MEASUREMENT OF SEMANTIC SIMILARITY BETWEEN WORDS: A SURVEY

Ankush Maind[1], Prof. Anil Deorankar[2] and Dr. Prashant Chatur[3]

[1]M.Tech. Scholar, Department of Computer Science and Engineering,
Government College of Engineering, Amravati, Maharashtra, India
ankushmaind@gmail.com
[2]Associate Professor, Department of Computer Science and Engineering,
Government College of Engineering, Amravati, Maharashtra, India
avdeorankar@gmail.com
[3]Head of Department, Department of Computer Science and Engineering,
Government College of Engineering, Amravati Maharashtra, India
chatur.prashant@gcoea.ac.in

ABSTRACT

Semantic similarity measures between words play an important role in community mining, document clustering, information retrieval and automatic metadata extraction. For a computer to decide the semantic similarity between words, it should understand the semantics of the given words. Computer is a syntactic machine, which cannot understand the semantics. So it always made an attempt to represent the semantics words as syntactic words. Today, there are various methods proposed for finding the semantic similarity between words. Some of these methods have used the information sources as precompiled databases like WordNet and Brown Corpus. Some are based on Web Search Engine. In this paper we have described the methods based on precompiled databases like WordNet and Brown Corpus as well as the web search engine. Along with this we have compared the all methods on the basis of performance and their limitation. From the study, Experimental result on Miller-Charles benchmark dataset show that the method by the Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka based on web search engine results outperforms all the existing semantic similarity measures by a wide margin, achieving a correlation coefficient of 0.87.

KEYWORDS

Semantic Similarity, WordNet, Brown Corpus.

1. INTRODUCTION

Semantic similarity is a central concept which finds great importance in various fields such as natural language processing (NLP), artificial intelligence (AI), cognitive science and psychology, both in the academic community as well as in industry. Accurate measurement of semantic similarity between words is essential for many tasks such as, information retrieval, document clustering [11], and synonym extraction [12], etc. The most popular way for people to compare two objects and acquire knowledge is the similarity between those two objects. For humans, it is easy to say if one word is more similar to a given word than another. For example, we can easily say that car is more similar to automobile than car is to apple. In fact, semantic similarity between words is defining a resemblance on relations. Obtaining semantic relation and similarity between words or concepts is required in many applications in psycholinguistics and NLP. Some of the

DOI : 10.5121/ijcseit.2012.2605

most popular semantic similarity methods are implemented and evaluated by using WordNet [10] as the underlying reference ontology.

Semantically related words of a particular word are listed in manually created lexical ontology such as WordNet [10]. In WordNet, a synset contains a set of synonymous words for a particular sense of words. However, semantic similarity between words changes over time and across domains. For example, apple is associated with computers on a web. However, this meaning of apple is not listed in most general-purpose dictionaries. A client, who searches for apple on the web, might be interested in this meaning of apple as a computer and not apple as a fruit. Always new words are constantly being created as well as new senses are assigned to existing words. Manually maintaining ontology to capture these new words and senses is complex task and it also costly [8].

In this paper, section 2 describes detail about the Information Resources such as database "WordNet" and "Web search engine" on which many researchers have done researches. Section 3 describes methods for semantic similarity measurement between Words in details based on ontology as well as web search engine. Section 4 gives detail about the comparisons of all the semantic similarity methods. The paper concludes in Section 5 that based on the benchmark data set.

2. INFORMATION RESOURCES

Information Resources are very important factor for measuring the semantic similarity between words. From the starting work of semantic similarity measurement between words many researcher have used WordNet as Information Resource and recently some have used Web search engine.

2.1. WordNet

WordNet [10] is a lexical database for the English language. WordNet was created and being maintained at cognitive Science Laboratory of Princeton University under the direction of psychology professor George A. Miller. Differing from other traditional lexicons, it groups words into sets of synonyms is called synsets, it provides short and general definitions, and records the many semantic relations between these synonym sets. WordNet is particularly well suited for semantic similarity measurement, since it organizes nouns and verbs into hierarchies of IS-A relations. Figure 1 illustrates a fragment of the WordNet2.1 IS-A hierarchy.

