Comparative Analysis of Similarity Measures for Sentence Level Semantic Measurement of Text

Sazianti Mohd Saad
Product Quality and Reliability Engineering, MIMOS Berhad
Technology Park Malaysia, 57000, Kuala Lumpur, Malaysia
sazianti.saad@mimos.my

Siti Sakira Kamarudin
School Of Computing
Universiti Utara Malaysia,
Sintok, 06010, Kedah, Malaysia
sakira@uum.edu.my

***Abstract*—The accuracy of similarity measurement between sentences is critical to the performance of several applications such as text mining, question answering, and text summarization. This paper focuses on calculating semantic similarities between sentences and performing a comparative analysis among identified similarity measurement techniqnes. Comparison between three popular similarity measurements which are Jaccard, Cosine and Dice similarity measures has been conducted. The performance of each identified measurement was evaluated and recorded. In this paper, we use a large lexical database of English known as WordNet to calculate the word-to-word semantic similarity. The result of this research concludes that the Jaccard and Dice performs better in measuring the semantic similarity between sentenees.**

***Index Terms*—Semantic Similarity, Similarity Measnrement, Sentence Similarity**

I. Introduction

The similarity between sentences becomes important in several applications of natural languages such as text summarization, example-based machine translation, automatic question-answering, information extraction and text clustering. The fundamental function in applications such as text mining and text summarization that usually operates at the sentence or sub-sentence level is to measure the similarity between sentences [1].

Sentences are a complete set of words in itself. It typically contains a subject and predicate, conveying a statement, question, exclamation, request, command or suggestion typically. Thus, sentences can represent different meaning although it consist the same words. This happens if the words are put together in a different way.

Naturally, human usually use natural languages to express their needs in order to retrieve information. Therefore they tend to create the query by using set of sentenees that based on daily use of language. This becomes a critical task in order to determine the similarity between sentences that have large impact in many text applications. [2-4].

Existing work in this area have attempted to compute text similarity by analyzing co-occurring words and word statistics in a probabilistic model. Despite of its inherent simplicity, such methods are efficient in handling long texts, which usually include a number of co-occurring words, however these method produces lower performance result for short text such as a single sentence.

In this research, we use different similarity measurement techniques to compute similarity between sentences and perform a comparative analysis among those similarity measurement techniques. We performed a comparative analysis to identify the suitable text similarity measurement for sentences levels. Comparison between three popular similarity measurements which are Jaccard, Cosine and Dice similarity measures was conducted. In addition, the performance of each identified measurement was evaluated and recorded. The next section briefly introduces the identified similarity measures. Section III presents related works. Section IV explains the proposed method followed by experimental results in section V. Section VI concludes the article.

II. SIMILARITY MEASURES

Similarity measure is the distance between various data points [5]. Similarity measures are also used in measuring similarity between sets based on the intersection of the two sets. Similarity measures are also known as a function that computes the degree of similarity between a pair of text objects. In summary, similarity is an amount that reflects the strength of relationship between two data.

There are several types of similarity measures such as Dice coefficient [6], Jaccard Similarity [4], Cosine Similarity [4], Euclidean distance [7] and others. Research conducted in [6] shows the importance of these similarity measures. Similarity measure can represent the similarity between two sentences and make it possible to rank the retrieved information in the order of presume importance. There are three types of well-known similarity measures that have been selected to be included in this research:

1) Dice Similarity Measure [8].

$$S_{A,B} = \frac{2\,|words_A \cap words_B|}{|words_A| + |words_B|} \quad (1)$$

978-1-4799-1506-4 ©2013 IEEE

2) Jaccard Similarity measure [8].

$$S_{A,B} = \frac{|words_A \cap words_B|}{|words_A \cup words_B|} \quad (2)$$

3) Cosine Similarity measure [8].

$$S_{A,B} = \frac{|words_A \cap words_B|}{\sqrt{|words_A| \, |words_B|}} \quad (3)$$

III. RELATED WORKS

The techniques for discovering the similarity between long texts or document have centered on analyzing shared words. These techniques are only available to deal with long texts because they contain adequate co-occurring words that express very similar meanings [9]. This will pose a tough challenge for computational method as the information in short texts is very limited.

