Information Processing and Management 54 (2018) 408–432

Contents lists available at ScienceDirect

Information Processing and Management

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

Unmasking text plagiarism using syntactic-semantic based natural language processing techniques: Comparisons, analysis and challenges

Vani K.[a], Deepa Gupta[b,*]

[a] Department of Computer Science and Engineering, Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, Amrita University, India
[b] Department of Mathematics, Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, Amrita University, India

ARTICLE INFO

Keywords:
Natural language processing
Plagiarism detection
Syntactic-semantic
POS tagging
Chunking
Semantic role labelling

ABSTRACT

The proposed work aims to explore and compare the potency of syntactic-semantic based linguistic structures in plagiarism detection using natural language processing techniques. The current work explores linguistic features, viz., part of speech tags, chunks and semantic roles in detecting plagiarized fragments and utilizes a combined syntactic-semantic similarity metric, which extracts the semantic concepts from WordNet lexical database. The linguistic information is utilized for effective pre-processing and for availing semantically relevant comparisons. Another major contribution is the analysis of the proposed approach on plagiarism cases of various complexity levels. The impact of plagiarism types and complexity levels, upon the features extracted is analyzed and discussed. Further, unlike the existing systems, which were evaluated on some limited data sets, the proposed approach is evaluated on a larger scale using the plagiarism corpus provided by PAN[1] competition from 2009 to 2014. The approach presented considerable improvement in comparison with the top-ranked systems of the respective years. The evaluation and analysis with various cases of plagiarism also reflected the supremacy of deeper linguistic features for identifying manually plagiarized data.

1. Introduction

The rapid evolution of information technology (IT) and the easy access of information through various means such as search engines, online databases and digital libraries, on the negative side have embarked unethical practices in academic and research domains. The unethical use and publishing of a third party's content or work as one's own is termed as '*plagiarism*', which is defined as a serious intellectual and academic transgression.[2] Plagiarist tries to manipulate and breach the information content either by simple copy–paste or by paraphrasing and intelligently obfuscating the content. The text can be manipulated intelligently through various ways such as synonym replacements, restructuring, idea adoptions, translations, summarizations, etc. Existing studies and surveys reveal that the effectiveness of available plagiarism detection tools usually drops with the increase in plagiarism complexity (Alzahrani, Salim, & Abraham, 2012; Chong et al., 2013; Clough, 2000; Marsh, 2004; Maurer, Kappe, & Zaka, 2006; Mozgovoy, Kakkonen, & Cosma, 2010; Potthast, Barrón-Cedeño, Eiselt, Stein, & Rosso, 2010; Vani & Gupta, 2016). These studies in turn suggest the need for syntactic002Dsemantic document processing mechanisms and effective similarity metrics for unmasking plagiarism

* Corresponding author.
E-mail addresses: k_vani@blr.amrita.edu (V. K.), g_deepa@blr.amrita.edu (D. Gupta).
[1] http://pan.webis.de.
[2] http://dictionary.reference.com/browse/plagiarism.

https://doi.org/10.1016/j.ipm.2018.01.008
Received 19 January 2017; Received in revised form 13 December 2017; Accepted 21 January 2018

cases beyond copy–paste and simple manipulations. For tackling higher levels of manipulations, natural language processing (NLP) techniques can be utilized. The proposed work focuses to explore this aspect, by exploiting the syntactic-semantic linguistic information extracted using NLP techniques for document processing and similarity metric computations. Mainly plagiarism detection is divided into two formal tasks, which are extrinsic detection and intrinsic detection (Potthast, Barrón-Cedeño et al., 2010). The former compares the suspicious document against a source data base or a reference collection, which is supposed to contain the sources of suspected case. This can be a domain specific data base or some offline/ online databases or the entire World Wide Web (WWW). In intrinsic detection task, author's writing styles are used for plagiarism identification, in the absence of any source collection. In a plagiarism detection system (PDS) with extrinsic detection, the major stages include: pre-processing, candidate retrieval, detailed or passage level analysis and post processing steps (Alzahrani et al., 2012; Potthast, Barrón-Cedeño et al., 2010). In the pre-processing stage, the source and suspicious documents are subjected to some basic processing, which may include sentence segmentation, tokenization, punctuation removal, lowercasing, etc. This may also include shallow NLP techniques, viz., lemmatization, stemming, etc. In the second stage termed as candidate or source retrieval, the near duplicate sources related to the suspicious document at hand are retrieved, which forms the candidate set. This is followed by the main stage wherein detailed analysis of a suspected document is done against its candidate set to identify the suspicious-source plagiarism fragment pairs. Finally, post-processing steps are carried out. In the proposed work, the main focus is given on the exploration of different features for improving the effectiveness of detailed analysis stage in extrinsic text plagiarism detection.

