

---

# Word Similarity In WordNet

Tran Hong-Minh<sup>1</sup> and Dan Smith<sup>1</sup>

School of Computing Sciences  
University Of East Anglia  
Norwich, UK, NR4 7TJ  
`mtht{djs}@cmp.uea.ac.uk`

**Summary.** We present a new information theoretic approach to measure the semantic similarity between concepts. By exploiting advantages of distance (edge-base) approach for taxonomic tree-like concepts, we enhance the strength of information theoretic (node-based) approach. Our measure therefore gives a complete view of word similarity, which cannot be achieved by solely applying node-based approaches. Our experimental measure achieves 88%, correlating with human rating.

## 1 Introduction

Understanding concepts expressed in natural language is a challenge in Natural Language Processing and Information Retrieval. It is often decomposed into comparing semantic relations between concepts, which can be done by using Hidden Markov model and Bayesian Network for part of speech tagging. Alternatively, the knowledge-based approach can also be applied but it has not been well explored due to the lack of machine readable dictionaries (such as lexicons, thesauri and taxonomies) [12]. However, more dictionaries have been developed so far (e.g., Roger, Longman, WordNet [6, 5] and etc.) and the number of research on this trend has been increased. The task of understanding and comparing semantics of concepts becomes understanding and comparing such relations by exploiting machine readable dictionaries.

We propose a new information theoretic measure to assess the similarity of two concepts on the basis of exploring a lexical taxonomy (e.g., WordNet). The proposed formula is domain-independent. It could be applied for either generic or specified lexical knowledge base. We use WordNet as an example of the lexical taxonomy.

The rest of the paper is organized as follows. In Section 2 we give an overview of the structure of a lexical hierarchy and use WordNet as a specific example. In the following section, Section 3 we analyze two approaches (such as distance (edge) based and information theoretic (node) based) for measuring the similarity degree. Based on these analysis we present our measure

which combines both advantages of the two approaches in Section 4. In Section 5 we discuss our comparative experiments. Finally we outline our future work in Section 6.

## 2 Lexical Taxonomy

A taxonomy is often organized as a hierarchical and directional structure, in which nodes present for concepts (Noun, Adjective, Verb) and edges present for relations between concepts. The hierarchical structure has seldom more than 10 levels in depth. Although hierarchies in the system vary widely in size, each hierarchy covers a distinct conceptual and lexical domain. They are also not mutually exclusive as some cross-references are required.

The advantage of the hierarchical structure is that common information to many items need not to be stored with every item. In the other word, all characteristics of the superordinate are assumed to be characteristic of all its subordinates as well. The hierarchical system therefore is called inheritance system with possibly multiple inheritance but without forming circular loops. Consequently, nodes at deeper levels are more informative and specific than nodes that are nearer to the root. In principle, the root would be semantically empty. The number of leaf-nodes is obviously very much more than the number of upper nodes.

In a hierarchical system, there are three types of nodes, such as, concept nodes indicating nouns (a.k.a Noun node), attribute nodes representing adjectives and function nodes standing for verbs. Nodes are linked together by edges to give a full information about concepts. A node and a set of nodes linked with by incoming edges make it distinguished.

Edges represent the relations between nodes. They are currently categorized into some popular types (such as, is-a, equivalence, antonymy, modification, function and meronymy). Among them, the IS-A relation, connecting a Noun node and another Noun node, is the dominant and the most important one. Like the IS-A relation, the meronymy relation connecting two Noun nodes together also has an important role in the system. Besides the two popular relations, there are four more types of relations. The antonymy relation (e.g. man-woman, wife-husband), the equivalence relation connects synonyms together. The modification indicates attributes of a concept by connecting a Noun node and an Adjective node and the function relation indicates behaviour of a concept by linking a Verb to a Noun. In Table 1 the characteristics of such relations are briefly summarized.

