Journal of King Saud University – Computer and Information Sciences (2015) xxx, xxx–xxx

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE

Uncovering highly obfuscated plagiarism cases using fuzzy semantic-based similarity model

Salha M. Alzahrani [a,*], Naomie Salim [b], Vasile Palade [c]

[a] College of Computers and Information Technology (CIT), Taif University, Taif, Saudi Arabia
[b] Faculty of Computer Science and Information Systems, University of Technology Malaysia, Johor, Malaysia
[c] Department of Computer Science, University of Oxford, UK

Received 13 August 2014; revised 24 October 2014; accepted 9 December 2014

KEYWORDS

Feature extraction;
Fuzzy similarity;
Obfuscation;
Plagiarism detection;
Semantic similarity

Abstract Highly obfuscated plagiarism cases contain unseen and obfuscated texts, which pose difficulties when using existing plagiarism detection methods. A fuzzy semantic-based similarity model for uncovering obfuscated plagiarism is presented and compared with five state-of-the-art baselines. Semantic relatedness between words is studied based on the part-of-speech (POS) tags and WordNet-based similarity measures. Fuzzy-based rules are introduced to assess the semantic distance between source and suspicious texts of short lengths, which implement the semantic relatedness between words as a membership function to a fuzzy set. In order to minimize the number of false positives and false negatives, a learning method that combines a permission threshold and a variation threshold is used to decide true plagiarism cases. The proposed model and the baselines are evaluated on 99,033 ground-truth annotated cases extracted from different datasets, including 11,621 (11.7%) handmade paraphrases, 54,815 (55.4%) artificial plagiarism cases, and 32,578 (32.9%) plagiarism-free cases. We conduct extensive experimental verifications, including the study of the effects of different segmentations schemes and parameter settings. Results are assessed using precision, recall, *F*-measure and granularity on stratified 10-fold cross-validation data. The statistical analysis using paired *t*-tests shows that the proposed approach is statistically significant in comparison with the baselines, which demonstrates the competence of fuzzy semantic-based model to detect plagiarism cases beyond the literal plagiarism. Additionally, the analysis of variance (ANOVA) statistical test shows the effectiveness of different segmentation schemes used with the proposed approach.

1. Introduction

Plagiarism detection (PD) in natural language texts is one example of NLP applications that are linked with approaches from related fields, such as information retrieval (IR), data mining (DM), and soft computing (SC). PD research has focused on finding patterns of text that are illegally copied

* Corresponding author.
E-mail address: s.zahrani@tu.edu.sa (S.M. Alzahrani).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2014.12.001

Please cite this article in press as: Alzahrani, S.M. et al., Uncovering highly obfuscated plagiarism cases using fuzzy semantic-based similarity model. Journal of King Saud University – Computer and Information Sciences (2015), http://dx.doi.org/10.1016/j.jksuci.2014.12.001

from others. The easiest and common way to commit plagiarism is to copy and paste texts from digital resources. This is called literal plagiarism and is easy to spot by current PD methods. Unlike literal plagiarism, obfuscated plagiarism can be hardly seen because plagiarized texts are changed into different words and structure, or maybe into a different language.

Obfuscated plagiarism cases can be in the form of paraphrasing the original texts using different syntactical structures and lexical variations such as synonyms, antonyms, hypernyms, etc., but with no citation given to the original text. Plagiarism can be also hidden when the text is translated from one language to another with no credit to the original version, which is called cross-language plagiarism. Another form is summarized plagiarism, wherein long texts are briefed into shorter forms, which exclude details and keep the most important ideas in the source text, but with no accreditation given to the original source. In these exemplar forms of plagiarism, the texts are changed but ideas in the original texts remain unchanged. Appropriating an idea in whole or in part, with superficial modifications and obfuscations, in order to hide their sources without giving credit to its originator, is called idea plagiarism (Roig, 2006; Bouville, 2008).

Traditional techniques for PD depend on document similarity models such as duplicate detection (Elhadi and Al-Tobi, 2008, 2009) and bag-of-words related models (Barrón-Cedeño et al., 2009, 2010, 2009). Applications of document similarity, however, achieve the retrieval of a set of documents which have global similarity (at the document-level) with the query document from some source archive. The purpose of PD is not achieved yet via the document similarity, and a further detailed comparison between the query document and its candidate list should be carried out to report the local similarity (at the sentence-level, for instance). Exact and approximate string matching has been commonly used to compare two documents in-detail and find plagiarism. The documents are segmented into small comparison units such as character *n*-grams (Grozea et al., 2009), word *n*-grams (Barrón-Cedeño et al., 2009), or sentences (Alzahrani, 2009; Yerra and Ng, 2005; Zechner et al., 2009). An exhaustive matching is carried out, whereby matched *n*-grams (or sentences) that are adjacent to each other are combined into passages. Such methods are effective with verbatim plagiarism, yet not working with plagiarized texts that are literally different.

A recent literature review on the field of PD research (Alzahrani et al., 2012) has shown that there is a need for effective and efficient algorithms to find patterns of plagiarism that are semantically, but not literally, the same with original texts. Most of the current PD methods fail to detect obfuscated plagiarism cases because the similarity metrics of compared texts are computed without any knowledge of the linguistic and semantic structure of the texts (Ceska, 2007). Just a few methods have been developed based on a partial understanding of texts, e.g., when the words are replaced by synonyms, antonyms and hypernyms (Yerra and Ng, 2005). For example, Alzahrani and Salim (2010) presented a method to compute the similarity score between sentences based on the words and their synonyms. The method may be helpful to detect semantically similar texts, but should be further enhanced because not all synonyms relate to every meaning.

Recently, sentence similarity measures based on the semantic relatedness of their words have attracted researchers in different areas and for different applications, such as knowledge-based systems (Lee, 2011), text clustering (Shehata et al., 2010), text categorization (Luo et al., 2011), and text summarization (Binwahlan et al., 2010). A study by Lee (2011) proposed a semantic-based sentence similarity measure wherein two sentences can be compared based on a semantic space composed of a noun vector and a verb vector. A cosine similarity was computed between the noun vectors of two sentences and between the verb vectors of the sentences, which is further combined into a single similarity score. In Li et al. (2006), a sentence similarity measurement was presented based on the syntactic structures, semantic ontology and corpus statistics. Fernando and Stevenson (2008) presented a method to detect paraphrases of short lengths. A joint similarity matrix was constructed based on joint words from compared texts, wherein the similarity values between word pairs were calculated using different semantic similarity metrics.

In this paper, we propose a deep word analysis, in accordance with the WordNet lexical database (Miller, 1995), to detect similar, but not necessarily the same, passages. We focus on highly obfuscated plagiarism cases which are rephrased into another text without proper attribution to the original text. Unlike existing PD methods, which extract bag-of-words features (such as *n*-grams) without use of semantic features, we implemented a feature extraction method (FEM) which maintains the part-of-speech (POS) semantic spaces of the texts before further chunking of the text. Text segmentation is thereafter done using different schemes including word 3-gram, word 5-gram, word 8-gram with 3-word overlapping, and sentences. The purpose of using different segmentation schemes is to investigate which one works better along with the semantic features in the text. A fuzzy semantic-based approach is presented based on the assumption that words (from two compared texts) have a fuzzy (approximate or vague) similarity with fuzzy sets that contain words of the same meaning from a certain language. To fuzzify the relationship of word pairs (from text pairs), we proposed a WordNet-based semantic similarity metric as a fuzzy membership function. The fuzzy relationship between two words ranges between 1, for words that are identical or have the same meaning (i.e. synonyms), and 0 for words that are totally different (i.e., do not have any semantic relationship). A fuzzy inference system was constructed to evaluate the similarity of two texts and infer about plagiarism.

Experimental work was conducted on 99,033 various cases composed of handmade/simulated plagiarism cases, artificial plagiarism cases constructed automatically from some text documents and inserted into another, and plagiarism-free cases. Results of PD on those cases were assessed using precision, recall, F-measure and granularity averaged over 10-fold cross-validation data. The proposed approach was evaluated statistically against different state-of-the-art baselines using paired t-tests, which demonstrate the effectiveness of this approach to detect highly obfuscated plagiarism cases.

The remainder of this paper is organized as follows. Section 2 presents related work on semantic similarity measures based on lexical taxonomies such as WordNet, and overviews of related PD methods. Section 3 describes the feature extraction methods used in this study. Section 4 presents the proposed model for PD based on a fuzzy semantic model. In section 5, we discuss the experimental design including the datasets, baselines, parameters setting, evaluation metrics, the 10-fold cross-validation approach, and statistical analysis.

Please cite this article in press as: Alzahrani, S.M. et al., Uncovering highly obfuscated plagiarism cases using fuzzy semantic-based similarity model. Journal of King Saud University – Computer and Information Sciences (2015), http://dx.doi.org/10.1016/j.jksuci.2014.12.001

Section 6 presents the results from the proposed approach using different sentence samples and two datasets, and discusses our results with the results obtained from different state-of-the-art baselines. Section 7 draws some conclusions on this work and outlines possible future research in this area.

2. Related work

2.1. Semantic similarity measures

In lexical taxonomies, such as the WordNet (Miller, 1995), *lexes* are arranged into "is-a" and "has-a" hierarchies wherein words with the same meaning are grouped together into a so-called *synsets* which are linked with more abstract/general words called *hypernyms*, and most specific words called *hyponyms*. Words usually have different *senses* (i.e., meanings) and, hence, may belong to different synsets. Based on such taxonomy, a word-to-word semantic similarity can be implemented as a relationship between words' synsets, as proposed in many research works (Leacock and Chodorow, 1998; Resnik, 1995; Lin, 1998; Jiang and Conrath, 1997; Wu and Palmer, 1994;, Hirst and St Onge, 1998; , Banerjee and Pedersen, 2003).

Part of word-to-word semantic similarity metrics assume a Directed-Acyclic-Graph (DAG) taxonomy that relates concepts within the same POS boundary via the *is-a* relationship. The *path* metric (Jiang and Conrath, 1997; Li et al., 2003), for example, measures the shortest path (i.e., number of hops) that connects two concepts (i.e., two word synsets) in the form of DAG taxonomy. The smaller the path the higher the semantic similarity between two words is. The *lch* metric (Leacock and Chodorow, 1998) relates the shortest path that connects two word synsets and the maximum depth from the root of the DAG taxonomy in which they occur, as shown in the following formula:

$$lch(w_1, w_2) = \log\left(\frac{path(w_1, w_2)}{2^{*} maxdepth}\right) \quad (1)$$

where $path(w_1,w_2)$ is as defined above, and *maxdepth* is the longest distance between the root and any leaf in the DAG taxonomy that contains both synsets. The *wup* metric (Wu and Palmer, 1994) relates the depth of the words' synsets in the DAG taxonomy and the depth of their least common subsumer (or the most specific ancestor), denoted as LCS. We will discuss this measure in detail in later parts of this paper.

Information content (IC) Fernando and Stevenson, 2008 is a measure that a concept c can be found in a standard textual corpus, which can be given by the following formula:

$$IC(c) = -\log(P(c)) \quad (2)$$

where $P(c)$ is the probability that c can be found in the corpus. The *res* metric (Resnik, 1995) defines a similarity score of two word synsets based on the IC of their LCS in the DAG taxonomy.

$$res(w_1, w_2) = IC(LCS(w_1, w_2)) \quad (3)$$

Besides, the *lin* metric (Lin, 1998) and *jcn* metric (Jiang and Conrath, 1997) are based on the IC of the LCS and that of the words' synsets as stated in (4) and (5), respectively.

$$lin(w_1, w_2) = \frac{2^{*} IC(LCS((w_1, w_2))}{IC(w_1) + IC(w_2)} \quad (4)$$

$$jcn(w_1, w_2) = 1 - \frac{IC(w_1) + IC(w_2) - 2^{*} IC(LCS(w_1, w_2))}{2} \quad (5)$$

Other word-to-word similarity metrics have been defined across the POS boundaries, such as *lesk* metric (Banerjee and Pedersen, 2003) and *hso* metric (Hirst and St Onge, 1998). These metrics are, in fact, semantic relatedness rather than similarity measures as stated in Corley and Mihalcea (2005), Budanitsky and Hirst (2006). The first incorporates information from the directions between the lexical chains of two word synsets, and the later measures the relationship of two words' synsets based on the overlap of their dictionary glosses.

Sentence similarity methods have been studied based on semantic similarity/relatedness of their words, as proposed by Mihalcea et al. (2006), Corley and Mihalcea (2005), Li et al. (2006), Lee (2011) and others. In Budanitsky and Hirst (2006), word similarity metrics have been categorized into knowledge- and corpus-based methods. Knowledge-based methods are based on semantic ontologies, WordNet for instance, that draw relationships between words. Such metrics include *path*, *lch* (Leacock and Chodorow, 1998), *wup* (Wu and Palmer, 1994), *res* (Resnik, 1995), *lin* (Lin, 1998), *jcn* Jiang and Conrath, 1997, *lesk* (Banerjee and Pedersen, 2003), and *hso* (Hirst and St Onge, 1998) metrics which we discussed previously. On the other hand, corpus-based methods implement the relationship between the words as derived from large (and standard) text corpora, such as the Penn Treebank Corpus, Brown Corpus, Project Gutenberg corpus, Wikipedia corpus and others. Examples of corpus-based measurements involve latent semantic analysis (LSA) (Mihalcea et al., 2006), and point-wise mutual information (PMI) (Turney, 2001). To compute the similarity of two texts, the study in Corley and Mihalcea (2005), Mihalcea et al. (2006) combined a local metric using one of the word-to-word similarity measures, and a global metric which is the IDF. The similarity between two texts T_1 and T_2 was defined as follows (Budanitsky and Hirst, 2006):

$$Sim(T_1, T_2) = \frac{1}{2}\left(\frac{\sum_{w \in T_1} maxSim(w, T_2) \times idf(w)}{\sum_{w \in T_1} idf(w)} + \frac{\sum_{w \in T_2} maxSim(w, T_1) \times idf(w)}{\sum_{w \in T_2} idf(w)}\right) \quad (6)$$

where $maxSim(w,T_2)$ is the maximum similarity score between each word w from T_1 and words in T_2 obtained by one of the knowledge- or corpus-based similarity metrics, and $idf(w)$ is the IDF obtained from the relation n_w/N, where n_w is the number of documents that contain the word w, and N is the total number of documents in a large text corpus.