There is a valid assumption for large-size text fragments such as document, where the more similar two texts are, the more words they have in common. The assumption does not hold for small-sized text fragments such as sentences, since two sentences may be semantically similar despite having few, if any, words in common. One approach to measure similarity between two sentences is based on representing the sentences in a reduced vector space consisting only of the words contained in the sentences.

There are a lot of researches have been conducted to evaluate the similarity between documents or long text [4], sentences [10] and short text [11]. In these cases, they used the same method that has been adopted from the approaches used for long text documents to evaluate the similarity measures for textual data. [12-14] have conducted research to determine the similarity between long texts or documents but there are less work related to the measurement of similarity between sentences or short text [15].

Research have been conducted in [12] by combining the information from multiple linguistic indicators to determine the semantic distance between pairs of the small textual units by presenting a new composite similarity metric over short passages. The potential features and the optimal combination selected via machine learning have been investigated in this research. In this method the syntactic information have been ignored, it only considers the semantic information.

A method for similarity measure between sentences that based on semantic information and word order has been presented in the study conducted by [10]. Their focus is to directly calculate the similarity between very short texts of the sentence length. Firstly, semantic similarity was derived from a lexical knowledge base and a corpus. Secondly, based on the number of different words and the number of word pairs in a different order, the word order similarity is measured. The overall sentence similarity is defined as a combination of semantic and word order similarity.

Semantic similarity measures have been proposed using a bag-of-word approach and the use of word specialization to calculate the similarity of the word [16]. The word in another sentence that has higher semantic similarity for each word in a sentence can be identified by the method. In order to combine the semantic similarity of each text segment in turn with respect to the other text segment, the metric has been used. However, the method always calculated and selects the higher similarity between words from two sentences. Therefore, many non-similar sentence pairs will be judged similar.

[7] presented the results of the study that identified similarity measure that was used for both Information Retrieval and Document Clustering. The similarity measures that have been used in [7] were Cosine similarity, Jaccard measure, and Euclidean measure. Based on their result, they indicate that the Cosine similarity measure is more efficient compared to others. In the study, they also mentioned that Cosine similarity measure is more efficient for text particularly.

[17] has conducted a study that identified the similarity measurement technique to be compared through SimReq Framework such as Consine similarity, Jaccard Similarity and Dice Similarity. Their study have been conducted using two different scenarios in which with removing of the repetition words from each and all requirements and without doing this. Based on their similarity measurement result, they concluded that Cosine similarity measure is more efficient to find out the similarity among requirement as compared to others.

[18] has conducted a survey on measurement of semantic similarity between words. They have described the methods based on precompiled databases like WordNet and Brown Corpus as well as the web search engine. Along with this they have compared all the methods on the basis of performance and their limitation. From the study, Experimental result on Miller-Charles benchmark dataset showed that the method based on page count outperforms all the existing semantic similarity measures by a wide margin, achieving a correlation coefficient of 0.87.

A number of researches have attempted to perform a comparative analysis on similarity measures for textual data, for example the work in [19] where 4 similarity measures were compared for web-paged clustering and the study performed in [4]. In [4], 5 measures were compared via empirical experiments i.e. Euclidean distance, cosine similarity, Jaccard coefficient, Pearson correlation coefficient and averaged Kullback-Leibler divergence. However, their work aims on document clustering where text was compared at document level.

Previous work [4] showed that the choice of similarity measure depends on the level of the compared text units. Text can be compared at document level, or sentence level, or

phrase level or even word level. Although previous studies proved that cosine distance performed well in text comparison at documents level, its performance decreases substantially when smaller text units are processed. It failed to perform well when the document is decomposed into sentences [20].

IV. THE PROPOSED COMPARATIVE ANALYSIS METHOD

This study aims to provide a focused comparative analysis of three popular text similarity measures applicable for sentence level comparison. We proposed to compare Jaccard, Dice and Cosine similarity measures in order to evaluate the performance on calculating sentence level similarity. This performance has been evaluated based on score from 0 to 1. The higher the score of each measurement, the more efficient the similarity measure is. We have measured the similarity of the sentences by comparing samples of sentences with same meaning but expressed in a different way and samples of sentence which are entirely different in meaning. Before calculating the similarity the sentences are transformed into Bag of Word representation [21]. The set of sentence that have been transformed was used in the calculation of each of the identified similarity measures to evaluate their performance.

Figure 1 illustrates the steps for measuring the sentence similarity between two sentences.

