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]

uses semantic model derived from collected corpus in computing semantic relatedness.

In sentence-to-sentence semantic relatedness measure, existing measures such as ESA and LSA aggregates the feature space terms in text in computing the semantic relatedness. Mihalcea et al. [9] represents texts in bipartite graph to calculate the semantic relatedness. The STS models proposed by Islam & Inkpen [10] adopts the second order co-occurrence point-wise mutual information (SOCPMI), a corpus-based word-to-word semantic relatedness to derive sentence-to-sentence semantic relatedness.

Jeon et al. [11] attempts to predict the quality of answers using 13-nontextual features in Maximum entropy learning model. They discovered that 1/3 of their dataset from Naver.com have some quality issues, and 10% of the total dataset were identified as bad answers. Jeon et al. [12] extended the study on price as factor for quality answers in fee-based Community Question Answering Services. Harper et al. [13] study the differences among quality answers in different Community Question Answering Services and observed how users accept the different quality criterion. Liu et al. [14] tried to predict the answer quality using non-contextual information. Besides Harper, Fichman [15] also studied the answer quality of different online Community Question Answering Services. Agichtein et al. [16] in the other hand, attempted the same thing using content features extracted from answers. Adamic et al. [17] explore the method to predict answer quality when question and answer are provided. Answer Quality prediction study done by Shah and Pomerantz [18] were based on information extracted from the question, answer and also user profile. Blooma et al. [19] also explores the methods to select best answers in Community Question Answering Services. Tian et al. [20] investigated different features from question and answers in the online Community Question Answering Service, stackoverflaw.com, to measure answer quality. Their study shows that the most significant factors in determining the answers quality were the amount of responses and the minimum similarities among the answers. However, it is noticed that most methods claim that non-contextual features such as author profile and length of answers provides better indicator of quality. This study still strongly believes that the best features to identify a good answer is still the relationship between the question and answer. Because only the relationship between question and answer can provides an independent perspective that only relevant to asker and the responder. That is the reason for this study to explore the use of semantic relatedness in identify relevant answers for a question.

III. Proposed Semantic Relatedness

This study proposed a method to measure text-to-text semantic relatedness. The proposed method adopts the use of word-to-word semantic relatedness in determine the text-to-text semantic relatedness.

To compute the semantic relatedness of two sentences, the proposed method received two sentences as input. The inputted sentences are then tokenised into word list *wordlist1* and *wordlist2*. A list of words, *matchlist* that exists in both *wordlist1* and *wordlist2* are compiled by comparing both of *wordlist1* and *wordlist2*. Besides that the words that existed in both *wordlist1* and *wordlist2* are removed from the list, so both *wordlist1* and *wordlist2* are now contents words that are independent entries. All three lists of words (*matchlist*, *wordlist1* and *wordlist2*) are the key elements for the next step to compute the semantic relatedness for the inputted sentences.

The next step is to identify the word-to-word semantic relatedness for the words in the inputted sentences. There are two parts in compiling the word-to-word semantic relatedness: the semantic relatedness for words that exist in both sentences and the semantic relatedness for words that only exists in one of the sentences.

After compiling the lists, the next step is to obtain the semantic relationships among the words in the sentences. In this study, Wu and Palmer (WUP) similarity measure based on WordNet resource is used to compile the word-to-word semantic relatedness. WUP measure considers the semantic relationship between two entities based on the depths of two concepts (s1, s2) using the Longest Common Substring (LCS) formula as shown in formula 1 below.

$$WUP_{(s1,s2)} = \frac{2 * (depth(LCS))}{depth(s1) + depth(s2)} \quad (1)$$

To compile the semantic relatedness for the words that exist in both inputted sentences is easy. Since each word is exists in both sentences, the semantic relatedness assigned to each word in the list should be the maximum figures as it is an exact match. For WUP measure, the maximum figure for two entities is 1.

Compiling the word-to-word semantic relatedness for the words that only exist in one of the sentences is a different process. First, a word is chosen from the *wordlist1*, and then the semantic relatedness for the chosen word and all the entries in *wordlist2* are compiled and compared. The figures with the highest semantic relatedness value are stored in the list. At the end of the process, a list of optimum semantic relatedness for the word exists in one of the sentences are compiled. This list of semantic relatedness will be used to compile the final sentence-to-sentence semantic relatedness.