In Fernando and Stevenson (2008)), a similarity matrix W of joint (distinct and non-stop) words between two candidate texts was proposed. Each text was represented as a binary vector with the entries: 1 if a word from joint word matrix is present and 0, otherwise. Each cell in similarity matrix W has an entry equal to a word-to-word similarity value obtained from knowledge-based metrics. The similarity score was computed as the mathematical product of the binary vectors from both texts and the similarity matrix, as follows:

$$Sim(T_1, T_2) = \frac{\vec{T}_1 W \vec{T}_2}{|\vec{T}_1||\vec{T}_2|} \quad (7)$$

where $\vec{T}_1$ and $\vec{T}_2$ are the binary vectors of texts T_1 and T_2, respectively, and W is the joint similarity matrix.

A study by Li et al. (2006) proposed a semantic similarity measure between sentences derived from the words' similarity and the words' order similarity. They proposed a word-to-word semantic similarity, which we referred to as *li* metric, that combines the shortest *path* between two words w_1 and w_2 and the *depth* of the their LCS in the taxonomy that has both words, as follows:

$$li(w_1, w_2) = e^{-\alpha \cdot path(w_1, w_2)} \times \frac{e^{\beta \cdot depth(LCS(w_1, w_2))} + e^{\beta \cdot depth(LCS(w_1, w_2))}}{e^{\beta \cdot depth(LCS(w_1, w_2))} - e^{\beta \cdot depth(LCS(w_1, w_2))}} \quad (8)$$

where $\alpha \in [0,1]$ and $\beta \in [0,1]$, are scaling parameters of the contribution of the *path* and *depth* metrics in the formula. Then, a joint word set was defined as the unification of unique, non-stop, and stemmed words from both texts T_1 and T_2. The value of an entry in the semantic vector s_1 for text T_1 was defined as below:

$$s_1(w_i) = li(w_i, \tilde{w}) \times IC(w_i) \times IC(\tilde{w}) \quad (9)$$

where *li* metric is evaluated as either 1 if the word is present in T_1 or the highest word-to-word semantic similarity found between the word w_i and any word in the candidate text T_2 as defined in (8), and IC is the information content of the words as defined in (2). The semantic vector s_2 for text T_2 was defined in a similar way, and the final sentence similarity score was computed as the Cosine similarity of the two vectors:

$$S_s(T_1, T_2) = \frac{s_1 \cdot s_2}{||s_1|| \cdot ||s_2||} \quad (10)$$

The order similarity (Li et al., 2006), on the other hand, means that a different word order may convey a different meaning and should be counted into the semantic similarity. If we have two candidate texts, for instance, T_1 = "A quick brown fox jumps over the lazy dog" and T_2 = "A quick brown dog jumps over the lazy fox", the joint word set $T = \{T_1 \cup T_2\}$ is {A, quick, brown, fox, jumps, over, the, lazy, dog}, wherein we can indicate the occurrence of each word by a unique number. Thus, the word order vectors from T_1 and T_2 can be given as r_1 = {1,2,3,4,5,6,7,8,9} and r_2 = {1,2,3,9,5,6,7,8,4}, respectively. The cosine similarity was obtained from the order vectors as shown below.

$$S_r(T_1, T_2) = 1 - \frac{||r_1 - r_2||}{||r_1 + r_2||} \quad (11)$$

The final similarity proposed in Li et al. (2006) combined both similarities in (10) and (11), as follows:

$$Sim(T_1, T_2) = \delta \cdot S_s(T_1, T_2) + (1 - \delta) \cdot S_r(T_1, T_2) \quad (12)$$

where δ is a scaling parameter $\in [0.5,1]$.

A recent study (Lee, 2011) reported a sentence similarity measure that implements a NOUN vector (NV) containing a joint noun set from two candidate texts T_1 and T_2, and VERB vector (VV) containing a joint verb set from T_1 and T_2. The value of an entry in NV vector (and VV vector, respectively) was defined as the highest *wup* similarity (Wu and Palmer, 1994) found between the corresponding noun and other nouns in the NV vector (and the corresponding verb and other verbs in the VV vector, respectively). Cosine similarity measurements were computed from both vectors as follows:

$$S_N(T_1, T_2) = \frac{NV_{T_1} \cdot NV_{T_2}}{||NV_{T_1}|| \cdot ||NV_{T_1}||} \quad (13)$$

$$S_V(T_1, T_2) = \frac{VV_{T_1} \cdot VV_{T_2}}{||VV_{T_1}|| \cdot ||VV_{T_1}||} \quad (14)$$

To find the final similarity score between two texts, the noun vector similarity SN and the verb vector similarity SV were integrated in a way similar to Eq. (12), as below

$$Sim(T_1, T_2) = \delta \cdot S_N(T_1, T_2) + (1 - \delta) \cdot S_V(T_1, T_2) \quad (15)$$

2.2. Plagiarism detection methods

Textual features applied for PD varied from lexical and syntactic features to semantic features. Table 1 shows a summary of the research works that have employed types of text features (Alzahrani et al., 2012).

Commonly, PD methods in textual documents have focused on chunking the texts and measuring the overlap between two documents (Alzahrani et al., 2012). A typical example of these approaches is to segment the texts into *N*-grams, and find the common ones using the Jaccard coefficient (16), Dice's coefficient (17), simple matching coefficient (18), or containment coefficient (19).

$$Jaccard(T_1, T_2) = \frac{|\{NGrams\}_{T_1} \cap \{NGrams\}_{T_2}|}{|\{NGrams\}_{T_1} \cup \{NGrams\}_{T_2}|} \quad (16)$$

$$Dice(T_1, T_2) = \frac{2|\{NGrams\}_{T_1} \cap \{NGrams\}_{T_2}|}{|\{NGrams\}_{T_1} \cup \{NGrams\}_{T_2}|} \quad (17)$$

$$Match(T_1, T_2) = |\{NGrams\}_{T_1}| - |\{NGrams\}_{T_1} \cap \{NGrams\}_{T_2}| \quad (18)$$

$$Contain(T_1, T_2) = \frac{|\{NGrams\}_{T_1} \cap \{NGrams\}_{T_2}|}{min(|\{NGrams\}_{T_1}|, |\{NGrams\}_{T_2}|)} \quad (19)$$

where $\{NGrams\}_{T_1}$ and $\{NGrams\}_{T_2}$ are the sets of *N*-grams generated from T_1 and T_2, respectively. In Yerra and Ng (2005), the authors adopted a sentence-based copy detection approach, namely the 3-least-frequent 4-grams. In their approach, sentences were divided into unique character 4-grams $\{g_1, g_2, \ldots, g_J\}$ and the frequency of each 4-gram was computed as follows:

$$f(g_i) = \frac{n_i}{\sum_{j=1}^{J} n_j} \quad (20)$$

where n_i is the number of occurrences of the *i*th 4-gram g_i, and J is the total number of distinct 4-grams in the sentence. Two sentences T_1 and T_2 were represented uniquely by their three least-frequent 4-grams, also called fingerprints. The fingerprints of sentences were matched using their representative fingerprints, and copied sentences were detected easily.

Nevertheless, plagiarism detection methods that incorporate partial understanding of the linguistic rules or the semantic relationships between two candidate texts have not been applied by most, if not all, plagiarism detectors (Alzahrani et al., 2012). A few research works have applied semantic-

Table 1 Text features applied in PD research.

–	Examples	Ref.
Lexical features	Character *n*-grams (fixed-length)	Grozea et al. (2009)
	Character *n*-grams (variable-length)	Yerra and Ng (2005)
	Word *n*-grams	Zechner et al. (2009), Koberstein and Ng (2006), Basile et al. (2009), Kasprzak et al. (2009); Alzahrani and Salim (2010)
Syntactic features	Chunks	Scherbinin and Butakov (2009)
	Part-of-speech and phrase structure	Elhadi and Al-Tobi, 2008, 2009; Ceska et al., 2007
	Word position/order	Li et al., (2006), Koroutchev and Cebrian (2006)
	Sentence	Alzahrani (2009), Yerra and Ng (2005)
Semantic features	Synonyms, hyponyms, hypernyms, etc.	Alzahrani (2009), Yerra and Ng (2005), Li et al. (2006), Alzahrani and Salim (2009), Alzahrani and Salim (2010)
	Semantic dependencies	Li et al. (2006), Muftah (2009)

based methods and reported positive results in comparison to *N*-gram matching methods (Turney, 2001). This is due to the ability of these methods to find plagiarism when plagiarized texts are reworded and rephrased. However, the time complexity of such methods has affected their implementation into practical tools. A method called *SVDPlag* was proposed based on Latent Semantic Analysis (LSA) of the Singular Value Decomposition (SVD) Ceska, 2008, 2009. The approach used feature extraction and reduction of *n*-grams from textual documents, where *n* was experimentally evaluated using different values between 1 and 8. The latent semantic associations between different *n*-grams were then incorporated into the document similarity model using LSA, which preserves the semantic associations between *n*-grams in the documents as in typical IR models (Manning et al., 2009). Sentence-based copy detection approach in Yerra and Ng (2005) was further improved using the fuzzy-set information retrieval (FIR) model reported in the literature (Ogawa et al., 1991; Bordogna and Pasi, 1993; Cross, 1994). FIR was capable to detect not only the same, but also similar sentences with superior results to 3-least-frequent 4-grams. The method was based on using fuzzy sets that contain words with the same or similar usage, which can be derived from documents in a large text corpus. Words that are related (and maybe similar) to each other normally occurred together in a number of documents; therefore, their correlation factors can be obtained as the ratio between the number of documents that have both words, and the number of documents that contain either or both words. Thus, Yerra and Ng (2005) proposed a *word-to-word correlation factor*, which we referred to as *yer* metric, which can be derived from the following formula (Yerra and Ng, 2005):

$$yer(w_1, w_2) = \frac{N(w_1, w_2)}{N(w_1) + N(w_2) - N(w_1, w_2)} \tag{21}$$

where $N(w_1,w_2)$ is the number of documents in a text collection that contain both words w_1 and w_2, $N(w_1)$ is the number of documents that contain w_1, and $N(w_2)$ is the number of documents that contains w_2. Sentences were compared based on the sum of the correlation factors of their words, and the *sentence-to-sentence similarity* was reported as a degree of membership between words in both sentences and the fuzzy sets. Another study by Pera and Ng (2011) used a different *word-to-word correlation* measurement, which we called *per* metric, for a sentence-based PD approach. The relationship between two words was derived from the formula (22) using 880,000 Wikipedia documents, and *sentence-to-sentence* similarity was obtained from the formula (23).

$$per(w_1, w_2) = \frac{\sum_{w_i \in V_1} \sum_{w_j \in V_2} (dis(w_i, w_j) + 1)^{-1}}{|V_1| \times |V_2|} \tag{22}$$

where V_1 is the set that includes the word w_1 and all of its *stem* variations in a text document D, V_2 is the set that contains the word w_2 and its stems, and $dis(w_i,w_j)$ is the distance (or the number of words) between w_i and w_j in D.

$$Sim(T_1, T_2) = \frac{\sum_{i=1}^{n} \min(1, \sum_{j=1}^{m} per(w_i, w_j))}{|T_1|} \tag{23}$$

where n and m are the number of words in T_1 and T_2, respectively.

2.3. Discussion

There are a number of semantic similarity methods which aim at comparing texts of short lengths, such as sentences, yet they are seldom used for PD applications. In fact, there are some situations in the academic society wherein we need to detect plagiarism activities that aimed to be hidden by the plagiarists via deriving similar content to the original source but with different words. Chunking (i.e., a method for splitting the text into small and scannable segments) and string matching, which are the dominant approaches used for PD, are awfully unsuccessful with obfuscated plagiarism cases. We suggest, therefore, the use of semantic similarity measurements for detection of literally-different plagiarism cases. In this regard, we address the problem of how to make a combination between chunking methods, which uses the semantic relationships of words, and fuzzy semantic-based PD. In this work, we modified the FIR model in Yerra and Ng (2005) to incorporate WordNet-based semantic similarity metrics rather than word correlation factors. We used FIR as a baseline to our approach and compared results from both on ground-truth annotated plagiarism corpora.

3. Feature Extraction Method (FEM)

In this study, we implemented two types of textual structures. The first aims at describing the text as word *k*-grams (also called *k*-shingles) where *k* is typically set before the experiments. In this context, we proposed the same settings that

achieved good results in previous research works, namely word 3-grams (Barrón-Cedeño et al., 2010), word 5-grams (Barrón-Cedeño et al., 2010; Alzahrani et al., 2012), and word 8-grams with 3-word overlapping (Alzahrani et al., 2012). The second aims at splitting the text into sentences using end-of-statement delimiters (i.e., full-stops marks, question marks, and exclamation marks). Sentence-based feature extraction methods have been applied widely in PD research (Alzahrani, 2009; Yerra and Ng, 2005; Zechner et al., 2009).

3.1. FEM framework

A feature extraction method (FEM) was used to characterize input texts in terms of the *lexicons* and *parts-f-speech* (POS) tags. The major components are shown in Fig. 1, and can be described as follows:

Tokenization – The input text is divided into tokens, whereby each token is marked as token [T], or end-of-sentence [E].

POS disambiguation (or tagging) – Before further pre-processing of the text, a POS tagger is employed to annotate parts of speeches according to the Pennsylvania Treebank POS tags (Marcus et al., 1993).

i. Lemmatization – A lemmatizer is applied on the extracted tokens, wherein a dictionary form (not necessarily the root form) is provided for each word with the assistance of WordNet (Miller, 1995). Thus, in this component, the tokens are changed to lemmas [L]. This would help, in later parts of this paper, to compare the semantic meaning of two sentences based on the semantic relatedness of their (lemmatized) words derived from the WordNet. Based on our experience from using "stemming" in a previous research work (Alzahrani and Salim, 2010), there could be a deficiency when using WordNet to provide the synsets of the words' stems, since WordNet is based on "lemmas" rather than "stems" which should help to find the appropriate synset in our model.

ii. Stop words removal – The most frequent English words such as "a", "an", "the", "is", "are", etc., are removed from the text. As a result, most of the conjunctions and interjections will be removed in this step. The stop words list has been obtained from the NLTK (nltk.org) project.

iii. Text segmentation – The resulting text is segmented into word 3-grams (W3G), word 5-grams (W5G), word 8-grams with 3-word overlapping (W8G3W), and sentences (S2S). These different segmentation schemes will be compared during the experimental work in terms of which approach can better handle obfuscated plagiarism cases along with the proposed fuzzy semantic-based similarity method.

iv. POS-related semantic space construction - The lemmas in each segment are categorized into the following tags: noun [N], verb [V], adjective [AJ] or adverb [AV]. In this regard, a transformation function is used to convert multiple Penn Treebank Tags into our tags. For instance, VB, VBD, VBN, VBG will be [V], and so on.

