# Information Content Measures of Semantic Similarity Perform Better Without Sense-Tagged Text

**Ted Pedersen**
Department of Computer Science
University of Minnesota, Duluth
Duluth, MN 55812
tpederse@d.umn.edu
http://wn-similarity.sourceforge.net

## Abstract

This paper presents an empirical comparison of similarity measures for pairs of concepts based on Information Content. It shows that using modest amounts of untagged text to derive Information Content results in higher correlation with human similarity judgments than using the largest available corpus of manually annotated sense–tagged text.

## 1 Introduction

Measures of semantic similarity based on WordNet have been widely used in Natural Language Processing. These measures rely on the structure of WordNet to produce a numeric score that quantifies the degree to which two concepts (represented by a sense or synset) are similar (or not). In their simplest form these measures use path length to identify concepts that are physically close to each other and therefore considered to be more similar than concepts that are further apart.

While this is a reasonable first approximation to semantic similarity, there are some well known limitations. Most significant is that path lengths between very specific concepts imply much smaller distinctions in semantic similarity than do comparable path lengths between very general concepts. One proposed improvement is to augment concepts in WordNet with *Information Content* values derived from sense–tagged corpora or from raw unannotated corpora (Resnik, 1995).

This paper shows that Information Content measures based on modest amounts of unannotated corpora have greater correlation with human similarity judgements than do those based on the largest corpus of sense-tagged text currently available.[1] The key to this success is not in the specific type of corpora used, but rather in increasing the number of concepts in WordNet that have counts associated with them. These results show that Information Content measures of semantic similarity can be significantly improved without requiring the creation of sense–tagged corpora (which is very expensive).

### 1.1 Information Content

Information Content (IC) is a measure of specificity for a concept. Higher values are associated with more specific concepts (e.g., *pitch_fork*), while those with lower values are more general (e.g., *idea*). Information Content is computed based on frequency counts of concepts as found in a corpus of text. The frequency associated with a concept is incremented in WordNet each time that concept is observed, as are the counts of the ancestor concepts in the WordNet hierarchy (for nouns and verbs). This is necessary because each occurrence of a more specific concept also implies the occurrence of the more general ancestor concepts.

When a corpus is sense–tagged, mapping occurrences of a word to a concept is straightforward (since each sense of a word corresponds with a concept or synset in WordNet). However, if the text has not been sense–tagged then all of the possible senses of a given word are incremented (as are their ancestors). For example, if *tree* (as a plant) occurs in a sense–tagged text, then only the concept associated

[1] These experiments were done with version 2.05 of WordNet::Similarity (Pedersen et al., 2004).

with tree as a kind of plant would be incremented. If the text is untagged, then all of the possible senses of *tree* would be incremented (such as the mathematical sense of tree, a shoe tree, a plant, etc.) In this case the frequency of all the occurrences of a word are divided equally among the different possible senses. Thus, if a word occurs 42 times in a corpus and there are six possible senses (concepts), each sense and all of their ancestors would have their frequency incremented by seven.[2]

For each concept (synset) $c$ in WordNet, Information Content is defined as the negative log of the probability of that concept (based on the observed frequency counts):

$$IC(c) = -logP(c)$$

Information Content can only be computed for nouns and verbs in WordNet, since these are the only parts of speech where concepts are organized in hierarchies. Since these hierarchies are separate, Information Content measures of similarity can only be applied to pairs of nouns or pairs of verbs.

## 2 Semantic Similarity Measures

There are three Information Content measures implemented in WordNet::Similarity: (res) (Resnik, 1995), (jcn) (Jiang and Conrath, 1997), and (lin) (Lin, 1998).

These measures take as input two concepts $c_1$ and $c_2$ (i.e., senses or synsets in WordNet) and output a numeric measure of similarity. These measures all rely to varying degrees on the idea of a least common subsumer (LCS); this is the most specific concept that is a shared ancestor of the two concepts. For example, the LCS of *automobile* and *scooter* is *vehicle*.

The Resnik (res) measure simply uses the Information Content of the LCS as the similarity value:

$$res(c_1, c_2) = IC(LCS(c_1, c_2))$$

The Resnik measure is considered somewhat coarse, since many different pairs of concepts may share the same LCS. However, it is less likely to suffer from zero counts (and resulting undefined values) since in general the LCS of two concepts will not be a very specific concept (i.e., a leaf node in WordNet), but will instead be a somewhat more general concept that is more likely to have observed counts associated with it.