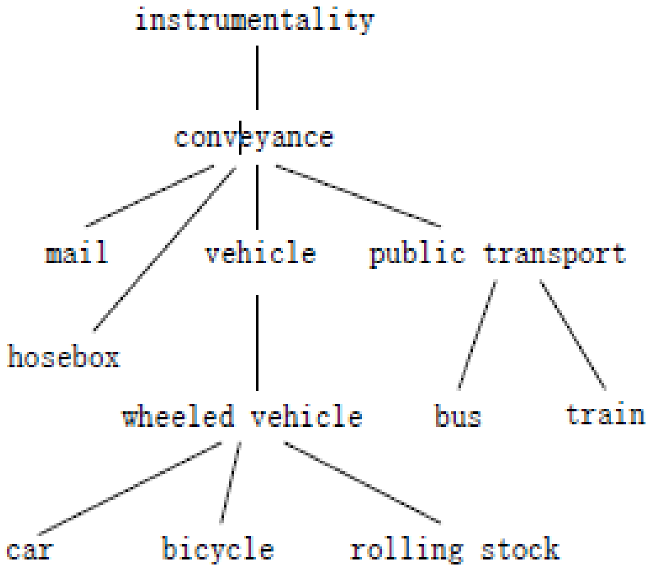

Figure 1. A fragment of WordNet2.1

2.2. Web Search Engine

Web search engines provide an efficient interface to the vast information. For
the measurement of semantic similarity between words many researcher have been used Web Search Engines Results as a resources. Page counts and Snippets [9] are two useful information sources provided by most web search engines results. Here, Page count of a given query is an estimate of the number of pages that contain the given query words and Snippets is a brief window of text extracted by a search engine around the query term in a documents, it provide useful information regarding the local context of the query. Semantic similarity measurements defined over snippets have been used in many task such as query expansion, personal name disambiguation [7], and community mining [8]. Processing snippets is also efficient as compare to downloading web pages, downloading web pages might be time consuming depending on the size of the pages.

Many researchers have been used snippets as an information source for the semantic similarity measurement between words, few have used only page count as information source and some have used combination of both.

3. SEMANTIC SIMILARITY MEASUREMENT METHODS

Semantic Similarity Measurement Methods are used to find the semantic similarity between words based on the information sources. Information sources such as Ontology like WordNet, Biomedical dictionary, Brown Corpus and another is Web Search Engine. Following are some methods based on both sources.

3.1. Traditional Ontology based methods

Ontology-based semantic similarity measurement methods are those use ontology source as the primary information source. They can be roughly grouped into three groups as follows

3.1.1. Distance based method

The distance based approach is a more natural and direct way of measuring semantic similarity between words using taxonomy. It estimates the distance (e.g. edge length) between nodes which correspond to the concepts being compared. In the given multidimensional concept space, the conceptual distance can conveniently measured by the geometric distance between the nodes representing the concepts.

For a hierarchical taxonomy, Rada et al. [1] given that the distance should satisfy the properties of a metric, properties such as positive property, zero property, symmetric property and triangular inequality. Therefore, in an IS-A semantic network, the simplest form of determining the distance between two elemental concepts nodes, A and B, is the shortest path that links A and B means the minimum number of edges that separate A and B.

Rada et al. [1] applied the distance method to a medical domain, and found that the distance function simulated well human assessments of conceptual distance. However, Richardson and Smeaton [16] had concerns that the measure was less accurate than expected when applied to a comparatively broad domain (e.g. WordNet taxonomy). They found that irregular densities of links between concepts result in unexpected conceptual distance outcomes. Also, without causing many serious side effects elsewhere, the value of depth scaling factor does not adjust the overall measure well because of the general structure of the taxonomy. In addition, we feel that the distance measure is highly depended upon the subjectively pre-defined network hierarchy. Since the main purpose of the design of the WordNet was not for semantic similarity computation purpose, few local network layer constructions may not be suitable for the direct distance manipulation.