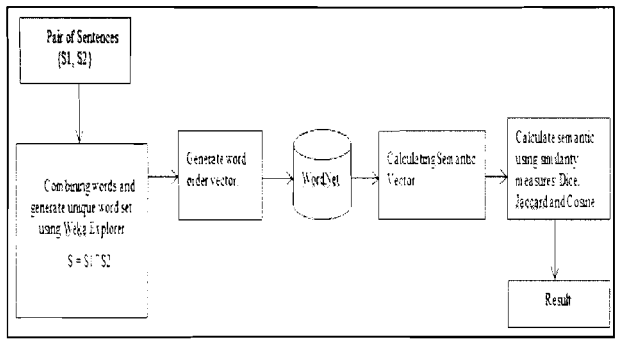

Fig. 1 Proposed Comparative analysis Method

As shown in figure 1, a joint distinct word set is formed for two compared sentences. The process begins by removing the stop words such as articles in the sentence. For example: "a", "an", "the" etc. The remaining words are transformed into bag of word vectors. Then, we use the WordNet [22] to find the synonyms for all words in the sentences and measure the semantic by weighting the overlap synonyms and comparing all the words in the sentence. Two set of vector is derived from the overlap synonym weight calculation. Next, the semantic similarity is calculated using the identified similarity measurement techniques; Dice, Jaccard and Cosine Similarity measures. Using the method proposed in Figure 1, the semantic vector of words is formed for the compared sentences. To illustrate the above method consider two sentences, S1 and S2, and a joint distinct word set S, is formed between S1 and S2 as follows:

$S_1 \cup S_2 = S$; where $S = (w_1, w_2, \ldots\ldots w_n)$; w_i are distinct.

For example, if we have the sentences:

> S_1 : *a soldier was killed Monday and another wounded when their convoy was ambushed in northern Iraq.;*
>
> S_2 : *On Sunday, a US soldier was killed and another injured when a munitions dump they were guarding exploded in southern Iraq;*

Then we will have:

> **$S = S_1 \cup S_2$ ={*soldier, killed, Monday, wounded, convoy, ambushed, northern, iraq , Sunday ,US, injured ,munitions, dump ,guarding ,exploded ,southern*}.**

The S set is derived by removing the stop word in both sentences and set S is represent as the semantic information for the compared sentences. A term matrix is constructed to derive the semantic information content of S_1 and S_2

$$S_1 = \begin{array}{c} S = \quad w_1 \quad w_2 \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad w_n \\ \begin{matrix} q_1 \\ q_2 \\ \vdots \\ \vdots \\ q_m \end{matrix} \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & \ldots & \ldots & \ldots & x_{1,n} \\ x_{2,1} & x_{2,2} & \ldots & \ldots & \ldots & \ldots & x_{2,n} \\ \vdots & \vdots & \ldots & \ldots & \ldots & \ldots & \vdots \\ \vdots & \vdots & \ldots & \ldots & \ldots & \ldots & \vdots \\ x_{m,1} & x_{m,2} & \ldots & \ldots & \ldots & \ldots & x_{m,n} \end{bmatrix} \end{array} \qquad (4)$$

Whereby $x_{i,j}$ represents the similarity measure between the i-th word in the compared sentences and j-th word of the joint set. The value of $x_{i,j}=1$, if q_i and w_i are the same word, whereas if $qi \neq wi$, $x_{i,j}=0$; while if the synonym of the word is the same , qi(synonym) = wi, $x_{i,j}=1$. The produced vector is applied in the similarity measurement for semantic similarity calculation. The result is compared with the benchmark record that was obtained from [3].

V. EXPERIMENTAL RESULT

We have conducted an experiment in order to perform the comparative analysis of similarity measurement techniques for sentences. The testing as illustrated in Table I is conducted using the 8 sample data obtained from [3]. We compare the first sentences of the list with the remaining 2 sentences. To assist discussion, we called the first sentence as ' S' and the remaining are 'S1' and 'S2'.Each of the two compared sentence('S1 and S2') is compared against the target sentence('S'). For Example we have:

> S_1: *'I am proud that I stood against Richard Nixon, not with him, 'Kerry said.*
> S_2 *'I marched in the streets against Richard Nixon and Vietnam War,' she said*
> **S={*proud,stood,richard,nixon,him,kerry,marched,streets, vietnam,war, said*}.**

Before representing the sentence in a vector format, we use WordNet [22] to find the synonyms for each of the word for S. Vector representation will consider the following condition:

if S = S_1, S_2 =1 ; while S = S_1, S_2 = 1 (synonyms), whereas , S ≠ S_1, S_2 =0

For Example:

S_1 = {1,1,1,1,1,1,0,0,0,1} , S_2 ={ 0,1,1,1,0,1,1,1,1,1}

After deriving the vectors, we applied the vectors in the identified similarity measures. In this paper we have computed the semantic similarity between sentences by considering the synonyms that are derived from WordNet [22].

