Paraphrase Detection using String Similarity with Synonyms

Lee JunChoi † **Cheah Yu-N‡**

†Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak, Kota Samarahan, Sarawak, Malaysia
†‡ School of Computer Science
Universiti Sains Malaysia, Penang, Malaysia
†jclee@fit.unimas.my
‡yncheah@cs.usm.my

Abstract

This paper presents an approach to enhance text similarity using synonyms for paraphrase detection. Paraphrase detection detects sentences or texts with similar meaning. Synonym of words can help in paraphrase detection. However, considering synonyms for all the text in the comparison posted extra computational task in the process. This study introduces a simpler approach in considering synonyms in text similarity. The proposed approach is able to be adopted in any term-based text similarity metrics for paraphrase detection. The proposed approach is evaluated using the Microsoft Research Paraphrase Corpus. The evaluation show better result compares to original similarity metrics and reasonable result compares to other selected paraphrase detection approaches in previous studies.

Keywords: paraphrases detection; string similarities; synonyms; cosine similarity; overlap coefficient; individual match ratio average.

1 Introduction

The effort to understand a text computationally started with simple lexical similarity. Paraphrase detection is another step after lexical similarity in text understanding, as paraphrase detection attempts to identify sentences with similar meaning. Paraphrase detection is not only a crucial process in text understanding but also in plagiarism detection.

Synonyms is word that shares same meaning. This linguistics feature also plays an important role in computational linguistics, as it allows computer to identify the meaning of text that presents in different variants. Therefore applying synonyms in text similarity calculation and paraphrase detection is one of the approach to improve these tasks.

This study present a novel approach in using synonyms to enhance the similarity metrics for paraphrase detection. The proposed approach provides a simpler way in considering synonyms in text similarity calculation. The study used paraphrase detection as the evaluation tools of the concepts as paraphrase detection requires the ability to recognise the text not only in lexical level, but also in context meaning.

This paper describe essential previous works for paraphrase detection in Section 2. Section 3 elaborates the proposed approach and how it being adopted in different term-based similarity metrics. A simple walk-through example is provided to demonstrate how the proposed paraphrase detection works in Section 4. The evaluation setup and process are explained in Section 5, while the result of the evaluation is presented in Section 6. Final Section of the paper is the conclusion and future work.

2 Previous Work

Early paraphrase detection started with simple lexical matching techniques [12], [13]. However, these approaches are limited to similarity of the lexical. The later study of paraphrase detection study utilise the similarity among words in the compare text to determine the similarity between the text. [9] computes the similarity of words using knowledge-based similarity measures that obtained from WordNet to detect paraphrases.

Later study also uses more advanced Information Retrieval (IR) approaches in detecting paraphrases. [6] uses Second Order Co-occurrence Pointwise Mutual Information