Semantically-informed distance and similarity measures for paraphrase plagiarism identification

Miguel A. Álvarez-Carmona, Marc Franco-Salvador, Manuel Montes-y-Gómez, Paolo Rosso, Luis Villaseñor-Pineda, Miguel A. Álvarez-Carmona[a,1,*], Marc Franco-Salvador[b], Manuel Montes-y-Gómez[a], Paolo Rosso[c], Luis Villaseñor-Pineda[a], Esaú Villatoro-Tello[d,**]

[a] *Language Technologies Lab., Computational Sciences Department, Instituto Nacional de Astrofísica Óptica y Electrónica (INAOE), Mexico.*
[b] *Symanto Research, Nuremberg, Germany*
[c] *PRHLT Research Center, Universitat Politècnica de València, Spain.*
[d] *Language and Reasoning Research Group, Information Technologies Dept., Universidad Autónoma Metropolitana Unidad Cuajimalpa (UAM), Mexico.*

Abstract

Paraphrase plagiarism identification represents a very complex task given that plagiarized texts are intentionally modified through several rewording techniques. Accordingly, this paper introduces two new measures for evaluating the relatedness of two given texts: a semantically-informed similarity measure and a semantically-informed edit distance. Both measures are able to extract semantic information from either an external resource or a distributed representation of words, resulting in informative features for training a supervised classifier for detecting paraphrase plagiarism. Obtained results indicate that the proposed metrics are consistently good in detecting different types of paraphrase plagiarism. In addition, results are very competitive against state-of-the art methods having the advantage of representing a much more simple but equally effective solution.

Keywords: Plagiarism identification, Paraphrase Plagiarism, Semantic similarity, Edit distance, Word2vec representation

1. Introduction

Text plagiarism means including other person's text as your own without proper citation [17]. Nowadays, because of the Web and text editing tools, it is

*Principal corresponding author
**Corresponding author
Email addresses: miguelangel.alvarezcarmona@ccc.inaoep.mx (Miguel A. Álvarez-Carmona), marc.franco@symanto.net (Marc Franco-Salvador), mmontesg@ccc.inaoep.mx (Manuel Montes-y-Gómez), prosso@dsic.upv.es (Paolo Rosso), villasen@ccc.inaoep.mx (Luis Villaseñor-Pineda), evillatoro@correo.cua.uam.mx (Esaú Villatoro-Tello)
[1] Address: Luis Enrique Erro No.1, Tonantzintla, Puebla 72840, Mexico

very easy to find and re-use any kind of information [1], causing the plagiarism practice to dramatically increase.

Traditional methods for plagiarism detection consider measuring the word overlap between two texts [13]. Using measures such as the Jaccard and cosine coefficients [9] resulted in a simple but effective approach for determining the similarity between the suspicious and the source texts [10, 21].

Likewise, measuring the similarity of texts by means of an edit-distance [12, 18, 5] or the Longest Common Subsequence (LCS) [9] resulted in effective approaches. In general, these approaches are very accurate on detecting verbatim cases of plagiarism (i.e., copy-paste), but they are useless to detect complex cases of plagiarism, such as *paraphrase plagiarism*, where texts show significant differences in wording and phrasing.

Detecting paraphrase plagiarism represents a challenging task for current methods since they are not able to measure the *semantic* overlap. Accordingly, some research works have tried to overcome this limitation by proposing the use of knowledge resources such as WordNet [15] for evaluating the semantic proximity of texts [3, 7, 16]. Although these methods have been widely applied for measuring the degree of paraphrases between two given texts, just [16] evaluates its relevance for plagiarism detection. More recently, [4, 11] discussed the use of semantic information without depending on any external knowledge resource. Particularly, they proposed using distributive representations, such as word2vec [14], in the task of plagiarism detection. The main drawback of these approaches is that they often need large training sets in order to learn accurate models.

This paper focuses on the detection of paraphrase plagiarism. It proposes two new measures for evaluating the relatedness of two given texts: a semantically informed similarity measure and a semantically informed edit distance. Both measures can extract the semantic information from WordNet and word2vec. On the top of these measures we trained a classifier for detecting paraphrase plagiarism. In short, the goal of this paper is threefold: *i*) to evaluate the effectiveness of the proposed measures, when using WordNet and word2vec, in the paraphrase plagiarism identification task; *ii*) to investigate the complementarity of both kind of measures for solving the posed task; and *iii*) to determine the effectiveness of the semantically informed measures on detecting specific types of (plagiarism) paraphrases.

