ISBN 978-83-60810-14-9
ISSN 1896-7094

A New Word Sense Similarity Measure in WordNet

Ali Sebti
Amirkabir university of technology, Intelligence Systems Laboratory[1], Tehran, Iran
Email: ali.sebti@aut.ac.ir

Ahmad Abodollahzadeh Barfroush
Amirkabir university of technology
Intelligence Systems Laboratory
Tehran, Iran
Email: ahmad@ce.aut.ac.ir

***Abstract*—Recognizing similarities between words is a basic element of computational linguistics and artificial intelligence applications. This paper presents a new approach for measuring semantic similarity between words via concepts. Our proposed measure is a hybrid system based on using a new Information content metric and edge counting-based tuning function. In proposed system, hierarchical structure is used to present information content instead of text corpus and our result will be improved by edge counting-based tuning function. The result of the system is evaluated against human similarity ratings demonstration and shows significant improvement in compare with traditional similarity measures.**

I. Introduction

SEMANTIC similarity is an important topic in natural language processing (NLP) and Information Retrieval (IR). It has also been subject to studies in Cognitive Science and Artificial Intelligence. Application areas of semantic similarity include word sense disambiguation (WSD) [19], information extraction and retrieval [2,22,24], detection and correction of word spelling errors (malapropisms)[3], text segmentation [10], image retrieval [21], multimodal document retrieval [20], and automatic hypertext linking [5], automatic indexing, text annotation and summarization [13].

To quantify the concept of similarity between words, some ideas have been put forth by researchers, most of which rely heavily on the knowledge available in lexical knowledge bases like WordNet.

There are mainly two approaches to compute semantic similarity. The first approach is making use of a large corpus or word definitions and gathering statistical data from these sources to estimate a score of semantic similarity, which we call text-based approach. The second approach makes use of the relations and the hierarchy of a thesaurus, such as WordNet, which we call structure-based approach.

In text-based approach, word relationships are often derived from their co-occurrence distribution in a corpus [7,6]. Gloss overlap, introduced by Lesk [12] and extended gloss overlap, introduced by Banerjee and Pedersen, are another instances of this approach. The latter is a measure that determines the relatedness of concepts proportional to the extent of overlap of their WordNet glosses [1]. Besides gloss vector measure of semantic relatedness, introduced by Pedersen and Patwardhan, is based on second order co–occurrence vectors in combination with the structure and content of WordNet, a semantic network of concepts [16].

In structure-based approach, first studies date back to Quilian's semantic memory model [17], where the number of hops between nodes of concepts in the hierarchical network specifies the similarity or difference of concepts. Wu and Palmer's semantic similarity measure was based on the path length between concepts located in a taxonomy [23]. Also, the similarity measure of Leacock and Chodorow is based on the shortest path length between two concepts in is-a hierarchy [11].

In combining two approaches, Resnik introduced a new factor of relatedness called information content (IC) [18]. The Similarity measures of Resnik, Jiang and Conrath [9] and Lin [14] all rely on the IC values assigned to the concepts in an is-a hierarchy, but their usage of IC has little differences. Using a different approach Hirst G. and St-Onge assign relatedness scores to words rather than word senses. They set different weights for different kinds of links in a semantic network, and uses those weights for edge counting [8].

In this paper, we first introduce a new method for computing IC of concepts in a hierarchical structure. We will show that this method only uses hierarchical structure and not corpus to determine IC. Furthermore, information content obtained from this method implicitly includes depth and branch factor of the concept from root to target concept. Then we use formula that is similar to Lin formula for measuring similarity. Then we analyze our result and comparing it with benchmark result and introduce an edge counting-based function for improving and overcome their problems. For adjusting our function's parameters we use genetic algorithm. Finally our combined similarity measure is evaluated against a benchmark set of human similarity ratings, and demonstrates that the proposed measure significantly outperformed traditional similarity measures.

In section 2 we describe WordNet, which was used in developing our method. Section 3 describes the extraction of our new information content metric from a lexical knowledge base. Section 4 presents the choice and organization of a benchmark data set for evaluating the similarity method, how to define a tuning function, experimental results and dis cussion about it. Finally, paper concludes in Section 5 that,

[*] Authors thanks Iran Telecommunication Research Center (ITRC) for their support during this research work

[1] http://ceit.aut.ac.ir/islab/

based on the benchmark data set, our measure outperforms existing measures.