In practice, one example of such lexical hierarchical systems is WordNet which is currently one of the most popular and the largest online dictionary produced by Miller *et al* from Princeton University in 1990s. It supports multiple inheritance between nodes and has the most numerous of relations implemented. WordNet hierarchical system includes 25 different branches rooted by 25 distinguish concepts. Each of such 25 concepts can be considered as the

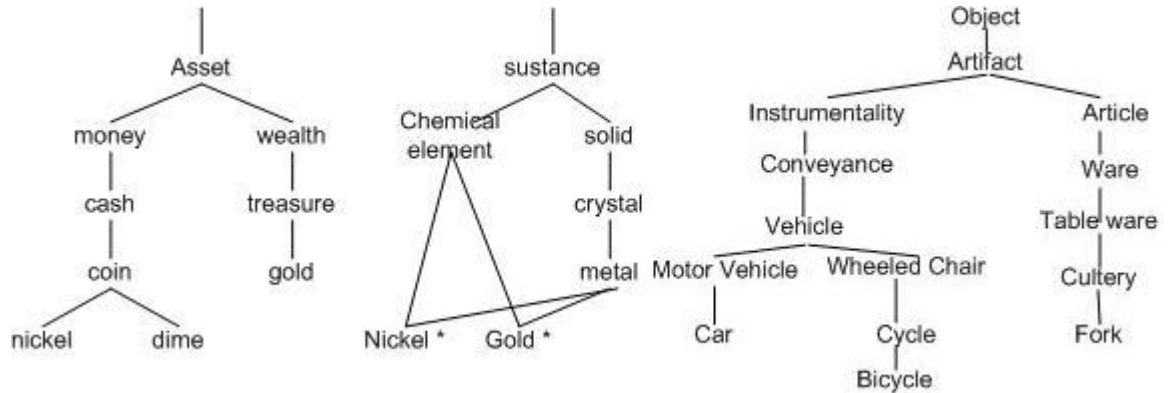
**Table 1.** Characteristic of relations in the lexical hierarchical system

	Is-A	Meronymy	Equivalence	Modification	Function	Antonymy
Transitive	✓	✓	✓	✓	✓	×
Symmetric	✓	×	✓	×	×	✓

beginners of the branches and regarded as a primitive semantic component of all concepts in its semantic hierarchy. Table 2 shows such beginners.

**Table 2.** List of 25 unique beginners for WordNet nouns

{act, action, activity}	{natural object}	{food}
{animal, fauna}	{natural phenomenon}	{group, collection}
{artifact}	{person, human being}	{location, place}
{attribute, property}	{plant, flora}	{motive}
{body, corpus}	{possession}	{shape}
{cognition, knowledge}	{process}	{state, condition}
{communication}	{quantity, amount}	{substance}
{event, happening}	{relation}	{time}
{feeling, emotion}		

**Fig. 1.** Fragments of WordNet noun taxonomy

Like many other lexical inheritance systems, the IS-A and the meronymy relations are fully supported in WordNet. Although the modification and the function relations have not been implemented, the antonymy and the synonym sets are implemented in WordNet. Figure 1 shows fragments of WordNet noun hierarchy.

With a hierarchical structure, similarity can be obtained not only by solely comparing the common semantics between two nodes in the system (informa-

tion theoretic based approach) but also by measuring their position in the structure and their relations (distance based approach).

### 3 Information Theoretic vs. Conceptual Distance Approach for Measuring Similarity

Based on different underlying assumptions about taxonomy and definitions of similarity (e.g., [10, 7, 3, 2, 1], etc.), there are two main trends for measuring semantic similarity between two concepts: node based approach (information content approach) vs. edge based approach (conceptual distance approach). The most distinguish characteristic of node-based approach is that the similarity between nodes is measured directly and solely by the common information content. Since taxonomy is often represented as a hierarchical structure — a special case of network structure — similarity between nodes can make use of the structural information embedded in the network, especially links between nodes. This is the main idea of edge-based approaches.

#### 3.1 Conceptual Distance Approach

The conceptual distance approach is natural, intuitive and direct to the problem of measuring the similarity of concepts in the hierarchical system with lexical labels presented in Section 2. The similarity between concepts is related to their differences in the conceptual distance between them. The more differences they have, the less similar they are. The distance between concepts is measured by the geometric distance between nodes presenting concepts.

