Weighted Edge: A New Method to Measure the Semantic Similarity of Words based on WordNet[1]

Liang Dong
School of Computing
Clemson University
Clemson, SC 29630, USA
+1-864-650-7580
ldong@cs.clemson.edu

Pradip K Srimani[1]
School of Computing
Clemson University
Clemson, SC 29630, USA
+1-864-656-5886
srimani@cs.clemson.edu

James Z. Wang
School of Computing
Clemson University
Clemson, SC 29630, USA
+1-864-656-7678
jzwang@cs.clemson.edu

Abstract

Recent study shows that humans are more sensitive to the semantic difference due to categorization than that due to the generalization/specification. We propose a new method to measure the semantic similarity of word pairs based on this discovery. Our method assigns an exponential decreasing weight on each edge along the WordNet hierarchy to measure weighted graph distance between two concepts; it then computes the semantic similarity by employing a set of non-linear transfer functions. Experiments show that this method produces results superior to existing distance-based methods using only hypernym relationship.

1 Introduction

Computing semantic similarity between words has always been a challenge in many areas such as artificial intelligence, natural language processing and information retrieval. It is difficult to model the human perspective on the semantic similarity of words due to two reasons: (1) polysemy and synonymy phenomena widely exist in natural language; (2) psychologists have demonstrated that the human perception of the similarity between words is subject to the context.

Previous studies can be broadly divided into three categories: (1) *distance-based* methods (Rada et al., 1989; Morris and Hirst, 1991; Wu and Palmer, 1994; Yang and Powers, 2005; Alvarez and Lim, 2007) – these are based on the shortest graph distance of words within knowledge resources (dictionaries, thesauri or encyclopedias); (2) *gloss-based* methods (Banerjee and Pedersen, 2003; Patwardhan and Pedersen, 2006; Hughes and Ramage, 2007) – these are based on the number of shared words ping with each other's gloss/definition; (3) *distribution-based* methods (Resnik, 1995; Jiang and Conrath, 1997; Lin, 1998; Resnik, 1999; Li et al., 2003; Sevilla et al., 2005) -- they measure the semantic similarity based on information content which is the frequency of occurrence in a corpus (Cilibrasi and Vitanyi, 2007; Gabrilovich and Markovitch, 2007; Wu et al., 2008; Agirre et al., 2009); they take advantage of concurrence statistics discovered from search engine.

We propose a simple yet effective distance-based method solely via hypernym relationship from WordNet (Fellbaum, 1998) without the need of gloss and corpus statistics. The method is based on an observation on the difference of human perception between *categorizing relation* and *inheriting relation* of word pairs. Our recent study shows that humans are more sensitive to the semantic difference due to the categorization than that due to the inheritance/specification. To capture this observation, we improve previous distance-based methods, by assigning non-linear weight on each edge by its depth in WordNet. Further, we introduce hyperbolic functions to transform weighted distances into similarity values. The proposed method has two advantages: (1) it distinguishes the difference between *categorizing relation* and *inheriting relation* of word pairs; (2) it solely relies on hypernym relationship within the WordNet, thus computationally more effective than those requiring extra computations of gloss or corpus statistics. Experimental studies show that our proposed method outperforms existing methods in distance-based category.

2 Weighted Edge Approach

2.1 Inheritance vs. Categorization

When two word-pairs have the same graph distance and their common ancestor are at the same specification level/depth, should any semantic

[1]Srimani's work supported by NSF grant # CCF-0832582

difference exist between them? What is the human perception of it? None of the existing studies has investigated these issues.

To answer these questions, and to further study how human beings judge the semantic similarity of words, we group together two word-pairs that share the same *Least Common Ancestor* (*LCA*) and have the same graph distance between the words in each pair respectively. In one word-pair, the words are both descendents of their *LCA*. We call it *categorization pair* since both words are separated by *LCA* in different categories. In another word pair, one word is descendant of another word. We call it *inheritance pair*.