The remainder of this paper is organized as follows. Section 2 describes the proposed semantically informed measures; Section 3 describes the used datasets and the experimental setup; Section 4 presents and discusses the obtained results. Finally, Section 5 depicts our conclusions and some future work directions.

2. Proposed semantically-informed measures

This section describes the two proposed measures for paraphrase plagiarism identification. Section 2.1 presents a modification of the Jaccard coefficient considering semantic information, whereas Section 2.2 describes our semantically informed version of the Levenshtein edit distance.

In order to illustrate the limitations of traditional measures and to motivated our proposed modifications, please consider the two sentences from Figure 1. Applying the traditional Jaccard measure it will result in a low similarity,

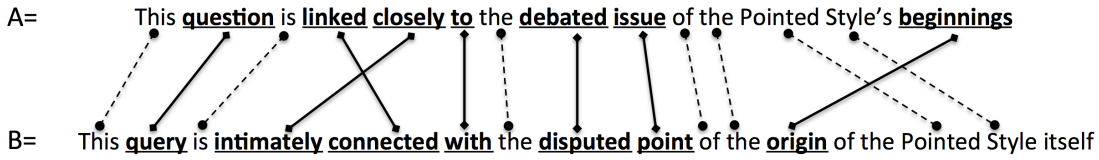

Figure 1: Example of two (A and B) semantically related sentences. Dotted lines connect exact matching words whilst solid lines connect semantically related words.

$\mathrm{J}(A, B) = 0.31$, since only 7 terms out of a total of 22 match exactly. Similarly, the classic Levenshtein edit distance will indicate that the sentences are very distant, $\mathrm{ED}(A, B) = 0.70$. Nevertheless, it is evident that these two texts are more similar than these results indicate; they contain several additional pair of terms (solid line connections) that are semantically related but not considered. Therefore, our proposal is to semantically enrich these measures by means of including the similarity degree of non-overlapped words.

2.1. Semantically-informed similarity measure

Let's assume A and B are two documents with vocabularies V_A and V_B, and that V'_A and V'_B indicate their non-overlapping words respectively. Their semantic similarity, based on the Jaccard coefficient, is computed as shown in Formula 1.

$$\mathrm{SJ}(A, B) = \frac{|V_A \cap V_B| + \mathrm{softmatch}(V'_A, V'_B)}{|V_A \cup V_B| - \mathrm{softmatch}(V'_A, V'_B)} \tag{1}$$

The function $\mathrm{softmatch}(X, Y)$ accounts for the maximum similarity values between words contained in the sets X and Y. For its computation we first measure the similarity $\mathrm{sim}(x, y)$ among all words $x \in X$ and $y \in Y$; this similarity could be measured using WordNet or word2vec. Then, we eliminate irrelevant relations, that is, we set $\mathrm{sim}(x, y) = 0$ if it is not the greatest similarity score for both x and y with any other term. Finally, we accumulate the resulting similarities as indicate by Formula 2.

$$\mathrm{softmatch}(X, Y) = \sum_{\forall x \in X, \forall y \in Y} \mathrm{sim}(x, y) \tag{2}$$

Continuing with the example from Figure 1, $V'_A = \{$question, linked, closely, to, debated, issue, beginnings$\}$ and $V'_B = \{$query, intimately, connected, with, disputed, point, origin$\}$. Using WordNet as semantic resource for computing word similarities as described in Section 3.2, $\mathrm{softmatch}(A', B') = 6.75$, resulting in $\mathrm{SJ}(A, B) = 0.90$, which in turn reflects a more realistic similarity than the initial estimated value.

