Sentence Similarity based on Relevance

Jin Feng
Beihang University
jesebel33@sina.com

Yiming Zhou
Beihang University
zhouyiming@buaa.edu.cn

Trevor Martin
University of Bristol
Trevor.Martin@bristol.ac.uk

Abstract

The issue about sentence similarity is essential to many areas of artificial intelligence. Although there are related studies on determining text similarity, fewer publications are about the similarity between short texts especially about the similarity between sentence pairs. In this paper, a novel method is proposed to estimate the sentence similarity with consideration of *direct relevance* and *indirect relevance* between sentence pairs. Analysis of experiment provides insights into the strengths and weakness of our proposed method.

Keywords: Relevance, Sentence Similarity, Information Retrieval, Information Processing, Uncertainty.

1 Introduction

Nowadays, people are surrounded by huge amount of information especially with the development of the internet. More and more techniques are developed to help people manage and process information. Many research themes in the field of artificial intelligence are emerging under this environment, for example, information retrieval, information extraction, information filtering, machine translation, question answering, and so on. And one of the key problems of these themes is the similarity which has close relationship with psychology and cognitive science. In this paper, we focus on the research of sentence similarity within the field of text retrieval.

Text retrieval is the basic research area of information retrieval, and currently most of the research about similarity used in text retrieval is on the text paragraph or whole document level. Many approaches are proposed to determine the text similarity. Some of those studies interpret the similarity from the pure mathematical perspective based on statistics or probability theory [10], some estimate it from the perspective of semantics contained in the paragraph or whole document using lexical resource [17, 18], and some other methods combine the ideas mentioned above to achieve the goal [16]. However, none of those approaches are suitable for short text retrieval especially on the sentence level. There are three drawbacks when we adopt those measures [22]. First, some methods represent a sentence as a high-dimensional vector which leads to the sparse data problem and computational inefficiency [20]. Second, some methods are not totally automatic and need active human involvement [9]. Third, some methods are domain-limited, and cannot be applied in general.

The measurement of sentence similarity is in essence an uncertainty problem. Because human judge the similarity between sentence pairs not only based on the sentences themselves but also considering some potential related information which is always situation and time dependent. How to capture this potential information is a problem because the working mechanism of the human brain is still unknown. But just as the great statistician

L. Magdalena, M. Ojeda-Aciego, J.L. Verdegay (eds): Proceedings of IPMU'08, pp. 832–839
Torremolinos (Málaga), June 22–27, 2008

Rao said: uncertain knowledge plus uncertainty measure equals useful knowledge [3]. It is a great challenge and a great chance.
In this paper, we propose a novel method to estimate sentence similarity. The approach measures sentence similarity considering the *direct relevance* and *indirect relevance* between sentences. We will introduce our method in section 2. In section 3, experiments are conducted to evaluate our method. Finally, we discuss the related work and future direction of this study.

2 The Proposed Method

Human usually make judgement on the similarity between two concepts with consideration of their direct relevance and indirect relevance. The direct relevance is the means by which a human can get the obvious coherence between two concepts, whereas the indirect relevance is the means by which a human can get some potential relatedness between two concepts. A sentence tends to be about a single topic whereas the whole document are usually concerned with a variety of topics. And a unique topic always can be treated as a complex concept. So in this paper we estimate the similarity between two sentences $s = f(s_1, s_2)$ with consideration of these two factors: direct relevance and indirect relevance, i.e., the sentence similarity is the function of direct relevance and indirect relevance and it can be written as follows:

$$s = f(s_1, s_2) = f(dr, indr) \qquad (1)$$

where dr is the direct relevance between sentence pairs, $indr$ is the indirect relevance between sentence pairs.
Furthermore, if we assume that these two factors are independent to each other, (1) can be rewritten as:

$$f(s_1, s_2) = f(f_1(dr), f_2(indr)) \qquad (2)$$

So in order to obtain the sentence similarity, first we have to compute $f_1(dr)$ and $f_2(indr)$ respectively.

