Identifying free text plagiarism based on semantic similarity

George Tsatsaronis
Norwegian University of Science and Technology
Department of Computer and Information Science
Trondheim, Norway
gbt@idi.ntnu.no

Iraklis Varlamis
Harokopio University of Athens,
Department of Informatics and Telematics
Athens, Greece
varlamis@hua.gr

Andreas Giannakoulopoulos
Ionian University,
Department of Audio and Visual Arts
Corfu, Greece
agiannak@ionio.gr

Nikolaos Kanellopoulos
Ionian University,
Department of Audio and Visual Arts
Corfu, Greece
kane@ionio.gr

Keywords: Omiotis, anti-plagiarism software, paraphrasing, semantic measures, plagiarism detection

Abstract

It is common knowledge that plagiarism in academia goes as back in time as research itself. However, in the last two decades this phenomenon of academic deception has turned into an academic plague. Undoubtedly, the rapid expansion of the Web and the vast amount of publicly available information and documents facilitate the unethical malpractice of computer-aided plagiarism, which in turn has inflated the problem. Anti-plagiarism techniques build upon technological solutions and especially the development of task-specific software. The role of anti-plagiarism software for text is to process a document and identify the pieces of text that have been reproduced from another source. This work presents a semantic-based approach to text-plagiarism detection which improves the efficiency of traditional keyword matching techniques. Our semantic matching technique is able to detect a larger variety of paraphrases, including the use of synonym terms, the repositioning of words in the sentences etc. We evaluate our methodology in a dataset comprising positive and negative plagiarism samples and present comparative results of both supervised and unsupervised methods.

1. Introduction

Among the many existing definitions of plagiarism, the one given by the *American Association of University Professors* seems to be the most comprehensive: "*taking over the ideas, methods, or written words of another, without acknowledgment and with the intention that they be taken as the work of the deceiver*" (as cited in Roig, 2006).

Plagiarism raises moral questions for both students and teachers. There is a vital need for tackling the problem immediately for various reasons, among which the most important are: (i) reputation and credibility of academic institutions; (ii) promotion of

the ideal of academic integrity; (iii) restoration of the feeling of fairness among students who do not commit plagiarism; (iv) emphasis on the benefits of independent learning and its role in building solid academic skills (Park, 2004).

Although there are many types of plagiarism, the ones that mostly concern academic writing, *viz.* textual plagiarism, are (i) plagiarism of ideas; (ii) improper paraphrasing; (iii) summarizing and paraphrasing (Roig, 2006). In this context, the adaptation of Omiotis (Tsatsaronis, et al., 2010), a semantic relatedness measure for text, can contribute to tackling the problem by identifying free text plagiarism based on conceptual similarity.

The idea behind our approach is that although textual similarity is a fast method for detecting text-plagiarism and has an acceptable performance in cases where the original text is copied as it is, it can easily be deceived when simple paraphrasing is employed. For this reason, the use of semantic relatedness will further improve results by solving ambiguous and tricky plagiarism cases.

In this task, we utilize two linguistic knowledge bases, the WordNet thesaurus and the Wikipedia electronic encyclopedia in order to find the meaning, which is hidden in keywords and measure the conceptual similarity between texts. WordNet's lexical database contains English nouns, verbs, adjectives and adverbs, organized in synonym sets (synsets). Synsets are connected with various links that represent semantic relations among them (i.e. hypernymy/hyponymy, meronymy/holonymy, synonymy/antonymy, entailment/causality, troponymy, domain/domain terms, derivationally related forms, coordinate terms, attributes, stem adjectives, etc.). Wikipedia is used in order to increase the coverage of our semantic relatedness measure to cases where proper names are abstracted (e.g. when the name of a city is replaced by the name of the respective country or the name of a person is replaced by its position or title etc.).

In the following section we briefly present the theoretical background of our work and the details of Omiotis. Subsequently we discuss our methodology and how Omiotis is adapted to the needs of plagiarism detection in free text. Finally, we present the findings of our work, which can be summarized in a single phrase: semantic relatedness can significantly improve the efficiency of anti-plagiarism tools for text.

