Paraphrase Detection using Semantic Relatedness based on Synset Shortest Path in WordNet

Jun Choi Lee
School of Computer Sciences
Universiti Sains Malaysia
11800, USM Penang, Malaysia
leejunchoi@gmail.com

Yu-N Cheah
School of Computer Sciences
Universiti Sains Malaysia
11800, USM Penang, Malaysia
yncheah@usm.my

***Abstract*— This study presents a sentence-to-sentence semantic relatedness measures for paraphrase detection. The proposed measures adopt the shortest path between synsets in WordNet as the core to measure the relatedness between two sentences. The interlinked synsets in WordNet are based on the conceptual-semantic relation between two synsets. Thus the distance between two synsets in WordNet can be used to measure the semantic relatedness between two synsets. This study derived a sentence-level semantic relatedness using this feature to detect paraphrasing among sentences. The performance of the proposed semantic relatedness in paraphrasing is evaluated based on the accuracy and F-measures of the proposed measures in identifying paraphrase in Microsoft Research Paraphrase Corpus. The proposed method achieved 71.1% in accuracy and 81.8% in F-measures. The performance of the proposed method is compared with 6 paraphrase detection methods which include Salient Semantic Analysis and Second-order Co-occurrence Pointwise Mutual Information. In the comparison, the proposed method achieved the fourth highest accuracy and the second highest F-measure compare to other methods. This is a reasonable performance for the proposed semantic relatedness in paraphrase detection.**

Keywords*— paraphrase detection, semantic relatedness, WordNet, synset shortest path.

I. INTRODUCTION

Paraphrase detection or paraphrase identification is the task to identify sentences with similar meaning by evaluating the similarity between two texts based on lexical and structural similarity. This process also serves as part of the computational evolution in Text Understanding studies.

Semantic relatedness is the measurement for the degree of relationships between two entities. This particular measurement posted significant differences compare to semantic similarity because semantic similarity identifies entities or texts that have similarity in terms of lexical or meaning, while semantic relatedness tries to identify how close two entities or texts are related to one another. The differences between Semantic Relatedness and Semantic Similarity were discussed by [7], [10], and [13].

WordNet is an electronic lexical database. It is formed by sets of synonyms which known as synset. Synsets in WordNet are interlinked with each other through conceptual-semantic and lexical relations. One of the approach to identify the similarity between two synsets is by measuring the path between two synsets. [5], [6].

This study presents a semantic relatedness measure that based on Synset Shortest Path in WordNet for paraphrase detection. The proposed method is then evaluated using a paraphrase detection evaluation based on the Microsoft Research Paraphrase (MRP) [3] Corpus. In the evaluation, the proposed semantic relatedness is compared with other paraphrase detection approaches.

This paper describes the natural of the text similarity, semantic relatedness and paraphrase detection in Section II. After that, Section III elaborates in details the proposed semantic relatedness approach and how it is used to identify paraphrases. The evaluation setup and process to evaluate the proposed semantic relatedness are explained in Section IV. While the evaluation results are shown and discussed in Section V. Finally the final conclusion and future works are discussed in the Conclusion Section.

II. RELATED WORKS

The early paraphrase detection approaches were based on simple lexical matching algorithms [11], [16]. These approaches are simple and easy to implement, however, the performance highly depended on the similarity of lexical and structure of the compared text. To overcome these limitations [7] employed knowledge-based similarity measures from WordNet to expand the detection beyond the limitation of lexical and structure similarity in the compared text. [12] applied bipartite graph and Term Frequency – Inverse Document Frequency (TF-IDF) in the efforts to compare and identify paraphrases. Besides graph based paraphrase detection approaches. [2] applied Natural Language Processing (NLP) to extracts numerous text features such as nouns, stemmed words from compared text to help in determining paraphrases. The approach does not yield high accuracy in the evaluation, but it provides a significant higher F-measure, which means the approach performed well in identifying positive cases in the evaluation.

Other paraphrase detection studies use Information Retrieval (IR) approaches such as Second Order Co-occurrence Pointwise Mutual Information (SOCPMI) [1] and Latent

978-1-5090-1636-5/16/$31.00 ©2016 IEEE