Both the Lin and Jiang & Conrath measures attempt to refine the Resnik measure by augmenting it with the Information Content of the individual concepts being measured in two different ways:

$$lin(c_1, c_2) = \frac{2*res(c_1,c_2)}{IC(c_1)+IC(c_2)}$$

$$jcn(c_1, c_2) = \frac{1}{IC(c_1)+IC(c_2)-2*res(c_1,c_2)}$$

All three of these measures have been widely used in the NLP literature, and have tended to perform well in a wide range of applications such as word sense disambiguation, paraphrase detection, and Question Answering (c.f., (Resnik, 1999)).

## 3 Experimental Data

Information Content in WordNet::Similarity is (by default) derived from SemCor (Miller et al., 1993), a manually sense–tagged subset of the Brown Corpus. It is made up of approximately 676,000 words, of which 226,000 are sense–tagged. SemCor was originally created using sense–tags from version 1.6 of WordNet, and has been mapped to subsequent versions to stay current.[3] This paper uses version 3.0 of WordNet and SemCor.

WordNet::Similarity also includes a utility (rawtextFreq.pl) that allows a user to derive Information Content values from any corpus of plain text. This utility is used with the untagged version of SemCor and with various portions of the English GigaWord corpus (1st edition) to derive alternative Information Content values.

English GigaWord contains more than 1.7 billion words of newspaper text from the 1990's and early 21st century, divided among four different sources: Agence France Press English Service (afe), Associated Press Worldstream English Service (apw), The New York Times Newswire Service (nyt), and The Xinhua News Agency English Service (xie).

This paper compares the ranking of pairs of concepts according to Information Content measures in WordNet::Similarity with a number of manually created gold standards. These include the (RG) (Rubenstein and Goodenough, 1965) collection of 65 noun

[2]This is the –resnik counting option in WordNet::Similarity.

[3]http://www.cse.unt.edu/˜rada/downloads.html

Table 1: Rank Correlation of Existing Measures

| measure | WS | MC | RG |
|---|---|---|---|
| vector | **.46** | **.89** | **.73** |
| lesk | .42 | .83 | .68 |
| wup | .34 | .74 | .69 |
| lch | .28 | .71 | .70 |
| path | .26 | .68 | .69 |
| random | -.20 | -.16 | .15 |

pairs, the (MC) (Miller and Charles, 1991) collection of 30 noun pairs (a subset of RG), and the (WS) WordSimilarity-353 collection of 353 pairs (Finkelstein et al., 2002). RG and MC have been scored for similarity, while WS is scored for relatedness, which is a more general and less well–defined notion than similarity. For example *aspirin* and *headache* are clearly related, but they aren't really similar.

## 4 Experimental Results

Table 1 shows the Spearman's rank correlation of several other measures of similarity and relatedness in WordNet::Similarity with the gold standards discussed above. The WordNet::Similarity vector relatedness measure achieves the highest correlation, followed closely by the adapted lesk measure. These results are consistent with previous findings (Patwardhan and Pedersen, 2006). This table also shows results for several path–based measures.[4]

Table 2 shows the correlation of jcn, res, and lin when Information Content is derived from 1) the sense-tagged version of SemCor (semcor), 2) SemCor without sense tags (semcor-raw), and 3) steadily increasing subsets of the 133 million word xie portion of the English GigaWord corpus. These subsets start with the entire first month of xie (199501, from January 1995) and then two months (199501-02), three months (199501-03), up through all of 1995 (199501-12). Thereafter the increments are annual, with two years of data (1995-1996), then three (1995-1997), and so on until the entire xie corpus is used (1995-2001). The afe, apw, and nyt portions of GigaWord are also used individually and then combined all together along with xie *(all)*.

[4] *wup* is the Wu & Palmer measure, *lch* is the Leacock & Chodorow measure, *path* relies on edge counting, and *random* provides a simple sanity check.

The size (in tokens) of each corpus is shown in the second column of Table 2 *(size)*, which is expressed in thousands (k), millions (m), and billions (b).

The third column (*cover*) shows what percentage of the 96,000 noun and verb synsets in WordNet receive a non-zero frequency count when Information Content is derived from the specified corpus. These values show that the 226,000 sense–tagged instances in SemCor cover about 24%, and the untagged version of SemCor covers 37%. As it happens the correlation results for semcor-raw are somewhat better than semcor, suggesting that coverage is at least as important (if not more so) to the performance of Information Content measures than accurate mapping of words to concepts.

A similar pattern can be seen with the xie results in Table 2. This again shows that an increase in WordNet coverage is associated with increased performance of the Information Content measures. As coverage increases the correlation improves, and in fact the results are better than the path–based measures and approach those of lesk and vector (see Table 1). The one exception is with respect to the WS gold standard, where vector and lesk perform much better than the Information Content measures. However, this seems reasonable since they are relatedness measures, and the WS corpus is annotated for relatedness rather than similarity.

