International Journal of Management, IT & Engineering
(ISSN: 2249-0558)

# CONTENTS

# Chief Patron



# Patron



# Editorial Board

# Editorial Advisory Board

**Mr. P. PRABHU**
Assistant Professor, Alagappa University, KARAIKUDI

**Mr. MANISH KUMAR**
Assistant Professor, DBIT, Deptt. Of MBA, DEHRADUN

**Mrs. BABITA VERMA**
Assistant Professor, Bhilai Institute Of Technology, DURG

**Ms. MONIKA BHATNAGAR**
Assistant Professor, Technocrat Institute of Technology, BHOPAL

**Ms. SUPRIYA RAHEJA**
Assistant Professor, CSE Department of ITM University, GURGAON

Title

# IDENTIFICATION OF PARAPHRASING IN THE CONTEXT OF PLAGIARISM

Author(s)

**Nidhi Kushwaha**
*C-DAC, NOIDA, INDIA*

**Deepak Kumar**
*C-DAC, NOIDA, INDIA*

**Dr. P. R. Gupta**
*C-DAC, NOIDA, INDIA*

## Abstract:

Paraphrasing is a very important form of processing for natural language processing (NLP). A characteristic property of natural language is that various expressions can exist to express a single concept. The aim of this project is to generate and identify the paraphrase which is type of plagiarism. For this firstly pre-processing of the text is necessary after that with the knowledge of their synonyms & related words generation of multiple text Natural language Generation with the help of WordNet (online dictionary) and SimpleNLG ( NLG) is necessary for generating the paraphrase texts from a single text. That means we have an original text and it's possible paraphrase texts to check the submitted paragraph is semantically similar or not.


**Keywords:** Paraphrasing, Plagiarism, WordNet, SimpleNLG**.**

## Introduction:

Plagiarism is the "theft of intellectual property" that has been around for as long as humans have produced work of art and research. Easy access to the Web, large databases, and telecommunication in general, has turned plagiarism into a serious problem for publishers, researchers and educational institutions. [1]

Plagiarize is to steal and pass off (the ideas or words of another) as one's own; [to] use (another's production) without crediting the source; [or] to commit literary theft [by] presenting as new and original an idea or product derived from an existing source.[2]

The Plagiarism that is written advisedly can only be recognized as such, if the reader knows the original work. However human intuition can suggest that a work has been copied, because of abnormal language use, stylistic Changes and grammatical errors. The conviction of the plagiarist, however, is much more difficult, because the intellectual search for a passage in a journal or a book is extremely time consuming and makes great demands on the seeker. Even if we concentrate on specialized databases plagiarism examination can be compared to looking for a needle in a haystack.

A Dictionary definition of a paraphrase is "a restatement of text, passage, or work" and giving the meaning in another form. Paraphrasing is most often done legitimately by news editors when reusing text from press agency sources. A paraphrase typically explains or clarifies the text that is being paraphrased. For example, "The signal was red" might be paraphrased as "The train was not allowed to proceed."[3]Paraphrase may attempt to preserve the essential meaning of the material being paraphrased. A Paraphrase is an alternative surface from in the same language expressing the same semantic content as the original form Paraphrase may occur at several levels. Individual lexical items having the same meaning are usually referred to as lexical paraphrase or, more commonly, synonyms, for example,<hot , warm> and <eat , consume>. Two sentences that represent the same semantic content are termed sentential paraphrase, for example, I finished my work, and I completed my assignment. [4]It is difficult to generate sometimes when the two sentences are completely different even if they are semantically related to each other.