Inheritance Pair			***Categorization Pair***	
baked-goods :: cookie	30	⇔	bread :: cake	19
beef :: food	48	⇔	meat :: chocolate	2
brownie :: cake	44	⇔	cookie :: fruit-cake	5
ground beef :: meat	24	⇔	pork :: mutton	25
apple pie :: pastry	42	⇔	pie :: puff	8
stove :: device	41	⇔	comb :: fan	8
engine :: machine	18	⇔	computer :: calculator	33
hunting dog:: canine	27	⇔	wolf :: fox	22
minicab :: car	29	⇔	jeep :: sedan	21
gold :: metal	37	⇔	aluminum :: zinc	14
Total	**340**			**157**

Table 1. 2-distant pairs with same *LCA*

Inheritance Pair			***Categorization Pair***	
apple pie :: food	44	⇔	cake :: beef	3
clementine :: fruit	36	⇔	apple :: almond	15
chicken :: food	47	⇔	octopus :: pastry	0
dynamo :: machine	45	⇔	engine :: abacus	4
abbey :: building	26	⇔	hostel :: mansion	23
tabloid :: medium	8	⇔	broadcasting :: journalism	43
laptop :: computer	51	⇔	workstation :: chatroom	0
American football :: athletic game	36	⇔	golf :: basketball	14
cliff diving :: sports	44	⇔	hunting :: swimming	6
collegiate dictionary :: book	41	⇔	atlas :: bestseller	7
Total	**378**			**115**

Table 2. 4-distant pairs with same *LCA*

We collect 20 groups of such word-pairs. The graph distance of the word-pair in the first 10 groups is 2, as shown in Table 1. The graph distance of the second 10 groups, shown in Table 2, is 4. Then we randomly stop people in Clemson University campus and ask them to judge which pair in each group is more similar semantically. 51 individuals finished the questionnaire anonymously. In Table 1 and Table 2, each row contains a group of word-pairs. The left pair is the inheritance pair and the right pair is the categorization pair. The number in the second column represents the number of people who think the inheritance pair is more similar semantically. The number in the last column represents the number of people who feel the categorization pair is more similar. For those who feel both pairs are semantically equal or who cannot tell which pair is more similar, no number is added to any column. The survey results in Table 1 show that in 68.41% of cases, people think the inheritance pairs are more similar, and in only 31.59% of cases, people think the categorization pairs are more similar. The results in Table 2 demonstrate that in 76.67% of cases, people think that the inheritance pairs are more similar, and in only 23.33% of cases, people feel the categorization pairs are more similar.

Our survey results have revealed an important factor in determining the semantic similarity of words. That is, in general, people are more sensitive to the semantic difference caused by categorization than caused by inheritance. This is more obvious when the graph distance of the words gets larger. This important factor has never been considered in any previous studies.

2.2 Weighted Distance Model

In this paper, the similarity of a word pair is measured by their maximum similarity of their synset/concept pairs (Yang and Powers, 2005). Previous studies by Rada (1989) and Resnik (1999) select the concept pair by choosing the shortest graph distance among all concept pairs of a word pair. In our method, we select the concept pair with the highest specification level/depth of *common ancestor* first. This adjustment is based on the observation that the *depth* of *common ancestor* is the most dominant factor in semantic similarity (Li et al., 2003).

We define *Graph Distance* ($Dist_g$)of a concept pair be the number of edges on the shortest path connecting them; and we define *Specification Level (SpecLev)* of a concept as its depth in the WordNet. If a concept is closer to the root, it has a lower *SpecLev* in the WordNet and represents a more general meaning. Vice Versa, higher *SpecLev* represent more specific meanings. We further define *LCA* of a pair be the con-

cept of that pair's least common ancestor along its path.

