International Journal of Engineering & Technology, 7 (2.14) (2018) 32-35

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper



Graph-based Representation for Sentence Similarity Measure : A Comparative Analysis

Siti Sakira Kamaruddin[1]*, Yuhanis Yusof[2], Nur Azzah Abu Bakar[3], Mohamed Ahmed Tayie[4], Ghaith Abdulsattar A.Jabbar Alkubaisi[5]

School of Computing, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia
**Corresponding author E-mail: sakira@uum.edu.my*

Abstract

Textual data are a rich source of knowledge; hence, sentence comparison has become one of the important tasks in text mining related works. Most previous work in text comparison are performed at document level, research suggest that comparing sentence level text is a non-trivial problem. One of the reason is two sentences can convey the same meaning with totally dissimilar words. This paper presents the results of a comparative analysis on three representation schemes i.e. term frequency inverse document frequency, Latent Semantic Analysis and Graph based representation using three similarity measures i.e. Cosine, Dice coefficient and Jaccard similarity to compare the similarity of sentences. Results reveal that the graph based representation and the Jaccard similarity measure outperforms the others in terms of precision, recall and F-measures.

Keywords: *Graph Based Representation, Latent Semantic Analysis, Text Representation, Text Similarity Measure, TF-IDF.*

1. Introduction

Almost 80% of the data that exist in the world today are textual data. Due to this fact, text mining has evolved to be an important research area. The ability to analyze the content of the accumulating textual database has become inevitable.

Effective text comparison is the key issue in analyzing textual data. Text comparison is performed in various text mining applications such as text clustering [1], text summarization [2], anomaly detection [3] etc. Text comparison is also the fundamental task in information retrieval where given a query, the contents of the query is compared with the information content to retrieve the most relevant information.

There are various methods exist in the literature for text comparison. The differences in the method depend on the text representation scheme used prior to text comparison. Text representation is an important task in any text analysis work because it is significant to transform the unstructured format of textual data into a more formal structure before any analysis are done on it. There are various text representation schemes proposed by researchers, among them are term frequency inverse document frequency (*tf-idf*) [4-7], Latent semantic analysis and graph based representation [3].

Since there are various ways to represent text, the similarity measure to compare text units also varies according to the representation schemes because one similarity measure may not be suitable for all representation schemes. Researchers have proposed a number of similarity measures that can be used to compare text units e.g. cosine similarity [2], dice coefficient, city block distance6, Chebychev dissimilarity distance, set difference [4, 8], and jaccard distance [3]. Among these, the cosine similarity which is based on geometric distance is a popular text similarity measure for text represented as bag of words, however it is arguable whether cosine similarity will produce acceptable results when text are represented with other representation schemes such as graph based representation.

In this paper, we perform a comparative analysis on three text representation schemes and three similarity measures for sentence level comparison. The rest of the paper is organized as follows; in section 2, some related work on comparative analysis are presented. Section 3 discusses the proposed comparative analysis method. Section 4 presents the results and discussion and section 5 concludes the paper.

2. Related work

Various work is reported in the literature on comparative analysis of similarity measures. For example, in [9] a comparative analysis was done on multimedia data. In [10] the comparison was done on ontology based representation of biomedical data while probability function is used to represent pattern recognition in [11]. However there are less work on comparing textual data. Among the work that had been performed on textual data comparison are reported in [12-14], however these work compared text at document level. In12 four similarity measures was tested on web pages. Similarly the work reported in [13] involve comparing five similarity measures for document clustering. The recent work reported in [14] focused on document level as well, where the comparison was done on Wikipedia articles. As was pointed out earlier, sentence level comparison poses different challenges.

In addition, it is hypothesized that the result of text comparison may be influenced by the applied text representation schemes. Furthermore, the level of text units i.e. documents, sentences, phrases or words might also influence the performance of different text representation schemes and different similarity measures.

This work investigates these issues and proposes a comparative analysis on three different text representation schemes and three different similarity measures.

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

3. Method

Figure 1 presents the framework of this work. The textual data are first pre-processed before it is represented into a more structural format. The three representation schemes that are investigated in this study are tf-idf, Latent Semantic Analysis (LSA) and graph based representation. Once represented into these three representation scheme, each represented sentence are compared with three similarity measures as shown in Figure 1 i.e. cosine, dice and jaccard similarity measures. The final step in the framework is to compare and analyse the produced results. We further explain each of the steps in detail.

