Proc. of the Second Intl. Conf. on Advances in Computer and Information Technology -- ACIT 2013

ISBN: 978-981-07-6261-2 doi:10.3850/978-981-07-6261-2_42

Semantic Similarity Measure for Graph-based Sentences

Siti Sakira Kamaruddin[1], Yuhanis Yusof[2], Nur Azzah Abu Bakar[3]
Computational Intelligence Group, School of Computing,
Universiti Utara Malaysia,
Sintok, Kedah, Malaysia.
[1]sakira@uum.edu.my, [2]yuhanis@uum.edu.my, [3]nurazzah@uum.edu.my

Abstract— **Graphical text representation method attempts to capture the syntactical structure and semantics of documents. As such, they are the preferred text representation approach for a wide range of problems namely in natural language processing, information retrieval and text mining. In a number of these applications, it is necessary to measure the similarity between knowledge represented in the graphs. In this paper, we present semantic similarity measure to compare graph based representation of sentences. The proposed method incorporates computational linguistic method to obtain syntactical information prior to representation with graph. Word synonyms are embedded in the graph representation to support semantic matching. In this paper, we present our idea and initial results on the feasibility of the proposed similarity measurement method.**

Keywords— semantic similarity measure, graph based text representation, sentence similarity, word synonyms

I. INTRODUCTION

Various text mining applications have been developed to effectively overcome the problem of information overload and discovering hidden knowledge in text. With the growing text databases, it is inevitable to efficiently retrieve the relevant knowledge from the text contents. Moreover, such a large collection of text has brought problems in storage, management and retrieval.

The fundamental issue in retrieving textual data is the effective comparison of text using text similarity measurements. This is the basic component in a query which is either automatically generated by an application or manually specified by a user. A given query is compared with the information content in the text database in order to retrieve the most similar information as per the query. Text comparison is extremely important for a variety of text mining tasks such as text classification, clustering and novelty detection. There are many similarity measures for text such as word overlap measures, term frequency- inverse document frequency *(tf-idf)* measures and linguistic measures [1]. Most of the work done in this area is based on popular IR models (e.g. variants of *tf-idf* models as reported in [2] which treats the document and the query as vectors of term weights. However, these methods represent words separately without considering the context in which the words were used.

A variety of graphical text representation and network languages were employed to induce structure into documents and to model the semantics of natural language [3-5]. In [5], the researchers presented a graph-based similarity measurement based on the link structure of a document-concept bipartite graph while [3] proposed a term based graph similarity measure. A setback of these methods is the true dependency of words in sentence which may include semantic relations was not taken into consideration. Representing sentences using graphs captures the syntactical and semantics of natural language, however, existing text similarity measure need to be enhanced to compare graph structures.

Inspired by the idea behind the work of [4] where a graph based similarity measure is proposed on parsed text, we propose a graph based semantic similarity measures applicable for sentence level comparison. A computational linguistics-based method specifically deep parsing is performed to obtain the sentence structure. The sentence structure is represented as graphs that capture the semantics of sentences with word synonyms embedded in the graphs. One distinguishable difference of the proposed work in this research is the modeling of syntactically well-formed sentences as opposed to modeling short sentences, documents, phrases or words. Sentence level comparison provides huge opportunity to detect information contents relations in the documents which is essential for a variety of application. The findings of this study will be very beneficial for text mining applications since the performance of such application relies on the choice of an appropriate measure.

II. RELATED WORKS

Previous work [6] showed that the choice of similarity measure depends on the representation method that was employed to represent the text. Another factor is the level of text units. Text can be compared at document level, or sentence level, or phrase level or even word level. For document based similarity, the high number of overlapping words enables the detection of similar document easily. This is not the case for smaller text sentences. Two sentences can be semantically similar even if the number of overlapping words is low since different terms are used in the sentence to convey the same meaning. In this section we discussed a

Proc. of the Second Intl. Conf. on Advances in Computer and Information Technology -- ACIT 2013

ISBN: 978-981-07-6261-2 doi:10.3850/ 978-981-07-6261-2_42

measures and graph based sentence similarity measures.

A. *Sentence Semantic Similarity*

[7] presented a sentence semantic similarity measure based on segmented comparison. In this work a sentence is split into the main trunk and other segments. Each segment is further split into smaller segments. These segments are assigned different weights. The calculation of similarity involves grammatical orders. A setback of this method is the parameters need to be identified through experiments.

A bag of words approach is proposed in the work of [8] and similarity between words are calculated by determining certain word specificity. The method proposed in this work always chooses the maximum similarity between words, hence it tends to report higher similarity score than the actual similarity. Therefore the accuracy of the result will be affected. Furthermore this work did not take into account the whole sentence or the order of words in the sentence.

