Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 46 (2015) 223 – 230

Procedia Computer Science

International Conference on Information and Communication Technologies (ICICT 2014)

An Improved SRL based Plagiarism Detection Technique using Sentence Ranking

Merin Paul[a,]*, Sangeetha Jamal[b]

[a]Department of Computer Science, Rajagiri School of Enginering and Technology, Kochi-39, India
[b]Assistant Professor, Department of Computer Science, Rajagiri School of Enginering and Technology, Kochi-39, India

Abstract

Plagiarism means intellectual theft which consists of turning someone else's work as your own. Plagiarism has become widespread in many fields like institutions, companies etc. This paper proposes a new technique which uses Semantic Role Labeling and Sentence Ranking for plagiarism detection. Sentence ranking gives suspicious and original sentence pairs through vectorizing the document. Then proposed method analyses and compares the ranked suspected and original documents based on the semantic allocation of each term in the sentence using SRL. It was found out that the application of sentence ranking in plagiarism detection method decreases the time of checking.

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)

Keywords: Plagiarism; Plagiarism detection; Semantic Role Labeling; Sentence Ranking; Semantic Similarity

1. Introduction

Plagiarism means a piece of writing that has been taken from a source without proper citation. Therefore it is an intellectual theft, which consists of turning someone else's work as your own. Plagiarism exists in many different scenarios and it causes an increasing challenge to publication industry, which affects academia and the publication industries in particular. Plagiarism detection in natural language documents is an important concept in the information processing field, and it is used to protect the author's intellectual property. Plagiarism originates from a Latin verb which means, 'to kidnap'. Therefore, if you plagiarize you're kidnapping and stealing others hard work and intellectual property, which is an act of academic and public dishonesty[13].

* Corresponding author. Tel. +91 9544583714
E-mail address: merin149@gmail.com

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)
doi:10.1016/j.procs.2015.02.015

Plagiarism occurs in various forms: submitting another's work exactly same without proper citation, paraphrasing text, reordering the sentences, using synonyms, or changing grammar, code plagiarism etc... Plagiarism mainly seen in academic institutions where academicians or researchers are requested to regularly update their work. Because of the availability of large amount of electronic documents they are tempted to copy the required content from these documents without properly citing its original owner. Therefore it is necessary for all the concerned persons to avoid and detect the plagiarism in the submitted work[14]. Plagiarism detection in text documents is an important field in information processing.

Plagiarism detection consists of searching of similar and more identical text between the documents[18]. It is a very complex task because most of the plagiarists will reuse the text from other source documents with aim of covering plagiarism by replacing words with synonyms, or by reordering the sentences[16]. There are many plagiarism detection methods that incorporate Natural Language Processing (NLP) techniques in their detection. These NLP techniques are applied to process the set of documents and also analysis the structure of the documents[17]. Plagiarism can involve changing the grammar, replacing the words with their synonyms, reordering sentences etc. In this case incorporating NLP techniques will be better than the other sophisticated methods. According to[15], applying NLP techniques for plagiarism could yield better accuracies through the detection of paraphrased texts. This paper mainly focuses on applying any new NLP technique such as Semantic Role Labeling for plagiarism detection could yield any better accuracy. And also focuses on the application of sentence ranking for reducing time of checking for plagiarism.

This paper proposed an improved method for plagiarism detection based on SRL by using sentence ranking for reducing the time of checking. The proposed method can detect near copy, synonym replacement, reordering the sentence and active or passive voice conversion. The rest of the paper is organized as follows: Section 2 details on the related work in plagiarism detection. Section 3 describes the architecture of our proposed system and also details about the various phases involved in the system. Section 4 gives a detailed explanation on the experimental setup and also presents the results that we have obtained. Section 5 concludes the paper.

2. Related Works

There are many plagiarism detection methods are available. Some of the plagiarism detection methods incorporate natural language processing techniques for plagiarism detection. These NLP techniques are applied to process the set of documents and also analysis the structure of the document[17]. Plagiarism can involve changing the grammar, replacing the words with their synonyms, reordering sentences etc. In this case incorporating NLP techniques will be better than the other sophisticated methods. There are many application areas of NLP such as part-of-speech tagging, morphological analysis, word sense disambiguation, anaphora resolution, co-reference resolution and discourse processing that help plagiarism detection[11]. According to[15], applying NLP techniques for plagiarism could yield better accuracies through the detection of paraphrased texts. In all plagiarism detection systems, pre-processing and candidate filtering are essential tasks. Ceska and Fox[12] showed that applying some pre-processing techniques can improve the accuracy of plagiarism detection. These include tokenization, stop-word removal, lemmatization, transforming numbers into dummy symbol and transforming all synonyms onto a unique word.