2.2. Semantically-informed edit distance

This new measure is based on the Levenshtein edit distance. It also computes the minimum number of operations permitted (generally substitution [S], deletion [D] and insertion [I]) for transforming text A to text B. However, different to the traditional version where each operation has unitary cost, our proposal accounts for the semantic similarity between substituted words and determines the impact of inserted/deleted words in the text. The proposed

semantically-informed edit distance between two texts A and B, of lengths $|A|$ and $|B|$ respectively, is given by $\mathrm{SED}_{A,B}(|A|,|B|)$ where:

$$\mathrm{SED}_{A,B}(i,j) = \min \begin{cases} \mathrm{SED}(i-1,j) + \mathrm{dist}(\tau, A_i) & \mathrm{D} \\ \mathrm{SED}(i,j-1) + \mathrm{dist}(\tau, B_j) & \mathrm{I} \\ \mathrm{SED}(i-1,j-1) + \mathrm{dist}(A_i, B_j) & \mathrm{S} \end{cases} \quad (3)$$

In this approach the substitution of a word x by a word y has a cost proportional to their semantic distance $\mathrm{dist}(x,y)$. This distance could be measured using WordNet or word2vec as described in Section 3.2. Similarly, the insertion or deletion of a word x has a variable cost, which is defined in function of its semantic distance to a predefined general word τ. The idea is that the greater $\mathrm{dist}(\tau,x)$, the more rare is the word x, and the more important its contribution of the meaning of the text.

Following with the example above, the new edit distance between texts A and B is small, $\mathrm{SED}(A,B) = 0.20$, because all words in bold face are substituted by semantically related words, for instance, "question" by "query" and "beginnings" by "origin". In addition, all removed words, such as "of", "the" and "itself" are very general and, therefore, their deletion do not have a considerable impact.

3. Experimental Setup

The proposed distance and similarity measures are especially suited to the task of paraphrase plagiarism identification. Accordingly, this section presents the datasets used for their evaluation as well as a description of their configuration for the task.

3.1. Datasets.

We used the P4PIN corpus[2] [19], a corpus specially built for evaluating the identification of paraphrase plagiarism. This corpus is an extension of the P4P corpus [2], which contains pairs of text fragments where one fragment represents the original source text and the other represents a paraphrased version of the original. In addition, the P4PIN corpus includes not paraphrase plagiarism cases, i.e., negative examples formed by pairs of unrelated texts samples with likely thematic or stylistic similarity. Table 1 shows two examples from this corpus, one case of paraphrase plagiarism and one of not-paraphrase plagiarism.

An important characteristic of this corpus is that each plagiarism case is labeled with a particular subtype of paraphrase. Authors of the P4P corpus [2] employed a paraphrases typology, which includes four general classes, two of them with four sub-classes, for a total of nineteen types of paraphrases. For our purposes, we took two classes from the most general categorization level, and the four subclasses from the second categorization level as described below:

- *Morphology-based changes* include inflectional changes (*e.g.,* affixes modification), modal verb modification (*e.g., might* → *could*) and derivation changes.

[2]Available at: `http://ccc.inaoep.mx/~mmontesg/resources/corpusP4PIN.zip`

Table 1: Examples of paraphrase-plagiarism and not-paraphrase-plagiarism in the P4PIN corpus. Underlined words represent common words between the original and the suspicious document; below each column appears the percentage of common words between text fragments.

	Paraphrase plagiarism example	**Not-paraphrase plagiarism example**
Original	I pored through these pages, and as I perused the lyrics of The Unknown Eros that I had never read before, I appeared to have found out something wonderful: there before me was an entire shining and calming extract of verses that were like a new universe to me.	The fact that an omnipresent God exists is the one universal factor that governs the laws of nature. God has set in place the laws of the universe for His own purposes.
Suspicious	I dipped into these pages, and as I read for the first time some of the odes of The Unknown Eros, I seemed to have made a great discovery: here was a whole glittering and peaceful tract of poetry which was like a new world to me.	The laws of nature are the art of God. Without the presence of such an agent, one who is conscious of all upon which the laws of nature depend, producing all that the laws prescribe. The laws themselves could have no existence.
Common words	57.4%	54.8%

- *Lexicon-based changes* comprise modifications such as synthetic and analytic reconstruction, spelling and format change, polarity substitutions and converse substitutions; in general these types of changes alter only one lexical unit within a sentence preserving the original meaning.
- *Syntax-based modifications* cause structural alterations in a sentence, allowing to have the same meaning but redirecting the main focus to different elements within the sentence; paraphrase types included in this category are: diathesis alterations, negation switching, ellipsis, coordination changes and subordination with nesting changes.
- *Discourse-based modifications* alter the sentences' form and order; they include changes in punctuation marks, modifications in the syntactic structure, modality changes as well as some direct or indirect style alternations.
- *Semantic-based changes* consider modifications involving substitution of some elements within a sentence that results in lexical and syntactical modifications without interfering with the original meaning of the sentence. Semantic-based changes represent the highest level of modifications.
- *Miscellaneous-based changes* recollect all types of modifications that do not correspond to specific linguistic paraphrase phenomena, such as addition, deletion or changing the order of lexical units.