Paraphrase is act as crime if one use the others idea without giving proper credit to its originator. Main aim of this project is to investigate the techniques for detecting paraphrase pairs. One way of doing this would be to use WordNet (Miller 1995) [5], SimpleNLG (Albert Gatt and Ehud Reiter 2009) to create paraphrase paragraphs manually. This has the advantage that we know that these really are paraphrases and we use these for checking similarity percentage with the newly submitted paragraph. This project is divided into two parts firstly to generate paraphrase text & then identify the paraphrasing in text. Paraphrase generation means to generate semantically equivalent paragraphs. To check paraphrased paragraph means that generate similarity score & obtained result the paragraph is paraphrased or not. Paraphrase recognition is basically defined as a task in which we try to determine if two sentences express the same content. Despite the lexical & syntactic transformations, determine the semantic equivalence of two sentences comes under it. The relevance of both the task in natural language processing is well known by a wide diversity of applications in which is used. Paraphrase recognition is applied in applications such as text summarization systems in which combination & reduction of sentences as well as lexical & syntactic transformation is essential. Information retrieval systems in which to make use of diverse ways to express information requests is very useful since search engine can be able to show much more information for the user. Text-to-text generation task is heavily relying on paraphrase detection. Paraphrase identification has been previously explored, by (Kozareva &

Montoyo 2006), (Mihalcea, Corley, & Strapparava 2006) [10] by knowledge-based similarity measures. The paraphrase relation between two texts is related to the entailment relation. Textual entailment is the task of deciding, given two text fragments, whether the meaning of one is entailed (can be inferred) from another (Dagan,Glickman, & Magnini 2005)[11]. Other approach which is based on the previous proposed approach uses minimal knowledge resources, as compared to other approach of entailment. The entailment approach (Rus et al. 2008) only uses lexical, syntactic, synonymy and antonymy information. The synonymy and antonymy information is extracted from a thesaurus, i.e. WordNet (Miller 1995), an online dictionary. Other enhanced approach by (Vasile Rus,M.McCarthy,Mihai C. Lintean 2008)[12] incorporates negation handling based on anatonymy relations in WordNet. Our approach is different from this approach in all these approach Microsoft Reasearch Paraphrase Corpus(Dolan, Quirk, & Brockett 2004) is used for the evaluation of the result, but in our approach firstly multiple text is generated with the help of WordNet & SimpleNLG , then the result is evaluated through the use of this generated texts.

## WordNet-An Online Dictionary:

WordNet is a semantic English dictionary as a traditional dictionary. It provides rich semantic relations of words including synonyms, antonym, and so on with which words are linked together to form a network. Using that semantic analysis to map user inputted words to domain terms helps increase recognition rate. WordNet is general lexical database which has been developed in the Princeton University for more than 10 years and can be used as a basic lexicon with which a domain lexicon is build for a specific application area. This not only gets quality domain lexicon but also reduces efforts of building a lexicon.

It has provided a very good vocabulary organization structure with which it is very easy to construct and expand a domain lexicon. Controlled natural language (CNL) is a subset of natural language, gained by restricting the grammar and vocabulary in order to reduce or eliminate ambiguity and complexity with knowledge of linguistic, logic, knowledge category theory, psychology, and information theory. The proposed WordNet approach enhances usability of CNL. WordNet is organized by semantic relations. In our proposed system this WordNet is very useful for generating multiple synonyms from a given word and provide an output to the

SimpleNLG, which is a Natural Language Generator. Tokenized words are the input of the WordNet and synonyms are the output of the WordNet which is the input of SimpleNLG.

## SimpleNLG – A Natural language Generator:

SimpleNLG is a realisation engine which grew out of recent experiences in building large-scale data-to-text NLG systems, whose goal is to summarise large volumes of numeric and symbolic data (Reiter 2007). Sublanguage requirements and efficiency are important considerations in such systems. Although meeting these requirements was the initial motivation behind SimpleNLG, it has since been developed into an engine with significant coverage of English syntax and morphology, while at the same time providing a simple API that offers users direct programmatic control over the realisation process. SimpleNLG is a java library that provides interfaces offering direct control over the realisation process, that is, over the way phrases are built and combined, inflectional morphological operations, and linearization. It defines a set of lexical and phrasal types; [7] corresponding to the major grammatical categories, as well as ways of combining these and setting various feature values. In this project SimpleNLG[7] is useful in the way the words are organized in the sentence.