**Definition 1** *Given two concepts  $c_1$  and  $c_2$  and  $\text{dist}(c_1, c_2)$  as the distance between  $c_1$  and  $c_2$ , the difference between  $c_1$  and  $c_2$  is equal to the distance  $\text{dist}(c_1, c_2)$  between them [3].*

**Definition 2** *The distance  $\text{dist}(c_1, c_2)$  between  $c_1$  and  $c_2$  is the sum of weights  $\text{wt}_i$  of edges  $e_i$  in the shortest path from  $c_1$  to  $c_2$  :*

$$\text{dist}(c_1, c_2) = \sum_{\text{wt}_i \in \{\text{wt}_i \text{ of } e_i \mid e_i \in \text{shortestPath}(c_1, c_2)\}} (\text{wt}_i) \quad (1)$$

As being a distance, Formula (1) should satisfy the properties of a metric [10], such as zero property, positive property and triangle inequality. However, the symmetric property may not be satisfied,  $\text{dist}(c_1, c_2) \neq \text{dist}(c_2, c_1)$ , as different types of relations give different contributions into the weight of the edge connecting two nodes. For example, regarding the meronymy type of relation, a aggregative relation may have different contributions with a part-of-relation, though they are reverse relation of each other.

Most of contributions to the weight of an edge come from the characteristics of the hierarchical network, such as local network density, depth of a node in the hierarchy, type of link and strength of link:

- Network density of a node can be the number of its children. Richardson *et al*[8] suggest that the greater density the closer distance between parent-child nodes or sibling nodes.
- The distance between parent-child nodes is also closer at deeper levels, since the differentiation at such levels is less.
- The strength of a link is based on the closeness between a child node to its direct parent, against those of its siblings. This is the most important factor determining the weighting to assign to an edge, but determining the optimal weighting is an open issue.

There are studies on conceptual similarity by using the distance approach with above characteristics of the hierarchical network (e.g., [1, 11]). Most research focus on proposing an edge-weighting formula and then applying Formula (1) for measuring the conceptual distance.

For instance, Sussna [11] considers depth, relation type and network density in his weighting formula as follows:

$$\text{wt}(c_1, c_2) = \frac{\text{wt}(c_1 \rightarrow_r c_2) + \text{wt}(c_2 \rightarrow_{r'} c_1)}{2d} \quad (2)$$

in which

$$\text{wt}(x \rightarrow_r y) = \max_r - \frac{\max_r - \min_r}{n_r(x)} \quad (3)$$

where  $\rightarrow_r$ ,  $\rightarrow_{r'}$  are respectively a relation of type  $r$  and its reverse.  $d$  is the deeper of  $c_1$  and  $c_2$  in the hierarchy.  $\min_r$  and  $\max_r$  are respectively the minimum and maximum weight of relation of type  $r$ .  $n_r(x)$  is the number of relation type  $r$  of node  $x$ , which is viewed as the network density of the node  $x$ . The conceptual distance is then given by applying Formula (1). It gives a good result in a word sense disambiguation task with multiple sense words. However, the formula does not take into account the strength of relation between nodes, which is still an open issue for the distance approach.

In summary, the distance approach obviously requires a lot of information on detailed structure of taxonomy. Therefore it is difficult to apply or directly manipulate it on a generic taxonomy, which originally is not designed for similarity computation.

### 3.2 Information Theoretic Approach

The information theoretic approach is more soundly based. Therefore it is generic and applied on many taxonomies without regarding their underlying structure. In a conceptual space, a node presents a unique concept and contains a certain amount of information. The similarity between concepts is related to the information in common of nodes. The more commonality they share, the more similar they are.

Given concept  $c_1$ , concept  $c_2$ ,  $IC(c)$  is the information content value of concept  $c$ . Let  $w$  be the word denoted concept  $c$ . For example, in Figure 1, word `nickel` has three senses:

- “a United States coin worth one twentieth of a dollar” (concept `coin` )
- “atomic number 28” (concept `chemical element` )
- “a hard malleable ductile silvery metallic element that is resistant to corrosion; used in alloys; occurs in pentlandite and smaltite and garnierite and millerite” (concept `metal` ).