In summary, the P4PIN corpus has 2236 instances, where 75% are not-plagiarism cases and 25% are plagiarism cases.

In order to get more insight on the relevance and robustness of the proposed measures we also evaluated them in the paraphrase identification task.[3] For this

[3]Although similar, paraphrase plagiarism identification differs from paraphrase identification in that the former is done with the intention of hiding the text-reuse (i.e., the plagiarism act)

purpose we used the well-known MSRP corpus [8], which contains pairs of sentences labeled as "mean the same thing" (paraphrase) or not (not-paraphrase) [8]. This corpus is divided in two partitions, a training set having 4,076 sentences pairs and a test set containing 1,725 examples; in both partitions, 67% of the instances are plagiarism examples and the remaining 33% are not-plagiarism cases. Contrary to the P4PIN, the MSRP corpus is not labeled by paraphrase sub-types.

3.2. Semantic word similarity

Both proposed measures rely on the calculus of the semantic similarity or distance between pairs of words ($\text{sim}(x, y)$ or $\text{dist}(x, y)$). For the sake of simplicity we defined $\text{dist}(x, y) = 1 - \text{sim}(x, y)$.

We used two different approaches for computing the word similarity. On the one hand, we used WordNet as knowledge source and applied the WUP similarity measure [20]. This measure calculates the semantic relatedness of two given words x and y by considering the depths of their synsets in the WordNet taxonomy (s_x and s_y), along with the depth of their most specific common synset (mcs) as described by Formula 4.

$$\text{sim}(x, y) = \frac{2 * \text{depth}(\text{msc})}{\text{depth}(s_x) + \text{depth}(s_x)} \tag{4}$$

On the other hand, we used the word2vec representation, and measured the similarity of words by means of the cosine function. In particular, we used the continuous Skip-gram model [14] of the word2vec toolkit[4] to generate the distributed representations of the words from the complete English Wikipedia. We considered 200-dimensional vectors, a context window of size 10, and 20 negative words for each sample.

3.3. Classification process

Once computed the similarity (or edit distance) between the suspicious and source texts, the next step is to determine whether or not the pair of texts are a case of plagiarism. When using the semantically-informed similarity measure, if the similarity score is greater than some threshold β_s, then the instance is classified as "plagiarism" otherwise the result is "not-plagiarism". On the other hand, when using the semantic-informed edit distance, if the distance score is greater that some threshold β_d, then the instance is labeled as "not-plagiarism" otherwise the result is "plagiarism".

For the experiments done with the P4PIN corpus we carried out a ten-fold cross-validation strategy. We considered as classification threshold (β_s or β_d) the one that maximizes the classification performance at training. For the MSRP corpus we used the given training and test partitions. The classification threshold is defined from the training partition. In all the experiments we used the macro F_1-measure as main evaluation measure.

[4] `https://code.google.com/archive/p/word2vec/`

Table 2: F_1 results in the identification of paraphrase and paraphrase plagiarism, using the traditional and the proposed similarity and distance measures. Suffix W2V means word2vec and WN indicates WordNet.

Corpus	J	SJ-W2V	SJ-WN	ED	SED-W2V	SED-WN
P4PIN	0.90	**0.91**	0.80	0.87	0.90	0.82
MSRP	0.80	**0.81**	0.73	0.75	**0.81**	0.76

4. Experimental Results

This section presents the results of several experiments aimed to assess the effectiveness of the proposed measures in the task of paraphrase plagiarism identification, as well as to analyze their complementarity and their appropriateness for identifying plagiarism cases using different categories of paraphrases.

4.1. Relevance of considering semantic information

To assess the relevance of considering semantic information in the calculation of the similarity/distance between two texts, we carried out the following set of experiments: *i*) using the original Jaccard coefficient (**J**), ; *ii*) using the original edit distance (**ED**); *iii*) using the proposed semantically-informed measures with WordNet (**SJ-WN** and **SED-WN**) and with word2vec (**SJ-W2V** and **SED-W2V**).