2.1 Direct Relevance between Sentences

From a general perspective, the direct relevance information between two concepts usually provides us with specific details about the concepts and has most significant impact on the determination of similarity between them. Towards our specific task, the similarity between sentence pairs, we treat the semantic similarity between two sentences as an indicator of the direct relevance. Because humans always make judgement about the similarity between two sentences mainly based on their semantic relations, so it is appropriate using the semantic similarity between two sentences as an indication of the direct relevance between sentences.
Words are the basic units making up a sentence, and the semantic similarity between sentences are always closely related to the semantic similarity between words.

2.1.1 Semantic Similarity between Words

Usually the semantic similarity between word pairs is represented by the semantic similarity between their related concepts. Several methods about the semantic similarity measurement between words have been proposed. Although all of these approaches are lexical resources based, the exact resource they are based is not the same, some are dictionary based [12], some are thesaurus based [7, 15], and others are wordnet based. All of the methods intend to use lexical resource as a network or directed graph and measure the semantic similarity based on properties of the network or graph.
WordNet[5] is a machine-readable lexical database which is organized by meanings and developed at Princeton University. The words in Wordnet are classified into four categories, nouns, verbs, adjectives and adverbs respectively. WordNet groups these words into sets of synonyms called synsets, provides short definitions, and records the various semantic relations between these synonym sets. Synsets are interlinked by means of conceptual-semantic and lexical relations.

The most common relationships include Hyponym/Hypernym (i.e., is-a relationships) and Meronym/Holonym (i.e., part-of relationships). Nouns and verbs are organized into hierarchies based on the hyponymy/hypernym relation between synsets while adjectives and adverbs are not. Figure 1 is part of the hierarchy of WordNet.

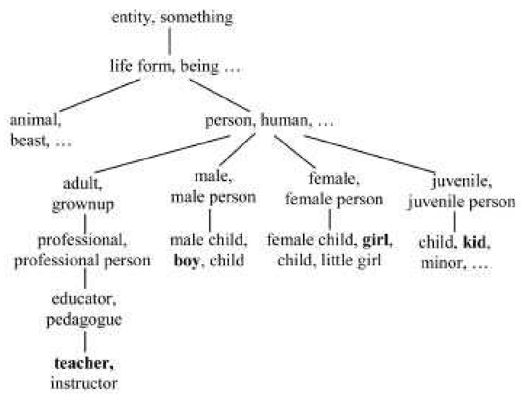

Figure 1: Part of WordNet

Several approaches have been proposed based on and tested on wordNet. Basically, we can classify these methods into four categories: edge counting methods [14, 21], information content methods [8, 13], feature based methods [1] and hybrid methods [16].
In this paper, we estimate the semantic similarity between two words adopting the method proposed in [23], which estimate the semantic similarity between two words based on the depth of the two words in WordNet and that of their least common subsumer (LCS). The LCS does not necessarily feature in the shortest path connecting the two senses, as it is by definition the common ancestor deepest in the taxonomy, not closest to the two senses.
Given two words, w_1 and w_2, the semantic similarity $s(w_1, w_2)$ is the function of their depth in the taxonomy and the depth of their least common subsumer. If d_1 and d_2 are the depth of w_1 and w_2 in WordNet, and h is the depth of their least common subsumer in WordNet, the semantic similarity [23] can be written as:

$$s(w_1, w_2) = \frac{2.0 * h}{d_1 + d_2} \qquad (3)$$

2.1.2 Semantic similarity between Sentences

In this paper we measure semantic similarity between sentence s_1 and s_2 by first considering two conditional similarities, more specifically, we use conditional similarities $s(s_1|s_2)$ and $s(s_2|s_1)$ instead of computing $s(s_1, s_2)$ directly. The method is explained below.