TABLE I. SEMANTIC SIMILARITY BETWEEN SENTENCES

Sample	*Cosine Similarity*	*Jaccard Similarity*	*Dice Similarity*	*Bench Mark[3]*
Sample1	0.961	0.923	0.960	0.857
Sample2	0.738	0.583	0.737	0.583
Sample3	0.667	0.500	0.667	0.929
Sample4	0.966	0.933	0.966	0.975
Sample5	0.481	0.313	0.476	0.375
Sample6	0.589	0.417	0.588	0.381
Sample7	0.668	0.500	0.667	0.386
Sample8	0.645	0.471	0.64	0.398

Based on Table I, the same 8 samples have been experimented using the identified text similarity measure to measure the semantic similarity. For Sample 1, the result for Cosine is 0.961, Jaccard is 0.923 and Dice is 0.960. For Sample 2, the result of Cosine is 0.738, Jaccard is 0.583 and Dice is 0.737. The results for Sample 3-8, for Cosine, Jaccard and Dice are as depicted in Table I. From the results in Table I it can be concluded that the nearest result to benchmark is Dice similarity. Figure 2 shows the relatedness of the 3 similarity with the benchmark value.

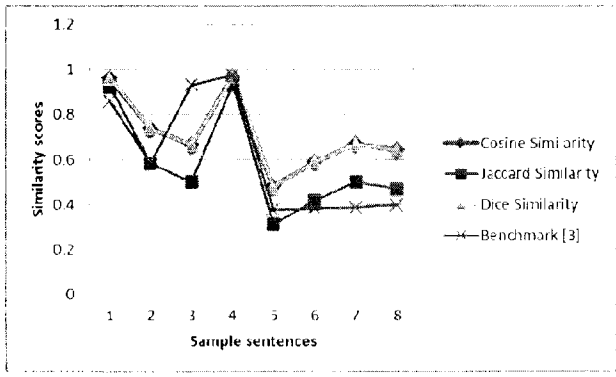

Fig. 2 Similarity scores graph

As can be learned from Figure 2, Cosine and Dice performs similar to benchmark for sentences with same meaning (sample 1-4) however for sentences that are dissimilar in meaning (sample 5-8) Jaccard's score are similar to benchmark. To confirm the findings we have calculated the correlation between the three similarity measures with the benchmark scores as shown in Table II.

TABLE II. CORRELATION RESULT SEMANTIC SIMILARITY BETWEEN SENTENCES

Method	*Correlation*
Cosine & benchmark	0.769
Jaccard & Benchmarks	0.771
Dice & Benchmarks	0.771

Based on Table II, we found that Dice and Jaccard similarity is more correlated to the benchmark scores. From the correlation result, we can conlude that Jaccard and Dice is more suitable to measure sentence level semantic similarity.

VI. CONCLUSION

Semantic sentence similarity is important in many applications such as information retrieval, information extraction and ontology learning. This paper has presented an approach based on bag of words in order to provide semantic similarity. This approach compares pair of sentences by first finding the similarity measures among words. The word-synonyms are derived from WordNet [22]. The obtained word synonyms are then used to construct the semantic vectors. Lastly the sentences similarity is calculated and compared using three well known similarity measures i.e.; Dice, Jaccard, and Cosine.

By conducting an experiment on implementing this approach, we have concluded that Jaccard and Dice performs better in measuring the semantic similarity between sentences. However more testing and evaluation works need to be conducted particularly involving real test data and human experts.

REFERENCES

[1] G. Ekran, D. Radev, "LexRank: Graph-based Lexial Centrality as Salience in Text Summarization. Journal of Art. Int. Research, 22. 2004. pp 457-479.