On the other hand [9] proposed a method that models the word order, semantic information and parts of speech of sentences. A Dynamic Time Wrapping (DTW) technique is also proposed in their work where the distance between sequences of word is taken into account. Another work that combines the word order information in calculating the semantic similarity is the work by [10]. Here both semantic and syntactic information is modeled and is proven to obtain better results. Hence the modeling of both syntactic and semantic information is more promising as the related work discussed here suggests.

B. *Graph based semantic similarity*

A number of semantic similarity measures are proposed in the literature for sentences which are represented in a graph based structure. [3] proposed a measure that incorporates paths in the graph structure and the depth of the nearest common ancestor. Their work focuses on protein terms represented using gene ontology and the proposed similarity measure is tightly related to gene ontology.

[5] proposed a unified graph model for document representation. Their method exploits Wikipedia as background knowledge and synthesizes both document representation and similarity computation. Their work compares text at the document level and it might not render accurate results for sentences.

In [13], a graph-based similarity measures was presented. Their work focused on extracting word synonyms from text which has undergone parsing prior to representation. Parsed text provides better opportunity to capture syntactical structures. Furthermore the connected dependency between parsed text enables text to be represented semantically using graph based structure. The work presented in [4 & 13] has inspired us to represent graph based structure from parsed text. The next section details our proposed method.

The proposed method contains 3 steps. It starts with sentence parsing to obtain the syntactic information. Next the syntactic information is used to construct a graph representation of the sentence. A graph based similarity measure is then formalized to calculate the similarity of sentences. In this section we elaborate these three steps.

A. *Sentence Syntactical Structure*

The first step in the proposed method is to obtain the syntactical structure of sentences. To achieve this purpose, a parsing tool is used. Typically the parsing technique ranges from simple part-of-speech tagging to more advanced techniques such as the stochastic approach. In this study Link Grammar Parser (LGP) [11] is used to parse the identified sentence to obtain sentence structure. LGP is a formal grammatical system to produce syntactical relations between words in a sentence. The parser is able to determine the syntactical structure of sentences by dividing sentences (S) into noun phrases (NP), verb phrases (VP), preposition phrases (PP), and adverb phrases (ADVP). LGP is used in this work because [12] reported that it provides a much deeper semantic structure than the standard context-free parsers. Fig. 1 shows the linguistic structure produced after parsing the example text using LGP.

```
Raw sentence:
This risk refers to volatility in the net profit income

Parsed sentence:
   [S [NP This risk NP]
      [VP refers
         [PP to
            [NP volatility NP] PP]
               [PP in
                  [NP the net profit income NP] PP] VP] S]
```

Figure 1. Example of a sentence structure.

B. *Graph-based sentence representation*

The construction of sentence graph involves using the syntactic information from the parsed text combined with the list of relation transformations. Some of the rules that were used are:

- the nouns, verbs, adjectives are generally represented as concepts, while
- the prepositions are represented as the relationship between the concepts
- additional relations are defined such as agent, object, and attribute.

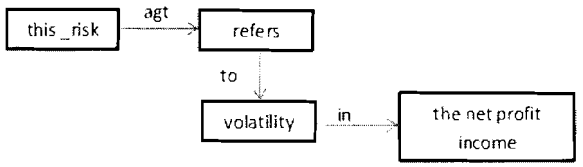
Generally a directed simple graph $G = (V,E)$ consists of V, a nonempty set of vertices, and E, a set of ordered pairs of distinct elements of V called edges where $V = (V_1, V_2 \ldots)$ and $E = \{e_1,e_2,e_3 \ldots e_n\}$ and $e_i = (V_i, V_j)$.

Proc. of the Second Intl. Conf. on Advances in Computer and Information Technology -- ACIT 2013

ISBN: 978-981-07-6261-2 doi:10.3850/ 978-981-07-6261-2_42

The notation used in this work is a directed graph $G= (V,E)$ where its vertex set $V = \{V_1, V_2 .. \}$ i.e. set of concepts in the sentence such that every edge $e_i \in E$ in the graph connects a concept V_i and a concept V_j. Therefore, $e_i = (R,V_i,V_j)$ where $R = \{R_1, R_2\}$ i.e. a set of relation in the sentence.

The example sentence structure from Fig. 1 consists the following set of vertices, V = {this risk, refers, volatility, the net profit income}, and the following set of edges E = {(agent, this risk, refers), (to, refers, volatility), (in, volatility, the net profit income) where the relation set, R = (agent, to, in). Fig. 2 shows the constructed graph in both graphical and linear form.