Let  $s(w)$  be the set of concepts in the taxonomy that are senses of word  $w$ . For example, in Figure 1, words `nickel`, `coin` and `cash` are all members of the set  $s(\text{nickel})$ . Let  $\text{Words}(c)$  be the set of words subsumed by concept  $c$ .

**Definition 3** *The commonality between  $c_1$  and  $c_2$  is measured by the information content value used to state the commonalities between  $c_1$  and  $c_2$  [3]:*

$$IC(\text{common}(c_1, c_2)) \quad (4)$$

**Assumption 1** *The maximum similarity between  $c_1$  and  $c_2$  is reached when  $c_1$  and  $c_2$  are identical, no matter how much commonality they share.*

**Definition 4** *In information theory, the information content value of a concept  $c$  is generally measured by*

$$IC(c) = -\log P(c) \quad (5)$$

where  $P(c)$  is the probability of encountering an instance of concept  $c$ . For implementation, the probability is practically measured by the concept frequency.

Resnik [7] suggests a method of calculating the concept probabilities in a corpus on the basis of word occurrences. Given  $\text{count}(w)$  as the number of occurrences of a word belonging to concept  $c$  in the corpus,  $N$  as the number of concepts in the corpus, the probability of a concept  $c$  in the corpus is defined as follows:

$$P(c) = \frac{1}{N} \times \sum_{w \in \text{Words}(c)} \text{count}(w) \quad (6)$$

In a taxonomy, the shared information of two concepts  $c_1$  and  $c_2$  is measured by the information content value of the concepts that subsume them. Given  $\text{sim}(c_1, c_2)$  as the similarity degree of two concepts  $c_1$  and  $c_2$  and  $\text{Sup}(c_1, c_2)$  as the set of concepts that subsume both  $c_1$  and  $c_2$ , the formal definition of similarity degree between  $c_1$  and  $c_2$  is given as follows:

$$\text{sim}(c_1, c_2) = \begin{cases} \max_{c \in \text{Sup}(c_1, c_2)} IC(c), & c_1 \neq c_2, \\ 1, & c_1 = c_2. \end{cases} \quad (7)$$

The word similarity between  $w_1$  and  $w_2$  is formally defined:

$$\text{sim}(w_1, w_2) = \max_{c_1 \in S(w_1), c_2 \in S(w_2)} [\text{sim}(c_1, c_2)] \quad (8)$$

When applying the above formulas to a hierarchical concept space, there are some slight specifications. A set of words  $\text{Words}(c)$ , which is directly or indirectly subsumed by the concept  $c$ , is considered as all nodes in the sub-tree rooted by  $c$ , including  $c$ . Therefore, when we move from the leaves to the root of the hierarchy, Formula (6) therefore gives a higher probability to encounter a concept at the upper level. The probability of the root obviously is 1. Consequently, the information content value given by Formula (5) monotonically decreases in the bottom-up direction and the information content value of the root is 0. Those means that concepts at the upper levels are less informative and the characteristic of lexical hierarchical structure discussed in Section 2 is qualified.

In a lexical hierarchical concept space,  $\text{Sup}(c_1, c_2)$  contains all superordinates of  $c_1$  and  $c_2$ . For example, in Figure 1 **coin**, **cash**, **money** are all member of  $\text{Sup}(\text{nickel}, \text{dime})$ . However, as analysis above, only  $\text{IC}(\text{coin})$  gives the highest information content value. The similarity computed by using Formula (7)  $\text{sim}(\text{nickel}, \text{dime})$  therefore is equal to the information content value of its direct superordinate,  $\text{IC}(\text{coin})$ . So the direct superordinate of a node in a hierarchy (e.g. **coin** is the direct superordinate of **nickel** and **dime**) is called the minimum upper bound of the node. Similarly for a multiple inheritance system, the similarity between concepts  $\text{sim}(c_1, c_2)$  is equal to the maximum information content value among those of their minimum upper bound. For example, in Figure 1,

$$\text{sim}(\text{nickel}^*, \text{gold}^*) = \max[\text{IC}(\text{chemicalelement}), \text{IC}(\text{metal})]$$

To conclude, unlike the distance approach, the information theoretic approach requires less structural information of the taxonomy. Therefore it is generic and flexible and has wide applications on many types of taxonomies. However, when it is applied on hierarchical structures it does not differentiate the similarity of concepts as long as their minimum upper bounds are the same. For example, in Figure 1,  $\text{sim}(\text{bicycle}, \text{fork})$  and  $\text{sim}(\text{bicycle}, \text{tableware})$  are equal.

