# Plagiarism Detection Scheme Based on Semantic Role Labeling

Ahmed Hamza Osman[1, 2], Naomie Salim[1], Mohammed Salem Binwahlan[1], Ssennoga Twaha[1], Yogan Jaya Kumar[1] and Albaraa Abuobieda[1, 2]

[1] Faculty of Computer Science and Information Systems,
Universiti Teknologi Malaysia, 81310,
Skudai, Johor, Malaysia.
[2] Faculty of Computer Studies,
International University of Africa, 2469,
Khartoum, Sudan.
**ahmedagraa@hotmail.com**

***Abstract*—Nowadays, many documents are available on the internet and are easy to access. Due to this wide availability, users can easily create a new document by copying and pasting. Plagiarism occurs when the content is copied without permission or citation. This paper introduces a plagiarism detection technique based on the Semantic Role Labeling (SRL). The technique analyses and compares text based on the semantic allocation for each term inside the sentence. SRL is superior in generating arguments for each sentence semantically. In addition, experimental results on PAN-PC-09 data sets showed that our method outperforms the modern methods for plagiarism detection in terms of Recall, Precision and F-measure.**

***Keywords* — Plagiarism Detection; Semantic Similarity; Semantic Role; Arguments**

## I. INTRODUCTION

Plagiarism defined as "unacknowledged copying of documents or programs" [1]. It can occur in many sectors. For example, companies may look for competitive advantage, and academicians need to advance their institutions by searching for quick ways for publishing their works. Most empirical studies and analysis were undertaken by the academic community to deal with student plagiarism. In order to discriminate plagiarized documents from non-plagiarized documents, a correct selection of text features is a key aspect. There are many types of plagiarism mentioned by [2], such as copy and paste, redrafting or paraphrasing of the text, plagiarism of idea, and plagiarism through translation from one language to another.

The matching algorithms are also dependent on the text's lexical structure rather than semantic structure. Therefore, it becomes difficult to detect the text paraphrased semantically. The big challenge is to provide plagiarism checking with appropriate algorithm in order to improve the percentage of finding result and time checking. The important question for the plagiarism detection problem in this study is whether it is possible to apply new techniques such as Semantic Role Labeling to handle plagiarism problems for text documents.

Several plagiarism detection tools use character matching or string matching method to detect the plagiarized text. However, most of the current softwares and techniques are less effective in detecting a plagiarized text because these tools tend to compare the suspected text with original text using characters matching, some with chunks while others by words. This leads to exhaustive search which takes a long time in the matching process. The matching algorithms are working depending on the text lexical structure rather than semantic structure. Therefore it becomes difficult to detect the text paraphrased semantically. One of the objectives of this study is to propose new semantic techniques for plagiarism detection based on Semantic Role Labeling. The proposed method does not analyze the content of a text document as text syntax only, but also captures the underlying semantic meaning in terms of the relationships among its terms.

Semantic Role Labeling (SRL) is one of the Natural Language Processing techniques that are used in many fields such as text summarization [3], text clustering [4] and text categorization [5]. In this paper, we proposed a new plagiarism detection method based on SRL. The proposed method can detect copy paste plagiarism, rewording or synonym replacement, changing of word structure in the sentences, modifying the sentence from passive voice to active voice and vice versa. SRL was used to analyze the sentences semantically and WordNet thesaurus was used to extract the concepts or synonymies for each word inside the sentences. The rest of the paper is organized as follows: Section 2 provides a description of the related work in plagiarism detection. In Section 3, a full description of the underlying idea involved in our method is covered. Section 4 discusses the experimental design used in our proposed method. Corpus and dataset, including similarity detection and results discussion of the proposed approach, are presented in Section 5, whereas Section 6 concludes the paper.

## II. RELATED WORKS

This section mainly discusses some of recently proposed plagiarism detection techniques.

Reference [6] considers text comparison based on word n-grams. With reference to this, the suspected text is split into two sets of tri-grams to be compared. The amount of

978-1-4673-1090-1/12/$31.00 ©2012 IEEE

common tri-grams is considered in order to detect potential plagiarism cases. Reference [7] considers the sentence as the comparison unit in order to compare local similarities. It differentiates among exact copy of sentences, word insertion, word removal and rewording.