Results from Table 2 show that the proposed semantically informed approaches, based on both the Jaccard and the Levenshtein edit distance measures, obtained better or equal F_1 results than the approaches using the original measures. This particularly happens when word2vec is used as word similarity function (SJ-W2V and SED-W2V). We attribute these results to the coverage of the semantic resources. Table 3 shows a comparative analysis of the vocabulary coverage for both WordNet and word2vec resources within each evaluated corpus. These results indicate that WordNet has lower coverage value than word2vec. Thus, results from Table 3 highlight the limitations of using an external resource such as WordNet.

Table 3: Comparative analysis of the vocabulary coverage.

Corpus	WordNet	word2vec
P4PIN	79.52%	91%
MSRP	79.1%	98%

4.2. Complementary of the proposed measures

The proposed measures are similar in that both consider semantic information and, therefore, both can identify related texts even when they do not contain exactly matching words. However, they differ from each other in the way they compute the relatedness of texts. On the one hand, the similarity measure focuses on the *content overlap*, whereas, on the other hand, the distance measure emphasizes the *word order*. Accordingly, this section presents an experiment aimed to analyze the complementarity of the two measures.

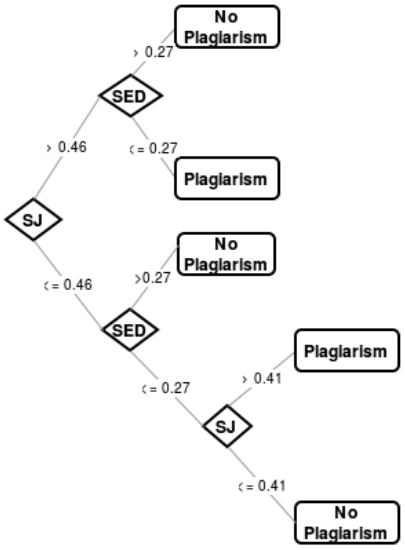

Figure 2: Decision tree of the combined approach on the P4PIN corpus.

Table 4: F_1 results from the combination of the semantically-informed similarity and distance measures. The SOA column indicates the state-of-the-art performance reported for each dataset.

Corpus	**SJ-W2V**	**SED-W2V**	***Combined***	**SOA**
P4PIN	0.90	0.91	**0.93**	0.92 [19]
MSRP	0.81	0.81	**0.83**	0.85 [6]

The experiment reported in this section combines the best results from the previous section (i.e., SJ-W2V and SED-W2V). For the combination we used a supervised classification approach, where the scores obtained from both measures were used as features. We considered several learning algorithms, such as SVM, Naïve Bayes and J48, but we only report the results obtained by J48 because they outperformed the others as well as allow us to understand the classification criteria (refer to Figure 2).

Table 4 shows the results from this experiment. It can be noticed that the results obtained by the combined approach clearly outperform the results from the approaches using the proposed measures individually. Hence, our preliminary conclusion is that these two measures are in fact complementary to each other. Additionally, this table shows the state-of-the-art results for the two used datasets. As noticed, the results from our combined approach are close to the reference results, nonetheless, ours is a much more simple approach (for example, [6] reports a recursive neural network using syntax-aware and multi-sense word embeddings).

Table 5: F_1 results in several paraphrase categories using different similarity and distance measures. The SOA column shows state-of-the-art results reported in [19]. In [19] character n-grams are used for representing the documents and measuring their similarity.

Paraphrases categories	*Jaccard*		*Levenshtein*		***Combined***	**SOA**
	J	**SJ-W2V**	**ED**	**SED-W2V**		[19]
Morphological	0.85	0.88	0.85	0.86	**0.92**	0.90
Lexical	0.90	0.91	0.88	0.89	**0.93**	0.92
Syntactical	0.88	0.89	0.85	0.87	**0.93**	0.91
Discourse	0.86	0.87	0.86	0.89	**0.92**	0.89
Semantic	0.77	0.78	0.73	0.80	**0.83**	0.77
Miscellaneous	0.89	0.89	0.85	0.87	**0.92**	0.90

4.3. Robustness on different paraphrase categories

The plagiarism examples from the P4PIN corpus are categorized according to their paraphrases types, namely: *morphology, lexicon, syntax, discourse, semantic* and *miscellaneous* changes [2] (refer to Section 3.1). The experiments reported in this section aim at measuring the robustness of the proposed semantically-informed measures against different paraphrase practices. Table 5 shows the obtained results.