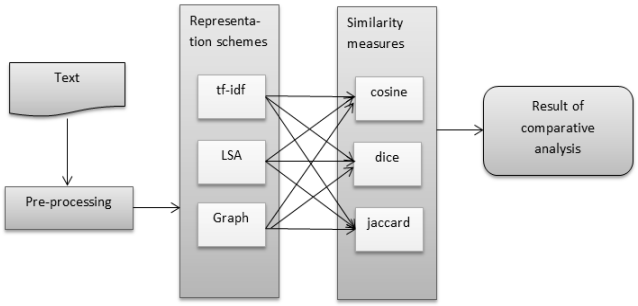


Fig. 1: The framework of comparative analysis

The data for this work are sample sentences obtained from the MS paraphrase dataset as applied in [16]. The data is first pre-processed where the processes such as tokenization, stop word removal, transformation to lower case, stemming, tagging, and parsing are performed. In tokenization, the sentences are broken down into individual terms. Then the stop words such as prepositions and determiners are removed from the data. Then, all uppercase letters are transformed into its lower case. In the stemming process, the derived words such as "fishes" "fishing" and "fished" are stemmed to its root word "fish". Tagging is the process to assign part of speech tags to each word. Part of speech includes noun, verb, prepositions, adjective etc. Then the parsing process identifies the structure of the sentences. It should be noted that not all representation schemes need all the pre-processing steps. For example, the tagging and parsing steps are needed for the graph based representation but not for tf-idf and LSA representations.

Text Representation Schemes

Three text representation schemes are investigated in this work. They are *tf-idf*, LSA and Graph based representation. This section briefly explains each of the representation schemes.

tf-idf. For the tf-idf representation, the pre-processed text will be represented as vectors of weighted terms. The dimension of the vector depends on the number of terms in the compared sentences. The weighting of the terms is calculated using Equation 1.

$$w_{i,j} = tf_{i,j} \times idf_i \quad (1)$$

The weight of term *ti* in a sentence *dj* is obtained by multiplying *tfij* and *idfi* where $tf_{i,j} = \frac{f_{i,j}}{max_h f_{h,j}}$ is the frequency of the terms *ti* in the sentence *j*. The denominator is the maximum number of occurrence of all terms in sentence *dj*. And the Inverse Document Frequency, $idf_i = log_2 \frac{C}{n_i} + 1$, where c is the size of the sentence and *ni* is the number of sentence that contains *ti* .

Latent Semantic Analysis. Using Latent Semantic Analysis (LSA), the relations that exists between terms can be represented using related concepts. The relationship between words are obtained with a factor called Singular Value Decomposition (SVD), this representation scheme represents the term document matrix using three matrices as shown in Equation 2.

$$X = U \times S \times V^T \quad (2)$$

Where *U* is a *(u × n)* matrix, *S* is a *(n × n)* matrix and *VT* is a (*v × n)* matrix with n 'latent semantic' dimensions.

Graph based representation. In graph based representation the textual data will be transformed into a graph of words. In the pre-processing step, besides tokenization, stop word removal, transformation into lower case, and stemming, additional steps are needed such tagging and parsing. Parsed sentences produces syntactic parse tree where the structure of the sentences can be obtained. In order to capture the semantics, we performed word sense disambiguation to identify the canonical form of words (different words that conveys the same meaning). The reason for this step is to enable the similarity measure to detect similar sentences but was written using different words that are synonym. Once the canonical form of words were identified the sentence structure are then converted into a graph based representation following the notation in Equation 3.

$$G = (V, E) \quad (3)$$

Where *V* is a nonempty set of vertices, and *E*, a set of ordered pairs of distinct elements of *V* called edges where *V* = (*Vi, Vj* …) and *E* = {*e1,e2,e3….. ek*} and *ei* = (*Vi, V*j).

Similarity Measures

Three text similarity measures are investigated in this work. They are dice coefficient, jaccard distance and cosine similarity measures. This section briefly introduces each of the similarity measures.