Most of the people hide plagiarism by replacing words with their synonyms. This makes most of the plagiarism detection methods fail. For synonymy recognition, they present three solutions which exploits WordNet thesaurus. WordNet groups' nouns, verbs, adjectives and adverbs into sets of cognitive synonyms called synsets[11]. WordNet synset is mapped into an interlingual index (ILI) which acts as a unique identifier. Therefore the algorithm searches for an equivalent word in WordNet and if a match is found, then the corresponding ILI is return. The second solution is based on a Naive Bayes classifier. This classifier selects the best matching word with respect to the adjacent words. The third solution is word generalization in which it replaces various words by a more general specific word. They also showed that these various pre-processing techniques have different effects in the process of plagiarism detection, some improves accuracy and some decrease time requirements. They also showed that applying various combinations of these preprocessing techniques allows gaining the benefit of each one.

Deeper NLP techniques are used to investigate the structure of texts rather thann their superficial information[8]. For example, Mozgovoy et al.(2007)[9] suggested parse trees to find the structural relations between documents. In their previous work, they proposed an approach technique that improves existing natural language-oriented plagiarism detection software for Russian languages. The techniques include tokenization, generalization of words into their hierarchical classes such as substituting the word fox with animal , and extraction of functional words and argumentative words for matching. After tokenization system firstly creates a suffix array from this tokenized collection of files. A suffix array is a lexicographically sorted array of all suffixes of a given string which allows to quickly finding a file, containing any given substring[10].

Chow and Salim[2] introduced a plagiarism detection system for detection both cross language and semantics plagiarism. They use Bahasa Melayu as the input language of the submitted document. The proposed method works by generating predictes of original and suspected documents using Stanford parser. Then calculate the degree of similarity between these predicates using Wordnet thesaurus. A similar semantics based method using SRL was introduced in[1]. The method transforms the suspected and original document into arguments depending on the location of each term in the sentence using SRL. Then proposed method compares and analyses the arguments of suspected sentences with the similar arguments of original sentences.

3. Proposed method

This paper proposes a framework for plagiarism detection which is very reliable and takes less time for reporting plagiarism in text documents. The proposed method uses Semantic Role Labeling for determining the semantic roles of each constitute term of a sentence based on its verb or also called predicate. This is determined by understanding the semantic meaning of the term occurring in that sentence. It is a sentence level semantic parsing or also called shallow parsing of the sentence which determine the object and subject of a sentence. It depends on the delineation of cases that determines how: “who” did “what” to “whom” at “ when” and “where”. Therefore it becomes clear that main objective of SRL is to determine the semantic roles of each term based on the semantic relationship between their predicates and terms[1]. The method also uses sentence ranking method for enhancing the plagiarism detection in which it retrieves source and suspicious sentence pairs.

The proposed method has five main steps, which are:

I. Pre-processing
II. Candidate Retrieval
III. Sentence Ranking
IV. Semantic Role Labeling
V. Similarity Detection

3.1. Pre-processing

Pre-processing is the first step in the plagiarism detection method which is one of the key step in Natural Language Processing. This step comprised two sub-steps, which were text segmentation and stop word removal. Text segmentation is the simplest type of pre-processing step. This pre-processing step segment the text into meaningful units like sentences or words. Here we choose for sentence segmentation in which the document is segmented into sentences for line-by-line processing Then we segment these sentences into words or tokens for further processing. In the process of stop word removal, some of the English words that are most frequently used does not contribute any meaning to the content. Removal of such words can improve accuracy and time requirements for comparisons by saving memory space and thus by increasing the speed of processing. For example, functional words such as articles, pronouns prepositions, and determiners such as *the, and,* and *a.*

3.2. Candidate Retrieval

The aim of this process is to identify a subset of source documents from a document collection (D_S) for a

suspicious document. This process includes the exhaustive comparison of the suspicious document with many number of source documents to identify any local similarities. This step is very important in the process of plagiarism detection because any source documents missed will no further examined in the next remaining stages. The care should also be taken when determining candidate documents which do not really become source document. Therefore the big challenge is to provide an algorithm for determining the candidate documents with high recall and low precision.