Enhancing sentence Since usually sentences tend to be short they must be enhanced. We only enhance the sentence with the synonyms of first sense of words appearing in the sentence. There are two reasons for our decision. The first is to avoid introducing noisy words, because we use the semantic similarity as an indication of direct relevance between sentences and commonly the direct relevance needs to be discriminative. The second is for simplicity because most of the words use their first sense in context.
In this paper, we only take nouns, verbs, adjectives and adverbs into consideration. Given two sentences, s_1 and s_2, denoted as:

$$S_1 = \{w_1, w_2, \ldots, w_n\}$$

$$S_2 = \{q_1, q_2, \ldots, q_m\}$$

The word sets S_1 and S_2 contain all the distinct words from s_1 and s_2. Furthermore the sentences s_1 and s_2 can be represented as:

$$S_1 = \{\mathrm{NN}_1, \mathrm{VB}_1, \mathrm{ADJ}_1, \mathrm{ADV}_1\}$$

$$S_2 = \{\mathrm{NN}_2, \mathrm{VB}_2, \mathrm{ADJ}_2, \mathrm{ADV}_2\}$$

where $\mathrm{NN_i}$, $\mathrm{VB_i}$, $\mathrm{ADJ_i}$, $\mathrm{ADV_i}$ $(i = 1, 2)$ are the word sets of nouns, verbs, adjectives and adverbs of s_1 and s_2 respectively.
The enhanced sentence is comprised of the words in sentence and their corresponding synonyms. Take s_1 for example, its corresponding enhanced sentence denoted as s_1' is the set of words w_k $(k = 1, \ldots, n)$ in sentence s_1 and the synonyms of w_k.

$$s_1' = \{\mathrm{NN}_1', \mathrm{VB}_1', \mathrm{ADJ}_1', \mathrm{ADV}_1'\}$$

where NN_1',VB_1',ADJ_1', ADV_1' separately stand for the union set of NN_1, VB_1, ADJ_1

and ADV_1 and their corresponding synonym words set.
The same operation can be performed on sentence s_2. So after this preprocess we get two enhanced sentences s'_1 and s'_2.

Computing conditional similarity The computation of conditional similarity can be summarized into the following three steps. Take $s(s_1|s_2)$ for example.
Step1. For each word w_i in NN'_1 compute its relative significance on sentence s_2, which is estimated by the formula:

$$sig(w_i|s_2) = \begin{cases} \frac{\sum_{q_k \in \mathrm{NN}'_2} s(w_i,q_k)}{wordpairnumber} & |\mathrm{NN}'_1|>0, |\mathrm{NN}'_2|>0 \\ \eta & |\mathrm{NN}'_1|>0, |\mathrm{NN}'_2|=0 \end{cases} \quad (4)$$

where w_i and q_k have the same Part-Of-Speech (POS) type. $s(w_i, q_k)$ is determined using formula (3). *wordpairnumber* records the frequency of word pairs used in calculation. The value of η is empirically set to 0.05. The same operation is carried on VB'_1,ADJ'_1 and ADV'_1. So we can get the significance of each word in sentence s'_1 on condition of sentence s'_2.
Step2. Compute the relative significance of NN'_1, VB'_1,ADJ'_1 and ADV'_1 . Take the relative significance of NN'_1 for example, the computation of the other three subword sets is the same.

$$sig(\mathrm{NN}'_1|s_2) = \sum_{w_i \in \mathrm{NN}'_1} \frac{1}{|\mathrm{NN}'_1|} sig(wi|s_2) * IC(w_i) \quad (5)$$

where $|\mathrm{NN}'_1|$ is the size of NN'_1, $IC(w_i)$ is the information content of word w_i and is determined by the following formula:

$$IC(w) = 1 - \frac{\log(n+1)}{\log(N+1)} \quad (6)$$

where n is the frequency of word w in the corpus, N is the total number of words in the corpus. These are increased by 1 for smoothing.
Step3. Compute conditional similarity $s(s_1|s_2)$ by:

$$\begin{aligned} s(s_1|s_2) = \quad & \alpha(sig(\mathrm{NN}'_1|s_2) + sig(\mathrm{VB}'_1|s_2)) \\ & +\beta(sig(\mathrm{ADJ}'_1|s_2) + sig(\mathrm{ADV}'_1|s_2)) \quad (7) \end{aligned}$$

where α indicates the significance of NN'_1 and VB'_1, β indicates the significance of ADJ'_1 and ADV'_1. We empirically set $\alpha = 1.0$ and $\beta = 0.8$.