[2] Liu, X.Y., Zho Y. M. and Zheng, R. S., "Sentence Similarity based on Dynamic Time Warping". The International Conference on Semantic Computing. 2007 pp. 250-256.

[3] Li, L., Xia, H., Hu B.Y., Wang J., and Zhou, Y. M. "Measuring Sentence Similarity from Different Aspects", Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, 2009, pp. 2244-2249.

[4] Huang, A., "Similarity measures for text document clustering", Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand, 2008, pp. 49–56.

[5] Patidar A. K., Agrawal, J. and Mishra, N., "Analysis of Different Similarity Measure Functions and their Impacts on Shared Nearest Neighbor Clustering Approach", International Journal of Computer Applications, Vol.40., No.16, February 2012, pp.1-5.

[6] Cha, S.H., "Comprehensive survey on distance/ similarity measures between probability density functions", International

Journal of Mathematical Models and Methods in Applied Science, 1 (4), 2007, pp. 300–307.

[7] Subhashini, R. and Kumar, V.J.S., "Evaluating the Performance of Similarity Measures Used in Document Clustering and Information Retrieval", First International Conference on Integrated Intelligent Computing, 2010.

[8] Natt-och-Dag, J., Regnell, B., Carlshamre, P., Andersson, M., and Karlsson, J., "Evaluating Automated Support for Requirements Similarity Analysis in Market-Driven Development", 7th Int. Workshop on Requirements Engineering: Foundation for Software Quality, June 4-5 2001, Interlaken, Switzerland.

[9] Bates, M., "Subject Access in Online Catalogue: a Design Model", J. American Society for Information Science 11,1986, pp. 357-376.

[10] Li, Y., McLean, D., Bandar, Z. A., O'Shea, J. D., and Crockett, K., "Sentence similarity based on semantic nets and corpus statistics", IEEE Transactions on Knowledgeand Data Engineering Vol 18, No. 8, 2006, pp. 1138–1150.

[11] Abdalgader, K. and Skabar, S."Short-Text Similarity Measurement Using Word Sense Disambiguation and Synonym Expansion", Jiuyong Li \(Ed.\), AI' Advances in Artificial Intelligence. 3rd Australasian Joint Conference Adelaide, Australia, Proceedings Springer-Verlag Berlin Heidelberg 2010.

[12] Hatzivassiloglou, V., Klavans, J., and Eskin, E., "Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning", Proceeding of Empirical Methods in natural language processing and Very Large Corpora,(1999).

[13] Landauer, T and Dumais, S., "A solution to platos problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge". Psych. Rev. 104, 2, 1997, pp. 211–240.

[14] Maguitman, A., Menczer, F, Roinestad, H. , and Vespignani,A.,. "Algorithmic detection of semantic similarity", In Proceedings of the 14th International World Wide Web Conference, 2005..

[15] Foltz, P., Kintsch, W., and Landauer, T. The measurement of textual coherence with latent semantic analysis. Disc. Proc. 25, 2–3, 1998, pp. 285–307.

[16] Mihalcea, R., Corley, C., and Strapparava, C., "Corpus-based and knowledge-based measures of text semantic similarity" Proceeding of the Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, 2006.

[17] Muhammad Ilyas and Josef Kung, A comparative analysis of similarity measurement techniques through SimReq framework. FIT '09 Proceedings of the 7th International Conference on Frontiers of Information Technology, 2009.

[18] Ankush, M., Prof. Anil, D., and Dr.Prashant, C., "Measurement of Semantic Similarity Between Words: A Survey". International Journal of Computer Science, E ngineering and Information Technology (IJCSEIT)), Vol .2, No. 6, December 2012.

[19] Strehl, A., Ghosh, J., & Mooney, R., "Impact of similarity measures on web-page clustering", Workshop on Arti_cial Intelligence for Web Search (AAAI 2000), 2000, pp. 58-64.

[20] Allan, J., Wade, C., & Bolivar, "A. Retrieval and novelty detection at the sentence level", In Proceedings of the 26th annual 1101 international ACM SIGIR conference on research and development in information retrieval, 2003, pp. 314–321..

[21] Bag of Words Model. In Wikipedia :Retrieved March, 9, 2013, from http://en.wikipedia.org/wiki/Bag-of-words_model

[22] Pedersen, T., Patwardhan, S. and Michelizzi, J.,. "WordNet::Similarity - Measuring the Relatedness of Concepts". 2004.