II. WordNet

WordNet is the product of a research project at Princeton University which has attempted to model the lexical knowledge of a native speaker of English [4]. In WordNet each unique meaning of a word is represented by a synonym set or *synset*. Each synset has a gloss that defines the concept of the word. For example the words *car*, *auto*, *automobile*, and *motorcar* is a synset that represents the concept define by gloss: *four wheel Motor vehicle, usually propelled by an internal combustion Engine*. Many glosses have *examples* of usages associated with them, such as *"he needs a car to get to work."*

In addition to providing these groups of synonyms to represent a concept, WordNet connects concepts via a variety of semantic relations. These semantic relations for nouns include:

- Hyponym/Hypernym (IS-A/ HAS A)
- Meronym/Holonym (Part-of / Has-Part)
- Meronym/Holonym (Member-of / Has-Member),
- Meronym/Holonym (Substance-of / Has-Substance)

Figure 1 shows a fragment of WordNet taxonomy.

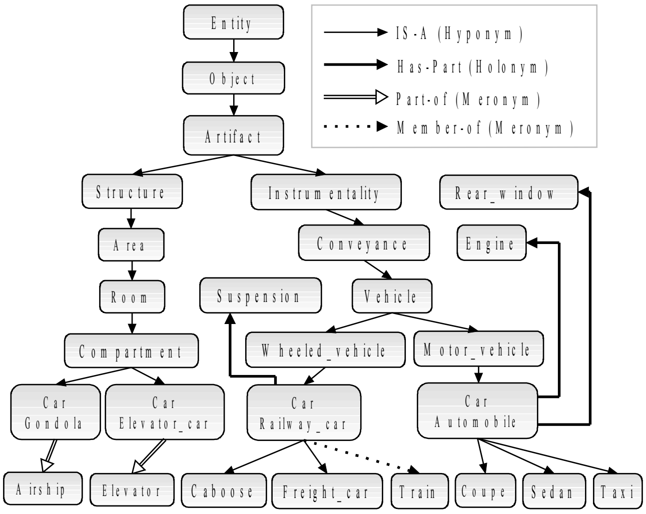

Fig. 1 fragment of WordNet taxonomy

III. The new Information Content Metric

A. Previous information content based approaches

Many researchers consider statistical figures to compute IC value. They assign a probability to a concept in taxonomy based on the occurrence of target concept in a given corpus. The IC value is then calculated by negative log likelihood formula as follow:

$$IC(c)=-\log(p(c)) \tag{1}$$

Where c is a concept and p is the probability of encountering c in a given corpus. Philip Resnik [18] used this formula to compute semantic similarity between concepts. Basic idea behind the negative likelihood formula is that the more probable a concept appears, the less information it conveys, in other words, infrequent words are more informative then frequent ones.

Resnik showed that semantic similarity depends on the amount of information that two concepts have in common, this shared information is given by the Most Specific Common Abstraction (MSCA) that subsumes both concepts. Therefore we must first discover the MSCA and then shared information is equal to the IC value of the MSCA. If MSCA does not exist then the two concepts are maximally dissimilar. Formally, Resnik semantic similarity is defined as:

$$sim_{res}(c_1,c_2)=\max_{c\in S(c_1,c_2)} ic_{res}(c) \tag{2}$$

where *S(c₁,c₂)* is the set of concepts that subsume c_1 and c_2. Another information theoretic similarity metric that used the same notion of IC was that of Lin [23], expressed by:

$$sim_{lin}(c_1,c_2)=\frac{2\times sim_{res}(c_1,c_2)}{(ic_{res}(c_1)+ic_{res}(c_2))} \tag{3}$$

Jiang and Conrath [9] also proposed a new measure of semantic distance that its corresponding semantic similarity can be obtained from the reverse of it. Common version of their distance metric is:

$$\begin{aligned} dist_{jcn}(c_1,c_2)=&(ic_{res}(c_1)+ic_{res}(c_2)) \\ &-2\times sim_{res}(c_1,c_2) \end{aligned} \tag{4}$$

B. Our new information content metric

Our method of obtaining IC values is based on the assumption that the taxonomic structure of WordNet is organized in a meaningful and principled way, where concepts in higher depths and having more sibling concepts in the taxonomy structure are more informative and their IC values are bigger. Our method includes implicitly these two parameters that figure 2 represent this method for computing IC value for a fragment of concepts in WordNet.