In determine the text-to-text semantic relatedness, the proposed method uses average word-to-word semantic relatedness as the final value. The average word-to-word semantic relatedness is compile by finding the optimum word-to-word pair semantic relatedness measures for each word in the comparing sentences and then divide the summation of all the word-to-word semantic relatedness with the number of words in the comparing sentence.

IV. Evaluation

An evaluation was conducted to validate the performance of the proposed semantic relatedness measure using paraphrases detection. This evaluation was conducted by using the benchmark in paraphrasing detection, an evaluation conducted in evaluating other semantic relatedness measure such as SSA, ESA and LSA.

A. Dataset

The proposed approach for paraphrase detection was evaluated using the Microsoft Research Paraphrase Corpus (MRPC) [21]. The corpus consists a corpus with 4076 sentence pairs as training corpus and another corpus with 1725 sentence pairs as evaluation set for the study. The text was extracted from news source on the web. Each of the sentence pairs were judged by human annotators to determine the semantic equivalent of the sentences. The human annotators tasked to evaluate the sentences were linguistic consultant from an independent company, the Butler Hill Group. Two annotators were assigned to a pair of sentences to determine the semantic equivalence of the pairs. If exist disagreement between the annotators, a third annotator will provide the final judgement. The training corpus consists 2753 or 67.5% pair sentences which tagged as semantically equivalent, while the test corpus consists 1147 or 66.5% pair sentences are semantically equivalent.

In this study, the corpus with 1725 sentence pairs is used to evaluate the accuracy of the proposed method. This is because this corpus is used in other study to determine the accuracy of their methods in detecting paraphrase, while the larger corpus set is used as training corpus.

B. Threshold Value

A threshold value is required in determine if the sentences are paraphrase. A set of experiments is conducted using the training corpus in the MRP Corpus to determine the threshold value for each similarity measure in determining paraphrase.

In each of the experiment, a threshold value is chosen. For each pair of sentences in the training dataset, the semantic relatedness measure is calculated. If the similarity value is larger than the threshold value, the sentence pair is consider paraphrases. This decision is then compared with the pre-defined tag in the dataset to determine if the decision is correct. The correctness is stored for accuracy measurement later. The process is repeated by incrementing the threshold value from 0.1 to 0.9. The accuracy for each threshold value are calculated using the formula 2 below;

$$Accuracy = \frac{Correct}{Total\ Number\ of\ Comparison} \quad (2)$$

TABLE I. PROPOSED SEMANTIC RELATEDNESS MEASURE ON VARIOUS THRESHOLD VALUES

Threshold	Accuracy of the Proposed Semantic Relatedness Measure
0.1	67.54%
0.2	67.54%
0.3	67.57%
0.4	67.91%
0.5	69.46%
0.6	**71.17%**
0.7	69.26%
0.8	57.48%
0.9	41.98%

In the results, the thresholds that provide the highest accuracy for both proposed similarity measure is 0.6. The overall result of the threshold evaluation is shown in Table I above.

A computer program was written for the evaluation purposes. The program read the sentence pairs in training corpus of the MRP Corpus. The proposed semantic relatedness measure for the sentence pair is calculated. Then based on the threshold value determined in the previous section, each sentence pair is tagged as paraphrase or not paraphrase. The decision was then compared with the pre-determined tags in the corpus. The number of sentence pair which being tagged correctly are recorded, and the accuracy and F-measure were calculated at the end of the experiment. The formula for accuracy is shown in formula 3 above. F-measure is an accuracy test in information retrieval study that considers the precision and recall in compute the accuracy score. Precision (P) is relevancy of the retrieved data, where the score is the number of correct positive results (True positive results) divided by the number of all positive results (True Positive + False Positive results). The precision formula is shown in formula 3 below.

$$P = \frac{TP}{TP + FP} \quad (3)$$

While Recall (R) is the successfulness of retrieving the correct data compare to the query data, where the score is the number of correct results returned (True Positive results) divided by number of correct results that should have been returned (True Positive + False Negative results) . The formula for recalls is shown in the formula 4 below.

$$R = \frac{TP}{TP + FN} \quad (4)$$

The F-Measure used in the study derived from the precision and recall values. The formula for F1 Measure is shown in the formula 5 below.

$$F = 2.\frac{P.R}{P + R} \quad (5)$$

To evaluate the performance of the proposed semantic relatedness, this study compares the accuracy results of the proposed method with some known semantic relatedness measures. The methods chosen were ESA, LSA and SSA. The results for these relatedness measures on the same dataset are well published. For this study, we compare the accuracy and the F-measure in determining the performance of the proposed method.