3.1.2. Information content based method

Resnik [2] pointed out the node based approach to determine the conceptual similarity is called the information content based method. In a multidimensional space upon which a node represents a unique concept consisting of a certain amount of information, and an edge always represents a direct association between two concepts, the similarity between two words is the extent to which they share information in common. Considering this in hierarchical concept space, this common information “carrier” can be identified as a specific concept node that subsumes both of the two words in the hierarchy. In simple word, this super-class should be the first class upward in this hierarchy that subsumes both classes. The semantic similarity value is defined as the information content value of this specific super-ordinate class. The information content value of a class is then obtained by estimating the probability of occurrence of this class in a large text corpus.

The information content method requires less information on the detailed structure of taxonomy. It is not sensitive to the problem of varying link types. However, it is still dependent on the skeleton structure of the taxonomy because it ignores information on the structure. Normally it generates a coarse result for the comparison of words. It means, it does not differentiate the similarity values of any pair of concepts in a sub-hierarchy as long as their “smallest common denominator” is the same.

Lin [12] calculates semantic similarity using a formula derived from information theory. Lin’s modification consisted of normalizing by the combination of information content of the compared concepts and assuming their independence. Similarity measure by Lin [12] takes an information-content approach based on three assumptions. Firstly, the more similar two concepts are, the more this concept will have in common. Secondly, the less two concepts have in common, the less similar this are. Thirdly, maximum similarity occurs when two concepts are identical.

3.1.3. Distance and Information Content based method

Jiang and Conrath [3] presented an approach for measuring semantic similarity between words and concepts. It combines lexical taxonomy structure with corpus statistical information so that the semantic distance between nodes in the semantic space constructed by the taxonomy can be better quantified with the computational evidence derived from a distributional analysis of corpus data. In short, this method is a combined approach which inherits the edge-based approach of the edge counting scheme, which is then enhanced by the node-based approach of the information content measurement.

Jiang-Conrath measure gives semantic distance rather than similarity or relatedness. This distance measure can be converted to a similarity measure by taking the multiplicative inverse of it.

3.2. Web Search Engines based methods

By using the existing ontology for measuring semantic similarity between words there is a limitation of new words. So to overcome this limitation many researchers have worked on web. Because updated information source is only web. Following are some web search engine based approaches to measure semantic between words.

3.2.1. Snippets based Method

Determining the similarity of short text snippets, such as search queries those works poorly with traditional document semantic similarity measures. Sahami and Heilman [5] address this problem by introducing a novel method for measuring the similarity between short text snippets by leveraging web search results to provide greater context for the short texts. They have defined such a similarity kernel function, that mathematically analyze some of its properties, and provide examples of its efficacy. They have also shown the use of this kernel function in a large-scale system for suggesting related queries to search engine users.

Sahami and Heilman [5] measured semantic similarity between two queries using snippets returned for those queries by a search engine. For each query, they have collected snippets from a search engine and represent each snippet as a TF-IDF weighted term vector. Each vector is L2 normalized and centroid of the set of vectors is measured. Semantic similarity between two queries is then they have defined as the inner product between the corresponding centroid vectors. Chen et al. [13] proposed a double-checking model using text snippets returned by a web search engine to compute semantic similarity between words. For two words P and Q, they have collected snippets for each word from a web search engine. Then, on the basis of this they count the occurrences of word P in the snippets for word Q and the occurrences of word Q in the snippets for word P. These values are combined nonlinearly to compute the similarity between P and Q. This is given by (1).

$$CODC(P,Q) = \begin{cases} 0 & if\ f(P@Q) \\ \exp\left(log\left[\frac{f(P@Q)}{H(P)}\ X\ \frac{f(Q@P)}{H(Q)}\right]^{\alpha} \right) & otherwise \end{cases} \tag{1}$$

Here, f (P@Q) denotes the number of occurrences of P in the top-ranking snippets for the query Q in Google, H (P) is the page count for query P, and α is a constant in this method, which was experimentally set to the 0.15. This method completely depends on the search engine's ranking algorithm. Although two words P and Q might be very similar, they have not assume that one can find Q in the snippets for P, or vice versa, because a search engine considers many other factors besides semantic similarity, such as publication date (novelty) and link structure (authority) when ranking the result set for a query.