The direct relevance between sentences The amount of information contained in a sentence is sensitive to the length of the sentence, which has an impact on the direct relevance between sentences. We define the direct relevance between sentence s_1 and s_2 as:

$$f_1(dr) = \frac{|s'_1|}{|s'_1| + |s'_2|} s(s_1|s_2) + \frac{|s'_2|}{|s'_1| + |s'_2|} s(s_2|s_1) \quad (8)$$

where $|s'_1|$ and $|s'_2|$ are the length of enhanced sentence s'_1 and s'_2.

2.2 Indirect Relevance between Sentences

The indirect relevance between two concepts always supplies the potential relationships, for example, reasoning and referring. So we think that the indirect relevance between sentences provides information about the potential relationships between sentence pairs which is important on the estimation of similarity between sentence pairs. We believe that words in sentence have potential relationship especially the nouns and verbs, but the relationship is always situation dependent and hard to compute with WordNet. Because every sentence is a unique words sequence and this sequence indicates the potential relationship among those words. So we assume that the relation between two words sequences can be an indication of the indirect relevance between sentences. And in this paper, we adopt the Needleman-Wunsch algorithm [19] to estimate the indirect relevance between sentence pairs.
The Needleman-Wunsch algorithm was commonly used in bioinformatics to aligns protein or nucleotide sequences. It was used to find the best global alignment of any two sequences. It is an example of dynamic programming, and was the first application of dynamic programming to biological sequence comparison. The main idea of the Needleman-Wunsch algorithm is to align

the two sequences with gaps to achieve the greatest number of matches by using dynamic programming to efficiently implement a recursion. In order to make the Needleman-Wunsch algorithm appropriate to words sequence, we make some modifications.
Given two input strings X and Y with length of M and N respectively, a two-dimensional matrix called F is allocated. The entry $F_{i,j}$ is the score of the optimal alignment of $\mathrm{x}_{[1...\mathrm{i}]}$ and $\mathrm{y}_{[1...\mathrm{j}]}$. The algorithm proceeds to fill the matrix from top left to bottom right in order to get $F_{M,N}$, which is a measurement of the similarity between two sequences. The formula is as follows:

$$F_{i,j} = \max \begin{cases} F_{i-1,j-1} + s(\mathrm{x_i}, \mathrm{y_j}) \\ F_{i-1,j} - d \\ F_{i,j-1} - d \end{cases} \quad (9)$$

and

$$F_{0,0} = 0$$

where $s(\mathrm{x_i}, \mathrm{y_j})$ is the similarity of words $\mathrm{x_i}$ and $\mathrm{y_j}$. Here, we set $s(\mathrm{x_i}, \mathrm{y_j})$ as:

$$s(\mathrm{x_i}, \mathrm{y_j}) = \begin{cases} \theta s(\mathrm{x_i}, \mathrm{y_j}) & \text{if } s(\mathrm{x_i}, \mathrm{y_j}) \geq \zeta \\ -1 & \text{if } s(\mathrm{x_i}, \mathrm{y_j}) < \zeta \end{cases} \quad (10)$$