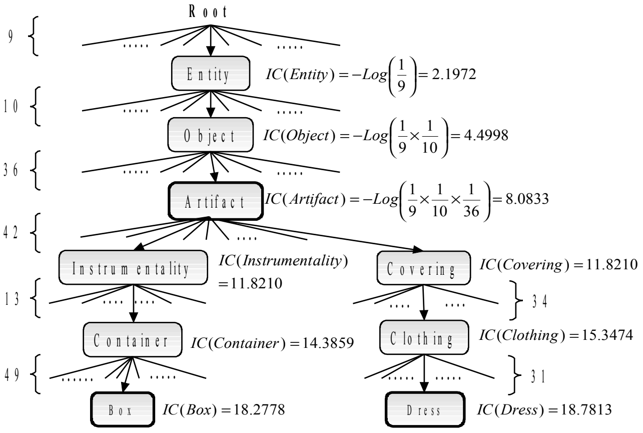
Root
9
.....
.....
Entity
$IC(Entity)=-Log\left(\frac{1}{9}\right)=2.1972$
10
Object
$IC(Object)=-Log\left(\frac{1}{9}\times\frac{1}{10}\right)=4.4998$
36
Artifact
$IC(Artifact)=-Log\left(\frac{1}{9}\times\frac{1}{10}\times\frac{1}{36}\right)=8.0833$
42
Instrumentality
$IC(Instrumentality)=11.8210$
Covering
$IC(Covering)=11.8210$
13
34
Container
$IC(Container)=14.3859$
Clothing
$IC(Clothing)=15.3474$
49
31
Box
$IC(Box)=18.2778$
Dress
$IC(Dress)=18.7813$

Fig. 2 example of computing our IC metric for some concepts

For better understanding of this method we show in equation 5 how IC value of *Box* is computed according to figure 2:

$$IC(Box) = -Log\left(\frac{1}{9}\times\frac{1}{10}\times\frac{1}{36}\times\frac{1}{42}\times\frac{1}{13}\times\frac{1}{49}\right) = 18.2778 \quad (5)$$

IV. Implementaion

A. Semantic similarity measure

To evaluate the effect of our Information Content Metric on semantic similarity, we first select an existing semantic similarity measure. For this purpose we use Lin semantic similarity measure. This approach makes the implementation easier with less complexity. Lin's formula is shown in equation 3.

B. Benchmark data

In accordance with previous research, we evaluated the results by correlating our similarity scores with that of human judgments provided by Miller and Charles [15]. In their study, 38 undergraduate subjects were given 30 pairs of nouns and were asked to rate similarity of meaning for each pair on a scale from 0 (no similarity) to 4 (perfect synonymy). The average rating for each pair represents a good estimate of how similar the two words are. This benchmark data is used by many researchers in semantic similarity subject [1,16].

C. Edge counting-based tuning function

For beginning our analysis, we first compute semantic similarity between pairs of words with Lin similarity and our similarity approach. As said before, our semantic similarity formula is the same as Lin formula. The difference between them is the method of computing IC value . Then, we draw our obtained result and Lin result and human judgments scores in a diagram. These results are showed in figure 3. As shown in figure 3, in some pairs of words our method is more accurate and in some pairs Lin similarity measure is better. We then decide to improve our accuracy in pairs that our method is less accurate. Therefore, the Next step is how we determine these pairs of words automatically. In others words we must define a new feature for pairs of words that discriminate pairs of words that our similarity method about them is less accurate toward Lin method.