3.2.2. Page counts based Method

Cilibrasi and Vitanyi [6] proposed a distance metric between words using only page counts retrieved from a web search engine. This proposed metric is named Normalized Google Distance (NGD) and is given by (2)

$$NGD(P,Q) = \frac{\max\{\log H(P), \log H(Q)\} - \log(P,Q)}{\log(P,Q) - \min\{\log H(P), \log H(Q)\}} \qquad (2)$$

Here, P and Q are the two words between which distance NGD (P, Q) is to be computed, H (P) denotes the page count for the word P, and H(P, Q) is the page count for the query P and Q. NGD is fully based on normalized information distance, which is defined using Kolmogorov complexity. Because NGD did not take into account the context in which the words co-occur, because it suffered from the some drawbacks.

3.2.3. Snippets and Page counts based Method

Bollegala, Matsuo and Ishizuka [9] have proposed an automatic method to estimate the semantic similarity between words or entities using web search engines. Because of the huge numerous documents and the high growth rate of the web, it is more time consuming to analyze each document separately. Generally, web search engines provide an efficient interface to this vast information by providing Page counts and Snippets are two useful information sources. They have used page count and snippets as an information sources for measuring the semantic similarity between words. Page count of a query is an estimate of the number of pages that contain the query words. Generally, page count may not necessarily be equal to the word frequency because the queried word might appear many times on same page. Snippets, a brief window of text extracted by a search engine around the query term in a document, provide useful information related to the local context of the query term. They have proposed a method that considers both page counts and lexical syntactic patterns extracted from snippets that they have shown experimentally to overcome the problems of using the page count or snippets only.

4. COMPARISONS OF SEMANTIC SIMILARITY MEASUREMENT METHODS

As discussed above, useful information in measuring word similarity on the basis of WordNet includes the path length of the two words, depth of the subsumer, information content of a concept and other parameters [4]. So to remove the limitation of the ontology based approaches web search engine based approaches have been proposed. There is not a standard for evaluation of lexical similarity. But we have compare result of different researchers with Miller and Charles (MC) 28 pairs of nouns Benchmark Dataset [15] that occurred in WordNet for finding the performance of methods. Because the performance of each semantic similarity measure can be depends upon the correlation result of each measure with the (MC) 28 pairs of nouns Benchmark Dataset [15]. The measure which is having the correlation result high this approach will be the best as compare to other.

On the basis of the result calculated according to the approaches by different researchers, following Table 1 and Table 2 shows the similarity methods and their respective correlation result. Table 1 shows the correlation result of the methods based on Ontology WordNet.

Table 1- Reference Results on Dataset based on WordNet

Similarity Method	**Correlation**
MC	1
Wu and Palmer[14]	0.8030
Resnik	0.8140
Lin	0.8340
Jiang and Conrath[3]	0.8363

Following Table 2 summarizes the experimental results on MC data sets. All the approaches in Table 4 are based on Web search engine. Here, MC - Miller and Charles, CODC is approach by Chen, SH means Sahami and Heilman, NGD is Normalized Google Distance by Vityani and BMI means Bollegala, Matsuo and Ishizuka.

Table 2- Reference Results on Dataset based on Web Search Engines

Similarity Method	**Correlation**
MC	1
CODC	0.69
SH	0.58
NGD	0.21
BMI	0.87

Figure 2 and Figure 3 shows the graphical representation of similarity methods and correlation against MC's Dataset [15] for the Table 1 and Table 2 respectively.

Figure 2 gives the graphical representation of similarity methods based on WordNet with their correlation result. From the Figure 2 and Table 1 approach by Jiang and Conrath perform better having correlation 0.8360. Figure 3 gives the graphical representation of similarity methods based on Web Search Engines with their correlation result. From the Figure 3 and Table 2 approaches by Bollegala, Matsuo and Ishizuka (BMI) perform better having correlation 0.8700. So from this figure we can say that the correlation of similarity method by Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka perform well among all the methods. This method is also web based so there is no limitation of new words.