2. Background

While it is the proliferation of personal computing and advances in IT which exaggerated plagiarism and increased the need for anti-plagiarism techniques, at the same time it has provided an extended corpus of data for anti-plagiarism applications, which now have a wider basis against which they apply the comparison algorithms. In the case of free text plagiarism (The Higher Education Academy – Information and Computer Sciences, 2009), each sentence is considered a building block of the text, so the comparison between two documents is based on the comparison of their sentences.

Many known algorithms, independently of the used back bone techniques such as text pre-processing with natural language processing techniques (White & Joy, 2004), statistical methods for weighting of keywords (Brin, et al., 1995), or even use of semantic and syntactic knowledge (Aslam & Frost, 2003), build on the pair-wise similarities between sentences from the two texts and provide an average similarity score for the documents under comparison.

Advances in the cross-section of informatics and linguistics, namely in the field of computational linguistics, along with a large amount of text-based resources nowadays publicly available in digital formats (e.g., online encyclopedias) allow

software to become more ingenious in the fight against plagiarism (Lukashenko, et al., 2007).

Semantic relatedness measures estimate the degree of relatedness or similarity between two concepts in a thesaurus. Such measures can be classified to dictionary-based, corpus-based and hybrid. Among dictionary-based measures, the measures of (Agirre & Rigau 1995) and (Leacock, et al., 1998) take into account factors such as the density and depth of concepts in the set, or the length of the shortest path that connects them, or even the maximum depth of the taxonomy. However, in most such measures, it is assumed that all edges in the path are equally important. Resnik's (1999) measure for pairs of concepts is based on the Information Content (IC) of the deepest concept that can subsume both. The measure combines both the hierarchy of the used thesaurus, and statistical information for concept occurrences measured in large corpora. Recent works include the measure of Patwardhan and Pedersen (2006), which utilizes the gloss words found in the word's definitions to create WordNet-based context vectors, and several Wikipedia-based measures (Gabrilovich & Markovitch, 2007) (Milne and Witten, 2008). We encourage the reader to consult the analysis in (Budanitsky and Hirst, 2006) for a detailed discussion on relatedness measures. Although any of the aforementioned measures of semantic similarity or relatedness could fit a semantic-aware anti-plagiarism method, in this work, we use the Omiotis measure of semantic relatedness between text segments (Tsatsaronis, et al., 2010), which was shown to provide the highest correlation with human judgments among the dictionary-based measures of semantic relatedness. For the cases where one of the words does not exist in WordNet, we use the Wikipedia-based measure of Milne and Witten (2008), since among the offered Wikipedia-based alternatives, this is the fastest, and provides very high correlation with human judgements.

Omiotis works in three steps in order to measure the semantic relatedness between two text segments (e.g. segment A and B in Figure 1). First, it finds the implicit semantic links between the words of the two segments. For this reason, it takes all the words in each segment (w_{A1}, … w_{Ai}, w_{B1}, … w_{Bj}) and finds all the possible concepts for each word ($c_{A1,1}$… $c_{A1,m}$ etc. as depicted in the two bottom layers of Figure 1). Second, it creates a semantic network that connects all the concepts of the two segments through WordNet or Wikipedia (middle and top layers of Figure 1). In the third step, it finds the shortest path connecting a word in the first segment to a word in the second segment, via one of its concepts and WordNet or via Wikipedia if the words do not appear in WordNet but exist in Wikipedia (depicted with thick line in Figure 1).