G_1
({[this_risk*c1],[refers*c2],[volatility*c3],[the_net_profit_income*c4]},{[agt,c1,c2],[to,c2,c3],[in,c3,c4]})

Figure 2. Example of a sentence graph structure.

C. *Graph Semantic Similarity*

To perform semantic matching of the graphs, the generated graphs were embedded with concept synonyms. A predefined dictionary extracted from *Wordnet* [14] is referred to accomplish this purpose. To perform this module, each concept in the concept list V is compared to the words in the *Wordnet* entry. The synonym corresponding to each matching concept is retrieved and embedded as a synonym list in the concept list set. The original graphs were enhanced with additional embedding of concept synonyms in the concept list set, V of the graph notation. As a result the graph G_1 in Fig. 2 is enhanced into graph GS_1 as shown in Fig. 3.

GS_1 ({[this_risk, danger, endangerment, hazard, jeopardy, peril*c1],[refers, cites, concerns, denotes, mentions, pertains, relates*c2],[volatility, unpredictability*c3],[the_net_profit_income, the financial gain, the earnings, the benefits*c4]},{[agt,c1,c2],[to,c2,c3],[in,c3,c4]})

Figure 3. Example of a synonym embedded sentence graph structure.

With the synonym embedded concepts set in the proposed graph structure, two sentences which use two different terms to describe the same meaning will be measured as similar. Once all the graphs have been embedded with the concept synonyms, graph comparison can be performed. Given two synonym embedded graphs GS_1 and GS_2, the graph semantic sentence similarity $G\text{-}Sim\ (GS_1, GS_2)$ is calculated as

$$G\text{-}Sim\ (GS_1, GS_2) = \frac{|(GS_1 \cap GS_2)|}{|(GS_1 \cup GS_2)|} \quad (1)$$

Where GS_1 and GS_2 is the graph to be compared. The interval for value of the similarity measure, $G\text{-}Sim\ (GS_1, GS_2)$ is [0-1] that is, 1, if both graphs are similar and 0 if both graphs are dissimilar.

IV. EVALUATION

In this section we present an initial experiment on selected sentences to evaluate the proposed method. This experiment acts as a pilot study to measure the feasibility of the proposed method in correctly identifying the similarity between sentences. In this experiment, 3 samples consisting 6 sentences were carefully selected. All the sentences were annotated by human expert who determines the semantic similarity of the selected sentences. Table 1 presents the extracted sentences used in this experiment.

TABLE I. SAMPLE SENTENCES

Graph ID	Sample Sentences
$GS_{1,1}$	Basic earnings per share of the Bank are calculated by dividing the net profit for the financial year by the weighted average number of ordinary shares in issue during the financial year
$GS_{1,2}$	Basic loss per share of the Bank is calculated based on the net loss attributable to the ordinary shareholders of RM1,296,789,000 and the weighted average number of ordinary shares outstanding during the year
$GS_{2,1}$	Total assets and liabilities transferred were approximately RM1.836 billion and RM1.837 billion respectively.
$GS_{2,2}$	Total assets dropped 8% to RM14.61 billion in comparison to RM15.85 billion recorded in the previous financial year.
$GS_{3,1}$	The total number of shares to be offered shall not exceed 10% of the issued and paid up share capital of the Company at any point of time during the tenure of ESOS
$GS_{3,2}$	The total number of shares to be offered under the ESOS shall not exceed 10% of the issued and paid-up share capital of the Company at any point in time during the duration of the share option schemes.

The first pair of sentences ($GS_{1,1}$ and $GS_{1,2}$) is selected because they contain similar syntactical structure but semantically both sentences are dissimilar. The second pair of sentences ($GS_{2,1}$ and $GS_{2,2}$) is selected because they are syntactically and semantically different except for some identical word usage. The third pair of sentences ($GS_{3,1}$ and $GS_{3,2}$) is syntactically and semantically similar. The human expert has recognized first and second pair as dissimilar while the third pair as similar. The selected sentences were parsed, transformed into graphs and measured with the proposed similarity measurement.

To evaluate the performance of the proposed method, the calculated similarity score is compared to the human decisions. Beside that a simple word overlap measure based on dice coefficient was calculated for comparison purpose as well. The result of the comparison is shown in Table II.