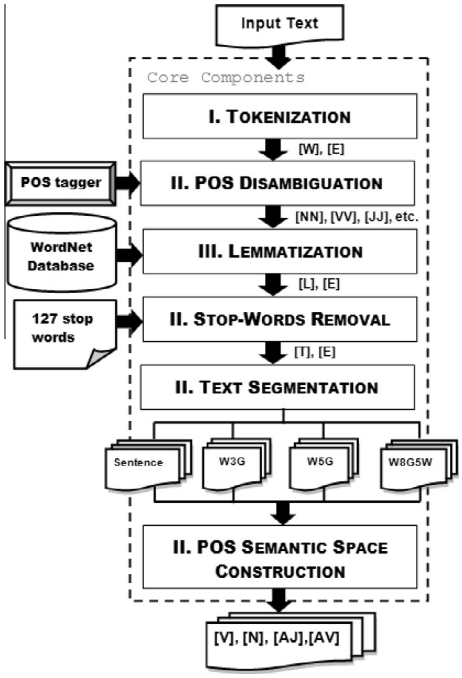


Figure 1 Feature extraction method (FEM) based on different segmentation settings and POS-related semantic space.

3.2. An example

In this section, let's consider the following raw text extracted from a corpus called PAN-PC-11 (Potthast et al., 2011) recently used by a benchmark PD evaluation Lab[1] (the datasets will be discussed in Section 5.2):

Raw Text:

Oh isn't she sweet! She said, thinking that she should present her with some kind of special gift. Floating above the little one's head she declared the child will marry whoever she chooses and live happily ever after.

We applied the FEM which maintains the lexical and syntactical features proposed for this study. Table 2 shows the results obtained from different pre-processing steps including: (I) tokenization process, wherein the text is splatted into tokens, and end-of-sentence delimiters; (II) POS disambiguation; (III) lemmatization, wherein tokens are converted into lemmas (dictionary forms); and (IV) stop words removal.

Table 3 shows the segmentation process into different structures involving sentences, W3G, W5G, and W8G3W (column 2), and the resulting POS-related semantic spaces (column 3) for each segment, whereby we maintained the original POS tag associated with each term during the POS disambiguation process on the input text. The outputs from the FEM algorithm will be used as different comparison schemes in the PD approach, and the POS semantic spaces will help to find the appropriate meaning of each word in the semantic-based metric.

[1] Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN) workshops, http://pan.webis.de/.

Table 2 Text Tokenization, Lemmatization, POS Disambiguation, and Stop-word Removal.

Features	Details	Input Text
Tokens	[T] tokens, [E] end-of-sentence	Oh[T] isn't[T] she[T] sweet[T][E] she[T] said[T] thinking[T] that[T] she[T] should[T] present[T] her[T] with[T] some[T] kind[T] of[T] special[T] gift[T][E] Floating[T] above[T] the[T] little[T] one's[T] head[T] she[T] declared[T] the[T] child[T] will[T] marry[T] whoever[T] she[T] chooses[T] and[T] live[T] happily[T] ever[T] after[T][E]
POS tags	Penn Treebank POS tags	oh/UH is/VBZ not/RB she/PRP sweet/JJ ./. she/PRP said/VBD thinking/VBG that/IN she/PRP should/MD present/VB her/PRP with/IN some/DT kind/NN of/IN special/JJ gift/NN ./. floating/NNP above/IN the/DT little/JJ one/CD '/'' s/JJ head/NN she/PRP declared/VBD the/DT child/NN will/MD marry/VB whoever/RP she/PRP chooses/VBZ and/CC live/VB happily/RB ever/RB after/IN ./.
Lemmas	[L] lemma, [E] end-of-sentence	oh[L] be[L] not[L] she[L] sweet[L][E] she[L] say[L] think[L] that[L] she[L] should[L] present[L] her[L] with[L] some[L] kind[L] of[L] special[L] gift[L][E] floating[L] above[L] the[L] little[L] one[L] head[L] she[L] declare[L] the[L] child[L] will[L] marry[L] whoever[L] she[L] choose[L] and[L] live[L] happily[L] ever[L] after[L][E]
Stop-words removed	SW list [W] word, [E] end-of-sentence	sweet[W][E] say[W] think[W] present[W] kind[W] special[W] gift[W][E] floating[W] little[W] head[W] declare[W] child[W] marry[W] whoever[W]choose[W] live[W] happily[W] ever[W][E]

4. Fuzzy semantic-based string similarity model for plagiarism detection

In this paper, we proposed a deep word analysis between two input texts utilizing their POS-related semantic spaces. Semantic relatedness between two words can be defined based on the "*is-a*" relationship from WordNet lexical taxonomies (Miller, 1995). Accordingly, the semantic relationship between two texts can be defined as the aggregation of different fuzzy rules that are based on the words' semantic similarity. According to Yerra and Ng (2005), "matching two sentences can be approximate or vague, which can be modeled by considering that each word in a sentence is associated with a fuzzy set that contains the words with the same meaning, and there is a degree of similarity (usually less than 1) between words (in a sentence) and the fuzzy set" (p. 563). We adapted the fuzzy-set IR system in Yerra and Ng (2005) into a fuzzy semantic-based model, and we used the former as a baseline (see Section 5.2 for more details). The model is based on the semantic relatedness between words as a degree of membership on one side, and the fuzzy rule-based comparison of two candidate texts on the other side.

4.1. General framework

Fig. 2 shows the general framework of this model. Two input texts (might be of document size) are used in the feature extraction method. The resulting features from the texts are used as inputs to the fuzzy inference system, whereby a semantic similarity measurement is modeled as a membership function. After the evaluation of the rules, the outputs are aggregated into a single value which can be interpreted as a similarity score between input texts. Parts of texts that are highly similar will be highlighted and displayed to the user. The system should be able to infer about literal plagiarism as well as obfuscated plagiarism cases.

4.2. Word-to-word semantic similarity

Word-to-word relationships can be based on different assumptions: words are identical, words are in the same synset (i.e. synonyms), words are not in the same synset but their synset contains at least one common word, words have at least one shared hypernym, words are different. In this regard, various semantic similarity metrics of words have been proposed with regard to their relationship in the WordNet lexical database, as discussed previously in Section 2.1. In this paper, we used Wu & Palmer (1994) measure Wu and Palmer, 1994 which has become very popular (Lee, 2011; Lin et al., 1998). This metric combines the *depth* of the least common subsumer (LCS) of two word synsets and the *depth* of each word in their lexical taxonomy as shown in Fig. 3. The formula can be expressed as follows:

$$wup(w_1, w_2) = \frac{2 \times depth(LCS(w_1, w_2))}{depth(w_1) + depth(w_2)} \quad (24)$$

where w_1 and w_2 are two word concepts (in the form of synsets), $depth(x)$ is the total number of edges from the root of the DAG taxonomy to the concept x.

Table 3 Text Segmentation Into Sentences and Word k-Grams.

Structure	Segments	POS-related semantic space
Sentences	#1: sweet #2: say think present kind special gift #3: floating little head declare child marry whoever choose live happily ever	#1: [AJ] #2: [V] [V] [V] [N] [AJ] [N] #3: [N] [AJ] [N] [V] [N] [V] [AV] [V] [V] [AV] [AV]
W3G	#1: sweet say think #2: say think present #3: think present kind #4: present kind special ...	#1: [AJ] [V] [V] #2: [V] [V] [V] #3: [V] [V] [N] #4: [V] [N] [AJ] ...
W5G	#1: sweet say think present kind #2: say think present kind special #3: think present kind special gift #4: present kind special gift floating ...	#1: [AJ] [V] [V] [V] [N] #2: [V] [V] [V] [N] [AJ] #3: [V] [V] [N] [AJ] [N] #4: [V] [N] [AJ] [N] [N] ...
W8G3W	#1: sweet say think present kind special gift floating #2: special gift floating little head declare child marry #3: declare child marry whoever choose live happily ever	#1: [AJ] [V] [V] [V] [N] [AJ] [N] [N] #2: [AJ] [N] [N] [AJ] [N] [V] [N] [V] #3: [V] [N] [V] [AV] [V] [V] [AV] [AV]

Structures used include sentences and word k-grams. Resulting segments will serve as different comparison schemes in the PD system. POS-related semantic spaces will assist to find the proper synset of each term (e.g., present[V] has a different meaning from present[N]).

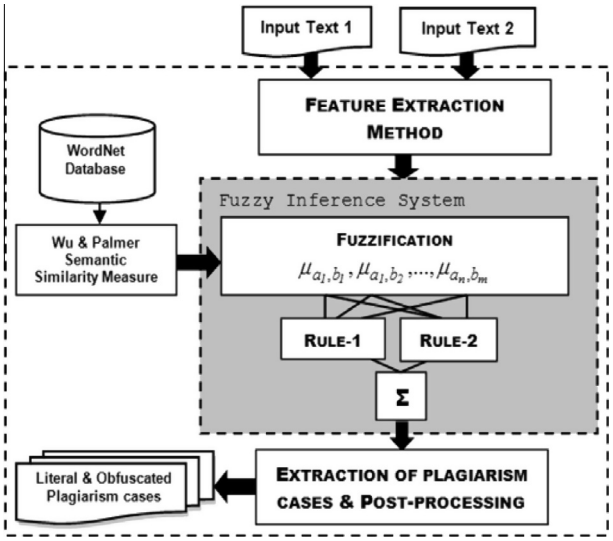


Figure 2 General framework of fuzzy semantic-based model for text similarity and plagiarism detection.

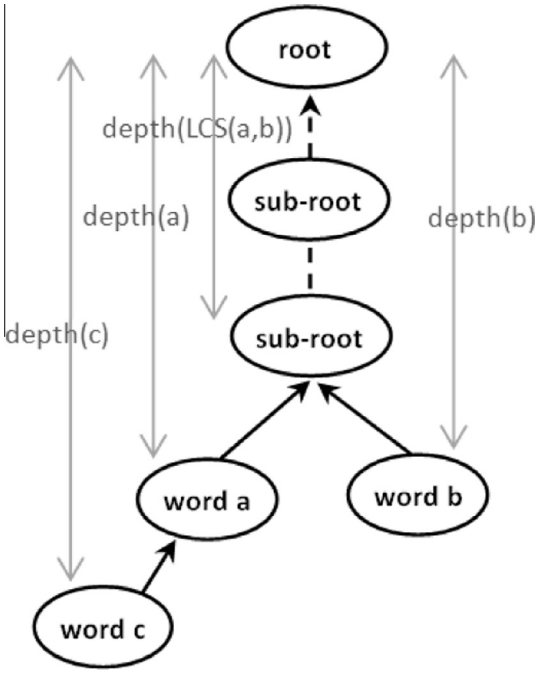


Figure 3 Directed-Acyclic-Graph (DAG) for WordNet lexical taxonomy.

To correctly use this formula, we utilized the POS semantic spaces to be able to find the appropriate synsets of the words from WordNet database. To illustrate, let's consider the word w_1 = "present" which can be a noun, verb, adjective or adverb, and the word w_2 = "gift" which can be a noun or verb as can be seen in the semantic ontology that represent both words in Fig. 4. Wu and Palmer similarity (Wu and Palmer, 1994) between two words can only be computed if they have the same POS tags; for instance, "present" and "gift" are semantically similar if they are nouns, but have no semantic similarity if "present" is verb, but "gift" is noun. Moreover, the similarity between two words of the same POS will vary based on different senses of both words. Using the NLTK (Edward and Steven, 2002), we computed different values between "gift" and different synsets of "present" wherein POS = [N] for both words:

```
['gift'],['present','nowadays'] = 0.3333
['gift'],['present'] = 0.9333
['gift'],['present', 'present_tense'] = 0.26667
```

However, in this research, we do not employ any word sense disambiguation approach to avoid additional complexities. We assumed the highest Wu & Palmer similarity between words' synsets with the same POS. Accordingly, we consider the wup similarity in the example of "present" and "gift" is 0.9333, where POS = [N] for both.

4.3. Fuzzy inference system for plagiarism detection

We proposed a fuzzy system for PD that uses as inputs a group of words[2] $\{a_1, a_2, \ldots, a_n\}$ in a text A taken from a source

[2] Words from this time and onwards refer to the non-frequent, lemma forms of the original words in the text.