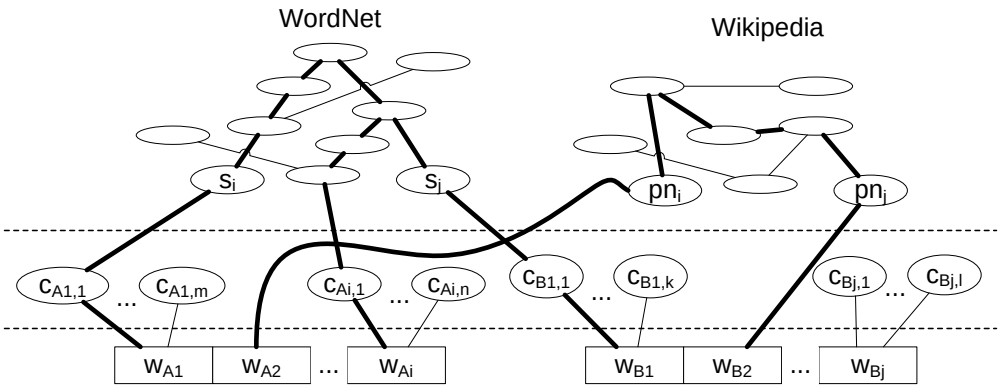

Figure 1. Semantic Network creation for text segments

The semantic relatedness between two words, one from each segment (e.g. w_{A1} and w_{B1} as depicted in the example of Figure 1) is measured on the path that connects the two words. Since the edges that connect two words differ in type and meaning and the vertices in a path from one word to another differ in specificity, Omiotis examines all the paths that connect two words, and thereafter computes the weights of these paths by considering: (a) the length of the semantic path; (b) the intermediate nodes' specificity, denoted by the node depth in the thesaurus' hierarchy; and (c) the weight of the semantic edges that compose the path, which depends to the edge type and is analogous to the edge type's frequency of occurrence in the thesaurus O. Eventually, the semantic relatedness for a pair of concepts corresponds to the maximum path weight.

To quantify the degree to which two text segments relate to each other, we build upon the semantic relatedness of their words but also consider the lexical similarity between segments. This is because segments may contain overly-specialized terms (e.g., an algorithm's name) that are inadequately (if at all) represented in WordNet or Wikipedia. We begin with the estimation of the terms' importance weights as these are determined by the standard TF-IDF weighting scheme. Thereafter, we estimate the lexical similarity, denoted as $\lambda_{i,j}$ between terms w_{Ai} (i.e. the ith word in segment A) and w_{Bj} (i.e. the jth word in title B) based on the harmonic mean of the respective terms' TF-IDF values, given by:

$$\lambda_{i,j} = \frac{2 \cdot TF\_IDF(a_i, A) \cdot TF\_IDF(b_j, B)}{TF\_IDF(a_i, A) + TF\_IDF(b_j, B)}$$

Having computed the lexical similarity between title terms α_i and b_j, we estimate their semantic relatedness, i.e. $SR(\alpha_i, b_j)$. Our next step is to find for every word α_i in title A the corresponding word b_j in title B that maximizes the product of lexical similarity and semantic relatedness values:

$$x(i) = \underset{j \in [1,|B|]}{\arg\max}(\lambda_{i,j} \cdot SR(a_i, b_j))$$

where x(i) corresponds to that term in title B, which entails the maximum lexical similarity and semantic relatedness with term α_i from title A. Consequently, we aggregate the lexical and semantic relevance scores for all terms in title A, with reference to their best match in title B denoted as shown in equation 3:

$$\zeta(A, B) = \frac{1}{|A|}\left(\sum_{i=1}^{|A|} \lambda_{i,x(i)} \cdot SR(a_i, b_{x(i)})\right)$$

We repeat the process for the opposite direction (i.e. from the words of B to the words of A) to cover the cases where the two titles do not have an equal number of terms. Finally, we derive the degree of relevance between titles A and B by combining the values estimated for their terms that entail the maximum lexical and semantic relevance to one another, given by the following equation:

$$Omiotis(A, B) = \frac{[\zeta(A, B) + \zeta(B, A)]}{2}$$

In order to improve the scalability of Omiotis, we have precomputed and stored all SR values between every possible pair of WordNet synsets in an RDBMS. This is a one-time computation cost, which dramatically decreases the computational complexity of Omiotis, making it scalable and fast.