The comparison results of the proposed semantic relatedness and the chosen methods were shown in Table II below. From the results, it is noticed that the proposed semantic relatedness measure outperformed the ESA and LSA measure. However it performed less compared to the SSA measure. Although the proposed method does not shows the highest accuracy in the evaluation, but the result of the

proposed semantic relatedness measure is considerable as it outperformed two of the most popular methods in the study.

TABLE II. ACCURACY COMPARISON BETWEEN PROPOSED SEMANTIC RELATEDNESS AND SELECTED SEMANTIC RELATEDNESS

Semantic Relatedness Measure	Accuracy	F Measure
ESA	67.0%	79.3%
LSA	68.8%	79.9%
Proposed Method	70.6%	80.4%
SSA	72.5%	81.4%

V. IDENTIFYING RELEVANT ANSWERS IN CQA

This study believes that good quality answers must have high relevancy towards its question. Therefore it is possible to identify the good and relevant with semantic relatedness. Since the proposed semantic relatedness measure is reasonable accurate in measuring the semantic relatedness between sentences, this study is using the proposed semantic relatedness to identify relevant answers in the Online Community Question Answering Services.

The suitability of such measure in identifying relevant answers in Online Community Question Answering Services were partially proven through a semantic relatedness distribution test using data from Yahoo! Answers. For the test, 5091 questions and their best answer were collected from Yahoo! Answers, a popular Online Community Question Answering Services. The semantic relatedness between the question and their best answers were calculated and classified into 10 categories using the proposed semantic relatedness, which explained in Section III.

The distribution of the proposed semantic relatedness for the collected question their best answers are plotted in Figure 1 below. From the Figure, it is noticeable that the graph is favor towards the right, means that most of the semantic relatedness value are above 0.5. The graph also shows that nearly 20% of question and their best answers have semantic relatedness between 0.7 and 0.6. If assume a relevant answer should have semantic relatedness above 0.5 value, the proposed semantic relatedness is able to identify almost 64% of the question and best answers in the collected dataset. The results may consider a bit lows compare to well-known machine learning methods in identifying relevant answers. However, using semantic relatedness in identifying relevant answers have the advantage of purely relying on the semantical data to identify relevant answers.

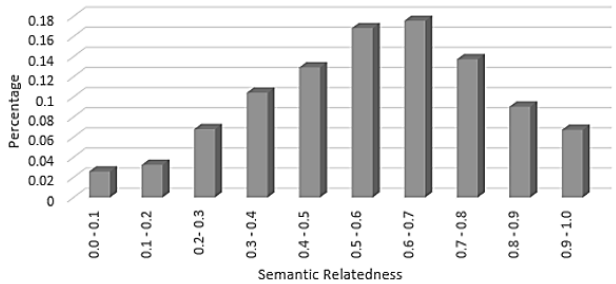

Figure 1: Proposed Semantic Relatedness Distribution on Yahoo! Answers Dataset

VI. CONCLUSION

This study proposed a sentence-to-sentence semantic relatedness measure that based on WUP similarity in WordNet. The performance of the proposed semantic relatedness measure was evaluated using a popular paraphrasing detection test, and the result shows reasonable accuracy compares to known semantic relatedness.

At the end of this paper, the study also shows that the proposed semantic relatedness is able to identify relevant answers from a collection of Yahoo! Answers dataset. This conclusion is based on the distribution of the proposed semantic relatedness measure between the question and their best answers in the collected dataset. However, this conclusion only proofed that the proposed semantic relatedness is able to identify relevant answers from Online Community Question Answering Services. More comprehensive process are needed to determine the best answers among the relevant answers.

The results in this study points to several future research directions for applying semantic relatedness measures in online community question answering services. The first research direction to improve the proposed semantic relatedness, currently the proposed semantic relatedness is calculated using the WUP similarity from WordNet as the word-to-word semantic relatedness. The limitation of semantic resources in WordNet will affects the ability of the proposed method in calculating more comprehensive semantic measures. For that, one of the solution is to find a semantic resources that are more effective in representing the real world context.