To reflect the true human perception in measuring the semantic similarity, we use *Specification Level Difference (SLD)* of concept-pair (c_1, c_2) to model the difference between categorization and inheritance:

$$SLD(c_1, c_2) = |\, SpecLev(c_1) - SpecLev(c_2) \,| \quad (1)$$

A small *SLD* means categorization factor is dominant, thus a relatively lower similarity value; and vice versa. Given a concept pair (c_1, c_2), with its *LCA* concept c_{lca}, the graph tance $Dist_g(c_i, c_j)$ is:

$$Dist_g(c_i, c_j) = SLD(c_i, c_{lca}) + SLD(c_j, c_{lca}) \quad (2)$$

From Equations (1) and (2), the three *SpecLev* of c_i, c_j, c_{lca} can represent both graph distance and *Specification Level Difference*. We then propose a non-linear weighted edge method to combine these three *SpecLev*s into a single *Weighted Edge Distance* ($Dist_w$). We assign each hypernym edge in WordNet hierarchy a weighted value, which is a non-linear exponential decreasing value associated to its *SpecLev*, shown in Figure 1. A coefficient $\alpha \in (0,1]$ is set to represent the *weight decreasing rate* along the edge of WordNet hierarchy. The edge connecting two concepts at *SpecLev* k and $k+1$ has a weighted value α^k.

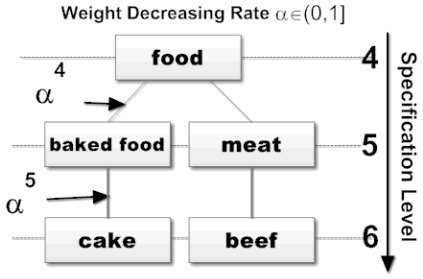

Figure 1. Non-linear Weight on the Edge

Using our weighted edge model, the *Weighted Edge Distance* is the sum of all the edge weight along its shortest path.

$$Dist_w(c_i, c_j) = \sum_{m=SpecLev(c_{lca})}^{SpecLev(c_i)-1} a^m + \sum_{n=SpecLev(c_{lca})}^{SpecLev(c_j)-1} a^n$$

Our weighted edge distance generalizes the traditional graph distance at $\alpha = 1$. When $\alpha \in (0,1)$, the edge value exponentially decreases with the increase of *SpecLev* in the hierarchy.

As illustrated in Figure 2, given two concept-pairs (c_1, c_2) and (c_3, c_4), which have the same *graph distance* and share the same *LCA*, but $SLD(c_1, c_2) = 0$ and $SLD(c_3, c_4) = 2$. We have:

$$Dist_w(c_1, c_2) - Dist_w(c_3, c_4) = \alpha^k - \alpha^{k+1} = \alpha^k(1-\alpha) \geq 0, \alpha \in (0,1] \quad (4)$$

Concept pair (c_1, c_2) has a larger weighted edge distance than that of pair (c_3, c_4). This result conforms to our discovery that humans are more sensitive to the semantic difference caused by categorization than that caused by inheritance.

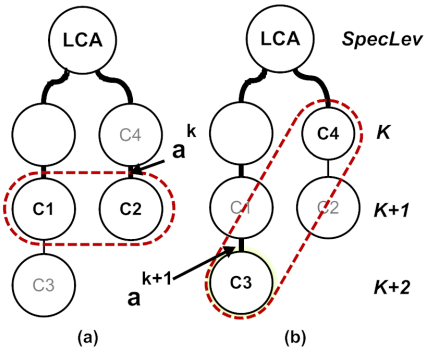

Figure 2. Weighted edge distance depicts difference between categorization and inheritance

2.3 New Transfer Function ϕ

New transfer functions are employed to calculate weighted edge distance to similarity value. We define the semantic similarity of word pair (w_i, w_j) be a function ϕ of its weighted edge distance $Dist_w$ of its max similar concept pair:

$$sim(w_i, w_j) = \phi(Dist_w(c_i, c_j)) \quad (5)$$

Two hyperbolic functions *Hyperbolic Secant* and *Hyperbolic Tangent Cardinal* are used as our transfer functions. Both hyperbolic functions are monotonically decreasing functions with range from 0 to 1.