## 4 A Measure for Word Similarity

We propose a combined model for measuring word similarity which is derived from the node-based notion by adding the structural information. We put the depth factor and link strength factor into the node-based approach. By adding such structural information of the taxonomy the node-based approach can exploit all typical characteristics of a hierarchical structure when it is applied on such taxonomy. Moreover, such information can be tuned via parameters. The method therefore is flexible for many types of taxonomy (e.g., hierarchical structure or plain structure).

**Definition 5** *The strength of a link is defined to be  $P(c_i|p)$ , the conditional probability of encountering a child node  $c_i$ , given an instance of its parent node  $p$ . Using Bayesian formula, we have:*

$$P(c_i|p) = \frac{P(c_i \cap p)}{P(p)} = \frac{P(c_i)}{P(p)} \quad (9)$$

The information content value of a concept  $c$  with regarding to its direct parent  $p$ , which is a modification of the Formula (5), is given:

$$IC(c|p) = -\log P(c|p) = -\log \left[ \frac{P(c)}{P(p)} \right] = IC(c) - IC(p) \quad (10)$$

As we discussed in Section 2, concepts at upper levels of the hierarchy have less semantic similarity between them than concepts at lower levels. This characteristic should be taken into account as a constraint in calculating the similarity of two concepts with depth concern. Therefore, the depth function should give a higher value when applied on nodes at lower levels.

The contribution of the depth to the similarity is considered as an exponential-growth function:

$$f_{c_1, c_2}(d) = \frac{e^{\alpha d} - e^{-\alpha d}}{e^{\alpha d} + e^{-\alpha d}}, \quad (11)$$

where  $d = \max(\text{depth}(c_1), \text{depth}(c_2))$  and  $\alpha$  is a tuning parameter. The optimal value of the parameter is  $\alpha = 0.3057$ , based on our numerous experiments.

Function (11) is a monotonically increasing function with respect to depth  $d$ . Therefore it satisfies the constraint above. Moreover, by employing an exponential-growth function rather than an exponential-decay function, it is an extension of Shrepard's Law [9, 2], which claims that exponential-decay function are a universal law of stimulus generalisation for psychological science.

Then, the function given in Formula (7) is now a function of the depth and the information content with the concern of the strength of a link as follows:

$$\text{sim}(c_1, c_2) = \begin{cases} \max_{c \in \text{Sup}(c_1, c_2)} (IC(c|p) \times f_c(d)), & c_1 \neq c_2, \\ 1, & c_1 = c_2. \end{cases} \quad (12)$$

## 5 Experiments

Although there is no standard way to evaluate computational measures of semantic similarity, one reasonable way to judge would seem to be agreement with human similarity ratings. This can be assessed by measuring and rating the similarity of each word pair in a set and then looking at how well its ratings correlate with human ratings of the same pairs.

We use the human ratings done by Miller and Charles [4] and revised by Resnik [7] as our baseline. In their study, 38 undergraduate subjects are given 30 pairs of nouns and were asked to rate the smilarity of meaning for each pair on scale from 0 (dissimilar) to 4 (synonym). The average rating of each pair represents a good estimate of how similar the two words are.