## APPROACH & IMPLEMENTATION:

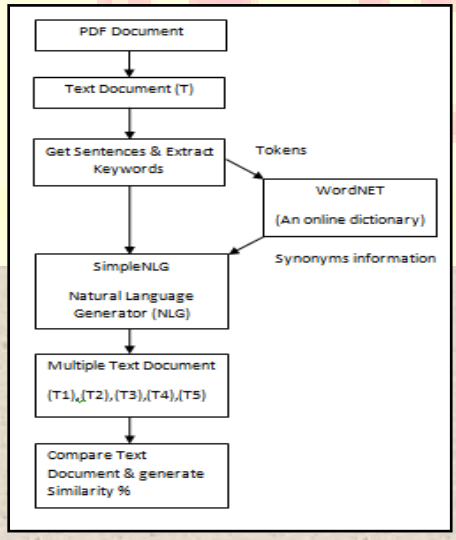


**Fig.1 Proposed Approach**

As one can see in the Fig.1, Input of this project is a research paper i.e. a PDF document. This PDF document must be converted into text document so that pre-processing of the data can be done on it. Pre-processing means prepare data for the further processing by removing stop words like-be, is, I etc. After pre-processing text document is tokenized. Tokenized words are then converted into their root words with the help of RitaWordNet[9]. These tokenized words are the input of the WordNet. WordNet also provide Part of Speech information of each word. WordNet gives synonyms information that in future works as the input of the Natural language Generator. There are some words that have no synonym information present in the WordNet, and then in that case these words are act as their original form for the input of the SimpleNLG[13]. Using this information SimpleNLG generates multiple text documents that are not same syntactically but similar to each other in semantics. Now we have one original text and its paraphrase texts, in this project only five different texts is generated. If we have to check a research paper is paraphrase or not then we have to check the similarity percentage between the folder that have the six documents (one original and five paraphrase texts) and the research paper that should be checked. Similarity percentage below 30 is not a paraphrase document.

## RESULT:

To measure the semantic similarity of the text similarity tool [8] is used. It compares the text to text, text & folder & generates the similarity percentage. Output of the text comparison is given in the Fig.2. It describes the processing time, Wordnet Lookup time etc., which will provide us information for performance calculation. Extracting the Abstract & Conclusion information can be give more accurate result in this similarity checking because both of this are different in different research papers.

In Fig.3 one can see that total similar paper gives 100% similarity score, Partial similar paper gives 91.42% similarity score and 0.15% result is for totally different similarity result.

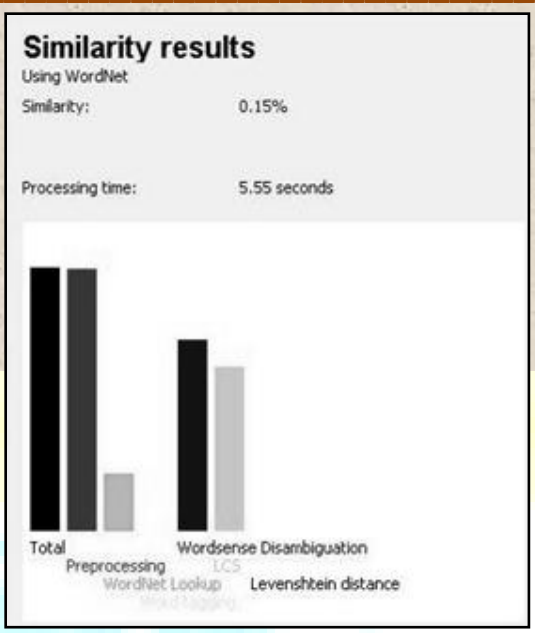


**Fig.2 Similarity score**

| Types of Files | Source File | Matched File | Similarity Check | % Match |
|---|---|---|---|---|
| Same Files | Abstract1.txt<br>Conclusion1.txt | Abstract1.txt<br>Conclusion1.txt | Complete Similarity | 100% |
| Similar Files | Abstract1.txt<br>Conclusion1.txt | Abstract1_syn.txt<br>Conclusion1_syn.txt | Mostly Similar | 91.42% |
| Different Files | Abstract1.txt<br>Conclusion1.txt | Abstract2.txt<br>Conclusion2.txt | No Similarity | 0.15% |