Reference [8] introduced a plagiarism detection system from Stanford Digital Library Project named COPS (copy protection system), which detects document overlap relying on string matching and sentences. Its main drawback is that it fails to consider individual words and takes the whole sentence as one part. The shortcomings of COPS was solved by [9], who then developed a new method called Stanford Copy Analysis Method (SCAM) to improve the COPS using Relative Frequency Model (RFM) to stand out subset copies. RFM is an essential asymmetric similarity measure for plagiarism detection. The main advantage of SCAM is that it can find the overlapping similarity between the parts of sentences, but many terms can be misleading in documents sharing comparison. The method proposed by [10] and adopted by [11] is called the Longest Common Subsequence (LCS). LCS is one of the techniques used in ROUGE, which is a well-known summary evaluation method. Given two sequences X and Y, the longest common subsequence (LCS) for X and Y is the common subsequence with maximum length [12]. Reference [13] proposed a new algorithm for plagiarism detection. It is capable of detecting sophisticated obfuscation (such as paraphrasing, reordering, merging, and splitting sentences) as well as direct copying.

## III. SEMANTIC ROLE LABELING (SRL)

Semantic structures or semantic frames were introduced by [14] where common frames were used for common roles and themes such as FrameNet proposed by Baker [15] and PropBank proposed by [16]. A statistical system is trained on the data from the FrameNet project to automatically assign semantic roles [17]. Reference [18], [19] and Palmer [20] followed this approach by improving sets of features and machine learning methods.

SRL is a process to identify and label arguments in a text. The basic idea is that the sentence level semantic analysis of text determines the object and subject of a text. It can be extended to the characterization of events such as determination of "who" did "what" to "whom", "where", "when", and "how". The predicate of a clause (usually a verb) establishes "what" took place, and other parts of the sentence express the other arguments of the sentence (such as "who" and "when"). The primary task of semantic roles labeling is to identify what semantic relation holds among a predicate and its associate participants or properties, with these relations drawn from a pre-defined list of possible semantic roles for that predicate or class of predicate. The typical labels used in SRL are Agent, Patient and Location for the entities participating in an event. Those labels can be extended to more specific arguments such as time and place in some text.

## IV. PLAGIARISM DETECTION USING SRL

Plagiarism detection using semantic role labeling aims to detect the semantic similarity between a sentence and possible semantic similarity between two sentences. In this Section, we discuss the idea of our proposed method. We first pre-processed suspected documents and original documents using text segmentation, stop words removal and stemming. Then, SRL was used to transform the sentences into arguments based on location for each term in the sentences. The verbs of the sentences play an important role in the process, and the analysis of the sentences rely on the verbs of the sentences as mentioned in the literature review. Extracted concepts from the arguments were represented as a graph. All the arguments extracted from the text were grouped in the nodes according to the argument type. Each node contains similar extracted arguments. Each group was named by the argument name such as Arg0, Arg1, V, Time, Location, etc. This step is called Argument Label Group (ALG). Then, we extracted all the concepts for each term in the argument groups using WordNet thesaurus. This step is called Semantic Term Annotation (STA). All the concepts that were extracted by WordNet thesaurus were collected in one node known by Topic Signature Node. The advantage of a Topic Signature Node is that it quickly guides us to capture the suspected parts from the documents. The following Fig. 1 shows general architecture for our proposed method.