Here we are using n-gram and Jaccard coefficient similarity. The value of *n* is typically 2, 3 or 4. The suspected document and original document is transferred into a set of n-grams. The text comparison is performed by considering the amount of common n-grams between the documents. By performing Jaccard similarity coefficient between two documents *A* and *B* to find out the common n-grams. The Jaccard similarity coefficient equation is stated as follows:

$$J(A,B) = \frac{|S(A) \bigcap S(B)|}{|S(A) \bigcup S(B)|} \quad (1)$$

where S(A) and S(B) represents the set of trigrams in suspected and original documents. A threshold value is set for Jaccard coefficient in order to take a subset of candidate document from a collection of source documents which are likely sources of plagiarism regarding suspicious document.

3.3. Sentence Ranking

Ranking is a central part of many information retrieval problems, such as document retrieval, text summarization, sentiment analysis etc...In our method, sentence ranking is used to rank sentences in the suspicious and original document to retrieve original and suspicious sentence pairs. The process is based on cosine similarity between the sentences. This process gives a set of source and suspicious sentence pairs for performing the remaining steps in order to avoid the processing of unwanted sentences, thus reducing the execution time. In this we discuss the idea of sentence ranking. We first maintain a dataset that contains 'n' unique terms from the original file. Then we convert the suspected and original sentences into n vectors (t_1, t_2 ,… t_n) where t_i has a value 1 if term *i* is present in the dataset otherwise it has value 0. Then we perform a vector matching approach between each sentence of suspected document with the sentences of original document. For this we use a cosine similarity measure to measure cosine angle between two sentences. The valueof cosine angle rangesfrom 0 to 1 and the closer the angle is to 1.0 the higher the similarity between the query and sentence vector. Thus we obtain a set of cosine similarity values for each suspected sentences with the sentences in the original document. From this set we take the sentence pairs having maximum similarity value among others.

3.4. Semantic Role Labeling

Semantic Role Labeling (SRL) also called shallow semantic parsing, which is one of the Natural Language Processing technique. This process detects the semantic relationship associated betweenthe verb or predicate of a sentence and its constituent terms. SRL consists of identifying semantic argumentsassociated with a verb in a sentence and their classification into different roles. For example, consider the sentence *"Rachel mentioned Kate*", the verb of this sentence is *mentioned* , *Rachel* is the speaker (subject) and *Kate* is the patient (object). Semantic Role Labeling is based on Fillmore's frame semantics and it adds a layer of semantic roles to the syntactic trees of thePenn Treebank. This common NLP task find its application in question answering (Q&A) systems, machine translation, text mining.

The objective of SRL is to identify and label the semantic roles of each term in a sentence. Therefore it is a sentence level semantic analysis of text, which determines the object and subject of a sentence for identifying the semantic roles of each term. It depends on the delineation of cases that determines how: "who" did "what" to "whom" at " when" and "where". Therefore it becomes clear that, the primary task of SRL is to determine the

semantic roles of each term based on the semantic relationship between their predicates and terms[1]. During the role labeling process, each word in the sentence is labeled with their corresponding roles depending on their position in the sentence. The typical labels used in SRL are Agent, Patient, Time and Location for the entities participating in the sentence [1].

Plagiarism detection using SRL aims to detect the semantic similarity between the ranked sentences. This process proceeds through 3 main steps. First pre-processing the suspected and original documents using sentence segmentation and stopword removal. In the second step, using SRL the sentences are transformed into arguments depending on the position of each term in the sentence. In the third step, these extracted arguments are grouped into a node of similar argument type. Then the comparison is made between these suspected and original argument label group. There are many tools are available for performing semantic role labelling. SENNA is such an example.

3.5. Similarity Detection

In this stage, sentence-based similarity analyses between ranked suspected and original sentences were performed. Sentences in suspected documents were compared with each sentence in the candidate documents according to the arguments of the sentences. Here we detect not only the arrangement similarity between sentences, but also possible semantic similarity between two sentences. For this we use Wordnet taxonomy as a core tool for the calculation of similarity values. Wordnet Taxonomy returns a path similarity score denoting how similar two words are depending on the shortest path between these two words in the taxonomy. This score ranges from 0 to1. For example consider the sentences given below:

Tom painted the entire house (Original Sentence)
The entire house was painted by Tom (Suspected Sentence)

Figure.1 and 2 illustrate the analysis for suspected sentence and original sentence using SRL in the example given above.