These results indicate that the proposed measures (using word2vec as semantic resource) consistently improve the performance results of the traditional variants. They also indicate that paraphrases from the *semantic* category are the harder to identify. This performance was expected, since semantic changes involve lexical and syntactical modifications. Additionally, these results outperform the state-of-the-art in all categories, evidencing that the supervised combined approach is the best option for identifying plagiarism regardless of the type of paraphrase.

4.4. On the complexity of corpora

In order to provide a deeper analysis on the obtained results, we decided to investigate the level of complexity of the employed corpora. Through this analysis we aim to figure out under which circumstances our proposed semantically informed metrics perform the better.

For determining the level of complexity of a given corpus C we propose the following straightforward measure (refer to Formula 5), which assesses the lexical concordance (LC) across both plagiarism and not-plagiarism examples.

$$\mathrm{LC}(C) = \frac{|C_{\mathrm{neg}}| - \mathrm{O}(C_{\mathrm{neg}}) + \mathrm{O}(C_{\mathrm{pos}})}{|C|} \tag{5}$$

where C_{neg} and C_{pos} represent the negative and positive partitions of corpus C respectively. Accordingly, $\mathrm{O}(C_x)$ represents the accumulated similarity between all pairs of documents contained in the x partition of the corpus C and it is obtained using the Formula 6, where $\mathrm{J}(A, B)$ represents the Jaccard coefficient between the pair of documents A and B.

$$\mathrm{O}(C_x) = \sum_{\forall (A,B) \in C_x} \mathrm{J}(A, B) \tag{6}$$

The closer the value of lexical concordance to zero means the corpus is more complex, whilst the closer to one indicated an easier corpus. For example, in a low complexity corpus ($\text{LC}(C) \to 1$) the positive instances are merely verbatim cases and the negative examples are completely unrelated text chunks.

Table 6 shows the LC values for the MSRP and P4PIN collections. It can be noticed that MSPR is more complex than P4PIN (see first two rows from Table 6). Additionally, in the P4PIN corpus we observe that the more complex paraphrase category is the semantic category, whereas the easier is the lexical one.

As a final experiment we analyze the influence of the complexity of the collections over the performance of the proposed semantic enriched measures. In particular we analyzed the correlation between the LC value of each category of the P4PIN corpus and the F_1 improvement of the proposed approach over the baselines. For this analysis we applied the Spearman Correlation Coefficient.

Table 6: Lexical concordance values of the employed corpora

Corpus	**LC value**
P4PIN	0.76
MSRP	0.56
Paraphrase types	**LC value**
Lexical	0.41
Discourse	0.41
Miscellaneous	0.39
Syntactical	0.39
Morphological	0.38
Semantic	0.29

Table 7 shows the obtained correlation results, indicating some very interesting insights from the proposed measures. On the one hand, there is a strong correlation between the complexity of the corpus and the performance of our combined method. Given the correlation is negative, it indicates that the more complex is the corpus (the smallest the LC value), the greater is the advantage of our method over SOA results; in other words, our proposed method performs consistently better when the corpus has a high complexity level. A similar situation occurs when employing our semantically informed edit distance (SED) approach; it especially outperforms the ED results for the complex paraphrase categories. On the other hand, the correlation results indicate that the improvement of SJ-W2V over J is not related to the corpus complexity.

Table 7: Correlation analysis

Compared methods	r
SJ-W2V *vs.* J	-0.0377
SED-W2V *vs.* ED	-0.8771
Combined vs. SOA	-0.8985

5. Conclusions and future work

We have introduced an approach for paraphrase plagiarism detection which proposes the inclusion of semantic information to traditional similarity and edit distance measures. The aim of the proposed semantically-informed measures is to allow assessing the relatedness between suspicious and source texts even when they do not contain exactly matching words.

We hypothesized that using the proposed semantically-informed measures, a method for paraphrase plagiarism identification would be more accurate in solving the task. Performed experiments indicate that our proposed method obtained state-of-the-art results, especially when distributed word representations are considered as a semantic resource. Additionally, experiments demonstrated that the information provided by the two semantically-informed measures is complementary to each other, resulting in useful features for a supervised classifier to learn whether or not the pair of texts are a case of plagiarism. Further, we investigated the degree of robustness of the proposed measures against different subtypes of paraphrase plagiarism. Obtained results showed that the proposed approaches, either individually or combined, are able to improve the performance of traditional techniques for the distinct paraphrase plagiarism categories, particularly for those with higher complexities. Finally, it is important to highlight that obtained results are competitive to those reported in recent research works, but, in contrast, the proposed approach represents a much more simple method.