One of the standard and widely used data sets for evaluation of plagiarism detection approaches is the PAN corpus. PAN is an international competition held in plagiarism detection research domain since 2009 and it had released data for detailed or passage level plagiarism analysis from 2009 to 2014. The existing systems were usually evaluated on PAN data of some specific years, whereas the proposed approaches are evaluated on a larger pool of data using the PAN data from 2009 to 2014 (Potthast, Barrón-Cedeño et al., 2010; Potthast et al., 2012, 2013, 2014; Potthast, Eiselt, Barrón-Cedeño, Stein, & Rosso, 2011; Potthast, Stein, Eiselt, Barrón-Cedeño, & Rosso, 2009). In PAN 2009 to 2011, data released was for both candidate retrieval and detailed analysis tasks, which constitutes the two major modules of extrinsic PDS. Thus in these tasks, initially candidate retrieval must be done to identify the near duplicate candidate pairs, which will be followed by the detailed analysis stage. While from 2012 to 2014, the data for these two modules were released as two separate tasks, viz., source retrieval and text alignment tasks. The former refers to online candidate source retrieval while latter is for passage level analysis, where the suspicious-source document pairs are already known. In the proposed work, the data for extrinsic plagiarism detection from 2009 to 2011 and the data for text alignment task from 2012 to 2014 are used, since our main focus is on passage level detections. For entailing fair comparison with the PAN systems reported in 2009 to 2011, a candidate document detection approach is implemented for the respective years. A two-phase approach that uses Vector Space Model (VSM) with information retrieval (IR) ranking approach in first phase and trigram comparisons in second is employed (Vani & Gupta, 2017a), which will be followed by the proposed detailed detection approach.

The main objective of this work is to explore the efficacy of syntactic-semantic linguistic structures, viz., POS tags, chunks and semantic roles for detailed passage level detections. A combined syntactic-semantic similarity metric that utilizes the linguistic information and extracts semantic concepts from WordNet[3] is utilized. Using these document features, the proposed approach also intend to filter out semantically irrelevant comparisons and hence increasing the detection efficiency. Another contribution of current work is the analysis of the dependency of NLP techniques, upon the level and type of plagiarism imposed in the document. This analysis is important, since the efficacy of any NLP based approach in plagiarism detection cannot be solely determined by the performance values (accuracy, *F*-measure) reported. Further, to analyze the performance variations with different degrees of simulated (manual) plagiarism cases, imposed by paraphrasing, evaluation is conducted on P4P corpus (Barrón-Cedeño, Vila, Martí, & Rosso, 2013). The major objectives focused in the proposed work are summarized as follows:

- Proposing a plagiarism detection approach that utilizes the syntactic-semantic linguistic information using NLP techniques for document processing. Three NLP techniques, viz., Part of Speech (POS) tagging, chunking and semantic role labelling (SRL) are compared and analyzed. Two more adjuncts, viz., chunking with POS tagging and SRL with POS tagging are explored which paves stricter comparisons.
- Proposing a combined syntactic-semantic similarity metric and utilizing the linguistic information for availing semantically meaningful comparisons. Thus, in this work the fragments that belong to same class, chunks or roles are only compared based on NLP technique utilized.
- Evaluating the proposed approach on a larger pool of diverse plagiarism cases provided by PAN across the years (2009 to 2014). This helps us to evaluate the approaches on plagiarized instances of various complexities and understand the performance variations and impacts.
- Performance analysis with different plagiarism types and obfuscation degrees to evaluate the behavioral variations and other dependencies.

In Section 2, the work reported by eminent researchers in this domain is presented. Section 3 describes the general workflow of the proposed PDS and the various modules in it. Statistics on the evaluation datasets and performance metrics are presented in Section 4. Section 5 describes the experimental results and the comparisons with different baselines. Section 6 reports the detailed analysis of the proposed approach based on various plagiarism conditions. In Section 7, the work is concluded and future insights are presented briefly.

[3] https://wordnet.princeton.edu/.