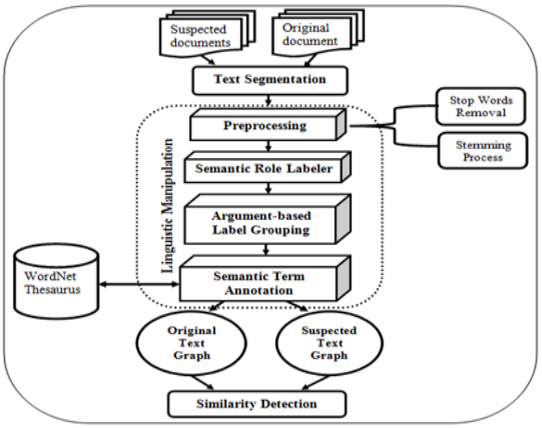


Figure 1. Proposed method of general architecture

Another situation for the plagiarist can be shown through the following example:

*John gave Mary the book (original sentence).*

*The book was given to Mary by John (suspected sentence)*

By using SRL the produced arguments are:

*John gave Mary the book*

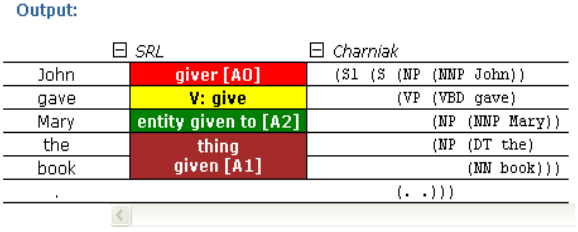


Figure 2. Analysis for original sentence using SRL

*The book was given to Mary by John*

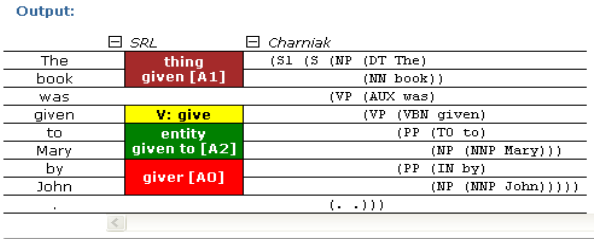


Figure 3. Analysis for suspected sentence using SRL

Fig. 2 and 3 illustrate the analysis for suspected sentence using SRL in the example above. We note that the structure of two sentences above may differ if the active versus passive voice or synonyms and antonyms are used. Actually, these sentences can be semantically the same. It was noted that the SRL captures the arguments (subject, object, verb and indirect object) for a sentence despite changing the places for the labels inside the sentences. This capturing supports our proposed method in plagiarism detection if comparison is applied based on the arguments of the sentence using SRL.

## V. EXPERIMENTAL DESIGN AND DATASET

Our experiments looked at the amount of detected plagiarized sentences from the original documents. The experiments were performed on 100 suspected documents, each plagiarized from one or more original documents according to the PAN-PC-09 dataset.

In this stage, sentence-based similarity analyses between tokenized suspected and original documents were performed. Sentences in suspected documents were compared with each sentence in the candidate documents according to the arguments of the sentences. We aimed to detect not only the arrangement similarity between sentences, but also possible semantic similarity between two sentences.

Similarity detection was conducted by comparing the original Topic Signature terms and suspected Topic Signature terms. If the two terms were found to be identical, we went directly to the argument label groups that contained these terms, and then determined the label group where they belonged, thus determining the possible sentences that may be plagiarized. This step compared the arguments of possible sentences that had been plagiarized with the corresponding arguments in original sentences. The argument label group guided us to the main arguments and each argument inside the group quickly guided us to the possible plagiarized sentence.

Some variables play an important role in the similarity calculation, such as the number of matched arguments and number of arguments which exist in the sentences. The first variable determines the similar arguments between the suspected document and original document while the second variable determines the argument that does not exist in the sentences. The similarity between the arguments of the suspected document and original document was calculated according to *Jaccard* coefficient measure [21] which is well-known as famous similarity measure between two sets. *Jaccard* coefficient [21] defined as a following equation:

$$Similarity\, C_i(ArgS_j, ArgS_k) = \frac{C(ArgS_j) \cap C(ArgS_k)}{C(ArgS_j) \cup C(ArgS_k)} \quad (1)$$

Where,

C($ArgS_j$)=concepts of the argument sentence in the suspected document; $C_i$ ($ArgS_k$) =concepts of the argument sentence in the original document;

We then calculated the similarity between the suspected document and original document based on the following equation:

TABLE I.
EVALUATION MEASURE OF THE PROPOSED METHOD

| Number of documents | Recall | Precision | F-measure |
|---|---|---|---|
| 100 | 0.818421 | 0.642406 | 0.719809 |

$$Total\ Simiarity(Doc\,1, Doc\,2) = \sum_{i=1,l} \sum_{j=1,m \atop k=1,n} SimC_i(ArgS_j, ArgS_k) \quad (2)$$

Where,

SimCi (*Arg Sj*, *Arg Sk*) is similarity between Arguments sentence j in suspected document containing concept i Arguments sentence k in original document containing concept i , l = no. of concepts, m= no. of Arguments sentence in suspected document, n = no. of Arguments sentence in original document.

## VI. RESULTS AND DISCUSSION

Our technique was tested according to the group of documents (5, 10, 20, 40, and 100). Those suspected documents were plagiarized with different ways of plagiarism such as simple copy and paste, changing some terms with their corresponding synonyms, and modifying the structure of the sentences (paraphrasing). We provided three general testing parameters that are commonly used in plagiarism detection as following:

$$Recall = \frac{Number\ of\ Detected\ Arguments}{Total\ Number\ of\ Arguments} \quad (3)$$

$$Precision = \frac{Number\ of\ Plagiarized\ Arguments}{Number\ of\ Detected\ Arguments} \quad (4)$$

$$F-Measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (5)$$

The proposed method is evaluated and compared with some plagiarism detection algorithms in PAN 2010 competition. The results from comparison are illustrated in Fig. 4.

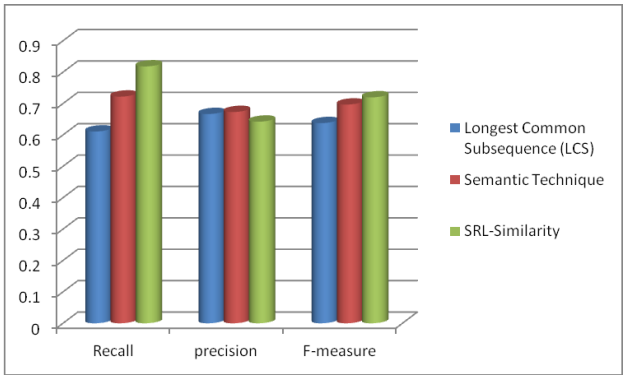


Figure 4. Comparison Results with Plagiarism Detection Techniques

Fig. 4 demonstrates the comparison between SRL-based Similarity with Fuzzy Semantic-based String Similarity [22], Longest Common Subsequence (LCS) [23] and Semantic-based similarity [24]. The figure demonstrates the results of similarity between the suspected and original documents for each set of documents. We found that all the scores that were obtained by our proposed method have good results than other method. The results from the comparison show that the proposed method achieved better results in terms of recall , precision and f-measure.

## VII. CONCLUSIONS

Semantic role labeling can be used for plagiarism detection by extracting argument of sentences and comparing the arguments. Tests were carried out using PAN-PC-09 standard dataset for plagiarism detection. The proposed methods were found to achieve better performance compared to fuzzy semantic-based string similarity, longest common subsequence (LCS) and semantic-based similarity.

## ACKNOWLEDGMENT

The researcher is sponsored by IDF and Ministry of Science Technology and Innovation under the university research grant vote number 01H74, Universiti Teknologi Malaysia.

## REFERENCES

[1] M. L. A Joy, M., "Plagiarism in Programming Assignments," *IEEE Transactions on Education* vol. 42, pp. 129-133, 1999.

[2] F. K. Hermann Maurer , Bilal Zaka "Plagiarism - A Survey," *Journal of Universal Computer Science,* vol. 12, pp. 1050-1084 2006.

[3] L. Suanmali, Salim, N. and Binwahlan, M. S. . Jurnal Teknologi Maklumat, pp.105-155,( 2009). "Automatic Text Summarization Using Feature-Based Fuzzy Extraction," *Jurnal Teknologi Maklumat,* vol. 2, pp. 105-155, 2009.