3. Methodology

In order to test whether semantic similarity improves the performance of plagiarism detection algorithms, we perform a series of comparative experiments using a synthetic corpus comprising original documents and documents that plagiarize their content. Plagiarized content has been created artificially, using a computer program, called random plagiarist, which constructs plagiarism by modifying the original content by repositioning keywords, replacing keywords with synonyms or even translating text to another language. The length of the plagiarized passages differs per case and so does the degree of obfuscation, which ranges from 'none' to 'high'.

3.1 The plagiarism detection process

We consider that the plagiarism detection is done in two steps as depicted in Fig.2 (Stein et al, 2007). First, each document that is candidate for plagiarism is compared against all the original documents in the collection and the suspect documents are selected for the second step. More specifically, the documents are processed in segments, so that only the suspect segments inside each suspect document are retrieved. In the second step, we compare each suspect segment against its possible source and decide whether it is a plagiarism or not.

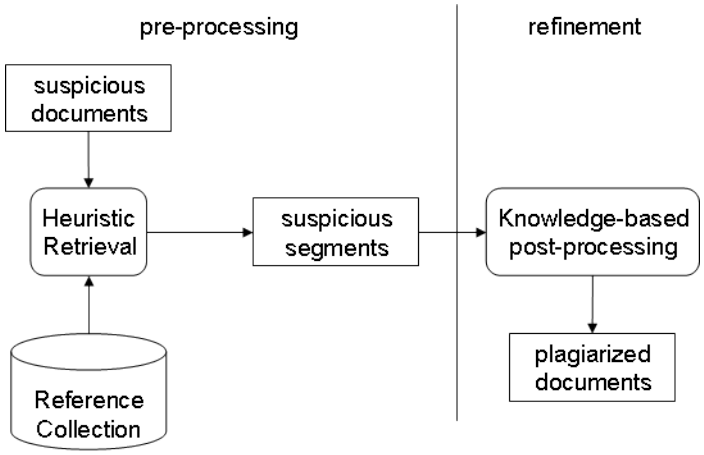

Figure 2. The plagiarism detection process

The first step is computationally demanding since it requires a lot of comparisons to be performed between text segments and is not so critical since it provides an indication on suspect segments. For this reason, a text based similarity approach can be employed. The decision in the second step is more critical, since we must determine whether the suspect text segment is a plagiarism or not based on the textual but also in the conceptual similarity between the suspect and the source segment. In this second step we employ Omiotis for measuring the similarity of text segment pairs.

3.2 Post-processing with Measures of Semantic Relatedness

For the second step, we present two approaches for plagiarism detection using Omiotis; one unsupervised and one supervised. In the first approach, Omiotis is used as a measure of semantic relatedness between text fragments, and high Omiotis scores indicate plagiarism between the examined texts. In the second approach we build upon the machinery of supervised learners, and more specifically on support vector

machines (SVM) and train a classifier in order to be able to identify plagiarisms in free text with higher accuracy than the unsupervised approach. In this case, Omiotis constitutes one of the features we are using to train the classifier.

3.2.1 Unsupervised Approaches

When we have to decide on whether a pair of text segments constitutes a plagiarism incident based on a similarity score, we should normally set a similarity threshold below which the pair is considered to be a plagiarism. As mentioned before, in the second step of the plagiarism detection process, we have a long list of suspect segment pairs and the respective similarity value for each pair. In this case, the definition of this threshold value is performed by using the complete set of all values and several statistical based techniques.

The first option is to rank all segment pairs based on the similarity value and then define the Cut-off threshold either as the average of all values or the median value in the set. Consequently all pairs that have a similarity value greater than the cut-off threshold are the plagiarism incidents whereas those that have values below threshold are considered clean.

