots,w_{i+k}$ as indicated by the following formula.

$$pos_O^r(w_i) = \underset{pos_O^r(w_i)}{argmax} \sum_{j=i+k\ldots i} P(w_j)$$

The optimization strategy requires multiple exploratory particles to handle the cases when a word in the search context has more than one possible $pos_O^r(w_i)$ value. In these cases, there is a bifurcation in the possible path of alignments to follow, and several particles are needed to explore all different paths. That is, during the exploration: each particle add the $P(w_i)$ value of the path that it follows after exploring the next k words; the particle getting the largest result is selected and the particles that did not follow the same *first* bifurcation are removed; then, the process continues determining the alignment of the following suspicious word w_{i+1}.

2.2 Determining of plagiarized sections

This module applies some heuristic functions on the $P(w_i)$ values in order to determine the plagiarized sections from the suspicious document. It first smoothes the $P(w_i)$ values computing the average over 40 words around w_i; the new values are represented by $P'(w_i)$. Then, it evaluates each word w_i and defines it as a plagiarized word if it satisfied the following three conditions:

I. All the words in Ψ_i have $P' > \alpha$
II. At least N words in Ψ_i have $P' > \beta$
III. At most M words in Ψ_i have $P' < \gamma$

where Ψ_i indicates a window of 21 words centered at w_i, and α, β, and γ are thresholds satisfying $\beta > \alpha > \gamma$.

Finally, all sequences of consecutive plagiarized words with length greater than 100 characters are determined as the plagiarized sections.

3 Experiments

For the experiments, we tuned the parameters using a set of 20 pairs of documents that contain examples of paraphrased plagiarism; the resulting values were: $\alpha = 0.01$, $\beta = 0.15$, $\gamma = 0.05$, $N = 4$, $M = 7$. Table 1 shows the results of the proposed method on the training and test collections. Details on these collections can be found at http://pan.webis.de/.

Table 1. Results of the proposed method on the PAN 2012 collections

Corpus	PlagDet	Precision	Recall	Granularity
Training	0.273	0.741	0.353	1.872
Test (competition)	0.309	0.537	0.349	1.577

4 Conclusions

This paper described the method developed by the Laboratory of Language Technologies from INAOE for the document detail comparison task at PAN 2012. The proposed method is based on a fuzzy alignment of a pair of documents. It mainly computes a plagiarism probability for each word of the suspicious document which, intuitively, indicates the degree of change needed in this document to produce an exact copy of the given source document. One important contribution of this work is the use of a search strategy guided by several exploratory particles that allow us to get semi-optimal alignments between the documents.

An initial analysis of the experimental results showed that our method was good in locating plagiarized sections in the suspicious documents but was not effective in determining the exact fragments from the source documents that were plagiarized. In addition, these results also showed that our method, although based on an expensive optimized search process, was faster that other approaches in completing the detection task.

Acknowledgements: This work was done under partial support of CONACYT (project grants 134186, 106013 and scholarship 258345/224483).

References

1. Martin Potthast, Andreas Eiselt, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. Overview of the 3rd International Competition on Plagiarism Detection. Working Notes of the Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse, PAN 2011. Amsterdam, Netherlands, September 2011.