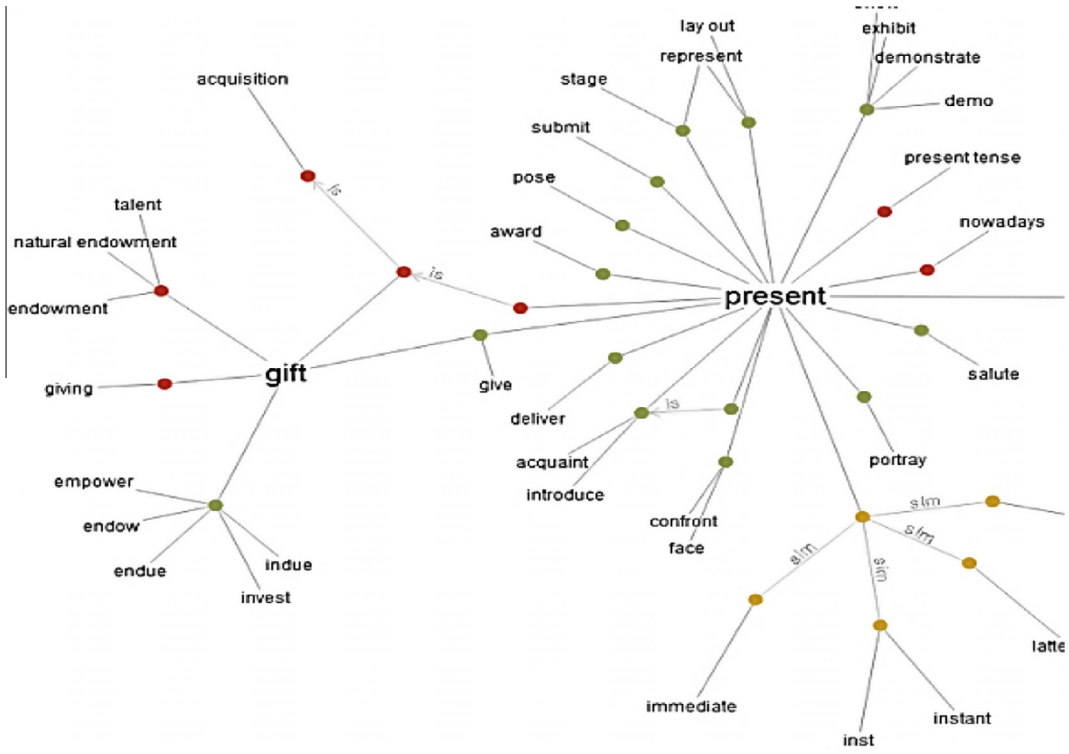


Figure 4 Semantic net of different senses of "gift" and "present"; two senses of these words are connected via "is-a" relationship.

document d_{source}, and a group of words $\{b_1, b_2, \ldots, b_m\}$ in a candidate text B taken from a suspicious document $d_{suspicious}$. Texts A and B are represented as features using the FEM method presented in Section 3. We can formulate two simple IF-THEN rules to examine two texts, as follows:

Rule 1:
IF (a_1 in A is matched/semantically similar with a word b_j in B)
AND (a_2 in A is matched/semantically similar with a word b_j in B)
...
AND (a_n in A is matched/semantically similar with a word b_j in B)
THEN A is similar to B

where b_j refers to any word that occurs in the candidate text B, $j \in [1,m]$, and m is the total number of word in B. Similarly, we can compare text B's words with regard to text A, as follows:

Rule 2:
IF (b_1 in B is matched/semantically similar with a word a_i in A)
AND (b_2 in B is matched/semantically similar with a word a_i in A)
...
AND (b_m in B is matched/semantically similar with a word a_i in A)
THEN B is similar to A

where a_i refers to any word that occurs in the text A, $i \in [1,n]$, and n is the total number of words in A.

As can be seen, such a fuzzy system has only two rules with n-AND conjunctions in the first rule, and m-AND conjunctions in the second one, where n and m refers the number of words in the text being compared to another. If the output of both checking rules is true, it is agreed that *A and B make a plagiarism case*. If the words in one text are neither matched nor *semantically* equivalent with words in the candidate text, this leads to the consequence that *A and B are totally different (i.e., plagiarism-free)*. That is, the consequence of the fuzzy rules can have only 2 values: true (1) and not true (0), and the fuzzy sets evaluation is done only on the antecedent; which means our rule system is similar to a Sugeno-style inference system (Sugeno, 1985). In these two "crisp" decisions (*plagiarism* vs. *plagiarism-free*), we could have various degrees of similarities between words in both texts and the fuzzy sets that contain words of the same meaning (i.e., sense). The similarity score between two texts could be interpreted based on a learning method as will be seen shortly.

4.3.1. Fuzzification

The word pairs from two input texts are considered the fuzzy *variables*. We considered Wu and Palmer (1994) similarity measure as the membership degree in the fuzzy system, which can be expressed as follows:

$$\mu_{a_i,b_j} = wup(a_i, b_j) \tag{25}$$

This relation evaluates the degree of (semantic) similarity between two words, which ranges from 0 (completely different when there is no shared hypernym between the words) to 1 (identical or synonymous).

4.3.2. Evaluation of rules

The if-then rule shown previously compares each word a_i in text A with all words in candidate text B and vice versa. To

evaluate the relationship of a word in one text with regard to words in the other text, we can use the fuzzy PROD operator as in the following formulas:

$$\mu_{a_1,B} = 1 - \prod_{b_j \in B; j \in [1,m]} (1 - wup(a_1, b_j))$$
$$\mu_{a_2,B} = 1 - \prod_{b_j \in B; j \in [1,m]} (1 - wup(a_2, b_j))$$
$$\cdots$$
$$\mu_{a_n,B} = 1 - \prod_{b_j \in B; j \in [1,m]} (1 - wup(a_n, b_j)) \quad (26)$$

We can also use the fuzzy MAX operator as follows:

$$\mu_{a_1,B} = MAX(wup(a_1,b_1), wup(a_1,b_2), \ldots, wup(a_1,b_m))$$
$$\mu_{a_2,B} = MAX(wup(a_2,b_1), wup(a_2,b_2), \ldots, wup(a_2,b_m))$$
$$\cdots$$
$$\mu_{a_n,B} = MAX(wup(a_n,b_1), wup(a_n,b_2), \ldots, wup(a_n,b_m)) \quad (27)$$

To evaluate the rule antecedent into a single value, we simply calculate the average sum, as follows:

$$\mu_{A,B} = (\mu_{a_1,B} + \mu_{a_2,B} + \cdots + \mu_{a_n,B})/n$$
$$\mu_{B,A} = (\mu_{b_1,A} + \mu_{b_2,A} + \cdots + \mu_{b_m,A})/m \quad (28)$$

Notice that, in general, $\mu_{A,B} \neq \mu_{B,A}$ if A and B are of different lengths.

4.3.3. Interpretation of the result

To decide whether or not there is a (degree of) plagiarism between two texts, a learning method should be introduced based on the similarities $\mu_{A,B}$ and $\mu_{B,A}$. We implemented the method in fuzzy-set IR (Yerra and Ng, 2005) to find whether two texts are plagiarized (PD) or not, as follows:

$$PD(A,B) = \begin{cases} 1 & \text{if} \quad MIN(\mu_{A,B}, \mu_{B,A}) \geqslant p \wedge |\mu_{A,B} - \mu_{B,A}| \leqslant v \\ 0 & \text{otherwise} \end{cases} \quad (29)$$

where p is called the *permission threshold* which is defined as the highest similarity value found between two texts for a human to say that these texts are semantically the same. On the other hand, v is called the *variation threshold*, which refers to the lowest *difference* of similarity values between two texts. The value of v can be used to lower the false positive detections. In other words, sentences that passed the permission threshold may not be similar if there is a "big" difference of $\mu_{B,A}$ and $\mu_{A,B}$. For example, the text similarity between A = "The book is authored by John" and B = "The book authored by John discussed best business practices", $\mu_{A,B} = 1$ since all words in A are found in B (i.e., A is subset of B after applying FEM) but $\mu_{B,A} = 0.77487$, so the difference $v = 0.225$, which allows us to not judge both sentences as similar even though their minimum similarity is "somehow" positive.

Despite sentences, it is not needed to find the minimum similarity nor the difference similarity with word k-grams as they are always of equal lengths, and hence $\mu_{A,B} = \mu_{B,A}$. Consequently, $PD(A,B)$ of word n-grams can be measured using (30).

$$PD(A,B) = \begin{cases} 1 & if \quad \mu_{A,B} \geqslant p \\ 0 & otherwise \end{cases} \quad (30)$$

4.3.4. An example

In this part, we demonstrate one example of a plagiarism case extracted from a plagiarism corpus called PAN-PC-11(Potthast et al., 2011). Notice that the first text was used to demonstrate the FEM in Section 3.2. The example includes the following raw texts:

Text A (Original):

Oh isn't she sweet! She said, thinking that she should present her with some kind of special gift. Floating above the little one's head she declared the child will marry whoever she chooses and live happily ever after.

Text B (Plagiarized):

What a darling!" She said; "I must give her something very nice. "She hovered a moment over the child's head, "She shall marry the man of her choice," she said, "and live happily ever after."

It can be observed that the second text is reworded from the first, but the meaning has remained almost unchanged. Texts A and B should pass the FEM and we should obtain text segments W3G, W5G, W8G3W, and S2S from both texts to be used as inputs to the fuzzy inference system. In this example, we considered sentences (S2S) but we will compare different segmentation schemes during the experimental work. A detailed analysis of both texts means that every sentence in A will be compared with every sentence in B. Here, we will consider a comparison of some sentence pairs. For example, we found that the sentences A_2 and B_2 are similar to some degree, and the sentences A_3 and B_3 are more similar, to a degree of 0.7856. Table 4 shows the details of the fuzzy similarity values obtained based on the proposed approach.

4.4. Detailed checking algorithm

A detailed checking should be carried out between source and suspicious texts in order to locate similar fragments. The final output of the algorithm is a list of segment pairs (A_i,B_j): $A_i \in A$, $B_j \in B$, which fulfill the condition of $PD(A_i,B_j)$.

Below we provide a pseudo code for the detailed checking algorithm used in this study:

```
Input Text A
Input Text B
Choose segmentation method {W3G, W5G, W8G3W, S2S}
Apply FEM for Text A
Apply FEM for Text B
For each segment Ai ∈ A
  For each Segment Bj ∈ B
    Input Ai and Bj to fuzzy inference engine
    Compute SIM(Ai,Bj)
    If PD(Ai,Bj) is true
      Output (Ai,Bj)
```

4.5. Post-processing

Because of using sentences/k-grams as comparison schemes, post-processing is required to merge subsequent sentences or k-grams detected as plagiarism into passages/paragraphs. The notion of *citation evidence*, which refers to the cited text, citation marker or the word/number used to link the cited text with one of the references and the reference phrase, has been

Table 4 Comparison of sentence similarity in a paraphrased plagiarism case.

Sentence pairs	$\mu_{A,B}$	$\mu_{B,A}$	MIN	DIFF
A_2 v.s. B_2	0.4857	0.5	0.4857	0.0143
A_3 v.s. B_3	0.7856	0.9075	0.7856	0.1219

Part of semantic similarity of word pairs in Sentences A_2 and B_2 are as follows: wup(say,say) = 1.0, wup(say,give) = 0.875, wup(say,something) = 0, wup(say,nice) = 0, wup(think,say) = 0.5714, wup(think,give) = 0.8, wup(think,something) = 0, wup(think,nice) = 0,. . ., wup(present,give) = 1.0; while in A_3 and B_3 are wup(float,hover) = 0.5714, wup(float,. . .) = 0,. . ., wup(declare,say) = 0.8571, . . .,wup(ever,ever) = 1.

used in PD research by Alzahrani et al. (2012). Similar texts that have no *citation evidence* can be judged as plagiarism while those with citation evidence should be excluded during the post-processing stage. Another exclusion should be made for small matches (*n*-grams where $n < 4$) that are surrounded by plagiarism-free texts as they are more likely to be unimportant and can be discarded by the plagiarism checker.

5. Experimental design

5.1. WordNet taxonomy

WordNet is an English dictionary that contains more hierarchical *lexes* which are arranged into groups called *synsets* (synonyms sets) Miller, 1995. Hierarchical taxonomies are constructed such that synsets that share a common property are organized under a shared *hypernym* which convey the meaning of that property. Synsets may also have some more specialized or composite lexes called *hyponyms*. POS tags used in WordNet are noun, verb, adjective and adverb, which required us to do some mapping (or simplification) of Treebank tags used in the POS disambiguation step (refer to the FEM algorithm, in Section 2, for more details) into WordNet tags.

5.2. Datasets

To evaluate the proposed method, we used a total of 99,033 ground-truth annotated cases extracted from different datasets, as shown in Table 5. Each case was defined as a quadruple $\rho = (Method, Obfuscation, S_{source}, S_{suspicious})$ where *Method* defines the method of construction used in each case which can be one of the following: *manual paraphrases*, *artificial paraphrases*, and *plagiarism-free*. Manual (also called handmade or simulated) plagiarism cases are constructed by humans who rewrite a source text in different words but maintain the same ideas in the source text and pretend neither to quote nor to use any citation evidence. Artificial plagiarism cases, on the other hand, are constructed automatically using plagiarism synthesizers (i.e., computer programs similar to automatic paraphrasers used to synthesize plagiarism from natural language sources texts). Texts are changed automatically by restructuring words/phrases/sentences, substituting words, and/or replacing words with synonyms. Plagiarism synthesizers, also called artificial plagiarists, are described in detail by Potthast et al. (2009, 2010a) and Alzahrani et al. (2012).

Obfuscation, on the other hand, refers to the degree of complexity (i.e. number of edit operations needed to convert one text into another) with regard to the original source. It can take one of the following values: *none* if no (or very few) changes were done in the suspicious text with regard to its original version, *low* if moderate number of words were altered, and *high* otherwise. In Table 5, we considered simulated plagiarism cases as highly obfuscated while artificial plagiarism cases can be of none, low or high obfuscation as annotated by the plagiarism synthesizer. Besides, in the quadruple ρ, S_{source} refers to the source text extracted from the source document d_{source} (i.e., original document in the test collection archives), and $S_{suspicious}$ refers to the suspicious text from $d_{suspicious}$ to be judged against plagiarism.

As can be seen in Table 5, the first two corpora, PAN-PC-11 (Potthast et al., 2011) and PAN-PC-10 (Potthast et al., 2010a,b), include 7645 manual paraphrases and 34,310 automatic paraphrases. In both datasets, the PAN's organization committee placed several human intelligent tasks (HITs) via the Amazon Mechanical Turk (Potthast et al., 2010a), whereby people were asked to rewrite/rephrase given source texts in their own words. PAN-PC-09 (Potthast et al., 2009b) involve 17,127 artificial cases but no simulated plagiarism cases were found. We ignored translated plagiarism cases found in the previous three corpora as well as verbatim plagiarism cases. Another 3,378 plagiarism cases were extracted from ALZAHRANI-PC (Alzahrani et al., 2012), constructed automatically using a plagiarism synthesizer software[3]. We ignored cases like translated and summarized plagiarism, as they are not within the scope of this study. Extracted plagiarism cases from ALZAHRANI-PC (Alzahrani et al., 2012) have three obfuscation degrees: none (i.e., exact copy), low (i.e., with small alterations such as words shuffling, removing or ordering), and high (i.e., deep word replacements with synonyms). We also used CLOUGH-PC (Clough and Stevenson, 2011) which contains 95 handmade cases synthesized from five Wikipedia articles. Multiple changes with regard to the source texts were given in about 76 cases. Microsoft paraphrase corpus (Dolan et al., 2004) include a total of 5,801 small-length paraphrase cases taken from different news sources. Two human raters judged each pair as semantically equivalent or not, and a third rater was consulted if the decisions made by former raters were different. Accordingly, 3900 were judged as paraphrased cases and 1901 as non-paraphrased cases. Finally, we included 30,677 plagiarism-free cases from ALZAHRANI-PC (Alzahrani et al., 2012), which would be useful to test the ability of PD methods to avoid false positives.