where $s(\mathrm{x_i}, \mathrm{y_j})$ can be computed using formula (3), ζ in our experiments is empirically set to 0.2, value of θ is dependent on the POS of x_i and y_j. If x_i and y_j are both nouns or verbs then the value of θ is 1.2, others is 1. θ is used to reflect the significance of nouns and verbs. d is a penalty for the gap between sequences, since we want to capture the words potential relationships based on the words sequence, here we set the penalty for gap to be average semantic similarity score between sentence pairs, i.e $d = -\frac{\sum s(w_1,w_2)}{totalwordpairs}$, and $s(w_1, w_2)$ is determined by formula (3).
Since we hope to get a value estimating the indirect relevance between sentences in interval $[0, 1]$, so we have to make some normalization. We measure the indirect relevance between sentences using the following formula:

$$f_2(indr) = \frac{F_{|s_1|,|s_2|}}{max(|s_1|, |s_2|)} \quad (11)$$

where $|s_1|$ and $|s_2|$ are the length of sentence s_1 and s_2.

2.3 Sentence Similarity

With the consideration of the two forementioned factors, more specifically, direct relevance and indirect relevance of sentence, we set the overall sentence similarity as:

$$\begin{aligned} s &= f(dr, indr) \\ &= \lambda f_1(dr) + (1-\lambda) f_2(indr) \\ &= \lambda(\tfrac{|s_1'|}{|s_1'|+|s_2'|} s(s_1|s_2) + \tfrac{|s_2'|}{|s_1'|+|s_2'|} s(s_2|s_1)) \\ &+ (1-\lambda)\tfrac{F_{|s_1|,|s_2|}}{max(|s_1|,|s_2|)} \end{aligned} \quad (12)$$

where $\lambda \in [0, 1]$ decides the relative contribution of direct and indirect relevance to the sentence similarity. Since we assume the direct relevance is more important than the indirect relevance, so $\lambda \in [0.5, 1]$.

3 Experiment Results

WordNet and Brown Corpus [2] are used in the implementation of our method. We use WordNet as the main semantic knowledge base to get lexical taxonomy information and derive statistics from the Brown Corpus. In our experiments, only "IS-A" relation is considered.

3.1 Experimental Materials

The goal of our project is to produce a measure of similarity between sentences. However, currently there is no suitable benchmark data sets for the evaluation of sentence similarity methods.
[22] collected human ratings for the similarity of pairs of sentences following existing design for word similarity measures. They used 65 noun word pairs whose semantic similarity was originally measured by Rubenstein and Goodenough [6] and these data has been used in many experiments in the intervening years.
[22] replaced these 65 noun word pairs with their definitions from the Collins Cobuild dictionary [11], which is constructed using

information from the Bank of English corpus. The dictionary contains 400 million words and more than one sense of a word was given, [22] chose the first sense in the list for the 65 noun word pairs for test. The complete sentence data set used in this study is available at http://www.docm.mmu.ac.uk/STAFF/D.McLean/SentenceResults.htm.

Each of the 65 sentence pairs was assigned a semantic similarity score calculated as the mean of the judgements made by the participants. The distribution of the semantic similarity scores was heavily skewed toward the low similarity end of the scale. Following a similar procedure to Miller and Charles [4], a subset of 30 sentence pairs was selected to obtain a more even distribution across the similarity range. This subset contains all of the sentences pairs rated 1.0 to 4.0 and 11 (from a total of 46) sentences rated 0.0 to 0.9 selected at equally spaced intervals from the list. These can be seen in Table 1, where all human similarity scores are provided as the mean score for each pair and have been scaled into the range $[0, 1]$.

Comparing the word-pair ratings from Rubenstein and Goodenough with the corresponding sentence-pair ratings, it is apparent that people perceive the semantic similarity of words differently from their definition. Take Midday & noon and Gem & jewel for example, their corresponding semantic similarity in Rubenstein and Goodenough are both 3.94, but the scores of sentence pairs are 0.96 and 0.65. A possible reason for this is that the judgements of human on similarity between sentence pairs are easy affected by the relationships of the words contained in sentences and some other background knowledge related with human being himself/herself. Additionally, this type of information is always situation and time dependent.