The second option is to rank the segment pairs using each different similarity score thus producing three different rankings and consequently merge the rankings in a single overall ranking. The highly ranked pairs are definitely plagiarisms whereas the low ranked ones are not. The problem here is to define the ranking threshold, the position below which segment pairs are not considered plagiarism. The method we employ as described in details in (Klementiev, et al., 2007), considers an initial ranking threshold k_i for each of the three rankings and a weighting scheme that initially assigns random weights in each of the three ranking. The algorithm iteratively adjusts the 3 weights and re-aggregates the different rankings until the aggregated ranking is stabilized.

3.2.2 Supervised Approaches

In this latter case, free text anti-plagiarism detection is addressed as a classification problem (i.e., a given pair of text fragments may or may not be a plagiarism example) and a classifier can be trained in order to evaluate the properties of each suspicious pair and decide whether this pair is actually a plagiarism. We experiment with several different classification algorithms, as given by the Weka Data Mining Suite (www.cs.waikato.ac.nz/ml/weka). In all the experiments we employ the scores given for each pair by the different similarity measures and perform a 10-fold cross validation that uses the 90% of the samples for training and the 10% for testing, repeating this experiment with a random initial split 10 times. In analogy to the combined rank unsupervised method, we test our classifiers when the three similarity scores for each pair are employed all in the same time. Once again we perform a 10-fold cross validation for this later case.

The evaluation of our methods, which is presented in the following section, shows our findings in the same order, using unsupervised approaches first for defining a cut-off threshold and supervised approaches next in order to classify a pair of segments as being a plagiarism or not.

4. *Findings*

For the evaluation of our methods, we employ a synthetic dataset comprising source and plagiarized documents. The dataset was released for the 1st International Competition on Plagiarism Detection[1] and comprises 20611 suspicious documents and 20612 source documents. In order to avoid the tedious task of locating suspicious text segments in the complete collection, we employed only the annotated segment pairs as released in the completion results and from those only the monolingual pairs, since our implementation is for the English language only. This resulted in almost 11.000 text segment pairs which have been tagged as plagiarism incidents and which are divided into 3 groups: 3400 pairs with high obfuscation, which are difficult to detect, 3400 pairs with low obfuscation and 4200 pairs with no obfuscation at all, which are the easiest to detect.

In order to have a counter set for our methods, we created 11.000 additional pairs of segments, which are definitely non-plagiarisms, so as to balance the data set between plagiarism cases and cases where plagiarism does not exist. The suspect segment of each pair was replaced by a randomly selected segment of the same size, from the same suspect document.

We produced 3 different similarity values for all the 22.000 pairs using a) cosine measure for textual similarity (using the TF-IDF weighting scheme for terms and the vector space representation), b) Omiotis and conceptual similarity using WordNet only as a knowledge base (we call this measure Omi), c) Omiotis and conceptual similarity using WordNet and Wikipedia as knowledge bases (we call this measure OmiWiki). In order to further evaluate the ability of each measure in detecting well hidden plagiarism (that uses paraphrasing, synonyms etc.) we present our results in each of the three obfuscation groups (none, low, high).

4.1. *Unsupervised Plagiarism Detection Evaluation:*

As a first step we evaluate the performance of the three measures, with respect to the similarity values they provide for all the pairs (both plagiarism and non-plagiarism pairs). The basic hypothesis is that any of the measures will identify high similarity between two text segments that constitute a plagiarism, and lower similarity in a different case. The aim of the unsupervised learning in this case is to identify a cut-off value (i.e., a threshold) for each measure, above which the respective pair should be deemed as a plagiarism case, and below which is judged as a non-plagiarism, but without using any training examples. In the following we present three different approaches of the problem in the unsupervised context. In the first approach (a), we try to identify the cut-off based on simple measurements, such as the mean and the median of the provided values for each measure. In an effort to combine the values of the three methods, we examine the problem from the information retrieval perspective (b) and we apply an unsupervised rank aggregation method to improve the overall performance.

a) Unsupervised learning of the plagiarism pairs

In the following two tables, we present the results in terms of Precision (P), Recall (R), and F-Measure (harmonic mean of precision and recall) for the cases where mean (Table 1) and median (Table 2) are used as cut-off values respectively. A first