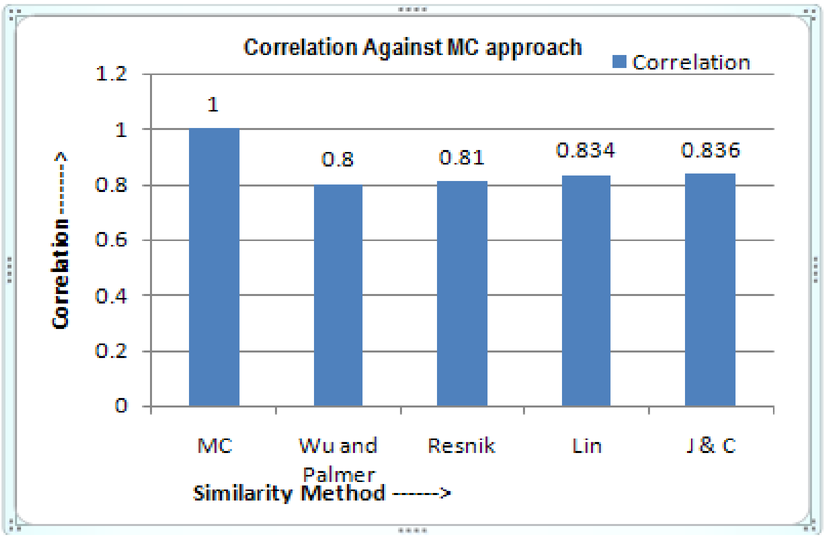

Figure 2. Representation of Correlation of each method in table 1 against MC's approach

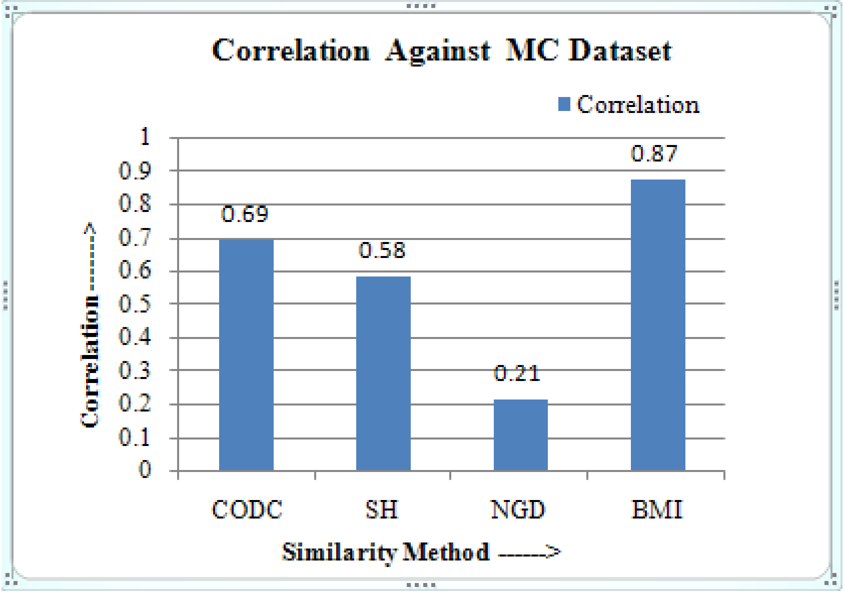

Figure 3. Representation of Correlation of each method in table 2 against MC's approach

5. CONCLUSIONS

This paper presented the similarity measurement of words. We argue that all information sources like shortest path length, depth, information content need to be properly processed in defining a similarity measure. But this are based on the WordNet and there is limitation of new words. Working of different researchers like Rada, Resnik, Lin, Jiang and Conrath and Miller Charles, we have presented in this paper. Along with this we have presented the approaches based on the Web search engine such as approach by Sahami and Helmand, Cilibrasi and Vitanyi, Chen, Bollegala, Matsuo and Ishizuka (BMI) to overcome the disadvantages of the ontology based approaches. From the result we can conclude that, correlation of dataset against Miller and Charles are varies with their technique. From the Table 1, Table 2 and Figure 2, Figure 3, we can say that the approach by the Bollegala, Matsuo and Ishizuka (BMI) gives the correlation 0.87 which is better than all approaches. So the approach by the Bollegala, Matsuo and Ishizuka perform the better as compare to other researchers approach.

In the future, for improving the existing result we would like to take more information resources or combination of them into account for measuring semantic similarity between words. So that result will be improved.