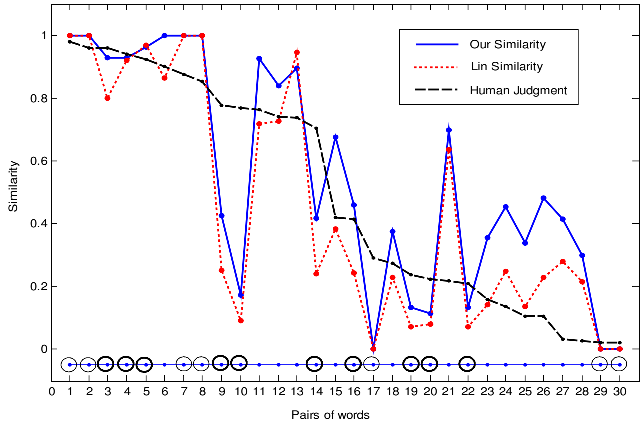

Fig 3 Compare our method with Lin and Human judgments

In figure 3 we show two types of circles: bold line circle and normal line circle. Bold line circles represent pairs of words that accuracy of our method is better than Lin and normal line circle shows that Lin and our method are the same. For other pairs, Lin method is more accurate. As said before in this step we extract a feature that determines inaccurate pairs. Table 1 shows our result, Human judgments, path of two words (concept) and depth of Lowest Common Subsumer (LCS) for two words. In this table if does not exist LCS for a pair, values of LCS depth and Path are -1.

Table 1
Result of our method, human judgment and three features

Pairs of words	HJ	Ours	LCS depth	Path	$(LCS_{depth}+1)$ /(path+1)
car -automobile	0.98	1	8	0	9
gem – jewel	0.96	1	6	0	7
Journey - voyage	0.96	0.93	5	1	3
boy – lad	0.94	0.93	4	1	2.5
coast – shore	0.92	0.96	4	1	2.5
asylum -madhouse	0.90	1	7	1	4
magician – wizard	0.87	1	4	0	5
midday - noon	0.85	1	7	0	8
furnace - stove	0.77	0.42	2	10	0.27
food – fruit	0.77	0.17	0	7	0.12
bird - cock	0.76	0.92	7	1	10
bird - crane	0.74	0.84	7	3	2
tool - implement	0.73	0.89	4	1	2.5
brother -monk	0.70	0.41	2	5	0.5
crane - implement	0.42	0.67	3	4	0.8
lad - brother	0.41	0.46	2	4	0.6
journey - car	0.29	0	-1	-1	10
monk - oracle	0.27	0.37	2	7	0.37
cemetery - woodland	0.23	0.13	0	9	0.1
food - rooster	0.22	0.11	0	13	0.07
coast - hill	0.21	0.69	3	4	0.8
forest - graveyard	0.21	0.13	0	9	0.1
shore - woodland	0.15	0.35	1	5	0.33
monk - slave	0.13	0.45	2	4	0.6
coast - forest	0.10	0.34	1	6	0.28
lad - wizard	0.10	0.48	2	4	0.6
chord - smile	0.03	0.41	3	10	0.36
glass - magician	0.02	0.29	1	7	0.25
noon - string	0.02	0	-1	-1	10
rooster - voyage	0.02	0	-1	-1	10

Our result shows the pairs of words that their LCS depth is little or path length of two concepts is large, our result is less accurate. These two conditions are combined with new feature that can be seen in the sixth column in table 1. Therefore if the new feature is low, pairs of words which our result is related to, is less accurate and hence is detectable. In figure 4 we show that, in new feature space a specific bund contains most inaccurate pairs of words.

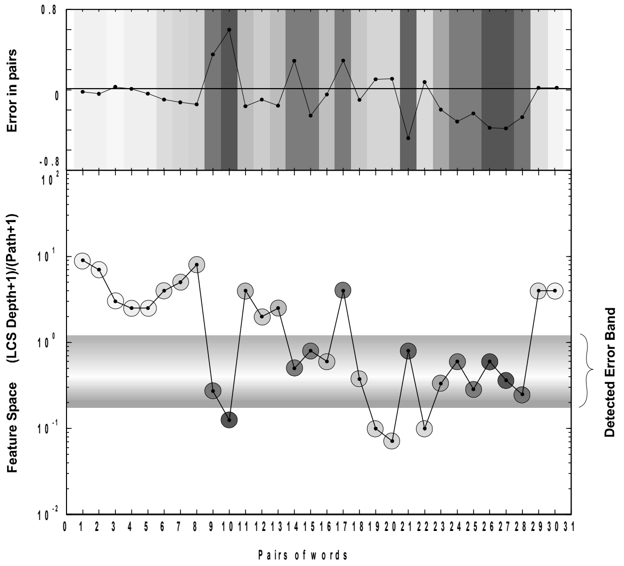

Fig 4 detecting error bund in new feature space

One considerable point is that, when new feature is too low, our similarity result is lower than human judgment and when not too low, our similarity result is higher than human judgment. This point persuades us to define a tuning function which modifies our result. Thus we define a function that its general shape is showed in figure 4. In equation 6 we show mathematical formula of this function.