5.3. Baselines

N-gram based approaches are considered the dominant PD methods, which generally use chunking and matching the overlap between textual documents. We adopted four PD methods,

[3] Please email the corresponding author to obtain the dataset.

Table 5 Details of plagiarism cases used in the study.

Datasets	Ref.	#Manual paraphrases	#Artificial paraphrases	Degree of obfuscation			#Plagiarism free	#Cases
				None	Low	High		
PAN-PC-11	Potthast et al. (2011)	4609	18,179	–	11,779	6400	–	22,788
PAN-PC-10	Potthast et al. (2010a,b)	3036	16,131	–	9750	6381	–	19,167
PAN-PC-09	Potthast et al. (2009b)	–	17,127	–	10,764	6363	–	17,127
ALZAHRANI-PC	Alzahrani et al. (2012)	–	3378	1120	1120	1138	30,677	34,055
CLOUGH-PC	Clough and Stevenson (2011)	76	–	19	19	57	–	95
MS-PARAPHRASE	Dolan et al. (2004)	3900	–	–	–	3900	1901	5801
Total instances		11,621 (11.7%)	54,815 (55.4%)	1139 (1.15%)	33,432 (33.8%)	24,239 (24.5%)	32,578 (32.9%)	99,033 (100%)

Datasets are grouped as follows: MANUAL-PARAPHRASE dataset (11,621 manually paraphrased cases, and 32,578 non-paraphrased cases), and ARTIFICIAL-PARAPHRASE dataset (54,815 artificially paraphrased cases, and 32,578 non-paraphrased cases).

which have been commonly used in existing plagiarism detectors, namely matching of word 3-gram, matching of word 5-gram (Kasprzak et al., 2009), matching of word 8-gram (Basile et al., 2009) with 3-word overlapping, and sentence-to-sentence matching (Alzahrani and Salim, 2010). In our experiments, we referred to these baselines as B1-W3G, B2-W5G, B3-W8G3W, and B3-S2S, respectively. Our proposed method is considered a modification of the former fuzzy-set IR approach in Yerra and Ng (2005); thus, we used it as another baseline for this study, referred to as B5-FIR. We used the *yer* metric in (21) as a membership function, and we used the Gutenberg text collection provided by the NLTK project[4] to compute this formula as a pre-processing step.

5.4. Stratified 10-fold cross-validation

There might be some criticism about the mixture of manual (handmade) and artificial plagiarism cases introduced in Section 5.2. One may think that artificial plagiarism cases are not as accurate as handmade cases, which is true in the sense that synonyms choice by artificial plagiarism synthesizers may not be as good as synonyms choice by humans. Similarly, maintaining the linguistic rules (e.g., grammar) by humans should be more accurate than by artificial synthesizers. Consequently, we preferred to separate the datasets into two groups:

- *Manual-Paraphrase* group (11,621 manual paraphrases, and 32,578 plagiarism-free cases).
- *Artificial-Paraphrase* group (54,815 artificial paraphrase, and 32,578 plagiarism-free cases).

In this study, a stratified 10-fold cross-validation was performed to obtain PD results on each dataset. Plagiarism cases with different degrees of obfuscation as well as plagiarism-free cases were divided equivalently into ten folds before cross-validation was performed. Tables 6 and 7 show the details of 10-fold cross-validation data obtained from manual dataset and artificial dataset, respectively. In the tables, the number of plagiarism and plagiarism-free cases is almost comparable between all folds in each dataset. Likewise, obfuscated plagiarism cases were stratified such that each fold contains cases with none, low and high obfuscation. Obfuscation was tagged in the artificial plagiarism cases during the construction by the

Table 6 Details of 10-fold cross-validation data for manual-paraphrase dataset.

Fold#	Obfuscation			Plagiarism-free	Total cases
	None	Low	High		
Fold1	58	278	828	3257	4421
Fold2	47	306	811	3257	4421
Fold3	46	291	827	3257	4421
Fold4	27	177	960	3257	4421
Fold5	8	70	1086	3257	4421
Fold6	15	65	1084	3257	4421
Fold7	4	86	1074	3257	4421
Fold8	7	64	1093	3257	4421
Fold9	15	67	1082	3257	4421
Fold10	15	73	1076	3265	4429

Table 7 Details of 10-fold cross-validation data for artificial-paraphrase dataset.

Fold#	Obfuscation			Plagiarism-free	Total cases
	None	Low	High		
Fold1	112	3785	1922	3257	8964
Fold2	112	3730	1977	3257	8964
Fold3	112	3708	1999	3257	8964
Fold4	112	3817	1890	3257	8964
Fold5	112	3839	1868	3257	8964
Fold6	112	3780	1927	3257	8964
Fold7	112	3745	1962	3257	8964
Fold8	112	1287	1045	3257	5589
Fold9	112	2849	2858	3257	8964
Fold10	112	2873	2834	3265	8972

[4] http://nltk.googlecode.com/svn/trunk/doc/book/ch02.html.

artificial plagiarists, but not tagged in the handmade cases (except Clough's dataset (Clough and Stevenson, 2011), which unfortunately has a limited number of cases). We presumed in Section 5.2 that manual cases can be considered highly obfuscated. In our opinion, it might still be convenient to count the percentage of exact words shared between texts (we used the relation δ = |set of common words|/|set of unified words| to compute these percentages). According to the computed percentages between text pairs in each ρ, we roughly stratified the manual cases as low ($\delta > 70\%$) and high ($\delta < 70\%$).

To perform the 10-fold cross validation, ten experiments were performed independently. In one experiment, we fine-tuned the algorithm (e.g., updated thresholds) on 9 folds such that better results can be obtained, while the remaining one fold was used as test data to report the final result. In a next experiment, we used a fold different from the one used in previous experiment to report the result. We repeated the experiments until all ten folds were involved as test data.

5.5. Evaluation measures

To evaluate the methods used in this study, we implemented precision (P_{plag}), recall (R_{plag}), the harmonic-mean (F_{plag}), granularity (G_{plag}), and plagiarism score ($\text{Score}_{\text{plag}}$) Alzahrani et al., 2012; Potthast et al., 2010a. Precision, recall and F-measure are defined in ().

$$P_{plag} = \frac{TP}{TP+FP}, \quad R_{plag} = \frac{TP}{TP+FN}, \quad F_{plag} = 2 \times \frac{P_{plag} \times R_{plag}}{P_{plag} + R_{plag}} \tag{31}$$

where TP refers to the number of correct plagiarism cases as defined in the quadruple ρ of each case, FP refers to the number of false detections of cases annotated as *plagiarism-free*, and FN refers to the number of plagiarism cases that are not detected as plagiarism. Further, granularity of ρ measures the ability of the detection algorithm to detect that case at once (Potthast et al., 2010a). To illustrate, methods that are based on small comparison units (e.g., sentences or *n*-grams) should be able to merge consequent small detections into coherent passages. In the meanwhile, PD methods should be able to ignore small detections that do not constitute much of the text. We used the Eq. (32) for granularity:

$$G_{plag} = \frac{N\rho_{detected} : \rho_{detected} \subseteq \rho_{annotated}}{N\rho_{annotated}} \tag{32}$$

where $N\rho_{\text{det ected}}$ is the number of true detections (i.e., intersects – partially or totally – with one plagiarism case), and $N\rho_{annotated}$ denotes the number of annotated cases in $d_{suspicious}$. Evaluation measures are combined into a single value (33), which can be used to make a quantitative comparison of PD algorithms.

$$Score_{plag} = \frac{F_{plag}}{\log_2(1 + G_{plag})} \tag{33}$$

5.6. Parameter setting

We conducted ad-hoc experiments to set up the ideal *permission threshold* value, referred to as p in (29) and (30), with four segmentation schemes. Then, another ad-hoc experiment was performed to choose the optimal *variation threshold,* called v in (29), for comparing sentences.

In both setups, we used the stratified 10-fold cross-validation data from the manual-paraphrase dataset (see Table 5 for details of plagiarism cases).

Fig. 5 shows plagiarism scores obtained based on (33) using four segmentation schemes; S2S (a), W3G (b), W5G (c), and W8G3W (d). In all experiments, we assigned p successive values from 0 to 1 with 0.05 increment in each run. To simplify, the figure shows only two folds (fold 2 vs. fold 5) since we noticed a similar behavior in the other folds. The optimum plagiarism score for S2S was obtained when $p \in [0.75,0.80]$. We can select $p = 0.78$ for S2S, accordingly. The best score for W3G was almost obtained when $p = 0.95$, which is reasonable, as we observed that the semantic similarity values between word 3-grams are always high and may lead to many false positives; hence, high threshold value is ideal. For W5G and W8G3W, we found that the best plagiarism results were obtained at the interval [0.80 and 0.85] in different folds but it is more solid when $p = 0.80$.

On the other hand, v was used to additionally reduce false detections in sentence-to-sentence matching (Yerra and Ng, 2005). We experimented different v values in the interval [0,0.3] with 0.01 increment in each run (it is not expected that sentences that passed p will have a *difference* between their similarities more than 0.2). In Fig. 6, we observed that when v equals 0.22, the plagiarism scores stabilize at the most optimal value (0.7883 in fold 2; 0.7712 in fold 5; and alike in other folds).

5.7. Statistical analysis

Results from the proposed method were compared statistically with the state-of the-art baselines discussed in Section 5.2. We examined the statistical significance using t hypothesis testing (Leech et al., 2008). To conduct statistical t-test, we set a null hypothesis that "the fuzzy semantic-based PD approach and the traditional PD method perform equally (i.e., the true mean difference is zero)", and work to gather evidence against this null hypothesis. The traditional PD method could be one of the baselines implemented in the study.

As the cross-validation technique yields 10-fold pairs of plagiarism score ($\text{Score}_{\text{plag}}$) values from compared algorithms, a paired t-test (Leech et al., 2008) was used to reject/do not reject the null hypothesis. To carry out the paired t-test on 10-fold cross-validation results ($k = 10$), we calculate the difference of the results obtained from two algorithms in each fold as $d_i = x_i - y_i$, *where* $i = 1,2,\ldots,k$, and x_i refers to $\text{Score}_{\text{plag}}$ value obtained from the traditional plagiarism detection method on the ith fold, and y_i refers to $\text{Score}_{\text{plag}}$ value obtained from the proposed algorithm on the ith fold. The mean difference was computed based on (34) and standard deviation of the mean differences across the k folds was computed as in (35).

$$\bar{d} = \left(\sum_{i=1}^{k} d_i\right)/k \tag{34}$$

$$\alpha = \sqrt{\sum_{i=1}^{k} (d_i - \bar{d})^2/(k-1)} \tag{35}$$

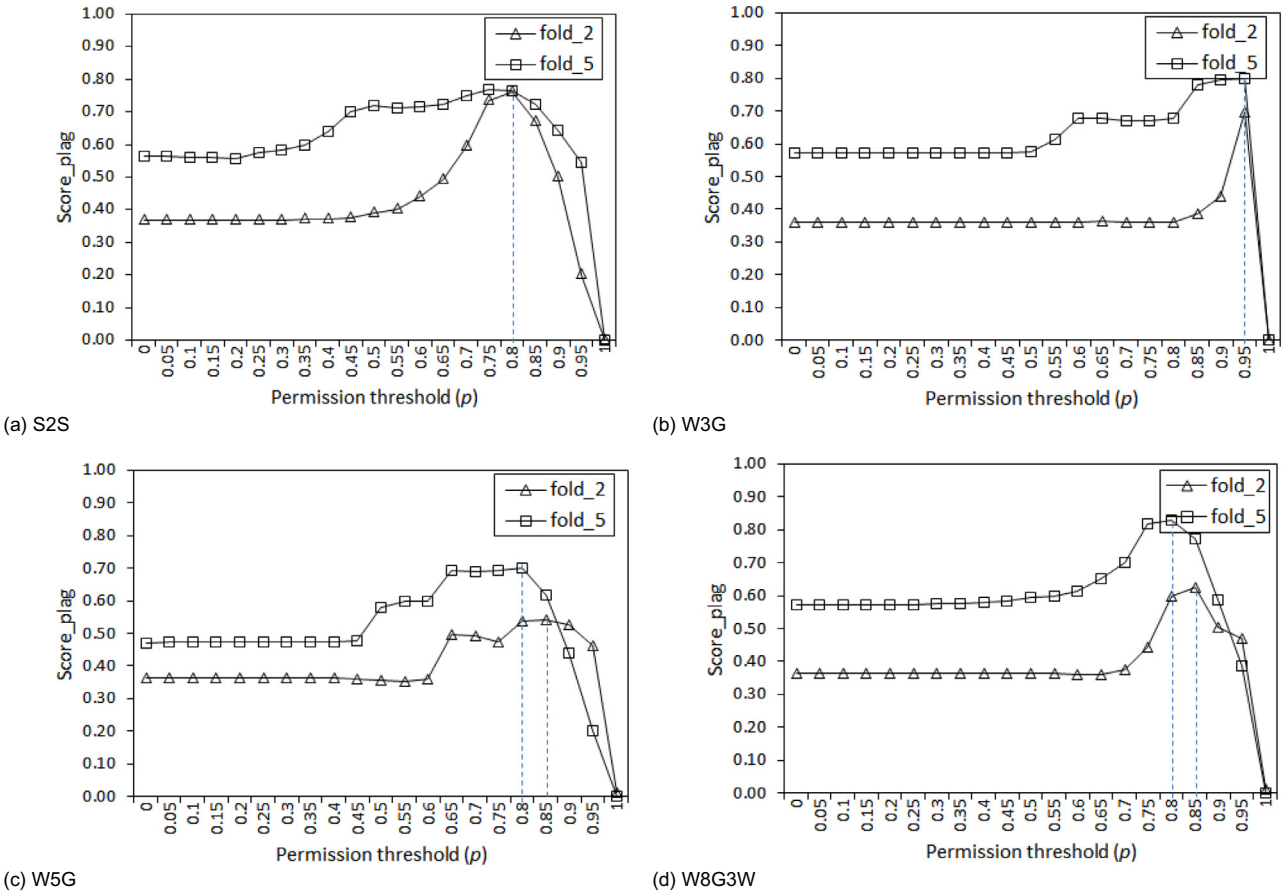


Figure 5 Plagiarism scores obtained with different permission thresholds in the interval [0,1] with 0.05 increment, and using four segmentation schemes; (a) sentences (S2S), (b) word 3-grams (W3G), (c) word 5-grams (W5G), and (d) word 8-grams with 3-word overlapping (W8G3W). The graphs show two different folds from the manual-paraphrase dataset.