Furthermore, we compare our similarity value with those produced by a simple edge-count measure and Lin’s [3]. We use WordNet 2.0 as the hierarchical system to exploit the relationships among the pairs. Table 3 shows that

**Table 3.** Results obtained evaluating with human judgement and WordNet 2.0

<i>word</i> <sub>1</sub>	<i>word</i> <sub>2</sub>	Human	<i>sim</i> <sub>edge</sub>	<i>sim</i> <sub>Lin</sub>	ours	<i>word</i> <sub>1</sub>	<i>word</i> <sub>2</sub>	Human	<i>sim</i> <sub>edge</sub>	<i>sim</i> <sub>Lin</sub>	ours
car	automobile	3.92	1.00	1.00	1.00	lad	brother	1.66	0.20	0.29	0.27
gem	jewel	3.84	1.00	1.00	1.00	journey	car	1.16	0.07	0.00	0.00
journey	voyage	3.84	0.50	0.69	0.92	monk	oracle	1.10	0.13	0.23	0.26
boy	lad	3.76	0.50	0.82	0.87	cemetery	woodland	0.95	0.10	0.08	0.07
coast	shore	3.70	0.50	0.97	1.00	food	rooster	0.89	0.07	0.10	0.26
asylum	madhouse	3.61	0.50	0.98	0.90	coast	hill	0.87	0.20	0.71	0.71
magician	wizard	3.50	1.00	1.00	1.00	forest	graveyard	0.84	0.10	0.08	0.13
midday	noon	3.42	1.00	1.00	1.00	shore	woodland	0.63	0.17	0.14	0.27
furnace	stove	3.11	0.13	0.22	0.32	monk	slave	0.55	0.20	0.25	0.31
food	fruit	3.08	0.13	0.13	0.73	coast	forest	0.42	0.14	0.13	0.38
bird	cock	3.05	0.50	0.80	0.85	lad	wizard	0.42	0.20	0.27	0.21
bird	crane	2.97	0.25		0.85	chord	smile	0.13	0.09	0.27	0.07
tool	implement	2.95	0.50	0.92	0.73	glass	magician	0.11	0.13	0.13	0.07
brother	monk	2.82	0.50	0.25	0.54	noon	string	0.08	0.08	0.00	0.00
crane	implement	1.68	0.20		0.80	rooster	voyage	0.08	0.05	0.00	0.00
correlation								1.00	0.77	0.80	0.88

our approach gives the results are the most correlative with human ratings of the same pairs. Our experimental measure achieves 88%, correlating with human results. Moreover, observing the Table 3 we also notice that the information theoretic approaches (Lin’s approach and ours) deliver better results than the simple distance based approach tested.

## 6 Conclusion

We have presented a review on two main trends of measuring similarity of words in a generic and hierarchical corpus. Based on such review we proposed a modification on the node based approach to capture the structural information of a hierarchical taxonomy. Therefore our approach give a complete view on similarity of words.

## References

1. J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *the International Conference on Research in Computational Linguistics*, 1997.
2. Y. Li, Z. A. Bandar, and D. McLean. An approach for measuring semantic similarity between words using multiple information sources. In *IEEE Transaction on Knowledge and Data Transaction*, volume 15, pages 871–882, 2003.
3. D. Lin. An information-theoretic definition of similarity. In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.
4. G. Miller and W. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.
5. G. A. Miller. Nouns in wordnet: A lexical inheritance system. *International journal of Lexicography*, 3(4):245–264, 1990.
6. G. A. Miller, C. Fellbaum, R. Beckwith, D. Gross, and K. Miller. Introduction to wordnet: An online lexical database. *International journal of Lexicography*, 3(4):235–244, 1990.
7. P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, pages 448–453, 1995.
8. R. Richardson and A. F. Smeaton. Using WordNet in a knowledge-based approach to information retrieval. Technical Report CA-0395, Dublin, Ireland, 1995.
9. S. RN. Toward a universal law of generalization for psychological science. 237(4820):1317–1323, September 1987.
10. R. Roy, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. In *IEEE Transactions on Systems, Man and Cybernetics*, volume 19, pages 17–30, 1989.
11. M. Sussna. Word sense disambiguation for free-text indexing using a massive semantic network. In *CIKM '93: Proceedings of the second international conference on Information and knowledge management*, pages 67–74, New York, NY, USA, 1993. ACM Press.
12. G. William A., C. Kenneth W., and Y. David. A method for disambiguating word senses in a large corpus. *Common Methodologies in Humanities Computing and Computational Linguistics*, 26:415–439, 1992.