Chapter 22
Hybrid Segmentation Prototype for Arabic Text-Based Documents: Towards Plagiarism Detection

Sonia Alouane-Ksouri
Université de Tunis El Manar, Tunisia

Minyar Sassi Hidri
Université de Tunis El Manar, Tunisia

ABSTRACT

The contribution of this work relates to the field of Arabic text-based document analysis for the detection of plagiarism. This analysis will be carried out according to the triadic computation model of document similarity. The authors propose a hybrid segmentation prototype for Arabic text-based documents that links different processing steps in order to generate the similarity rate between the documents of an Arabic corpus. It involves two segmentation systems and a morphological analysis in order to obtain a matrix representation adapted to the triadic similarity computation according to three abstraction levels: documents, sentences and words.

INTRODUCTION

The use of Internet and other networks makes for easy access to information, but it also makes plagiarism an easy operation for students. There are several types of plagiarism, from direct copying of sentences or passages out of a published text without citing the sources, to plagiarism of ideas, sources, and authorships (Stamatatos, 2011).

The detection of plagiarism is of most importance particularly to protect the intellectual properties of productions issued from teaching, research or industry (Debili & Achour, 1998; Bao & Malcolm, 2006).

DOI: 10.4018/978-1-5225-8057-7.ch022

Many language-sensitive prototypes for detecting plagiarism in natural language documents have been developed, particularly for English and French. Language independent tools exist as well, but are considered restrictive as they usually do not takeinto account specific language features.

Much research has been done, and tools proposed to find similar documents. They are mainly based onfinding identical so-called n-grams (Kent et al., 2009; Oberreuter et al., 2010; Stamatatos 2011; Jaballah et al., 2012; Menai, 2012) to detect copies.

The search for documents is carried out according to two approaches. The first one, which is based on the author's style, focuses on looking for a document presenting the same or a similar style, basing our search on the sentence structure, grammar or stylistic observation (Stamatatos, 2009).

The second approach is based on the content of the document (Eissen & Stein, 2006). It consists insearching for exact or very close similarities between documents through content analysis by comparing the characteristic features extracted from each one. The specific features often analyzed are terms, spelling mistakes, word patterns, or exact sequences. This approach can be used for identification and extraction of plagiarism in a multilingual setting.

Although plagiarism detection in Arabic natural language documents is important in schools and universities in Arab countries, yet there are no known techniques to detect plagiarism in Arabic scripts. Detecting plagiarism in Arabic documents is a particularly challenging task because of the complex linguistic structure of the Arabic language (Farghaly & Shaalan, 2009).

In this work, we are interested in analyzing Arabic text-based documents for the detection of plagiarism. We focus on the particularities of the Arabic language by proposing a hybrid segmentation prototype. It is based on a combination of the STAr (Belguith et al., 2005) and MORPH-2 (Kammoun et al., 2010) systems that would provide us matrix representations adapted to the triadic document similarity computing proposed in (Sassi-Hidri et al., 2014).

The rest of the paper is organized as follows: section 2 presents the particularities of Arabic text. Section 3 gives an overview of related work on Arabic text analysis. Section 4 briefly presents the triadic computing model of document similarity. In section 5, we present our hybrid segmentation prototype for Arabic text-based documents. Section 6 presents an extension of the segmentation model to detect plagiarism. Section 7 concludes the paper and highlights future work.

PARTICULARITIES OF ARABIC TEXT

In order to clearly identify this field of application, we give a brief overview of the particularities of an Arabic text: it is read and written from right to left, it lacks vowels and punctuation, the words are characterized by agglutination and the word order in the sentence by irregularity.

Lack of Vowels

Arabic text is edited in vowelized form when there are notes, in the form of diacritic signs above or below the letters. The problem is in the virtual systemic absence of vowelization in Arabic text, except in certain text (Qur'an, Hadith) or in literature.

The example of the non-vowelized word: 'بتك' possesses many possible vowelizations, and that represent different grammatical categories (Debili & Achour, 1998). Table 1 represents an example of a vowel.