Another research direction from this study is to use the proposed semantic relatedness in generating user acceptable answers using the relevant answers found in Online Community Question Answering Services like Yahoo! Answers and Stackoverflow.com.

ACKNOWLEDGMENT

This study is supported by Universiti Sains Malaysia and Universiti Malaysia Sarawak (UNIMAS).

REFERENCES

[1] J. Gracia, and M. Eduardo. "Web-based measure of semantic relatedness." *Web Information Systems Engineering-WISE 2008*. Springer Berlin Heidelberg, 2008. 136-150.

[2] Budan, I. A., and H. Graeme. "Evaluating WordNet-Based Measures of Semantic Distance." *Comutational Linguistics* 32.1 (2006): 13-47.

[3] S. Hassan. Measuring semantic relatedness using salient encyclopedic concepts. University of North Texas, 2011.

[4] Z. Wu and M. Palmer, Verbs semantics and lexical selection, Proceedings of the 32nd annual meeting of the Association for Computational Linguistics (ACL'94) (Las Cruces, New Mexico), 1994, pp. 133–138.

[5] M. Lesk, Automatic sense disambiguation using machine readable dictionaries, Proceedings of the 5th annual international conference on Systems documentation (SIGDOC'86)(Toronto, Ontario), ACM Press, 1986, pp. 24–26.

[6] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, Proceedings of the 14th International Joint Conference on

Artificial Intelligence (Montreal, Quebec, Canada), Morgan Kaufmann Publishers Inc., 1995, pp. 448–453.

[7] T.K. Landauer, P.W. Foltz, and D. Laham. "An introduction to latent semantic analysis." *Discourse processes* 25, no. 2-3 (1998): 259-284.

[8] E. Gabrilovich, and S. Markovitch. "Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis." In *IJCAI*, vol. 7, pp. 1606-1611. 2007.

[9] R. Mihalcea, C. Corley, and C. Strapparava. "Corpus-based and knowledge-based measures of text semantic similarity." In *AAAI*, vol. 6, pp. 775-780. 2006.

[10] A. Islam, and D. Inkpen. "Semantic text similarity using corpus-based word similarity and string similarity." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 2, no. 2 (2008): 10.

[11] J. Jeon, W. B. Croft, J. H. Lee, and S. Park. "A framework to predict the quality of answers with non-textual features." In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 228-235. ACM, 2006.

[12] G.Y. Jeon, Y.M. Kim, and Y. Chen. "Re-examining price as a predictor of answer quality in an online Q&A site." In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 325-328. ACM, 2010.

[13] F.M. Harper, D. Raban, S. Rafaeli, and J. A. Konstan. "Predictors of answer quality in online Q&A sites." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 865-874. ACM, 2008.

[14] Y. Liu, S. Li, Y. Cao, C.Y. Lin, D. Han, and Y. Yu. "Understanding and summarizing answers in community-based question answering services." In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pp. 497-504. Association for Computational Linguistics, 2008.

[15] P. Fichman."A comparative assessment of answer quality on four question answering sites." *Journal of Information Science* 37, no. 5 (2011): 476-486.

[16] E. Agichtein, C. Carlos, D. Debora, G. Aristides, and M. Gilad. "Finding high-quality content in social media." In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pp. 183-194. ACM, 2008.

[17] L.A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. "Knowledge sharing and yahoo answers: everyone knows something." In *Proceedings of the 17th international conference on World Wide Web*, pp. 665-674. ACM, 2008.

[18] C. Shah, and J. Pomerantz. "Evaluating and predicting answer quality in community QA." In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 411-418. ACM, 2010.

[19] M. J. Blooma, A. Y. Chua, & D. H. L. Goh. "Selection of the best answer in cqa services." In *Information Technology: New Generations (ITNG), 2010 Seventh International Conference on*, pp. 534-539. IEEE, 2010.

[20] Q.J. Tian, P. Zhang, and B. Li. "Towards Predicting the Best Answers in Community-based Question-Answering Services." In *ICWSM*. 2013.

[21] B. Dolan, Q. C. Quirk, and C. Brockett. "Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources." In *Proceedings of the 20th international conference on Computational Linguistics*, p. 350. Association for Computational Linguistics, 2004.