Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Candidate document retrieval for cross-lingual plagiarism detection using two-level proximity information

Nava Ehsan[a], Azadeh Shakery[a,b,*]

[a] School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran
[b] School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Iran

ARTICLE INFO

Article history:
Received 21 February 2015
Revised 11 April 2016
Accepted 18 April 2016
Available online xxx

Keywords:
Candidate document retrieval
Cross-language plagiarism detection
Text segmentation
Proximity-based retrieval

ABSTRACT

The rapid growth of documents in different languages, the increased accessibility of electronic documents, and the availability of translation tools have caused cross-lingual plagiarism detection research area to receive increasing attention in recent years. The task of cross-language plagiarism detection entails two main steps: candidate retrieval and assessing pairwise document similarity. In this paper we examine candidate retrieval, where the goal is to find potential source documents of a suspicious text. Our proposed method for cross-language plagiarism detection is a keyword-focused approach. Since plagiarism usually happens in parts of the text, there is a requirement to segment the texts into fragments to detect local similarity. Therefore we propose a topic-based segmentation algorithm to convert the suspicious document to a set of related passages. After that, we use a proximity-based model to retrieve documents with the best matching passages. Experiments show promising results for this important phase of cross-language plagiarism detection.

1. Introduction

Plagiarism refers to unauthorised use of text, code and ideas (Potthast, Barrón-Cedeño, Stein, & Rosso, 2011). In automatic cross-language plagiarism detection, the task is to retrieve plagiarized text written in language L that has originated from another document in a language other than L. With the rapid growth of documents in different languages, the increased accessibility of electronic documents, and the availability of translation tools, cross-language plagiarism has become a serious problem and its detection requires more attention.

Given a suspicious document s' and a set of potential source documents D, we should determine whether a fragment of the suspicious document, $s'_{f'} \in s'$, was borrowed from a source document. This task comprises two main steps: candidate retrieval and detailed analysis. Candidate retrieval entails the identification of source documents that contain suspicious fragments. Detailed analysis requires closer comparison of the subject document with each suspected source and retrieval of plagiarized fragments. In this paper we focus on the first step, candidate document retrieval. Since a second phase will follow this step to eliminate false positive matches, we are more interested in high recall than in high precision in this research.

[*] Corresponding author. Tel.: +00982182089722.
E-mail addresses: n.ehsan@ece.ut.ac.ir (N. Ehsan), shakery@ut.ac.ir (A. Shakery).

http://dx.doi.org/10.1016/j.ipm.2016.04.006
0306-4573/© 2016 Elsevier Ltd. All rights reserved.

In cross-language plagiarism detection the languages of source and suspicious documents differ. To date only a few approaches have been focused on cross-language plagiarism detection (Barrón-Cedeño, Gupta, & Rosso, 2013a). Most previous methods are based on translating the whole suspicious or source documents coupled with monolingual techniques (Barrón-Cedeño et al., 2013a). Document translation depends on the existence and quality of machine translators. Translating documents in languages with low quality translation tools may cause poor quality documents. In this paper we propose an approach for the candidate retrieval phase of cross-language plagiarism detection which only considers a set of representative words and phrases extracted from each document as its content representation, instead of using the whole text. Since documents are represented by some extracted words and phrases, this approach is insensitive to punctuation, extra white space, and permutation of the document context and requires less translation time rather than translating the entire document. Our approach is therefore less dependent on the quality of machine translation between two languages, and if there is not a high quality translation tool available, any other translation resources such as dictionaries, parallel or comparable corpora could be used for translating representative words. Thus, our approach is applicable in languages with even limited translation resources.

Since plagiarism usually happens in parts of the text, there is a requirement to segment the texts into fragments to detect local similarity. There are some previous works that break the document into constituent parts such as sections, paragraphs (Nawab, 2012), or a fixed number of sentences (Pereira, Moreira, & Galante, 2010). In this paper a topic-based text segmentation approach is proposed in order to break the document based on its topical structure. Thus, a set of topically related passages from the suspicious document are used to retrieve potential sources.

