A Winning Approach to Text Alignment for Text Reuse Detection at PAN 2014

Miguel A. Sanchez-Perez, Grigori Sidorov, and Alexander Gelbukh

Instituto Politécnico Nacional, Centro de Investigación en Computación, Juan de Dios Batiz, 07738 Mexico City, Mexico
masp1988@hotmail.com, sidorov@cic.ipn.mx, www.gelbukh.com

The task of monolingual text alignment consists in finding similar text fragments between two given documents. This task has numerous applications, such as in plagiarism detection, detection of text reuse, author identification, authoring aid, and information retrieval, to mention only a few.

Our method relies on a sentence similarity measure based on a tf-idf-like weighting scheme that permits us to keep stopwords without increasing the false positives rate. The similarity is computed representing individual sentences with a tf-idf vector space model (VSM), as if each sentence were, in terminology of VSM, a separate "document" and all sentences in the pair of suspicious and source document formed a "document collection." The idf measure calculated in this way is called isf measure (inverse sentence frequency), to emphasize that it is calculated over sentences and not documents.

We introduce a recursive algorithm to extend the matching sentences to maximal length passages. The algorithm consists in forming corresponding text fragments, called plagiarism cases, composed by several adjacent matching sentences, with possible gaps no greater than a given threshold. For those plagiarism cases that have insufficient similarity between the corresponding fragments, the algorithm is executed recursively in order to split them into smaller but higher-quality plagiarism cases.

We also introduce a novel filtering method to resolve overlapping plagiarism cases. We call two plagiarism cases overlapping if their fragments from the suspicious document share at least one sentence: each suspicious sentence can be plagiarized from only one source, and thus can only belong to one plagiarism case. Thus, of a set of overlapping cases we keep only one. For this, we introduce two quality measures for plagiarism cases, based on the similarity and the size of the corresponding text fragments.

In addition, we introduce dynamic adjustment of the parameters (most importantly, the maximum allowable gap size) depending on the type of plagiarism case at hand. We use two different sets of parameters for plagiarism cases with summary obfuscation and for all other types. We detect whether a particular case is of summary obfuscation type basing on the sizes of the corresponding fragments.

By the cumulative measure Plagdet, our approach outperforms the best-performing system of the PAN 2013 competition, obtaining Plagdet 0.90032 (precision 0.97853, recall 0.83369, granularity 1.00000), and resulted in the best-performing (on the first corpus, Plagdet 0.87818) and third best-performing (on the second corpus, Plagdet 0.89197) system according to the official results of PAN 2014.