Proc. of the Second Intl. Conf. on Advances in Computer and Information Technology -- ACIT 2013

ISBN: 978-981-07-6261-2 doi:10.3850/ 978-981-07-6261-2_42

TABLE II. COMPARISON RESULTS

Graph ID	Human	Word Overlap	*G-Sim*
$GS_{1\,1}$ & $GS_{1\,2}$	0	0.85	0.16
$GS_{2\,1}$ & $GS_{2\,2}$	0	0.32	0.13
$GS_{3\,1}$ & $GS_{3\,2}$	1	0.79	0.9

As can be seen in Table II, the similarity score of the proposed method, *G-Sim* is more correlated to the human decisions as compared to the word overlap method. The result revealed the following findings:

- For sentences which are syntactically similar, the proposed method is able to differentiate the semantics of the sentences.
- For sentences which are syntactically and semantically different the proposed method is able to differentiate both syntactic and semantic differences.
- For sentences which are syntactically and semantically similar, the proposed method is able to identify the similarities.
- Word overlap method failed to capture the similarity between semantically similar sentences.

This experiment is an initial experiment to evaluate the feasibility of the proposed graph based semantic similarity measure. For a complete evaluation the proposed method need to be compared with existing graph based semantic similarity measure as proposed in [13 & 15]. This will be our future work.

V. CONCLUSION

This paper presents a graph based semantic similarity measure for sentence comparison. The proposed method has proven that exploiting whole sentences and representing it with graph structure renders significant improvement. One important contribution in this work is the explicitly embedding of concept synonyms into the graphs. This enables the semantic matching of sentences. Although the performance of the proposed method is not fully evaluated on larger dataset, the result of initial experiment is presented. The experimental results show that the proposed method performs substantially better than the word overlap method. One of the reason for good performance is the proposed method incorporates linguistic computation to obtain the sentence structure prior to representing it as graphs. Hence it is able to capture both the syntactical and semantics of sentences. Our future work will be to enhance the proposed method and to evaluate its performance by experimenting on larger dataset and comparing it with state of the art graph semantic similarity measures.

ACKNOWLEDGMENT

We would like to thank the Universiti Utara Malaysia for funding the research under the LEADS grant scheme.

REFERENCES

[1] P. Achananuparp, X. Hu, and X. Shen, "The evaluation of Sentence Similarity Measures," Data Warehousing and Knowledge Discovery, Lecture Notes in Computer Science, vol. 5182, 2008, pp. 305-316.

[2] G. Salton, and C. Buckley, "Term weighting approaches in automatic text retrieval", Information Processing and Management vol. 24 no. 5, 1988, pp. 513-523.

[3] M. A. Alvarez, and C. Yan, "A Graph-Based Semantic Similarity Measure for the Gene Ontology", Journal Of Bioinformatics And Computational Biology, vol. 9, no. 6, Imperial College Press, 2011, pp. 681-695.

[4] E. Minkov and W. W. Cohen., "Learning Graph Walk Based Similarity Measures for Parsed Text", In Conference on Empirical Methods in Natural Language Processing (EMNLP), 2008.

[5] L. Zhang, C. Li,. J. Liu, and H. Wang, "Graph-Based Text Similarity Measurement *by*. Exploiting Wikipedia as Background Knowledge", World Academy of Science, Engineering and Technology, vol. 59, 2011, pp. 1548- 1553.

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

[7] Y. Liu, and Y. Liang, "A Sentence Semantic Similarity Calculating Method based on Segmented Semantic Comparison", Journal of Theoritical and Applied Information Technology, vol. 48, no. 2, 2013, pp 231- 235.

[8] R. Mihalcea, C. Corley, and C. Strapparava, "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.

[9] X. Liu, Y. Zhou, and R. Zheng, "Sentence Similarity based on Dynamic Time Warping", International Conference on Semantic Computing, ICSC2007, 2007, pp. 250-256.

[10] X. Liu, Y. Zhou, and R. Zheng, "Measuring Semantic Similarity within Sentences", Proceeding of ICMLC2008 Conference, Kunming, 2008.

[11] D. Sleator, and D. Temperley, "Parsing English with a link grammar", 3rd Int. Workshop of Parsing Technologies, 1993.

[12] F. M. Suchanek, G. Ifrim and G. Weikum, "Combining Linguistic and Statistical Analysis to Extract Relations from Web Documents", SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006.

[13] E. Minkov and W. W. Cohen, "Graph based similarity measures for synonym extraction from parsed text", In Workshop Proceedings of TextGraphs-7 on Graph-based Methods for Natural Language Processing (TextGraphs-7 '12). Association for Computational Linguistics, Stroudsburg, PA, USA, 2012, pp. 20-24.

[14] G. A. Miller, "WordNet: A Lexical Database for English", Communications of the ACM vol. 38, no.11, 1995, pp 39-41.

[15] L. Lin, X. Hu, B. Hu, J. Wang, "Measuring sentence similarity from different aspects", The Eighth International Conference on Machine Learning and Cybernetics, Baoding, Hebei, China, 2009, pp. 2244-2249.