3.2 Results and Discussion

The parameter λ for weighting the significance between direct relevance and indirect relevance is set to 0.85, because the indirect relevance plays a subordinate role in the similarity of sentence.

Table 1: Sentence Data Set Results.

W.P.	**H.S.**	$\lambda = 1.0$	$\lambda = 0.85$
Cord&smile	0.01	0.17	0.15
Autograph&shore	0.01	0.29	0.28
Asylum&fruit	0.01	0.35	0.31
Boy&rooster	0.11	0.42	0.40
Coast&forest	0.13	0.12	0.13
Boy&sage	0.04	0.34	0.36
Forest&graveyard	0.07	0.23	0.23
Bird&woodland	0.01	0.16	0.16
Hill&woodland	0.15	0.22	0.21
Magician&oracle	0.13	0.34	0.31
Oracle&sage	0.28	0.21	0.20
Furnace&stove	0.35	0.29	0.29
Magician&wizard	0.36	0.35	0.36
Hill&mound	0.29	0.18	0.18
Cord&string	0.47	0.56	0.50
Glass&tumbler	0.14	0.29	0.27
Grin&smile	0.49	0.48	0.43
Serf&slave	0.48	0.54	0.49
Journey&voyage	0.36	0.34	0.32
Autograph&signature	0.41	0.33	0.30
Coast&shore	0.59	0.32	0.31
Forest&woodland	0.63	0.26	0.25
Implement&Tool	0.59	0.27	0.25
Cock&rooster	0.86	0.97	0.92
Boy&lad	0.58	0.66	0.61
Cushion&pillow	0.52	0.28	0.29
Cemetery&graveyard	0.77	1.0	0.91
Automobile&car	0.56	0.5	0.45
Midday&noon	0.96	1.0	0.99
Gem&jewel	0.65	0.71	0.64

Table 1 shows the average human ratings in the column Human similarity (mean) (H.S.). Our algorithm results are listed in the third colum and fourth column with parameter $\lambda = 1.0$ and $\lambda = 0.85$. Table 2 shows the performance of the proposed similarity measurement. The Pearson correlation coefficient with the human ratings for the experiment with parameter $\lambda = 1.0$ and $\lambda = 0.85$ are 0.754 and 0.756 respectively, significant at the 0.01 level. Though the difference between the two correlation coefficients is slight, it shows that the indirect relevance has a subordinate

Table 2: Similarity Correlations

	correlation
$s(s_1, s_2)(\lambda = 1.0)$	0.754
$s(s_1, s_2)(\lambda = 0.85)$	0.756

affect on the the sentence similarity determination.

Most of the proposed methods are sensitive to the accuracy of POS tagging, but since we estimate the similarity by two conditional similarities, it weakens the sensitivity of our method. The performance of our method is sensitive to three factors. The first one is the process of sentence enhancement. Because we enhance a sentence by the synonyms of words appearing in the sentence, it is easy to introduce noisy words. The estimation between Forest and Woodland is an example. The second one is that we only used the first sense of words in computation which sometimes is not consistent with the reality and makes the performance bad, just as the estimation for Cock & rooster. The third one is that it is hard to capture the relationship between words which have different POS type within WordNet. But this information is very important to the estimation of direct and indirect relevance between sentence pair.

4 Conclusion and Future Work

This paper presented a novel method to measure sentence similarity with consideration of direct relevance and indirect relevance of sentence pairs. Furthermore an improved Needleman-Wunsch algorithm was proposed. The best alignment was used to estimate the indirect relevance between sentence pairs. Experimental results show that indirect relevance plays a subordinate role in determining sentence similarity and the performance of our method is acceptable.

In the future, we are interested in using fuzzy set theory in estimation of direct and indirect relevance between sentences.

References

[1] A. Tversky. Features of Similarity. Psychological Review, 84(4):327-352, 1977.