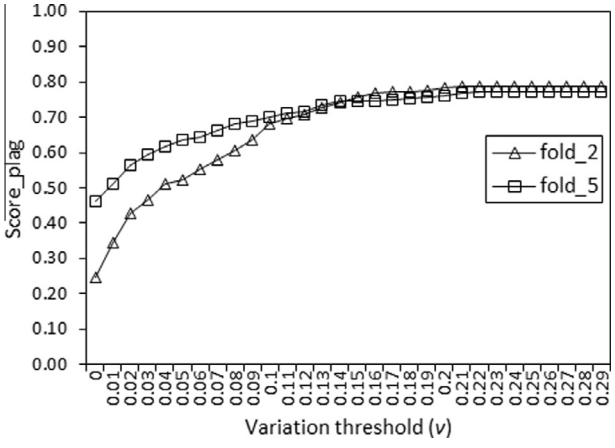


Figure 6 Plagiarism scores obtained with variation thresholds in the interval [0,1] with 0.02 increment, when using S2S segmentation scheme and permission threshold 0.78.

We used α to compute the standard error $SE(\bar{d}) = \alpha/\sqrt{k}$, and the t-statistic $T = \bar{d}/SE(\bar{d})$, which, under the null hypothesis, follows a normal distribution with $k-1$ degrees of freedom. Using t-distribution table,[5] we compared T to the tk-1 distribution to obtain the probability value, referred to as p-value, which answers the alternative hypothesis (i.e., we could decide to reject/do not reject the null hypothesis).

Besides, we conducted a statistical test to see whether or not there is a statistical difference between different segmentation schemes used in the proposed model. ANOVA (ANalysis Of VAriance) statistical test, which generalizes the paired t-test, was used to examine the statistical importance of the results obtained from several algorithms (Leech et al., 2008). We set a null hypothesis that "All segmentation schemes used with the proposed fuzzy semantic-based model; namely W3G, W5G, W8G3W and S2S, perform equally". Then, evidence against the null hypothesis was used – at least one of the segmentation schemes is significantly different.

6. Results and discussion

In this section, we initially present the results obtained from two sentence benchmarks to find out how similar/dissimilar two pair of texts might be using our approach. The majority of the experimental works investigates the effectiveness of the proposed approach on handmade versus artificial plagiarism datasets. Besides, we present the results from both datasets using four segmentation schemes to be extensively

[5] http://www.statsoft.com/textbook/distribution-tables/#t.

compared with well-known plagiarism detection methods found in the literature. We referred to the proposed approach in the experimental work as FS-W3G, FS-W5G, FS-W8G3W, and FS-S2S (FS refers to the fuzzy semantic-based method, and W3G, W5G, W8G3W, and S2S are the segmentation schemes introduced in Section 3.1). Five baselines as previously mentioned; four of them use typical string matching and overlapping approach namely B1-W3G, B2-W5G, B3-B8G3W, and B4-S2S, while the fifth one uses sentence-based FIR similarity, which we referred to as B5-FIR. We also present statistical analysis results and provide discussion on the statistical significance among different algorithms used in this study.

6.1. Results from sentence samples

Results from two sentence sample sets selected from different literature papers are presented here. Table 8 shows eight sentence pairs chosen from different papers and books on natural language understanding (Li et al., 2006). Li et al. (2006) stated that computed similarity values were found to be fairly consistent with the human intuition. We found sentence similarity values by our approach have a small difference with computed similarities given by Li et al. (2006) (the mean difference≈0.075, standard deviation ≈ 0.069, correlation coefficient[6] ≈ 0.948). There is a bit difference in Pair H, which could be due to reduction of the number of words by our FEM while every word was accounted in Li et al. (2006). Another possibility may be related to the POS-related semantic spaces where we compare words with the same POS tags, and hence we compared the words "dog", "animal" with the word "pet", and other words were discarded yielding about 0.5 minimum similarity.

As the previous evaluation was shown on very short-length sentence pairs, other medium-length sentences were designed by Lee (2011). Table 9 shows seven sentence triples, and the computed similarity scores of sentence pairs in each triple. As there was no human similarity scoring procedure in their study, we just presented our results alongside with Lee's results (Lee, 2011) and the judgment is left to the reader. Similarity results obtained from our approach have a correlation coefficient ≈ 0.867 with reported results in Lee (2011). Nevertheless, we did not pursue any further human rating on sentence pairs. This is because all plagiarism cases used in this study were extracted from standard and ground-truth annotated data used for paraphrase detection (Dolan et al., 2004) and plagiarism detection research, (Alzahrani et al., 2012; Potthast et al., 2010a,b, 2011; Potthast et al., 2009b; Clough and Stevenson, 2011). In each plagiarism case, a pair of texts is provided, which contains various sentences, up to a paragraph length, and annotated as plagiarism or plagiarism-free.

6.2. Results from fuzzy semantic-based method and baselines on different datasets

This section covers the experimental work that we carried out to validate the proposed model. To evaluate different

[6] Correlation coefficient measures the strength of linear relationship between two methods based on the covariance of the methods divided by their standard deviation.

Table 8 Experimental results on raw sentences of short lengths.

Sentence pairs			Similarity		Sentence pairs			Similarity	
			Li et al. (2006)	FS-S2S				Li et al. (2006)	FS-S2S
Pair A	Sentence A-1	I like that bachelor	0.561	0.649	Pair E	Sentence E-1	It is a red alcoholic drink	0.000	0.074
	Sentence A-2	I like that unmarried man				Sentence E-2	It is an English dictionary		
Pair B	Sentence B-1	I have a pen	0.000	0.000	Pair F	Sentence F-1	Canis familiaris are animals	0.362	0.391
	Sentence B-2	Where do you live				Sentence F-2	They are common pets		
Pair C	Sentence C-1	John is very nice	0.977	1.00	Pair G	Sentence G-1	It is a glass of cider	0.678	0.652
	Sentence C-2	Is John very nice?				Sentence G-2	It is a full cup of apple juice		
Pair D	Sentence D-1	It is a dog	0.623	0.737	Pair H	Sentence H-1	Dogs are animals	0.738	0.494
	Sentence D-2	It is a log				Sentence H-2	They are common pets		

Similarity of sentence pairs is computed based on fuzzy-semantic based approach (FS-S2S) and compared with the approach from Li et al. (2006) which is based on semantic nets and corpus statistics. The mean difference ≈ 0.075, standard deviation ≈ 0.069 and correlation coefficient ≈ 0.948.

Table 9 Experimental Results on Raw Sentences of Moderate Lengths.

Sentence Triples	Raw texts		
Triple A			
Sentence A-1	If she can be more considerate to others, she will be more popular		
Sentence A-2	She is not considerate enough to be more popular to others		
Sentence A-3	You are not supposed to touch any of the art works in this exhibition		
Similarity (Lee, 2011)	A-1 v.s. A-2 = 0.9125	A-1 v.s. A-3 = 0.01956859	A-2 v.s. A-3 = 0.02903207
Similarity (FS-S2S)	A-1 v.s. A-2 = 0.75	A-1 v.s. A-3 = 0.00	A-2 v.s. A-3 = 0.00
Triple B			
Sentence B-1	I won't give you a second chance unless you promise to be careful this time		
Sentence B-2	If you could promise to be careful, I would consider to give you a second chance		
Sentence B-3	The obscurity of the language means that few people are able to understand the new legislation		
Similarity (Lee, 2011)	B-1 v.s. B-2 = 0.9384236	B-1 v.s. B-3 = 0.4190409	B-2 v.s. B-3 = 0.3293912
Similarity (FS-S2S)	B-1 v.s. B-2 = 0.9333333	B-1 v.s. B-3 = 0.3575533	B-2 v.s. B-3 = 0.4857226
Triple C			
Sentence C-1	About 100 officers in riot gear were needed to break up the fight		
Sentence C-2	The army entered in the forest to stop the fight with weapon		
Sentence C-3	He thus avoided a pack of journalists eager to question him		
Similarity (Lee, 2011)	C-1 v.s. C-2 = 0.6952305	C-1 v.s. C-3 = 0.4072169	C-2 v.s. C-3 = 0.5830132
Similarity (Fuzzy-Sem)	C-1 v.s. C-2 = 0.8774377	C-1 v.s. C-3 = 0.7006131	C-2 v.s. C-3 = 0.6885147
Triple D			
Sentence D-1	Your digestive system is the organs in your body that digest the food you eat		
Sentence D-2	Stomach is one of organs in human body to digest the food you eat		
Sentence D-3	We had better wait to see what our competitors do before we make a move		
Similarity (Lee, 2011)	D-1 v.s. D-2 = 0.9187595	D-1 v.s. D-3 = 0.2684233	D-2 v.s. D-3 = 0.2639506
Similarity (FS-S2S)	D-1 v.s. D-2 = 0.7774170	D-1 v.s. D-3 = 0.2225959	D-2 v.s. D-3 = 0.2299756
Triple E			
Sentence E-1	I don't think it is a clever idea to use an illegal means to get what you want		
Sentence E-2	It is an illegal way to get what you want, you should stop and think carefully		
Sentence E-3	There is something wrong with the steel supporting member of the device		
Similarity (Lee, 2011)	E-1 v.s. E-2 = 0.5911233	E-1 v.s. E-3 = 0.2679752	E-2 v.s. E-3 = 0.1166667
Similarity (FS-S2S)	E-1 v.s. E-2 = 0.7180556	E-1 v.s. E-3 = 0.3418523	E-2 v.s. E-3 = 0.26703297
Triple F			
Sentence F-1	The powerful authority is partial to the members in the same party with it		
Sentence F-2	Political person sometimes abuse their authority that it is unfair to the citizen		
Sentence F-3	He reasoned that we could be there by noon if we started at dawn		
Similarity (Lee, 2011)	F-1 v.s. F-2 = 0.872057	F-1 v.s. F-3 = 0.1842038	F-2 v.s. F-3 = 0.1540446
Similarity (FS-S2S)	F-1 v.s. F-2 = 0.422338	F-1 v.s. F-3 = 0.3403922	F-2 v.s. F-3 = 0.2775399
Triple G			
Sentence G-1	The fire department is an organization which has the job of putting out fires		
Sentence G-2	An organization which has the job of putting out fires is the fire department		
Sentence G-3	The man wore a bathrobe and had evidently just come from the bathroom		
Similarity (Lee, 2011)	G-1 v.s. G-2 = 1.00	G-1 v.s. G-3 = 0.5586169	G-2 v.s. G-3 = 0.5586169
Similarity (FS-S2S)	G-1 v.s. G-2 = 1.00	G-1 v.s. G-3 = 0.4826319	G-2 v.s. G-3 = 0.4826319

Similarity of sentence triples is computed based on the proposed approach (FS-S2S) and compared with the semantic similarity measure by Lee (2011). The mean difference ≈ 0.118, standard deviation ≈ 0.106 and correlation coefficient ≈ 0.867 between results from both methods.

PD methods implemented in this study, we used precision, recall, and $\text{Score}_{\text{plag}}$ averaged over the ten-fold cross-validation data. Table 10 presents the results obtained from the baselines on manual and paraphrase datasets (top half of the table), and the results from the proposed methods on both datasets as well (bottom half of the table). Again, each row in the table shows the mean precision, mean recall, and mean $\text{Score}_{\text{plag}}$ when we performed the experiments on 10 folds. Fig. 7 visualizes the same results from manual and artificial datasets drawn in Table 10. The results are discussed in the following paragraphs.

The performance of the word *n*-gram-based string matching baselines, B1-W3G, B2-W5G, B3-W8G3W, and B4-S2S, was overall weak as the highest recall result obtained was 0.1156 on the manual paraphrases, and 0.1823 on the artificial paraphrases, using B1-W3G. Near-optimum precision achieved by these baselines is unsurprising since exact string matching can "precisely" detect plagiarism by copying parts from the source text and, therefore, no false positives would be expected using these approaches. Three of these baselines, B2-W5G, B3-W8G3W, and B4-S2S, have been used in our previous work (Alzahrani et al., 2012), yet their performance is even poorer in this paper because we used obfuscated plagiarism cases here (our previous dataset in Alzahrani et al. (2012) includes verbatim and near copy plagiarism cases as well).

Table 10 Results from baselines and fuzzy semantic-based method on manual vs. artificial datasets.