Dice coefficient. Dice coefficient measures the overlap to the average size of the two sets. In this work, the dice coefficient is calculated using Equation 4.

$$D_{a,b} = \frac{2\,|word_a \cap word_b|}{|word_a| + |word_b|} \quad (4)$$

Jaccard distance. Jaccard distance calculates the similarity of sentences by finding the fraction of intersection and union of words. In this work, the jaccard distance is calculated using Equation 5.

$$J_{a,b} = \frac{|word_a \cap word_b|}{|word_a \cup word_b|} \quad (5)$$

Cosine Similarity Measure. The cosine similarity measure is the most popular similarity measure for text. It relates to the overlap of the geometric average of the two sets. In this work, the cosine similarity measure is calculated using Equation 6.

$$C_{a,b} = \frac{|word_a \cap word_b|}{\sqrt{|word_a||word_b|}} \quad (6)$$

The result of all the above similarity measurement is a value between 0 and 1. 1 denotes that the sentences are completely similarity and 0 denotes that the sentences are completely dissimilar. As a rule of thumb, any value above 0.5 is considered similar and value less than 0.5 is considered dissimilar.

Comparative Analysis

In order to analyse the performance of the representation schemes on different similarity measures, the experiment was performed on 8 pairs of sentences obtained from MS Paraphrase test corpus dataset [15]. The chosen sentences are listed in Table 1 & 2. The sentences were given to human expert to judge the similarity and dissimilarity. As a result, the human expert has determined 4 pairs (pair 1-4) are similar sentences and the other 4 pairs (pair 5-8) are dissimilar sentences. The expert judgement is used as a benchmark to evaluate the automatic similarity calculation on these sentences.

Table 1: Pairs of similar sentences

Pair	Sentences
1	Taha is married to former Iraqi oil minister Amir Muhammed Rasheed, who surrendered to U.S. forces on April 28." "Taha's husband, former oil minister Amer Mohammed Rashid, surrendered to U.S. forces on April 28."
2	"On July 22, Moore announced he would appeal the case directly to the U.S. Supreme Court." "Moore of Alabama says he will appeal his case to the nation's highest court."
3	"Six Democrats are vying to succeed Jacques and have qualified for the Feb. 3 primary ballot." "Six Democrats and two Republicans are running for her seat and have qualified for the Feb. 3 primary ballot."
4	"Agriculture Secretary Luis Lorenzo told Reuters there was no damage to the vital rice crop as harvesting had just finished." "Agriculture Secretary Luis Lorenzo said there was no damage to the vital rice crop as the harvest had ended."

Table 2: Pairs of dissimilar sentences

Pair	Sentences
5	"A soldier was killed Monday and another wounded when their convoy was ambushed in northern Iraq." "On Sunday, a U.S. soldier was killed and another injured when a munitions dump they were guarding exploded in southern Iraq."
6	"Perkins will travel to Lawrence today and meet with Kansas Chancellor Robert Hemenway." "Perkins and Kansas Chancellor Robert Hemenway declined comment Sunday night."
7	"'I am proud that I stood against Richard Nixon, not with him,' Kerry said." "'I marched in the streets against Richard Nixon and the Vietnam War,' she said."
8	"The report by the independent expert committee aims to dissipate any suspicion about the Hong Kong government's handling of the SARS crisis." "A long awaited report on the Hong Kong government's handling of the SARS outbreak has been released."

The performance measures used in the experiment are accuracy, precision, recall and F-measures. These measures are calculated by determining the number of sentence correctly identified as similar or dissimilar compared to the decisions by human experts. In other words, using the human decisions as benchmark the number of true positive (TP) which is equivalent to actual similar sentences correctly identified as similar, true negative (TN) which is equivalent to actual dissimilar sentences correctly identified as dissimilar, false positive (FP) which is equivalent to actual similar sentences incorrectly identified as dissimilar, and false negative (FN) which is equivalent to actual dissimilar sentences incorrectly identified as similar are determined. Then, the accuracy is calculated as (TP + TN) / all data, precision is TP/(TP+FP), recall is TP/(TP+FN) and the F-measures as the harmonic mean of precision and recall, which is equal to 2TP/(2TP+FP+FN). The results are presented in the next section.

4. Results and discussion

Figure 2, 3 and 4 presents the graph of similarity measurements of the sample sentences using dice, jaccard and cosine similarity measures for each representation schemes i.e. *tf-idf*, LSA and graph based representation. As can be seen in Figure 2, for *tf-idf* based representation, the similarity score produced by the three similarity measurements are mostly below the 0.5 value threshold. Therefore we can conclude that using *tf-idf* representation for sentence level comparison is not advisable to detect similarity at the sentence level.