REFERENCES

[1] R. Rada, H. Mili, E. Bichnell & M. Blettner, (1989) "Development and application of a metric on semantic nets", IEEE *Transaction on. Systems, Man and Cybernetics,* Vol. 9, No. 1, pp17- 30.

[2] P. Resnik, (1995) "Using information content to evaluate semantic similarity in a taxonomy", *Proceeding of 14th International Conference on Artificial Intelligence*.

[3] J. Jiang & D. Conrath, (1997) "Semantic similarity based on corpus statistics and lexical taxonomy", *Proceeding of International Conference on Research in Computational Linguistics (ROCLING X).*

[4] D. Mclean, Y. Li & Z.A. Bandar, (2003) "An approach for measuring semantic similarity between words using multiple information sources", IEEE *Transactions on Knowledge and Data Eng.*, Vol. 15, No. 4, pp871-882.

[5] M. Sahami & T. Heilman, (2006) "A web-based kernel function for measuring the similarity of short text snippets", *Proceeding of 15th International World Wide Web Conference.*

[6] R. Cilibrasi & P. Vitanyi,(2007) "The google similarity distance", IEEE *Transactions on Knowledge and Data Eng.*, Vol. 19, No. 3, pp370-383.

[7] D. Bollegala, Y. Matsuo &M. Ishizuka,(2006) "Disambiguating personal names on the web using automatically extracted key phrases", *Proceeding of 17th European Conference on Artificial Intelligence*, pp. 553-557.

[8] D. Bollegala, Y. Matsuo, and M. Ishizuka, (2007) "Measuring semantic similarity between words using web search engines", *Proceeding of International Conference on World Wide Web*, pp757-766.

[9] D. Bollegala, Y. Matsuo &M. Ishizuka, (2011) "A web search engine-based approach to measure semantic similarity between words", IEEE *Transactions on Knowledge and Data Eng*., Vol. 23, No. 7, pp977-990.

[10] G. A. Miller, (1995) "WordNet: A lexical database for english", *Comm. ACM*, Vol. 38, No. 11, pp39-41.

[11] R. K. Srihari, Z. F. Zhang & A. B. Rao, (2000) "Intelligent indexing and semantic retrieval of multimodal documents", *Information Retrieval*, Vol. 2, pp245-275.

[12] D. Lin, (1998) "An information-theoretic definition of similarity", *Proceeding of International Conference on Machine Learning*.

[13] H. Chen, M. Lin, & Y. Wei,(2006) "Novel association measures using web search with double checking", *Proc. 21st International Conference on Computational Linguistics and 44th Ann. Meeting of the Assoc. for Computational Linguistics (COLING/ACL)*, pp1009-1016.

[14] Z. Wu & M. Palmer, (1994) "Verb semantics and lexical selection," in *In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp133–138.

[15] G. Miller & W. Charles, (1998) "Contextual correlates of semantic similarity", *Language and Cognitive Processes*, Vol. 6, No. 1, pp1-28.

[16] Richardson, R. & A.F. Smeaton, (1995), "Using wordnet in a knowledge-based approach to information retrieval", *School of Computer Applications*, Dublin City University, Ireland.

Authors

Ankush M. Maind has received his B.E. degree from Umrer College of Engineering, Nagpur, India in 2010 and doing his M.Tech at Government College of Engineering, Amravati, India. He has published one paper in international journal and one paper in international and national conference in the field of Data Mining and NLP. His area of research includes Data Mining, Web Mining, Knowledge and Data Mining.

A. V. Deorankar has received his M.E. degree in Electronics Engineering from Govt. College of Engineering, Amravati, India. He has published nine papers in national and seven papers in international journals. He has also patent of one paper. His area of research includes Computer Network, Web Mining. Currently he is working as an Associate Professor at Govt. college of Engineering, Amravati, Maharashtra, India.

P. N. Chatur has received his M.E. degree in Electronics Engineering from Govt. College of Engineering, Amravati, India and PhD degree from Amravati University. He has published twenty papers in national and ten papers in international journal. His area of research includes Neural Network, data mining. Currently he is head of Computer Science and Engineering department at Govt. College of Engineering, Amravati, Maharashtra, India.

.