Manual-paraphrase	Baseline method	P_{plag}	R_{plag}	G_{plag}	$Score_{plag}$	Std. deviation
State-of-the-art baselines	B1-W3G	0.9803	0.1156	1.0000	0.1939	0.1461 (0.2929)
	B2-W5G	0.9751	0.0448	1.0000	0.0820	0.0824 (0.2929)
	B3- W8G3W	0.5722	0.0078	1.0000	0.0153	0.0173 (0.2929)
	B4-S2S	0.8977	0.0306	1.0000	0.0588	0.0270 (0.2929)
	B5-FIR	**0.6920**	**0.8673**	**1.0000**	**0.7646**	**0.1050 (0.2929)**
Fuzzy semantic-based approach	**FS-W3G**	**0.8844**	**0.6948**	**1.0000**	**0.7740**	**0.0624 (0.1168)**
	FS-W5G	0.8007	0.6431	1.0000	0.7110	0.1490 (0.1168)
	FS-W8G3W	0.7594	0.7550	1.0000	0.7524	0.1663 (0.1168)
	FS-S2S	**0.9178**	**0.6933**	**1.0000**	**0.7850**	**0.0421 (0.1168)**
Artificial-paraphrase	Segmentation method					
State-of-the-art baselines	B1-W3G	0.9389	0.1823	1.0000	0.2553	0.2962 (0.3055)
	B2-W5G	0.7153	0.0589	1.0000	0.0990	0.1534 (0.3055)
	B3-W8G3W	0.4246	0.0072	1.0000	0.0140	0.0242 (0.3055)
	B4-S2S	0.7000	0.0110	1.0000	0.0214	0.0262 (0.3055)
	B5-FIR	0.5568	0.6289	1.0000	0.5907	0.0671 (0.3055)
Fuzzy semantic-based approach	FS-W3G	0.6924	0.7302	1.0000	0.7040	0.1078 (0.1502)
	FS-W5G	0.3836	0.4723	1.0000	0.4060	0.0787 (0.1502)
	FS-W8G3W	0.3684	0.6952	1.0000	0.4712	0.0607 (0.1502)
	FS-S2S	0.6975	0.6138	1.0000	0.6445	0.1010 (0.1502)

Highest results obtained by the state-of-the-art baselines and the proposed methods are shown in bold.
The first four columns give the mean precision, recall, granularity and score of plagiarism over all folds. The last column shows the standard deviation over 10 runs of cross-validation in each approach, as well as the standard deviation of the means over all approaches, in parentheses.

The fifth baseline B5-FIR showed superior performance in comparison to other baselines (mean $Score_{plag} = 0.7784$). Since this baseline used word correlation factors to measure similarity of words, it can detect sentences that have been reworded (Yerra and Ng, 2005).

On the other hand, fuzzy semantic-based approach showed encouraging results, as we obtained up to 0.9178 precision, 0.6933 recall, and 0.7850 $Score_{plag}$ using FS-S2S on manual-paraphrase dataset, and up to 0.6974 precision, 0.7302 recall, and 0.7040 $Score_{plag}$ using FS-W3G on artificial-paraphrase dataset. Our results were superior to the results obtained from B1-W3G, B2-W5G, B3-W8G3W, and B4-S2S baselines. In comparison with B5-FIR, it can be observed that precision results obtained by our approach using four segmentation schemes were basically higher than in B5-FIR. Experimental $Score_{plag}$ results obtained from FS-W3G and FS-S2S were slightly better than that obtained from B5-FIR. However, we cannot say whether or not the results from the proposed approaches, namely FS-W3G and FS-S2S, are significantly better than B5-FIR before conducting a statistical test, which we will present in the next section.

Further, we cannot tell which segmentation scheme works better with (or professionally can handle) obfuscated plagiarism cases used in the dataset; therefore, the analysis of variance test (ANOVA) will be conducted to compare results from different schemes. Roughly it can be seen that the utmost precision and $Score_{plag}$ was yielded using sentences FS-S2S over other methods, namely FS-W3G, FS-W5G, and FS-W8G3W.

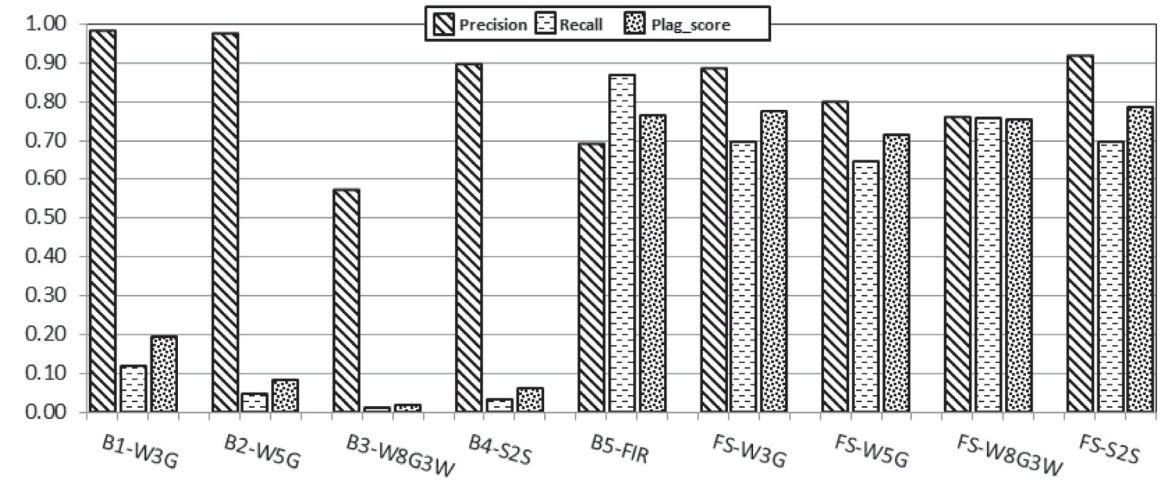
Finally, we noticed that manual and artificial datasets behaved differently. The accuracy of the results on handmade paraphrases was overall exceeding that on artificial paraphrases given the same segmentation scheme. It can be observed that FS-W3G showed the optimal performance in terms of $Score_{plag}$ when using artificial cases, while FS-S2S performed well with manual cases.

6.3. Statistical results

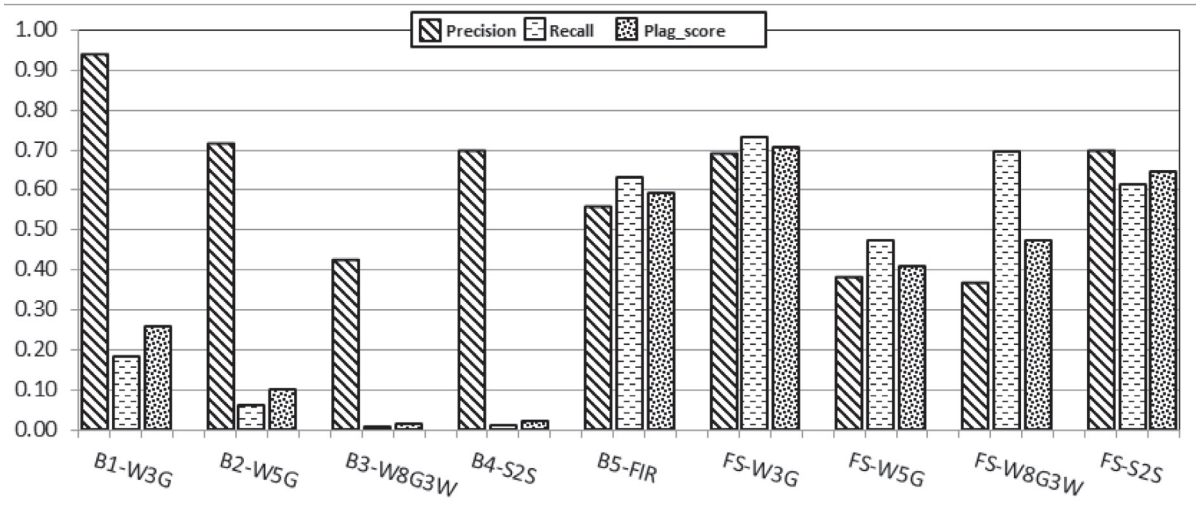
In this section, we present the result obtained from the dependent-sample (i.e., using the same 10 folds in both algorithms) paired t-test of the proposed model and former baselines. Here we included the statistical results from the *manual-paraphrase* dataset.

Table 10 shows that the standard deviation for each method over 10 runs of cross-validation was relatively small, which apparently means that there is a slight variance between the results obtained from each fold. This indicates that both datasets used for experiments were equally stratified and the methods behaved in a similar way over the 10 runs using both datasets. The standard deviation of the means over all PD methods, shown in parentheses, could indicate the performance variance among different methods using manual dataset on one hand, and using artificial dataset on the other hand. The higher standard deviation indicates that the results may possibly be unreliable, but further statistical analysis should be done.

Table 11 shows the statistical results between the proposed method using sentences as a segmentation scheme (FS-S2S) versus typical sentence-based string matching baseline (B4-S2S). The table shows that a paired t-test revealed a statistically reliable difference of two $Score_{plag}$ means from FS-S2S and B4-S2S, which led to reject the null hypothesis. The table shows that t-stastic $= -54.9077$ is greater than t-critical $= \pm 2.2622$, with 0.95 confidence level. Similar paired t-tests were conducted between the proposed methods and other string matching baselines namely B1-W3G, B2-W5G, and B3-W5G3W but not shown due to space limitations in this



(a) Manual dataset



(b) Artificial dataset

Figure 7 Recall, precision, and plagiarism score results from fuzzy-semantic-based method with four segmentation schemes (FS-W3G, FS-W5G, FS-W8G3W, FS-S2S) and baselines (B1-W3G, B2-W5G, B3-W8G3W, B4-S2S, B5-FIR).

Table 11 Statistical results from dependent-sample paired *t*-test of fuzzy semantic-based model using sentences (FS-S2S) and sentence matching baseline (B4-S2S).

Hypothesis test for the difference of two means from 10-fold cross-validation data				
Statistics	Two-tailed test	B4-S2S	FS-S2S	Difference
Hypothesis =	B4-S2S = FS-S2S	0.0327	0.6918	−0.6591
Alternative hypothesis =	B4-S2S ≠ FS-S2S	0.0224	0.8105	−0.7881
Alpha level =	0.05	0.0185	0.8045	−0.7860
Mean differences =	−0.7261	0.0636	0.7906	−0.7270
Standard deviation =	0.0418	0.0630	0.7488	−0.6859
Sample size =	10	0.0612	0.7630	−0.7018
t-Statistic =	−54.9077	0.0678	0.7756	−0.7078
t-critical value =	±2.2622	0.1006	0.8261	−0.7255
p-Value =	0.000000000001	0.0895	0.8086	−0.7190
Decision =	Reject hypothesis	0.0692	0.8302	−0.7610
Confidence interval for paired difference				
Confidence level		0.95		
Confidence interval		−0.7560 < μd < −0.6962		

Table 12 Statistical results from dependent-sample paired *t*-test of fuzzy semantic-based model (FS-S2S) and fuzzy IR baseline (B5-FIR).

Hypothesis test for the difference of the two means from 10-fold cross-validation data				
Statistics	Two-tailed test	B5-FIR	FS-S2S	Difference
Hypothesis =	B5-FIR = FS-S2S	0.7171	0.6918	0.0253
Alternative hypothesis =	B5-FIR ≠ FS-S2S	0.6408	0.8105	−0.1697
Alpha level =	0.05	0.5922	0.8045	−0.2124
Mean differences =	−0.0204	0.8927	0.7906	0.1021
Standard deviation =	0.1189	0.8497	0.7488	0.1009
Sample size =	10	0.9178	0.7630	0.1549
t-Statistic =	−0.5426	0.7168	0.7756	−0.0589
t-Critical value =	±2.2622	0.8104	0.8261	−0.0158
p-Value =	0.60056	0.7426	0.8086	−0.0660
Decision =	Do not reject hypothesis	0.7657	0.8302	−0.0645
Confidence interval for paired difference				
Confidence level		0.95		
Confidence interval		−0.1054 < μd < 0.06464		

paper. The same results were concluded, confirming that fuzzy semantic-based model, no matter what segmentation scheme has been used, showed statistically significant results in comparison with these baselines. Table 12 shows another paired sample *t*-test between B5-FIR and FS-S2S. The test failed to reveal a statistically reliable difference between the proposed method and this baseline. Thus, the null hypothesis that says that "both of fuzzy-set IR method denoted as B5-FIR and fuzzy semantic-based method denoted as FS-S2S behaved equally to detecting obfuscated plagiarism cases" is true.

Results from ANOVA parametric test is shown in Table 13 for the difference of the means of four segmentation schemes used in the proposed similarity model using 10-fold cross-validation data (i.e., sample size = 10). The test failed to reveal a statistically reliable difference among the segmentation schemes because *F*-stastic ≈ 0.7676 is less than *F*-critical = 2.8663 with 9 degrees of freedom.

6.4. Discussion

The purpose of using different baselines is to benchmark the performance of our model. Strictly speaking, PD methods that incorporate semantic understanding of the text have shown superior results with obfuscated, or paraphrased, texts plagiarized from other's contributions without proper acknowledgement. On the contrary, *n*-gram based matching have demonstrated good results in terms of precision and recall with literal plagiarism (Barrón-Cedeño et al., 2009, 2010; Basile et al., 2009; Kasprzak et al., 2009), which, in fact, are not the cases addressed in this study.

Although our proposed model achieved comparably good results compared with former model, namely fuzzy IR method in Yerra and Ng (2005), there are some differences to be discussed here. Through our experimental works, we found that B5-FIR needed considerable pre-processing time to construct

Table 13 Statistical results from ANOVA parametric test of four segmentation schemes used in the fuzzy semantic-based approach.

ANOVA Parametric test for the difference of the means from 10-fold cross-validation data			
t-Critical value =	8.598796653	Decision	Confidence interval
Test statistics for *W3G* vs. *W5G*	1.427055647	Do not reject hypothesis	−0.0916 < μd < 0.2174
Test statistics for *W3G* vs. *W8G3W*	0.167911064	Do not reject	−0.1329 < μd < 0.1761
Test statistics for *W5G* vs. *W8G3W*	0.615949995	Do not reject	−0.1132 < μd < 0.1959
Test statistics for *W3G* vs. *S2S*	0.043491997	Do not reject	−0.1435 < μd < 0.1655
Test statistics for *W5G* vs. *S2S*	1.968806615	Do not reject	−0.0806 < μd < 0.2284
Test statistics for *W8G3W* vs. *S2S*	0.382315759	Do not reject	−0.1219 < μd < 0.1871
Hypothesis =	Means of the four segmentation schemes are equal		
Alternative hypothesis =	At least one mean of one segmentation scheme is significantly different		
Alpha level =	0.05	Degree of freedom = 9	
Total mean =	0.755603604	Final decision: do not reject hypothesis	
F-Statistic =	0.767588513		
Critical *F*-value =	2.866265551		
p-Value =	0.519723312		

The top part of the table shows the t-critical value, t-statistics for segmentation schemes with each other, and the decision taken is either to reject the hypothesis if *t*-statistic > *t*-critical, or do not reject, otherwise. The last column shows the confidence interval between different pairs of segmentation schemes. The bottom part of the table summarizes the ANOVA test statistics between four segmentation schemes, wherein F-statistic is less than the F-critical with 9 degrees of freedom, and hence we do not reject the hypothesis (i.e., all segmentation schemes perform equivalently).