**Fig.3 Result analysis**

## CONCLUSION & FUTURE WORK:

The increasing volume of information available on-line to students, manually detecting plagiarism is becoming harder. These will emphasise the need of an automatic technique to detect similarities among paraphrase documents. To match the documents we firstly needs similar documents by which we can match them and able to say that they are paraphrased of original document. Paraphrased detection is not an easy task to detect. Various researches are in progress to make best-paraphrased detector. As one can say one sentence in various types they

all are may not look it but their semantically equivalent to each other. In this project I track this idea to make semantically equivalent sentences from the knowledge of their synsets which is fetch from the online dictionary, organized in hierarchal form, so that it can easy to fetch the information of the synsets of the words found in the particular sentence.

Also, with the use of SimpleNLG it is easy to generate paragraph that are semantically related to each other. Extraction of Abstract & Conclusion makes is more time efficient to generate text only that paragraph that comes under the headings respectively.

Matching similarity between the original & generated texts gives 90% score that means they are related to each other, the documents that are gives less than 50% similarity results are not related to each other. Before doing these much process the text documents should go through the pre-processing stage in which Stop word removal, Morphology of words is taken under consideration. Tokenized words & information of their synonyms and SimpleNLG is responsible for multiple paragraph generation. SimpleNLG which can be efficiently used as a NLG is not much processed in this project, because it can work efficiently only with the simple sentences.

In future other NLG can be used to generate text in more semantically processed form and it may give more accurate result to find paraphrased documents.

**REFERNCES:**


- Hermann Maurer, Frank Kappe, Bilal Zaka, **"Plagiarism-A Survey"**, Institute for Information Systems and Computer Media Graz University of Technology, Austria.
- O¨ zlem Uzuner, Boris Katz, and Thade Nahnsen, **"Using Syntactic Information to Identify Plagiarism",** Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory, Cambridge,2005.
- http://en.wikipedia.org/wiki/Paraphrase.
- Nitin Madnani, Bonnie J. Dorr, **"Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods"**, *University of Maryland, College Park, 16 December 2008.*

- Ho Li, Yong Shi, **"A WordNet Lexicon Model for Controlled Natural Language"**, *Institute of pattern recognition & Artificial Intelligence, Huazhong University of Science & Technology, Wuhan, China.*
- George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller, "**Introduction to WordNet: An On-line Lexical Database"**, *August 1993.*
- Albert Gatt and Ehud Reiter, **"SimpleNLG: A realisation engine for practical applications"**, *Department of Computing Science University of Aberdeen, March 2009.*
- Andreas Schmidt Jensen & Niklas Skamriis Boss, **"Textual Similarity: Comparing texts in order to discover how closely they discuss the same topics"**, *Kongens Lyngby 2008 IMM-BSc-2008-15.*
- http://www.rednoise.org/rita/wordnet/documentation/index.htm
- Mihalcea, R.; Corley, C.; and Strapparava, C. 2006. **"Corpus-based and knowledge-based measures of text semantic similarity",** in *Proceedings of the American Association for Artificial Intelligence (AAAI 2006).*
- Dagan, Gilckman and Magnani,B. 2005. **"The pascal recognising textual entailment challenge. In Proceedings of the recognising textual entailment."** *in Proceedings of the recognizing Textual Entailment Challenge Workshop.*
- McCarthy, Rus, Crossley, Bighman, and McNamara, D. 2007. **"Assessing entailer with a corpus of natural language."** in *proceedings of the Florida Artificial intelligence Research Society International Conference (FLAIRS). Menlo Park, CA: AAAI Press.*
- http://www.csd.abdn.ac.uk/research/simplnlg/doSimpleNlg V3.7 tutorial.