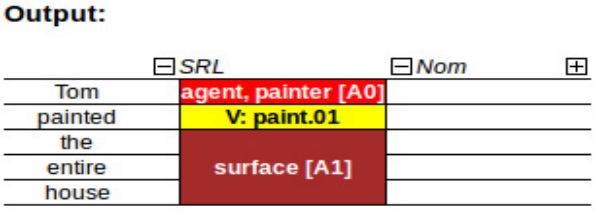

Fig. 1. Analysis for original sentence using SRL.

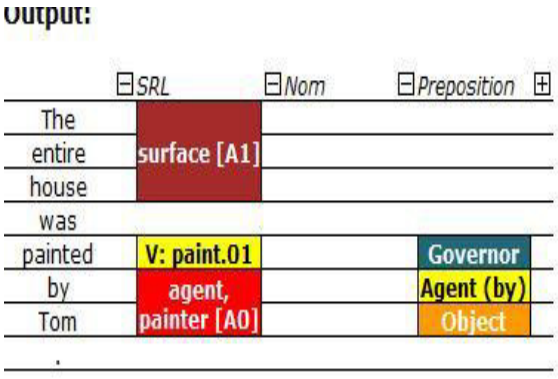

Fig. 2. Analysis for suspected sentence using SRL

plagiarism detection if comparison is applied based on the arguments of the sentence using SRL. We calculate the overall similarity between the original and suspected documents using the below equation:

$$Total\ Similarity\ (Doc1, Doc2) = \frac{\sum_0^n Sim(Args, Argo)}{n} \quad (2)$$

where $Sim(Arg_S ; Arg_O)$ gives the similarity between the arguments of the suspected document and original document was calculated using path similarity score return from the Wordnet taxonomy and 'n' is the total number of arguments in the suspected document.

4. Experimental Design and Results

The technique was tested on 100 documents .The 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). The experiments were performed on these 100 suspicious files each plagiarized from one or more original documents according to the Webis-CPC-11 corpus. In this experiment we looked only the amount of detected plagiarized sentences from the original documents.For this, sentence based similarity analysis between the ranked suspected and original documents were performed. Suspicious and original sentence pairs obtained from the sentence ranking were compared according to their arguments.

To evaluate the effectiveness of the proposed system, we use two metrics recall and precision. These are two general testing parameters that are commonly used in plagiarism detection. They are recall and precision.

$$\mathrm{Re}call = \frac{number\ of\ plagiarized\ paragraphs\ \mathrm{det}ected}{total\ number\ of\ plagiarized\ paragraphs} \quad (3)$$

$$\mathrm{Pr}ecision = \frac{number\ of\ plagiarized\ paragraphs}{number\ of\ plagiarized\ paragraphs\ \mathrm{det}ected} \quad (4)$$

Recall is defined as the percentage of paragraphs identified as plagiarized with respect to the total number of actual plagiarized paragraphs between two documents[11]. Precision is defined as the percentage of plagiarized paragraphs identified with respect to the total number of identified paragraphs [11].

The proposed method is evaluated and compared with existing SRL based method on the basis of their execution time and detection accuracy. The results are shown below.

Table 1. Performance evaluation of proposed method

	Recall	Precision	Execution time
SRL-based method	.89	.85	Takes more time
SRL with sentence ranking	.89	.90	Takes time less than SRL

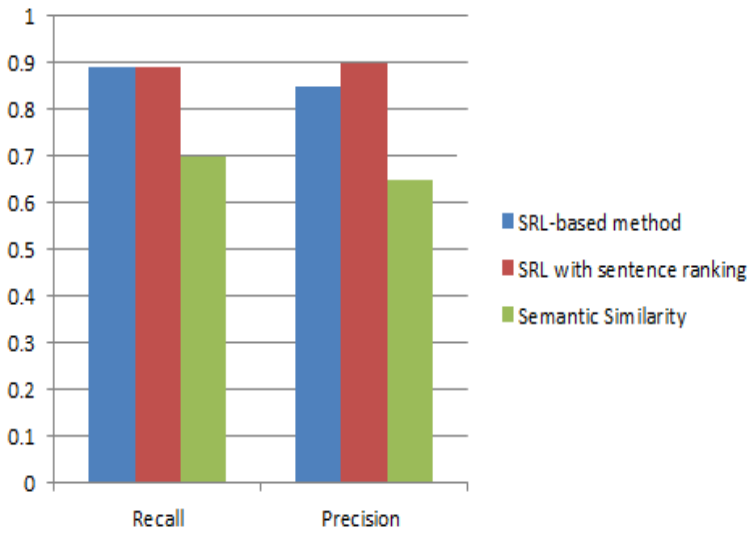

Fig. 3. Comparison results with plagiarism detection techniques

Figure gives the comparison between the proposed method with SRL-based similarity[1] and Semantics-based similarity[2]. From the obtained results itself it is clear that the proposed method yields better results than the other methods. Also it become clear that proposed method reduces the execution time for checking plagiarism.