3 Experimental Studies

3.1 The Benchmark Dataset

Checking the computed similarity against human judgments is a common practice in evaluating the similarity measurement techniques. Many previous studies (Jiang and Conrath, 1997; Lin, 1998; Resnik, 1999; Yang and Powers, 2005; Alvarez and Lim, 2007; Agirre et al., 2009) used either Rubenstein and Goodenough (1965) (RG set) or Miller-Charles (1991) (MC set) as the comparison baseline. In this paper, we conduct similar experiments using our proposed method on three different strategies and calculate the correlation between our computed similarities and

the human judgments. Similar to Li's method (2003), we train the optimal parameter (Weighted Decreasing Rate) using the training set D_1 (37 word pairs in RG set but not in MC set), run on the testing set D_0 to get the correlation with human judgments.

3.2 Experiments on Different Strategies

To ensure the computed similarities obtained by transfer function matches the human judgments as closely as possible; we need to find an optimal Weighted Decreasing Rate α. We propose three different strategies. For each strategy, we use the training set D_1 to obtain the optimal α value. We vary α from 0.05 to 1 with an increment of 0.05, and calculate the correlation between the computed similarities and human judgments on training set D_1. The α value that yields the highest correlation between computed similarities and human judgments is selected as the optimal parameter. Previous studies (Li et al., 2003; Yang and Powers, 2005) need to tune multiple parameters to achieve high performance. In our method, the Weighted Decreasing Rate is the only parameter to tune. The optimal parameter obtained by training set D_1 will be used to calculate the semantic similarity values for word-pairs in testing set D_0. In the end, we calculate the correlations between the computed similarity values and MC human judgments on these word-pairs.

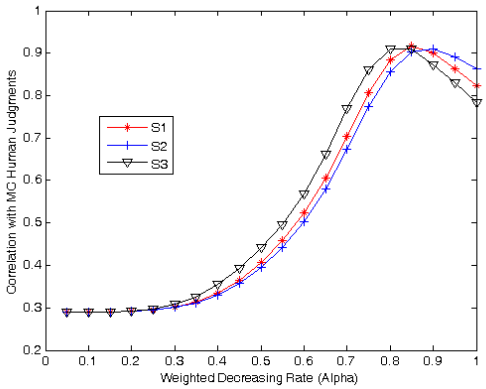

Figure 3. Correlations with human judgments with Weighted Decreasing Rate (α)

Strategy 1: This strategy uses *hyperbolic secant* (*Sech*) as transfer function:

$$\phi_1 = \operatorname{sec} h(x) = \frac{2}{e^x + e^{-x}}$$

$$sim_1(w_i, w_j) = \phi_1(Dist_w(c_i, c_j))$$

As shown S1 in Figure 3, the computed similarities arrive the highest correlation with human judgments on training set D_1 when $\alpha = 0.85$. Using this parameter to calculate the similarities of word-pairs in testing set D_0 we found their correlation with human judgments to be 0.8111.

Strategy 2: In this strategy, we employ the *Hyperbolic Tangent Cardinal* (*Tanhc*) function:

$$\phi_2 = \tanh c(x) = \begin{cases} \dfrac{e^x - e^{-x}}{(e^x + e^{-x}) \cdot x}, & x \neq 0 \\ 1, & x = 0 \end{cases}$$

$$sim_2(w_i, w_j) = \phi_2(Dist_w(c_i, c_j))$$

As shown S2 in Figure 3, when $\alpha = 0.9$, the computed similarities have the highest correlations with human judgments on training set D_1. Using this parameter to calculate the semantic similarities of word pairs in testing set D_0, we found their correlation with human judgments to be 0.8247. This confirms that a better non-linear function can improve the semantic similarity measure.

Strategy 3: The final strategy combines Strategy 1 and Strategy 2 by multiplying the non-linear functions used by these two strategies. The similarity is calculated as:

$$sim_3(w_i, w_j) = \phi_1(Dist_w(c_i, c_j)) \cdot \phi_2(Dist_w(c_i, c_j))$$

As shown S3 in Figure 3, when $\alpha = 0.85$, the computed similarities have the highest correlation with the human judgments on training set D_1. Using $\alpha = 0.85$, we calculate the similarities of word pairs in testing set D_0 and found their correlation with human judgments is 0.8350, which is the highest among all strategies tested. The correlations between the computed similarity values and the human judgments on testing set D_0 using four different strategies are summarized in Table 3.