As a final test of the hypothesis that coverage matters as much or more than accurate mapping of words to concepts, a simple baseline method was created that assigns each synset a count of 1, and then propagates that count up to the ancestor concepts. This is equivalent to doing add-1 smoothing without any text (add1only). This results in correlation nearly as high as the best results with xie and semcor-raw, and is significantly better than semcor.

## 5 Conclusions

This paper shows that semantic similarity measures based on Information Content can be significantly improved by increasing the coverage of the frequency counts used to derive Information Content. Increased coverage can come from unannotated text or simply assigning counts to every concept in WordNet and does not require sense–tagged text.

Table 2: Rank Correlation of Information Content Measures From Different Corpora

| | | | jcn | | | lin | | | res | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| corpus | size | cover | WS | MC | RG | WS | MC | RG | WS | MC | RG |
| semcor | 226 k | .24 | .21 | .72 | .51 | .30 | .73 | .58 | .38 | .74 | .69 |
| semcor-raw | 670 k | .37 | .26 | .82 | .58 | .32 | .79 | .65 | .38 | .76 | .70 |
| xie: | | | | | | | | | | | |
| 199501 | 1.2 m | .35 | .35 | .78 | .57 | .37 | .75 | .63 | .37 | .73 | .68 |
| 199501-02 | 2.3 m | .39 | .31 | .79 | .65 | .32 | .75 | .67 | .36 | .73 | .68 |
| 199501-03 | 3.8 m | .42 | .34 | .88 | .69 | .34 | .81 | .70 | .37 | .75 | .69 |
| 199501-06 | 7.9 m | .46 | .36 | .88 | .69 | .36 | .81 | .70 | .37 | .75 | .69 |
| 199501-09 | 12 m | .49 | .36 | .88 | .69 | .36 | .81 | .70 | .37 | .75 | .69 |
| 199501-12 | 16 m | .51 | .37 | .87 | .73 | .36 | .81 | .71 | .37 | .75 | .69 |
| 1995-1996 | 34 m | .56 | .37 | .88 | .73 | .36 | .81 | .72 | .37 | .75 | .69 |
| 1995-1997 | 53 m | .58 | .37 | .88 | .73 | .36 | .81 | .71 | .37 | .75 | .69 |
| 1995-1998 | 73 m | .60 | .37 | .89 | .73 | .36 | .81 | .72 | .37 | .75 | .69 |
| 1995-1999 | 94 m | .62 | .36 | .88 | .73 | .36 | .81 | .72 | .37 | .76 | .69 |
| 1995-2000 | 115 m | .63 | .36 | .89 | .73 | .36 | .81 | .71 | .37 | .76 | .70 |
| 1995-2001 | 133 m | .64 | .36 | .88 | .73 | .36 | .81 | .71 | .37 | .76 | .70 |
| afe | 174 m | .66 | **.36** | **.88** | **.81** | .36 | .80 | .78 | .37 | .77 | .79 |
| apw | 560 m | .75 | .36 | .84 | .78 | .36 | .79 | .78 | .37 | .76 | .79 |
| nyt | 963 m | .83 | .36 | .84 | .78 | .36 | .79 | .77 | .37 | .77 | .80 |
| all | 1.8 b | .85 | .34 | .85 | .79 | .35 | .80 | .78 | .37 | .77 | .79 |
| add1only | 96 k | 1.00 | .36 | .85 | .73 | .37 | .77 | .73 | .39 | .76 | .70 |

## Acknowledgements

Many thanks to Siddharth Patwardhan and Jason Michelizzi for their exceptional work on WordNet::Similarity over the years, which has made this and a great deal of other research possible.

## References

L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

J. Jiang and D. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings on International Conference on Research in Computational Linguistics*, pages 19–33, Taiwan.

D. Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the International Conference on Machine Learning*, Madison, August.

G.A. Miller and W.G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

G.A. Miller, C. Leacock, R. Tengi, and R. Bunker. 1993. A semantic concordance. In *Proceedings of the Workshop on Human Language Technology*, pages 303–308.

S. Patwardhan and T. Pedersen. 2006. Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. In *Proceedings of the EACL 2006 Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together*, pages 1–8, Trento, Italy, April.

T. Pedersen, S. Patwardhan, and J. Michelizzi. 2004. Wordnet::Similarity - Measuring the relatedness of concepts. In *Proceedings of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 38–41, Boston, MA.

P. Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, Montreal, August.

P. Resnik. 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.

H. Rubenstein and J.B. Goodenough. 1965. Contextual correlates of synonymy. *Computational Linguistics*, 8:627–633.