5. Conclusions

This paper presented an improved SRL-based plagiarism detection method with sentence ranking. This system was found to achieve better performance than SRL-based similarity and Semantics-based similarity. The proposed method detects copy paste, synonym replacement and active or passive voice conversion in less execution time and with better accuracy.

References

1. A.H. Osman, N. Salim M.S An improved plagiarism detection scheme based on semantic role labelling, in Journal of Applied Soft Computing Elsevier vol:12, p. 1493-1502, 2012.
2. Chow Kok Kent and N. Salim, Web Based Cross Language Plagiarism Detection, *Second International Conference onComputational Intelligence, Modelling and Simulation,* p. 199-204, 2010.
3. Naomie Salim, Ahmed Hamza Osman, Plagiarism Detection Scheme Based on Semantic Role Labeling, International conference march 2012.
4. C. J. Fillmore, The case for case. In Emmon Bach and Robert T, *Universals in Linguistic Theory. Holt, Rinehart, and Winston, NewYork,* p. 1–210, 1968.
5. 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.
6. M. Palmer, *et al.*, The Proposition Bank: An Annotated Corpus of Semantic Roles, *Comput. Linguist.,* vol. 31, p. 71-106, 2005.
7. A. Si, H. V Leong, and R. W. Lau. CHECK: A Document Plagiarism Detection System. In Proceedings of ACM Symposium for Applied Computing, pages 70-77, February 1997.
8. Chong, B. M., Specia, L., & Mitkov, R. (2010).A Study on Plagiarism Detection and Plagiarism Direction Identification Using Natural Language Processing Techniques. In Proceedings of the 4th international plagiarism conference. Newcastleupon- Tyne, UK.
9. Maxim Mozgovoy, Tuomo Kakkonen, and Erkki Sutinen. Using natural language parsers in plagiarism detection. In Proceedings of the Workshop on Spoken.
10. Maxim Mozgovoy, Vitaly Tusov, and Vitaly Klyuev. The Use of Machine Semantic Analysis in Plagiarism Detection. In Proceedings of the 9th International Conference on Humans and Computers, pages 72-77, Aizu-Wakamatsu, Japan, 2006.
11. Chien-Ying, C., Jen-Yuan, Y., & Hao-Ren, K. (2010). Plagiarism detection using ROUGE and WordNet. Journal of Computing, 2(3), 34-44.
12. Ceska, Z., & Fox, C. (2009). The Influence of Text Preprocessing on Plagiarism Detection. In Recent Advance in Natural Language

Processing, RANLP ’09.
13. Wadsworth, available at: http:// wadsworth.cengage.com /english_d/special_features/plagiarism/definition.html.
14. Asim M. El Tahir Ali, Hussam M. Dahwa Abdulla, Vaclav Snasel Survey of Plagiarism Detection Methods IEEE 2011 39-42.
15. Clough, P. (2003). Old and new challenges in automatic plagiarism detection. National Plagiarism Advisory Service, 391-407.
16. Fernando Sanchez-Veg, Esau Villatoro-Tello, Manuel Montesy- Gomez, Luis Villasenor-Pineda, Paolo Rosso. Determining and characterizing the reused text for plagiarism detection, Expert Systems with Applications 40 (2013) 1804-1813 ELSEVIER.
17. Chong, B. M., Specia, L., & Mitkov, R. (2010).Using natural language processing for automatic detection of plagiarism. In Proceedings of the 4th international plagiarism conference. Newcastle-upon-Tyne, UK.
18. Mein Paul & Sangeetha Jamal A Survey on Plagiarism Detection in Text Documents in Natioanl Conference on Communications and Computing (NCCC), March 2014, GEC Idukki.
19. SRL Demo. Retrive June 10, 2011, from http://cogcomp.cs.illinois.edu/demo/srl