[4] S. D. B. S.Murali Krishna, "An Efficient Approach for Text Clustering Based on Frequent Itemsets," *European Journal of Scientific Research* vol. 42, pp. 399-410, 2010.

[5] S. Shehata, *et al.*, "AN EFFICIENT MODEL FOR ENHANCING TEXT CATEGORIZATION USING SENTENCE SEMANTICS," *Computational Intelligence,* vol. 26, pp. 215-231, 2010.

[6] C. Lyon, J. A. Malcolm, and R. G. Dickerson, "Detecting short passages of similar text in large document collections," *Empirical Methods in Natural Language Processing,* 2001.

[7] N. Kang, Gelbukh, A., Han, S.-Y. . In: Sojka, P., Kopeˇcek, I., Pala, Keds. TSD . LNCS LNAI, vol. 4188, pp. . Springer, Heidelberg,( ). "PPChecker: Plagiarism pattern checker in document copy detection," *LNCS LNAI,* vol. 4188, pp. 661–667, 2006.

[8] S. Brin, *et al.*, "Copy detection mechanisms for digital documents," presented at the Proceedings of the 1995 ACM SIGMOD international conference on Management of data, San Jose, California, United States, 1995.

[9] S. Brin, *et al.*, "Copy detection mechanisms for digital documents," *SIGMOD Rec.,* vol. 24, pp. 398-409, 1995.

[10] J. W. Hunt and T. G. Szymanski, "A fast algorithm for computing longest common subsequences," *Commun. ACM,* vol. 20, pp. 350-353, 1977.

[11] Chow Kent and N. Salim, "Features Based Text Similarity Detection," *Journal of Computing,* vol. 2, pp. 53-57, 2010.

[12] C.-Y. Lin, "ROUGE: A Package For Automatic Evaluation Of Summaries," 2004.

[13] D. R. White and M. S. Joy, "Sentence-based natural language plagiarism detection," *J. Educ. Resour. Comput.,* vol. 4, p. 2, 2004.

[14] C. J. Fillmore, "The case for case. In Emmon Bach and Robert T," *Universals in Linguistic Theory. Holt, Rinehart, and Winston, New York,* pp. 1–210, 1968.

[15] C. F. Baker, *et al.*, "The Berkeley FrameNet Project," presented at the Proceedings of the 17th international conference on Computational linguistics - Volume 1, Montreal, Quebec, Canada, 1998.

[16] M. Palmer, *et al.*, "The Proposition Bank: An Annotated Corpus of Semantic Roles," *Comput. Linguist.,* vol. 31, pp. 71-106, 2005.

[17] D. Gildea and D. Jurafsky, "Automatic labeling of semantic roles," *Comput. Linguist.,* vol. 28, pp. 245-288, 2002.

[18] M. Surdeanu, *et al.*, "Using predicate-argument structures for information extraction," presented at the Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, Sapporo, Japan, 2003.

[19] S. S. Pradhan, Wayne H. Ward, Kadri Hacioglu, James H. Martin, and Dan Jurafsky. , pages , , "Shallow semantic parsing using support vector machines," *In Proceedings of NAACL-HLT 2004,* pp. 233–240, 2004.

[20] N. a. M. P. Xue, "Calibrating features for semantic role labeling," *n Proceedings of EMNLP 2004,* pp. 88–94, 2004.

[21] P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et des Jura," *Bulletin de la Société Vaudoise des Sciences Naturelles,* vol. 37, pp. 547–579, 1901.

[22] S. Alzahrani and N. Salim, "Fuzzy Semantic-Based String Similarity for Extrinsic Plagiarism Detection," *CLEF (Notebook Papers/LABs/Workshops)* 2010.

[23] N. S. Chow Kok Kent, "Features Based Text Similarity Detection," *Journal of Computing,* vol. 2, pp. 53-57, 2010.

[24] Chow Kok Kent and N. Salim, "Web Based Cross Language Plagiarism Detection," *Second International Conference on Computational Intelligence, Modelling and Simulation,* pp. 199-204, 2010