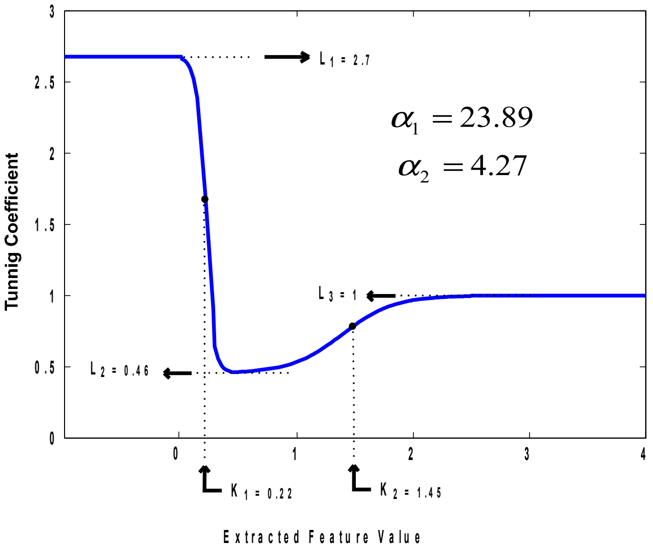

Fig 5 General shape of our tuning function

To determine the best value for parameters of this function we use genetic algorithm. Because of lack in data, for tuning parameters, we use Leave-one-out cross validation and then values of each parameters is average of the obtained values in 30 experiments. In Leave-one-out cross validation, for a dataset with N examples, perform N experiments. For each experiment use N-1 examples for training and the remaining example for testing. Figure 5 present the best value obtained by GA algorithm for l_1, l_2, l_3, k_1, k_2, α_1 and α_2. In table 2 we compare our final result with other methods. In all other similarity measures that use IC value, IC value is computed like Resnik's manner which was discussed in section III. For all of these experiments, Miller's benchmark data (30 pairs of words) is used. In order to make fair comparisons, we decided to use an independent software package that would compute similarity values using previously established strategies while allowing the use of WordNet 2.0. One freely available package is that of Siddharth Patwardhan and Ted Pederson [25]. This result shows that our similarity measure is comparable with other similarity measures.

$$f(x)=\begin{cases} x\le k_1-\frac{5}{\alpha_1} & l_1 \\ x>k_1-\frac{5}{\alpha_1} \quad and \quad x<k_1+\frac{5}{\alpha_1} & \frac{l_1-l_2}{1+\exp^{-\alpha_1(-x+k_1)}}+l_2 \\ x\ge k_1+\frac{5}{\alpha_1} \quad and \quad x\le k_2-\frac{5}{\alpha_2} & l_2 \\ x>k_2-\frac{5}{\alpha_2} \quad and \quad x<k_2+\frac{5}{\alpha_2} & \frac{l_3-l_2}{1+\exp^{-\alpha_2(x-k_2)}}+l_2 \\ x\ge k_1+\frac{5}{\alpha_2} & l_3 \end{cases} \qquad (6)$$

TABLE 2
COMPARE OUR METHOD WITH OTHERS RELATED WORK IN CORRELATION WITH HUMAN JUDGMENT

Similarity measure	correlation
Jiang and Conrath	0.695
Hirst St.Onge	0.689
Leacock Chodorow	0.821
Lin	0.823
Resnik	0.775
Wu and Palmer	0.803
Patwardhan and Pedersen	0.77
Our Similarity Measure	**0.87**

V. CONCLUSION AND FUTURE WORK

In this paper, we have introduced a new word sense similarity measure with a proper tuning function. For computing information content, we used hierarchical structure alone, instead of text corpus. Experimental evaluation against a benchmark set of human similarity ratings demonstrated that the proposed measure significantly outperformed traditional similarity measures. In future work, we intend to that use this similarity measure in real world applications such as word sense disambiguation. Also, our tuning function can be used with other previous similarity measures.

REFERENCES

[1] S. Banerjee and T. Pedersen. "Extended gloss overlaps as a measure of semantic relatedness". In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence* , pages 805–810, Acapulco, Mexico, 2003.