[1] http://www.uni-weimar.de/medien/webis/research/workshopseries/pan-09/competition.html

observation is the fact that Omiotis, as well as OmiWiki, provide a better ranking of the pairs compared to cosine, and this is more obvious from the results of the second table, where essentially the median value for the cosine is 0 (this is also why recall has a value of 1). Omiotis and OmiWiki provide more fine grained distinction while measuring the similarity between two text segments, and this is also shown from the fact that the overall F-Measure in all cases is better when Omiotis or OmiWiki is used, compared to Cosine. Another observation is the fact that in this data set the mean value acts as a better cut-off for all measures, compared to the median. These results show two things: (1) The distribution of the produced values is skewed, and in fact, after conducting analysis, it was found that it was right-skewed, meaning that there are a lot of small values and few large values for all measures. (2) Mean is a good cut-off value in the unsupervised case, since the respective precision in all cases is really high, as well as the overall F-Measure.

	All			None			Low			High		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Cosine	0.99	0.82	0.90	0.98	0.75	0.85	0.97	0.94	0.95	0.97	0.94	0.95
Omi	**0.99**	**0.85**	**0.92**	0.96	0.77	0.86	0.96	0.94	0.95	0.94	0.95	0.94
OmiWiki	0.99	0.84	0.91	**0.98**	**0.76**	**0.87**	**0.98**	**0.94**	**0.96**	**0.97**	**0.95**	**0.96**

Table 1: Precision (P), Recall (R), and F-Measure (F1) for all three used measures, in the four examined data sets (all, none obfuscation, low obfuscation, and high obfuscation), with the mean value used as a cut-off value.

	All			None			Low			High		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Cosine	0.92	0.89	0.90	0.29	1	0.45	0.24	1	0.39	0.25	1	0.40
Omi	**0.93**	**0.89**	**0.91**	0.49	0.84	0.62	**0.48**	**0.98**	**0.65**	**0.48**	**0.98**	**0.65**
OmiWiki	**0.93**	**0.89**	**0.91**	**0.51**	**0.85**	**0.63**	0.48	0.97	0.64	0.48	0.97	0.64

Table 2: Precision (P), Recall (R), and F-Measure (F1) for all three used measures, in the four examined data sets (all, none obfuscation, low obfuscation, and high obfuscation), with the medan value used as a cut-off value.

b) Evaluation from the information retrieval perspective

In an effort to combine in an unsupervised manner the values produced by the three measures, we examined the problem from the information retrieval perspective. In this case we consider that we rank all the pairs in the given data set (all, none, low, high), both plagiarism and non-plagiarism pairs, by their similarity value and we measure the interpolated precision at the 11 standard recall points, given that we are searching for the plagiarized pairs, and thus considering as relevant only those pairs. The detailed results show that Omiotis and OmiWiki outperform Cosine in the first 5 recall points, and in all four examined data sets. We summarize the interpolated precision/recall curves with the mean average precision (MAP) values for the different measures and datasets, in Table 3. As expected, Omiotis and OmiWiki provide a better ranking in the cases where there is some type of obfuscation, either low, or high. Essentially, one important conclusion from this analysis is the fact that simple keyword matching measures, such as cosine cannot detect easily the cases where the plagiarism has been altered with some synonyms, or few different expressions, while Omiotis and OmiWiki can capture better those cases. With this analysis, we can then proceed with combining the rankings of the measures, and identifying a proper cut-off for improving the overall performance.