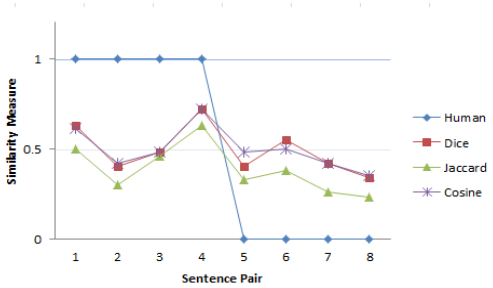


Fig.2. Similarity measures for *tf-idf* representation

In Figure 3, for the LSA based representation there is a slight improvement especially for the cosine similarity scores which were able to detect similar sentence but did not perform well for dissimilar sentences. In Figure 4 almost all the similarity score performed well for the graph based representation where similar sentence produced similarity score above 0.5 and dissimilar sentences produced similarity score below 0.5. This concludes that the graph based representation performs better for sentence level text comparison using all similarity measures.

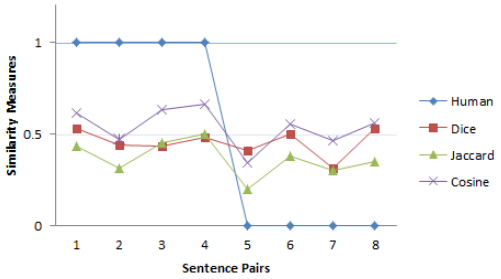


Fig.3. Similarity measures for LSA representation

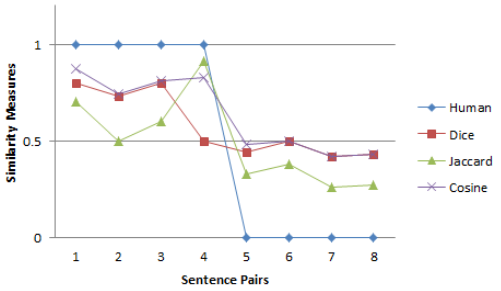


Fig.4. Similarity measures for graph representation

To prove our point further, we calculated the correlation scores for each similarity measures against the human benchmark. For every dissimilarity scores produced by similarity measure A_i, (i=1,2,……d) its correlation coefficient, r to the benchmark similarity Bj,(j=1,2,…….d) is given by $\frac{\Sigma(A_i-\overline{A})(\boldsymbol{B}_j-\overline{B})}{\sqrt{\Sigma(A_i-\overline{A})^2+\Sigma(\boldsymbol{B}_j-\overline{B})^2}}$ where is $\overline{A}$ the mean score of similarity score A and $\overline{B}$ is the mean score of benchmark similarity B. The correlation coefficient scores are shown in Table 3.

Table 2: Correlation of the similarity scores to the benchmark

Representation Schemes	Similarity Measures	Correlation to the benchmark
tf-idf	Dice	0.53
	Jaccard	0.68
	Cosine	0.55
LSA	Dice	0.24
	Jaccard	0.64
	Cosine	0.58
Graph	Dice	0.82
	Jaccard	0.85
	Cosine	0.97

From the correlation scores in Table 3, it can be perceived that the graph based representation produced the highest correlation coefficient for all similarity measures. The *tf-idf* based representation performs slightly better than the LSA based representation in terms of correlation to the benchmark. We further analyse the produced result by calculating the accuracy, precision, recall and F- measures as explained in the previous section.

Table 4: Presents the results

Representation Schemes	Similarity Measures	Performance Measures			
		Accuracy	Precision	Recall	F-measure
tf-idf	Dice	0.63	0.67	0.5	0.57
	Jaccard	0.75	1.0	0.5	0.67
	Cosine	0.63	0.67	0.5	0.57
LSA	Dice	0.38	0.33	0.25	0.29
	Jaccard	0.63	1.0	0.25	0.4
	Cosine	0.63	0.6	0.75	0.67
Graph	Dice	0.88	0.8	**1.0**	**0.89**
	Jaccard	**1.0**	**1.0**	**1.0**	**1.0**
	Cosine	0.88	0.8	**1.0**	**0.89**

Table 4 gives a clear picture on the performance of the evaluated representation schemes on each similarity measures. For the accuracy, precision, recall and F-measure, all the tested similarity measures performed well for graph based representation with scores above 0.8 or 80%. From these results, we can see that the best performing similarity measure for graph based representation is the Jaccard similarity measure with accuracy, precision, recall and F-measure scores of 1.0 or 100%. Jaccard similarity measure also produced the best results for *tf-idf* based representation with accuracy, precision, recall and F-measure scores of 0.75, 1.0, 0.5 and 0.67 respectively. However, for LSA, the cosine similarity score outperforms the Jaccard with accuracy, precision, recall and F-measure scores of 0.63, 0.6, 0.75 and 0.67 respectively.