In our proposed candidate retrieval process, after segmentation, we use a second level for considering proximity in retrieval of candidate documents. For each segment the word proximity is measured with positional language modelling (PLM) (Lv & Zhai, 2009). We believe this to be the first use of PLM in cross-language plagiarism detection.

We present the results of no-segmentation with a non-proximity-based language model as a baseline. According to the candidate document retrieval experiments, the segmentation technique increased F_2 measure about 0.11 (21% improvement) over the baseline. Accompanying the segmentation technique with the positional language model increased F_2 measure about 0.13 (25% improvement) over the baseline. These results are further compared with CL-CNG (Mcnamee & Mayfield, 2004) and a combination of translation and monolingual analysis. The proposed approach with text segmentation and using proximity-based retrieval outperforms these approaches with respect to F_2.

The rest of the paper is organized as follows: Section 2 outlines related work in cross-language plagiarism detection. Section 3 describes the candidate document retrieval process in which the text segmentation approach, representative word extraction, and retrieval model are explained. Finally the experimental framework and results are discussed in Section 4, and our conclusion and future work are reported in Section 5.

2. Related work

Plagiarism detection methods can be classified into two approaches, intrinsic and external (Potthast et al., 2012). Intrinsic detection methods are those that use style analysis to detect parts of the text that are inconsistent in terms of writing style (Meyer zu Eißen & Stein, 2006; Oberreuter & Velásquez, 2013). The aim of external plagiarism detection methods is not only finding the suspicious text, but also finding the source for the plagiarized text. In monolingual plagiarism detection, parts of the suspicious text could be an exact copy or a modified copy, and those parts should be large enough to be more than just a coincidence. In cross-language plagiarism, the suspicious document in language L originates from another document in a language other than L. The plagiarized fragment could be the exact translation or a paraphrased translation.

In order to detect cross-lingual plagiarism, either cross-language similarity is used or the document is translated and monolingual similarity is used. There are five kinds of cross-language similarity assessment models that have been proposed in the literature (Franco-Salvador, Gupta, & Rosso, 2013; Potthast et al., 2011), (1) syntax-based, (2) dictionary-based, (3) parallel corpus-based, (4) comparable corpus-based, and (5) multilingual semantic network based approaches.

Syntax-based methods rely on lexical similarities between languages. Cross-Language Character N-Gram (CL-CNG) is a syntax-based method (Mcnamee & Mayfield, 2004). In this model, documents are represented by overlapping character n-grams. Defining the alphabet $\sum$ and n, the texts will be coded into character n-grams. The resulting texts are compared by means of similarity measures. The model is useful for comparing multilingual documents without translation, and is applicable for languages with similar syntax (Potthast et al., 2011), but it is ineffective when the languages differ syntactically.

Pouliquen et al. propose a dictionary based method to find similar documents in a multilingual document collection (Pouliquen, Steinberger, & Ignat, 2006). They map the document content to a vector of descriptors from the Eurovoc thesaurus and measure the semantic similarity between the resulting vectors. In this work the authors assume that the documents are completely similar, whereas in plagiarism detection the similarity could happen in parts of the text only.

CL-ASA (Barrón-Cedeño, Rosso, Pinto, & Juan, 2008), LSI (Dumais, Letsche, Littman, & Landauer, 1997) and KCCA (Vinokourov, Cristianini, & Shawe-taylor, 2002) use parallel corpora in order to find cross-language similarity, while Cross-Language Explicit Semantic Analysis (CL-ESA) (Potthast, Stein, & Anderka, 2008), which is the cross-lingual generalization of ESA (Gabrilovich & Markovitch, 2007), uses comparable corpora for this purpose. CL-ASA (Barrón-Cedeño et al., 2008) and CL-CNG (Mcnamee & Mayfield, 2004) are compared in (Barrón-Cedeño et al., 2013a) for document-level retrieval when suspicious documents are entirely plagiarised from sources.

Please cite this article as: N. Ehsan, A. Shakery, Candidate document retrieval for cross-lingual plagiarism detection using two-level proximity information, Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.04.006