	Cosine	Omiotis	OmiWiki
ALL	**0,948216**	0,94637	0,946629
None	**0,875089**	0,852222	0,853184
Low	0,930807	**0,931353**	**0,931341**
High	0,929725	**0,93113**	**0,930869**

Table 3: Mean Average Precision (MAP) for all three used measures, in the four examined data sets (all, none obfuscation, low obfuscation, and high obfuscation), with none cut-off value used.

c) Unsupervised rank aggregation

In order to combine the rankings of the measures provided in (b), we are using an unsupervised rank aggregation method described in (Klementiev, Roth and Small, 2007). The results of the unsupervised rank aggregation using mean as a cut-off value are shown in Table 4. These results are directly comparable with the ones presented in Table 1. The results show that aggregating the values of the three measures may not improve the performance in the cases where the plagiarism does not contain obfuscation, but it improves the performance in cases where there is a degree of obfuscation in the plagiarized cases.

	All			None			Low			High		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Aggregation	0.999	0.69	0.81	0.988	0.76	0.86	0.99	0.94	0.97	0.98	0.95	0.97

Table 4: Precision (P), Recall (R), and F-Measure (F1) for the unsupervised rank aggregation, in the four examined data sets (all, none obfuscation, low obfuscation, and high obfuscation), with the mean value used as a cut-off value.

4.2. Supervised Plagiarism Detection Evaluation

In an effort to examine the use of supervised techniques for plagiarism detection, we employed two well known techniques, namely logistic regression and support vector machines. In all cases we used a 10-fold cross validation to evaluate the performance of the classifiers, and we examined the instances from four different perspectives: (a) the only feature used is the cosine value, (b) the only feature used is the Omiotis value, (c) the only feature used is the OmiWiki value, and (d) all values are used as features. The results are shown in Table 5 for the case of the logistic regression, and Table 6 for the case of SVM.

	All			None			Low			High		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Cosine	0.987	0.88	0.93	0.983	0.758	0.855	0.988	0.935	0.96	0.963	0.934	0.948
Omi	0.988	0.878	0.929	0.99	0.758	0.858	0.991	0.934	0.961	0.988	0.936	0.961
OmiWiki	0.992	0.878	0.931	0.991	0.759	0.859	0.993	0.934	0.962	0.989	0.934	0.96
All Features	0.989	0.879	0.93	0.987	0.758	0.857	0.992	0.935	0.962	0.989	0.938	0.962

Table 5: Precision (P), Recall (R), and F-Measure (F1) for all four used set of features, in the four examined data sets (all, none obfuscation, low obfuscation, and high obfuscation), with the mean value used as a cut-off value, and Linear Regression used as a classifier.

	All			None			Low			High		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Cosine	0.992	0.871	0.932	0.992	0.745	0.851	0.99	0.932	0.96	0.984	0.932	0.957
Omi	0.995	0.871	0.933	0.996	0.744	0.852	0.994	0.929	0.96	0.991	0.931	0.96
OmiWiki	0.996	0.87	0.933	0.997	0.744	0.852	0.995	0.928	0.96	0.992	0.932	0.961
All Features	0.995	0.873	0.934	0.995	0.775	0.871	0.993	0.933	0.962	0.991	0.937	0.963

Table 6: Precision (P), Recall (R), and F-Measure (F1) for all four used set of features, in the four examined data sets (all, none obfuscation, low obfuscation, and high obfuscation), with the mean value used as a cut-off value, and SVM used as a classifier.

Discussion-Conclusion

The evaluation of our semantic similarity measure using WordNet and Wikipedia resources showed improved performance against baseline statistical methods (stemming, tf/idf weighting and cosine), either supervised or unsupervised approaches are employed for determining the appropriate similarity thresholds.

Undeniably, using only the textual information and occurrence statistics is the first step in detecting plagiarism. Mainly because of the complexity of the semantic solution, this preprocessing is necessary in order to narrow the set of suspicious cases. However, the use of semantic relatedness is necessary to decide for the ambiguous cases of plagiarism.

Concerning the scalability of our approach we should mention that it can be embedded in any existing plagiarism detection software in order to improve its results, either it searches for plagiarism in a predefined corpus of essays or it uses the web as a database. As explained in 3.1, traditional matching techniques can be used to locate suspect fragments in the first step and our semantic method can be subsequently applied to refine results at sentence level. Our next step is to find an open source plagiarism detection software to employ in the first phase and combine it with our semantic-based plagiarism detection module in order to be able to apply it directly on collections that comprise suspicious documents.

