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 Astrofísica, Óptica y Electrónica (INAOE)

Luis Enrique Erro 1, Sta. María, Tonanzintla, Puebla, 72840, México

{fer.callotl,mmontesg,villasen}@inaoep.mx

This paper presents the method developed by the Laboratory of Language Technologies from INAOE-Mexico for the document detail comparison task carried out at the PAN 2012 plagiarism detection competition. The proposed method is supported on the premise that a strong alignment between the suspicious and source documents is a clear indicator of the existence of plagiarism. It is mainly based on a fuzzy alignment strategy that allows matching the words from both documents without considering any word order restriction.

The proposed 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 as 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, it measures the degree of change needed in the suspicious document to produce an exact copy of the source document. Given the computational cost of finding an optimal alignment between the documents, we proposed an optimization strategy that examines local context by multiple exploration particles. The second module of the method determines the plagiarized sections of the documents by applying some threshold functions on the weights computed by the first module. At the end, it defines the sequences of consecutive –highly probable- plagiarized words with length greater than 100 characters as the plagiarized sections.

Evaluation results on the test set of the PAN corpus showed that the method could detect 35% of the plagiarized words with accuracy greater than 50%. However, they also showed that it was not effective in determining the exact fragments from the source documents that were plagiarized. In addition, and surprisingly, evaluation results also indicated that our method, although based on an expensive optimized search process, was faster that other approaches in completing the detection task.