Please cite this article in press as: Alzahrani, S.M. et al., Uncovering highly obfuscated plagiarism cases using fuzzy semantic-based similarity model. Journal of King Saud University – Computer and Information Sciences (2015), http://dx.doi.org/10.1016/j.jksuci.2014.12.001

the word-to-word correlation factor tables. It also required allocation of disk space to save the tables. Not to mention the computational time (and programing difficulties) required to search for words and retrieve their correlation value.

Our proposed model, on the other hand, employed WordNet lexical database and Wu & Palmer similarity metric which have fruitfully eliminated the construction of correlation factor tables as a pre-processing step, and have dramatically reduced the time to compute semantic similarity of words. Another difference is that the precision of fuzzy semantic-based similarity method is noticeably better than the precision in fuzzy IR similarity model, which is directly linked to the reduction of false positives in the results obtained by the proposed model.

7. Conclusion and future work

This paper described a fuzzy semantic-based model for plagiarism detection based on fuzzy rules and semantic information from words in compared texts. Firstly, features were extracted from texts to implement *n*-gram/sentence segments and POS-related semantic spaces. Secondly, two fuzzy rules were evaluated to judge the similarity in compared texts wherein word-to-word semantic similarity was studied based on Wu and Palmer similarity measure. Using a dataset of more than 99,000 handmade and artificial plagiarism cases, the proposed model was evaluated based on four different segmentation schemes and compared with the state-of-the-art baselines. The results were statistically evaluated using 10-fold cross-validation data which led to concluding that the proposed model obtained a reliable and significant performance in comparison with different *n*-gram/sentence matching baselines, and comparable performance with the fuzzy-set sentence similarity model in Yerra and Ng (2005). Yet we believed that our approach might be consistently better since using the lexical taxonomies such as WordNet is efficient than word correlation factors obtained from large corpora. Future work will include experiments on other semantic word-to-word metrics such as *lch* (Leacock and Chodorow, 1998), *res* (Resnik, 1995), *lin* (Lin, 1998), *jcn* Jiang and Conrath, 1997, *lesk* (Banerjee and Pedersen, 2003), and *hso* (Hirst and St Onge, 1998), and integration of more semantic rules such as word-order similarity (Li et al., 2006) and semantic role labeling (Gildea and Jurafskyy, 2000).

References

Alzahrani, S., 2009. Plagiarism auto-detection in arabic scripts using statement-based fingerprints matching and fuzzy-set information retrieval approaches. MSc Thesis, Universiti Teknologi Malaysia, Johor.

Alzahrani, S.M., Salim, N., 2009. On the Use of Fuzzy Information Retrieval for Gauging Similarity of Arabic Documents. In: 2nd International Conference on the Applications of Digital Information and Web Technologies (ICADIWT'09). London Metropolitan University, UK, pp. 539–544.

Alzahrani, S.M., Salim, N., 2010. Fuzzy semantic-based string similarity for extrinsic plagiarism detection: Lab Report for PAN at CLEF'10. In: 4th International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN-10) in conjunction with CLEF'10, Padua, Italy.

Alzahrani, S.M., Salim, N., Abraham, A., 2012. Understanding plagiarism linguistic patterns, textual features and detection methods. IEEE Trans. Syst. Man Cybernet. C Appl. Rev. 42, 133–149.

Alzahrani, S., Palade, V., Salim, N., Abraham, A., 2012. Using structural information and citation evidence to detect significant plagiarism cases. J. Am. Soc. Inf. Sci. Technol. 63, 286–312.

Banerjee, S., Pedersen, T., 2003. Extended gloss overlaps as a measure of semantic relatedness. In: 18 International Joint Conference on Artificial Intelligence (IJCAI-03) August 9–15, Acapulco, Mexico. pp. 805–810.

Barrón-Cedeño, A., Rosso, P., 2009. On automatic plagiarism detection based on *n*-grams comparison. In: Advances in Information Retrieval. pp. 696–700.

Barrón-Cedeño, A., Basile, C., Degli Esposti, M. Rosso, P., 2010. Word length *n*-grams for text re-use detection. In: Computational Linguistics and Intelligent Text Processing. pp. 687–699.

Basile, C., Benedetto, D., Caglioti, E., Cristadoro, G., Esposti, M.D., 2009. A plagiarism detection procedure in three steps: Selection, Matches and "Squares". In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (Eds.), 25th Conference of the Spanish Society for Natural Language Processing, SEPLN'09. Donostia, Spain, pp. 19–23.

Binwahlan, M.S., Salim, N., Suanmali, L., 2010. Fuzzy swarm diversity hybrid model for text summarization. Inf. Process. Manage. (Accepted) 46, 571–588.

Bordogna, G., Pasi, G., 1993. A fuzzy linguistic approach generalizing boolean information retrieval: a model and its evaluation. J. Am. Soc. Inf. Sci. Technol. 44, 70–82.

Bouville, M., 2008. Plagiarism: words and ideas. Sci. Eng. Ethics 14, 311–322.

Budanitsky, A., Hirst, G., 2006. Evaluating WordNet-based measures of lexical semantic relatedness. Computat. Linguist. 32, 13–47.

Ceska, Z., 2007. The future of copy detection techniques. In: 1st Young Researchers Conference on Applied Sciences, YRCAS'07, Pilsen, Czech Republic. pp. 5–10.

Ceska, Z., 2008. Plagiarism detection based on singular value decomposition. In: Lecture Notes in Computer Science. pp. 108–119.

Ceska, Z., 2009. Automatic plagiarism detection based on latent semantic analysis, PhD Thesis. In: Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic.

Clough, P., Stevenson, M., 2011. Developing a corpus of plagiarised short answers. Lang. Resour. Evaluat. 45, 5–24, Special Issue on Plagiarism and Authorship Analysis.

Corley, C., Mihalcea, R., 2005. Measuring the semantic similarity of texts. In: ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment. Association for Computational Linguistics, Ann Arbor, pp. 13–18.

Cross, V., 1994. Fuzzy information retrieval. J. Intell. Inf. Syst. 3, 29–56.

Dolan, B., Quirk, C., Brockett, C., 2004. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In: 20th International Conference on Computational Linguistics. Association for Computational Linguistics, Geneva, Switzerland, p. 350.

Edward, L., Steven, B., 2002. NLTK: the Natural Language Toolkit. In: ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Association for Computational Linguistics, Philadelphia, PA, pp. 63–70.

Elhadi, M., Al-Tobi, A., 2008. Use of text syntactical structures in detection of document duplicates. In: 3rd International Conference on Digital Information Management, ICDIM'08, London, UK. pp. 520–525.

Elhadi, M., Al-Tobi, A., 2009. Duplicate detection in documents and webpages using improved longest common subsequence and documents syntactical structures. In: 4th International Conference on Computer Sciences and Convergence Information Technology, Seoul, Korea. pp. 679–684.

Fernando, S., Stevenson, M., 2008. A semantic similarity approach to paraphrase detection. In: 11th Annual Research Colloquium on

Computational Linguistics UK (CLUK) Oxford University Computing Laboratory, Oxford, UK.

Gildea, D., Jurafskyy, D., 2000. Automatic labeling of semantic roles. In: 38th Annual Conference of the Association for Computational Linguistics (ACL-00), ACL, Hong Kong. pp. 512–520.

Grozea, C., Gehl, C., Popescu, M., 2009. ENCOPLOT: pairwise sequence matching in linear time applied to plagiarism detection. In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (Eds.), 25th Conference of the Spanish Society for Natural Language Processing, SEPLN'09. Donostia, Spain, pp. 10–18.

Hirst, G., St Onge, D., 1998. Lexical chains as representation of context for the detection and correction malapropisms. In: Fellbaum (Ed.), WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press, pp. 305–332.

Jiang, J.J., Conrath, D.W., 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In: International Conference Research on Computational Linguistics (ROCLING X).

Kasprzak, J., Brandejs, M., Křipač, M., 2009. Finding plagiarism by evaluating document similarities. In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (Eds.), 25th Conference of the Spanish Society for Natural Language Processing, SEPLN'09. Donostia, Spain, pp. 24–28.

Koberstein, J., Ng, Y.-K., 2006. Using word clusters to detect similar web documents. In: Knowledge Science, Engineering and Management. pp. 215–228.

Koroutchev, K., Cebrian, M., 2006. Detecting translations of the same text and data with common source. J. Statist. Mech. Theory Experiment 2006, P10009.

Leacock, C., Chodorow, M., 1998. Combining local context with WordNet similarity for word sense identification. In: Fellbaum, C. (Ed.), WordNet: A Lexical Reference System and its Application. MIT Press, Cambridge, MA, pp. 265–283.

Lee, M.C., 2011. A novel sentence similarity measure for semantic-based expert systems. Expert Syst. Appl. 38, 6392–6399.

Leech, N.L., Barrett, K.C., Morgan, G.A., 2008. SPSS for Intermediate Statistics Use and Interpretation, 3rd ed. Lawrence Erlbaum Associates, New York.

Li, Y., Bandar, Z.A., McLean, D., 2003. An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans. Knowledge Data Eng. 15, 871–882.

Li, Y., McLean, D., Bandar, Z.A., O'Shea, J.D., Crockett, K., 2006. Sentence similarity based on semantic nets and corpus statistics. IEEE Trans. Knowledge Data Eng. 18, 1138–1150.

Lin, D., 1998. An information-theoretic definition of similarity. In: Shavlik, J.W. (Ed.), Fifteenth International Conference on Machine Learning (ICML '98). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 296–304.

Lin, D., 1998. An information-theoretic definition of similarity. In: 15th International Conference on Machine Learning, ICML '98, Madison, Wisconsin, USA. pp. 296–304.

Luo, Q., Chen, E., Xiong, H., 2011. A semantic term weighting scheme for text categorization. Expert Syst. Appl. 38, 12708–12716.

Manning, C.D., Raghavan, P., Schütze, H., 2009. Scoring, term weighting and the vector space model. In: Introduction to Information Retrieval. Cambridge University Press, pp. 109–133.

Marcus, M., Santorini, B., Marcinkiewicz, M.A., 1993. Building a large annotated corpus of English: the Penn Treebank. Computat. Linguist. 19.

Mihalcea, R., Corley, C., Strapparava, C., 2006. Corpus-based and knowledge-based approaches to text semantic similarity. In: American Association for Artificial Intelligence (AAAI 2006), Boston. pp. 775–780.

Miller, G.A., 1995. WordNet: a lexical database for English. Commun. ACM 38, 39–41.

Muftah, A.J.A., 2009. Document plagiarism detection algorithm using semantic networks. In: Faculty of Computer Science and Information Systems, MSc Thesis, Universiti Teknologi Malaysia, Johor.

Ogawa, Y., Morita, T., Kobayashi, K., 1991. A fuzzy document retrieval system using the keyword connection matrix and a learning method. Fuzzy Sets Syst 39, 163–179.

Pera, M.S., Ng, Y.-K., 2011. SimPaD: a word-similarity sentence-based plagiarism detection tool on Web documents. Web Intell. Agent Syst., IOS Press 9, 27–41.

Potthast, M., Stein, B., Eiselt, A. Barrón-Cedeño, A., Rosso, P., 2009. Overview of the 1st international competition on plagiarism detection. In: 25th Conference of the Spanish Society for Natural Language Processing, SEPLN'09, Donostia, Spain, pp. 1–9.

Potthast, M., Eiselt, A., Stein, B., BarrónCedeño, A., Rosso, P., 2009. PAN Plagiarism Corpus (PAN-PC-09). In: Webis at Bauhaus-Universität Weimar and NLEL at Universidad Politécnica de Valencia.

Potthast, M., Stein, B. Barron-Cedeno, A., Rosso, P., 2010. An evaluation framework for plagiarism detection. In: 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China.

Potthast, M., Eiselt, A., Stein, B., BarrónCedeño, A., Rosso, P., 2010. PAN Plagiarism Corpus (PAN-PC-10). In: Webis at Bauhaus-Universität Weimar and NLEL at Universidad Politécnica de Valencia.

Potthast, M., Stein, B., Barrón-Cedeño, A., Rosso, P., 2011. PAN Plagiarism Corpus (PAN-PC-11). In: Webis at Bauhaus-Universität Weimar and NLEL at Universidad Politécnica de Valencia.

Resnik, P., 1995. Using information content to evaluate semantic similarity in a taxonomy. In: Mellish, Chris S. (Ed.), . In: 14th International Joint Conference on Artificial Intelligence – Volume 1 (IJCAI'95), vol. 1. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 448–453.

Roig, M., 2006. Avoiding plagiarism, self-plagiarism, and other questionable writing practices: a guide to ethical writing, in, St. Johns University.

Scherbinin, V., Butakov, S., 2009. Using Microsoft SQL server platform for plagiarism detection. In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (Eds.), 25th Conference of the Spanish Society for Natural Language Processing, SEPLN'09. Donostia, Spain, pp. 36–37.

Shehata, S., Karray, F., Kamel, M., 2010. An efficient concept-based mining model for enhancing text clustering. IEEE Trans. Knowledge Data Eng. 22, 1360–1371.

Sugeno, M., 1985. Industrial Applications of Fuzzy Control. Elsevier Science Inc., New York, NY, USA.

Turney, P.D., 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In: 12th European Conference on Machine Learning. Springer-Verlag, London, UK.

Wu, Z., Palmer, M., 1994. Verb semantics and lexical selection. In: 32nd Annual Meeting of the Association for Computational Linguistics, New Mexico State University, New Mexico. pp. 133–139.

Yerra, R., Ng, Y.-K., 2005. A sentence-based copy detection approach for web documents. In: Fuzzy Systems and Knowledge Discovery. pp. 557–570.

Zechner, M., Muhr, M., Kern, R., Granitzer, M., 2009. External and intrinsic plagiarism detection using vector space models. In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (Eds.), 25th Conference of the Spanish Society for Natural Language Processing, SEPLN'09. Donostia, Spain, pp. 47–55.

Please cite this article in press as: Alzahrani, S.M. et al., Uncovering highly obfuscated plagiarism cases using fuzzy semantic-based similarity model. Journal of King Saud University – Computer and Information Sciences (2015), http://dx.doi.org/10.1016/j.jksuci.2014.12.001