5. Conclusion

It can be concluded that the best representation scheme for sentence level comparison is the graph based representations and the best similarity measure for sentence level comparison is the Jaccard similarity measures. We can also further conclude that the *tf-idf* representation scheme is not suitable to be used for sentence level comparison. The reason is because *tf-idf* only captures individual terms, therefore a better alternative representation that captures the semantics of words is the LSA representation, however if the LSA is used then, the best similarity measure is the cosine similarity measure. If the textual data are small in size than we advocate to go through the process of tagging and parsing to produce the proposed graph based representation that not only captures the structure of sentences but also the semantics. This will ensure a better text comparison performance.

Acknowledgement

We would like to thank the Malaysian Ministry of Higher Education for providing the funding for this research through the Exploratory Research Grant Scheme (ERGS) and Universiti Utara Malaysia through the Leads Research Grant Scheme.

References

[1] A. J. Mohammed, Y. Yusof, & H. Husni. Integrated Bisect K-Means and Firefly Algorithm for Hierarchical Text Clustering. J. Eng. Applied Sci, 100(3), (2016) 522-527.

[2] S. A., Waheeb & H. Husni. Multi-Document Arabic Summarization Using Text Clustering to Reduce Redundancy. International Journal of Advances in Science and Technology (IJAST), 2(1), (2014) 194-199.

[3] S. S. Kamaruddin,.A. A. Bakar, A. R. Hamdan, F.M. Nor, M.Z. Z. Nazri, Z. A. Othman, & G. S. Hussein. A text mining system for deviation detection in financial documents. Intelligent Data Analysis, 19(s1), (2015) S19-S44.

[4] Allan, J., Wade C. and Bolivar A., Retrieval and Novelty Detection at the Sentence Level. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, (2003) 314-321.

[5] Jacquenet, F. and Largeron, C., Using the structure of documents to improve the discovery of unexpected information. Proceedings of the 2006 ACM symposium on Applied computing table of contents, (2006) 1036-1042.

[6] Abouzakhar, N., Allison, B. and Guthrie, L., Unsupervised Learning-based Anomalous Arabic Text Detection. Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), (2008) 291-196.

[7] F. Jacquenet, and C. Largeron. Discovering unexpected documents in corpora. Knowledge-Based Systems 22: (2009) 421-429.

[8] Fernández, R. T. and Losada, D. E., Novelty Detection Using Local Context Analysis. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR'07, (2007) 813-814.

[9] Beecks, C. Uysal, M. S. and. Seidl, T., A comparative study of similarity measures for content-based multimedia retrieval. In Proc. IEEE International Conference on Multimedia & Expo, (2010) 1552-1557.

[10] Lee, W. N., Shah, N., Sundlass, K., Musen, M., Comparison of Ontology-based Semantic-Similarity Measures. AMIA Annu Symp Proceedings, (2008) 384-388.

[11] S. Cha, Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions, International Journal of Mathematical Models and Methods in Applied Sciences, vol. 1(4), (2007) 300-307

[12] Strehl, A., Ghosh, J., and Mooney, R., Impact of similarity measures on web-page clustering. In AAAI-2000: Workshop on Artificial Intelligence for Web Search, July (2000).

[13] Huang, A., Similarity Measures for Text Document Clustering. In Proceedings of the New Zealand Computer Science Research Student Conference (NZCSRSC'08), Christchurch, New Zealand (2007).

[14] J. Szymanski, Comparative Analysis of Text Representation Methods using Classification. Cybernetics and System 45(2). (2014).

[15] Dolan, W., Quirk, C., and Brockett, C., "Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources", Proceeding of the 20th International Conference on Computational Linguistics, (2004).

[16] Lin L., Hu X., Hu B., Wang J., "Measuring sen-tence similarity from different aspects", The Eighth International Conference on Machine Learning and Cybernetics. Bao-ding, Hebei, China, (2009) 2244-2249.