As future work we plan to study the sensitivity of our method to the coverage of the semantic resource, in particular we plan to evaluate our method using a word2vec representation trained over a larger corpus.

Acknowledgements: This work was partially supported by CONACYT under scholarship 401887, project grants 257383, 258588 and 2016-01-2410 and under the Thematic Networks program (Language Technologies Thematic Network project 281795). The work of the fourth author was partially supported by the SomEMBED TIN2015-71147-C2-1-P MINECO research project and by the Generalitat Valenciana under the grant ALMAMATER (Prometeo II/2014/030).

[1] A. Abdi, N. Idris, R. M. Alguliyev, and R. M. Aliguliyev. Pdlk: Plagiarism detection using linguistic knowledge. *Expert Systems with Applications*, 42(22):8936–8946, 2015.

[2] A. Barrón-Cedeño, M. Vila, M. A. Martí, and P. Rosso. Plagiarism meets paraphrasing: insights for the next generation in automatic plagiarism detection. *Computational Linguistics*, 39(4):917–947, 2013.

[3] S. Biggins, S. Mohammed, and S. Oakley. University of shefield: Two approaches to semantic text similarity. In *First Joint Conference on Lexical and Computational Semantics (SEM at NAACL 2012)*, pages 655–661, Montreal, Canada., 2012.

[4] A. Brlek, P. Franjic, and N. Uzelac. Plagiarism detection using word2vec model. *Text Analysis and Retrieval 2016 Course Project Reports*, page 4, 2016.

[5] K. Chatterjee, T. A. Henzinger, R. Ibsen-Jensen, and J. Otop. Edit distance for pushdown automata. *arXiv preprint arXiv:1504.08259*, 2015.

[6] J. Cheng and D. Kartsaklis. Syntax-aware multi-sense word embeddings for deep compositional models of meaning. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1531–1542. ACL, 2015.

[7] C. Courtney and R. Mihalcea. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment (EMSEE at NAALC 2005)*, pages 13–18, 2005.

[8] W. B. Dolan and C. Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proc. of IWP*, 2005.

[9] W. H. Gomaa and A. A. Fahmy. A survey of text similarity approaches. *International Journal of Computer Applications*, 68(13):13–18, 2013.

[10] T. C. Hoad and J. Zobel. Methods for identifying versioned and plagiarized documents. *Journal of the American society for information science and technology*, 54(3):203–215, 2003.

[11] S. Kim, W. J. Wilbur, and Z. Lu. Bridging the gap: a semantic similarity measure between queries and documents. *arXiv preprint arXiv:1608.01972*, 2016.

[12] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

[13] R. Lukashenko, V. Graudina, and J. Grundspenkis. Computer-based plagiarism detection methods and tools: an overview. In *Proceedings of the 2007 international conference on Computer systems and technologies*, page 40. ACM, 2007.

[14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[15] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[16] Y. Palkovskii, A. Belov, and I. Muzyka. Using wordnet-based semantic similarity measurement in external plagiarism detection. In *Notebook for PAN at CLEF'11*, 2011.

[17] A. Pandey, M. Kaur, and P. Goyal. The menace of plagiarism: How to detect and curb it. In *Emerging Trends and Technologies in Libraries and Information Services (ETTLIS), 2015 4th International Symposium on*, pages 285–289. IEEE, 2015.

[18] E. Stamatatos. Plagiarism detection using stopword n-grams. *Journal of the American Society for Information Science and Technology*, 62(12):2512–2527, 2011.

[19] J. F. Sánchez-Vega. *Identificación de plagio parafraseado incorporando estructura, sentido y estilo de los textos.* PhD thesis, Instituto Nacional de Astrofísica, Optica y Electrónica, 2016.

[20] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

[21] M. Zechner, M. Muhr, R. Kern, and M. Granitzer. External and intrinsic plagiarism detection using vector space models. In *CEUR Workshop Proceedings*, volume 502, pages 47–55, 2009.