References

Agirre E. and Rigau. G. (1995) ‘A proposal for word sense disambiguation using conceptual distance’. Proceedings of the International Conference on Recent Advances in NLP.

Aslam J. A. and Frost M. (2003) ‘An Information-theoretic Measure for Document Similarity’, Proceedings of the 26th International ACM/SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, July 28-August 01, pp. 449–450.

Brin S., Davis J. and Garcia M.H. (1995) ‘Copy Detection Mechanisms for Digital Documents’, Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, San Jose, California, May 22-25, pp. 398–409.

Budanitsky A. and Hirst, G. Evaluating wordnet-based measures of lexical semantic relatedness. Computational Linguistics, 32(1):13–47, 2006.

Gabrilovich E. and Markovitch, S. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In IJCAI, pages 1606–1611, 2007.

Leacock, C., Miller, G. and Chodorow, M. (1998). Using corpus statistics and wordnet relations for sense identification. Computational Linguistics, 24(1):147–165.

Klementiev, A., Roth, D., and Small, K. 2007. An Unsupervised Learning Algorithm for Rank Aggregation. In Proceedings of the 18th European Conference on Machine Learning (Warsaw, Poland, September 17 - 21, 2007). J. N. Kok, J. Koronacki, R. L. Mantaras, S. Matwin, D. Mladenič, and A. Skowron, Eds. Lecture Notes In Artificial Intelligence, vol. 4701. Springer-Verlag, Berlin, Heidelberg, 616-623.

Lukashenko R., Graudina V. and Grundspenkis J. (2007) 'Computer-Based Plagiarism Detection Methods and Tools: An Overview', Proceeding of the International Conference on Computer Systems and Technologies- CompSysTech'07, Rousse, Bulgaria, June 14-15, pp. IIIA.18-1 – IIIA.18-6.

Milne, D. and Witten, I.H. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In AAAI Workshop on Wikipedia and Artificial Intelligence, 2008.

S. Patwardhan and T. Pedersen. Using wordnet based context vectors to estimate the semantic relatedness of concepts. In EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together, 2006.

Park, C. (2004) 'Rebels Without a Clause: Towards an Institutional Framework for Dealing with Plagiarism by Students', Journal of Further and Higher Education, vol. 28, no 3, pp. 292-306. [Online] Available at: http://148.88.1.1/staff/gyaccp/rebels%20without%20a%20clause.pdf (Accessed: 30/11/09)

Resnik, P. (1999) Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research, 11:95–130, 1999.

Roig, M. (2006) "Avoiding plagiarism, self-plagiarism, and other questionable writing practices: A guide to ethical writing". [Online] Available at: http://ori.dhhs.gov/education/products/plagiarism/plagiarism.pdf (Accessed: 02/12/09).

Stein, B., zu Eissen, S. M., and Potthast, M. 2007. Strategies for retrieving plagiarized documents. In Proceedings of the 30th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Amsterdam, The Netherlands, July 23 - 27, 2007). SIGIR '07. ACM, New York, NY, 825-826.

The Higher Education Academy – Information and Computer Sciences (2009) [Online] Available at:

http://www.ics.heacademy.ac.uk/resources/assessment/plagiarism/research_freetext.html (Accessed: 10/12/09).

Tsatsaronis, G., Varlamis, I. and Vazirgiannis, M. (2010) 'Text Relatedness Based on a Word Thesaurus', Journal of Artificial Intelligence Research, JAIR, Volume 37, pages 1-39.

White, D. and Joy, M. (2004). 'Sentence-based Natural Language Plagiarism Detection'. *Journal on Educational Resources in Computing (JERIC)*, vol 4**,** no 4. [Online] Available at: http://portal.acm.org/citation.cfm?doid=1086339.1086341 (Accessed: 10/12/09).