[2] C. Buckley, J. Salton, J. Allen and A. Singhal, A. "Automatic query expansion using Smart: TREC 3". In *The third Text Retrieval Conference* , Gaithersburg, MD, 1995.

[3] A. Budanitsky and G. Hirst, "Semantic Distance in WordNet: An Experimental, Application-Oriented Evaluation of Five Measures". Proc. *Workshop WordNet and Other Lexical Resources, Second Meeting North Am. Chapter Assoc. for Computational Linguistics* , June 2001.

[4] C. Fellbaum, editor. "WordNet: An Electronic Lexical Database". *MIT Press*, Cambridge, USA, 1998.

[5] S. J. Green, "Building Hypertext Links by Computing Semantic Similarity". *IEEE Trans. Knowledge and Data Eng* , vol. 11, no. 5, pp. 713-730, Sept./Oct. 1999.

[6] G. Grefenstette. "Use of Syntactic Context to Produce Term Association Lists for Text Retrieval". *Proceedings of the 15th Annual International Conference on Research and Development in Information Retrieval* , SIGIR'92, 1992.

[7] D. Hindle. "Noun Classification from Predicate-Argument Structures". *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics* , ACL28'90, 268-275, 1990.

[8] G. Hirst and D. St-Onge. "Lexical chains as representations of context for the detection and correction of malapropisms". In *Fellbaum* , pp. 305–332, 1998.

[9] J. Jiang and D. Conrath. "Semantic similarity based on corpus statistics and lexical taxonomy". In *Proceedings of International Conference on Research in Computational Linguistics* , Taiwan, 1997.

[10] H. Kozima, "Computing Lexical Cohesion as a Tool for Text Analysis". *doctoral thesis, Computer Science and Information Math* , Graduate School of Electro-Comm., Univ. of Electro-Comm., 1994.

[11] C. Leacock and M. Chodorow. "Combining local context and WordNet similarity for word sense identification". *In Fellbaum*, pp. 265–283, 1998.

[12] M. Lesk. "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone". In *Proceedings of the SIGDOC Conference,* Toronto, 1986.

[13] C. Y. Lin, and E. Hovy. "Automatic evaluation of summaries using n-gram co-occurrence statistics". In *Proceedings of Human Language Technology Conference (HLT-NAACL)* , Edmonton, Canada, 2003.

[14] D Lin. "An information-theoretic definition of similarity". In *Proceedings of the 15th International Conference on Machine Learning* , Madison, WI, 1998.

[15] G. Miller and W. Charles. "Contextual correlates of semantic Similarity". *Language and Cognitive Processes* , 6, 1–28, 1991.

[16] S. Patwardhan and T. Pedersen. "Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts". In *Proceedings of Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together* , EACL, .2006.

[17] M. R. Quilian. "Semantic memory". *Semantic Information Processing* . pages 216–270, 1968.

[18] P. Resnik. "Using information content to evaluate semantic similarity". In *Proceedings of the 14th International Joint Conference on Artificial Intelligence* , pages 448–453, Montreal, 1995.

[19] P. Resnik, "Semantic Similarity in a Taxonomy: An Information-Based Measure and Its Application to Problems of Ambiguity in Natural Language". *J. Artificial Intelligence Research*, vol. 11, pp. 95-130, 1999.

[20] R. K. Srihari, Z.F. Zhang, and A.B. Rao, "Intelligent Indexing and Semantic Retrieval of Multimodal Documents". *Information Retrieval* , vol. 2, pp. 245-275, 2000.

[21] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early Years". IEEE *Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349-1380, Dec. 2000.

[22] O. Vechtomova and S. Robertson. "Integration of collocation statistics into the probabilistic retrieval model". In *22 nd Annual Colloquium on Information Retrieval Research* , Cambridge, England, , 2000.

[23] Z. Wu and M. Palmer. "Verb semantics and lexical selection". In *32nd. Annual Meeting of the Association for Computational Linguistics* . pages 133 –138, New Mexico State University, Las Cruces, New Mexico, 1994.

[24] J. Xu, and B. Croft. "Improving the effectiveness of information retrieval". *ACM Transactions on Information Systems* , 18(1):79-112, 2000.

[25] http://wn-similarity.sourceforge.net