Strategy	Sim_1 $\alpha = 0.85$	Sim_2 $\alpha = 0.9$	Sim_3 $\alpha = 0.85$
Correlation	0.8111	0.8247	0.8350

Table 3. Correlations between computed similarity values and human judgments on testing set D_0 using different strategies

4 Conclusion and Future work

We present a simple yet effective Weighted Edge method to measure the semantic similarity of word pairs solely using pure hypernym relationship in WordNet. This method is based on an observation that human perception are more sensitive to the semantic difference caused by categorization than specification. The experimental results outperform previous studies only employing hypernym relationship. Our future work will take extra meronym (part-of) relationship and

gloss factors into consideration. Besides, we will conduct future experiments on Similarity353 dataset.

References

Agirre, E., E. Alfonseca, et al. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Boulder, Colorado, Association for Computational Linguistics.

Alvarez, M., A. and S. Lim. 2007. A Graph Modeling of Semantic Similarity between Words. Proceedings of the International Conference on Semantic Computing, IEEE Computer Society.

Banerjee, S. and T. Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. *the Eighteenth International Joint Conference on Artificial Intelligence*, Acapulco, Mexico.

Cilibrasi, R. L. and P. M. B. Vitanyi. 2007. "The Google similarity distance." *Ieee Transactions on Knowledge and Data Engineering* 19(3): 370-383.

Fellbaum, C. 1998. WordNet: An Electronic Lexical Database and Some of its Applications. Cambridge, Mass, MIT Press.

Gabrilovich, E. and S. Markovitch. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis *Proceedings of The Twentieth International Joint Conference for Artificial Intelligence*.

Hughes, T. and D. Ramage. 2007. Lexical Semantic Relatedness with Random Graph Walks. *the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague.

Jiang, J. J. and D. W. Conrath. 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy *Proc. ROCLING X*.

Li, Y. H., Z. A. Bandar, et al. 2003. "An approach for measuring semantic similarity between words using multiple information sources." *Ieee Transactions on Knowledge and Data Engineering* 15(4): 871-882.

Lin, D. 1998. An Information-Theoretic Definition of Similarity. *In Proceedings of the 15th International Conference on Machine Learning*, Morgan Kaufmann.

Miller, G. A. and W. G. Charles. 1991. "Contextual Correlates of Semantic Similarity." *Language and Cognitive Processes* 6(1): 1-28.

Morris, J. and G. Hirst. 1991. "Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text." *Computational Linguistics* 17(1): 21-48.

Patwardhan, S. and T. Pedersen. 2006. Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. *EACL 2006 Workshop Making Sense of Sense---Bringing Computational Linguistics and Psycholinguistics Together*, Trento, Italy.

Rada, R., H. Mili, et al. 1989. "Development and Application of a Metric on Semantic Nets." *Ieee Transactions on Systems Man and Cybernetics* 19(1): 17-30.

Resnik, P. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*.

Resnik, P. 1999. "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language." *Journal of Artificial Intelligence Research* 11: 95-130.

Rubenste.H and Goodenou.Jb. 1965. "Contextual Correlates of Synonymy." *Communications of the Acm* 8(10): 627-&.

Sevilla, J. L., V. Segura, et al. 2005. "Correlation between gene expression and GO semantic similarity." *Ieee-Acm Transactions on Computational Biology and Bioinformatiocs* 2(4): 330-338.

Wu, L., S. H. Hua, et al. 2008. Flickr distance. *Proceeding of the 16th ACM international conference on Multimedia*, Vancouver, British Columbia, Canada

Wu, Z. and M. Palmer. 1994. Verbs semantics and lexical selection. Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Las Cruces, New Mexico, Association for Computational Linguistics.

Yang, D. and D. Powers, M. W. 2005. Measuring semantic similarity in the taxonomy of WordNet. Proceedings of the Twenty-eighth Australasian conference on Computer Science - Volume 38. Newcastle, Australia, Australian Computer Society, Inc.