Semantic Relatedness Measure for Identifying Relevant Answers in Online Community Question Answering Services

JunChoi Lee[1,2]
Faculty of Computer Science and Information Technology[1]
Universiti Malaysia Sarawak
Kota Samarahan, Malaysia
jclee@fit.unimas.my

Yu-N Cheah
School of Computer Sciences[2]
Universiti Sains Malaysia
Penang, Malaysia
yncheah@usm.my

***Abstract*—This study introduces a new sentence-to-sentence semantic relatedness measure. The proposed measure optimized the word-to-word semantic relatedness that based on the depth of two concepts in WordNet. The study used Microsoft Research Paraphrases Corpus to validate the accuracy of the proposed method in identifying sentences with high semantic similarity. The result shows the proposed methods performed well compare to other unsupervised methods. At the end of the study, this paper also shows that the proposed semantic relatedness is able to identify relevant answers in Online Community Question Answering Services.**

Keywords—semantic relatedness, sentences semantics, answer quality, paraphrase detection, online community question answering

I. INTRODUCTION

This paper presents a semantic relatedness measure that optimised the word-to-word semantic relatedness measure. The proposed semantic relatedness measure is for identify relevant answers in Online Community Question Answering Services. Answers in Online Community Question Answering Services are responses generated by public user. Not all user generated responses are good answer to a posted question. Some of the responses were spam, social greetings or unrelated text. There are many methods to predict the quality of the responses.

Semantic Relatedness posts different definitions and functions compared to Semantic similarity, as semantic similarity is aims to identify entities that have same or similar structural or meaning, while semantic relatedness is defining the degree of relationships between two entities. This differentiation between semantic Relatedness and Semantic Similarity was discussed in [1], [2], and [3]. For example, if one were calculating the Semantic Similarity between "fish" and "eagle", the Semantic Similarity returned would be zero, as two terms / entities does not have any resemblance in turn of structure or meaning. However the Semantic Relatedness between two words will give a different measures. WUP measures in WORDNET for 'Fish' and 'Eagle' is 0.7826. This is because there exists a high degree of relationship between two terms / entities, where Eagle feeds on Fish (predator and prey relationship).

This study believes that a good answers for a question must have very high degree of semantic relationship between the question and answers. Therefore it is possible to use semantic relatedness to identify good and relevant answers in Online Community Question Answering Services.

This study investigates the truth behind this claims by observing the relationships between Semantic Relatedness and best answers in Online Community Question Answering Services. This is done through observing the distribution of the semantic relatedness between questions and their best answer collected from Online Community Question Answering Services. This study will serves a preliminary study in using semantic relatedness to evaluate answer quality in online community question answering services.

This study also introduced a method to evaluate the semantic relatedness between two texts. The proposed method are further evaluated compare to other existing semantic relatedness measures. The proposed semantic relatedness is based on the WUP measures in WORDNET.

II. RELATED WORK

Semantic relatedness for text can be divided into word-to-word semantic relatedness and sentence-to-sentence semantic relatedness. Word-to-word semantic relatedness is to measure the semantic relation between two words or two terms. The word-to-word semantic relatedness have two distinct categories; knowledge-based and corpus-based semantic relatedness. Knowledge-based semantic relatedness likes Wu & Palmer (WUP) [4], Lesk [5] and Resnik [6] uses structured lexical resource such as WordNet to compute the semantic relatedness measure. Corpus-based semantic relatedness such as Latent Semantic Analysis (LSA) [7], Explicit Semantic Analysis (ESA)[8] and Salient Semantic Analysis (SSA) [3]