[2] Brwon Corpus Information. http://nltk.sourceforge.net/index.php/Main_Page, 2005.

[3] C. R. Rao. Statistics and Truth: Putting Chance to Work. 2nd ed., World Scientific, 1997.

[4] G. A. Miller and W. G. Charles. Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*, vol.6, no.1, pages 1-28, 1991.

[5] G. A. Miller. WordNet: A Lexical Database for English. *Comm.*ACM, vol.38, no.11, pages 39-41, 1995.

[6] H. Rubenstein and J. B. Goodenough. Contextual Correlates of Synonym. *Communications of the ACM*, vol.8, no.10, pages 627-633, 1965.

[7] Jarmasz, Mario and Stan Szpakowicz. Roget's Thesaurus and semantic similarity. In Proceedings of the International Conference on Recent Advances in Natural Language Processing, pages 212-219, Borovetz, Bulgaria, Sep. 2003.

[8] Jiang, Jay J. and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceeding of International Conference on Research in Computional Linguistics*, pages 19-33, Taiwan, 1997.

[9] J. L. McClelland and A. H. Kawamoto. Mechanisms of senntence processing: assigning roles to constituents. Parallel distributed processing: explorations in the microstructure of cognition, vol.2, pages 272-325, MIT Press, 1998.

[10] J. Ponte and W. B. Croft. A Language Modeling Approch to Information Retrieval. In *Proceedings on the 21st annual international ACM SIGIR conference*, pages 275-281, 1998.

[11] J. Sinclair, ed.. *Collins Cobuild English Dictionary for Advanced Learners*. Harper Collins Pub., 2001.

[12] Kozima, Hideki and Teiji Furugori. Similarity between words computed by spreading activation on an English dictionary. In Proceedings of 6^{th} Conference of the European Chapter of the Association for Computational Linguistics, pages 232-239, Utrecht, 1993.

[13] Lin, Dekang. An information-theoretic definition of similarity. In *Proceeding of the 15th International Conference on Machine Learning*, pages 296-304, Madison, WI, July, 1998b.

[14] Liu XiaoYing, Zhou Yiming, and Zheng Ruoshi. Sentence Similarity based on Dynamic Time Warping. International Conference on Semantic Computing, pages 250-256, Sept., 2007.

[15] Morris, Jane and Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics, 17(1):21-48, 1991.

[16] M. Rodriguez and M. Egenhofer. Determining Semantic Similarity Among Entity Classes from Different Ontologies. *IEEE Trans. On Knowledge and Data Engineering*, 15(2), pages 442-456, March/April, 2003.

[17] Rada R, Mili H, Bicknell E and Blettner M. Development and Application of a Metric on Semantic Nets. *IEEE Transactions on Systems, Man, and Cybernetics*, vol.19, issue 1, pages 17-30, 1989.

[18] Resnik, Philip. Using information content to evlaute semantic similarity. In *Proceeding of the 14th International Joint Conference on Artificial Intelligence*, pages 448-453, Montreal, Canada, August, 1995.

[19] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol.48(3), pages 443-453, 1970.

[20] T. K. Landauer, P. W. Foltz, and D. Laham. Introduction to Latent Semantic Analysis. Discourse Processes, volume 25, nos.2-3, pages 259-284, 1998.

[21] Y. H. Li, Z. Bandar, and D. McLean. An Approach for Measuring Semantic Similarity Using Multiple Information Sources. *IEEE Trans. Knowledge and Data Eng.*, vol.15, no.4, pages 871-882, July/Aug., 2003.

[22] Yuhua Li, David McLean, Zuhair A. Bandar, James D. O'Shea, and Keeley Crockett. Sentence Similarity Based on Semantic Nets and Corpus Statistics. *IEEE Transaction of Knowledge and Data Engineering*, volume 18, No.8, pages 1138-1150, August, 2006.

[23] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computional Linguistics*, pages 133-138, Morristown, NJ, USA, 1994.