Measuring Semantic Similarity in the Taxonomy of WordNet

Dongqiang Yang **David M.W. Powers**

School of Informatics and Engineering
Flinders University of South Australia
PO Box 2100, Adelaide 5001, South Australia

{Dongqiang.Yang|David.Powers}@flinders.edu.au

Abstract

This paper presents a new model to measure semantic similarity in the taxonomy of WordNet, using edge-counting techniques. We weigh up our model against a benchmark set by human similarity judgment, and achieve a much improved result compared with other methods: the correlation with average human judgment on a standard 28 word pair dataset is 0.921, which is better than anything reported in the literature and also significantly better than average individual human judgments. As this set has been effectively used for algorithm selection and tuning, we also cross-validate an independent 37 word pair test set (0.876) and present results for the full 65 word pair superset (0.897).

Keywords: semantic similarity, correlation, taxonomy.

1 Introduction

Word similarity is a widespread topic in natural language processing (NLP). It has been applied in a number of applications, including word sense disambiguation, detection of malapropisms, information retrieval, natural language learning etc.

The popular methodologies in measuring semantic relatedness with the help of a thesaurus can be classified into two categories: one uses solely semantic links (i.e. edge-counting), the other combines corpus statistics with taxonomic distance.

Generally speaking, similarity models in the taxonomy of WordNet, proposed by Wu and Palmer (1994), Leacock and Chodorow (1998), Jiang and Conrath (1997), and Lin (1997), can be abstracted into one of the following forms,

$$Sim(c1, c2) = 2\gamma \div (\alpha + \beta) \quad (1)$$

$$Sim(c1, c2) = 2\gamma - (\alpha + \beta) \quad (2)$$

where α, β and γ respectively denote attributes of concepts $c1$, $c2$ and the nearest common node (*ncn*) of $c1$ and $c2$ in the 'IS-A' hierarchy. The attribute can be viewed as the distance in the taxonomy or information content extracted from the outer corpus. Distance is typically assessed by counting the edges traversed from $c1$ to $c2$ via *ncn*, *dist*($c1$, $c2$). The idea of edge-counting goes back to Quillian's semantic memory model (Quillian 1967; Collins and Quillian 1969) where concept nodes are planted within the hierarchical network and the number of hops between the nodes specifies the similarity or difference of concepts. Indeed it can be traced back further to Zipf (1965) who proposed an access model to explain his well-known law.

The work of Hirst and St. Onge (1995) departs from this model in that they set different weights for different links in the semantic net in order to more closely model human performance.

Resnik (1995) also notes that having the links in the hierarchy of WordNet represent a uniform distance in the edge-counting measurement does not accurately reflect the semantic variability of a single link. He suggests judging the similarity of two items as information content of *ncn* using frequency statistics retrieved from a corpus not through the distance of edge-counting. However, Resnik still employs the structure of a conceptual net and one drawback is that the *ncn* for all concept pairs that have the same parent node is the same.

It is worth noting that none of these techniques makes use of either the part-whole (hol/meronym) or the synset (syn/antonym) information in WordNet.

In this paper we design two variant search algorithms to measure noun entity similarity taking account of syn/antonym, hyper/hyponym ('IS-A' link) and hol/meronym ('PART-OF' link). Since all hyper/hyponym and hol/meronym relationships are the same for synonyms, all members of a synset (synonym set) share the same neighbours and a shortest path is either a single identity or synonym link or contains no identity or synonym links.

Our work makes a basic assumption that WordNet can fully specify the superset-subset relationship through the taxonomy organization, although in practice this is achieved only to some degree. We also propose three different measures of noun-relatedness, which are evaluated for each algorithm.

2 Quantitative model

In this section, we propose a new model based on edge-counting, that is partly motivated by Hirst and St. Onge's approach.

We define our quantitative model in two search variants: bidirectional depth-limit search (BDLS) and uni-

directional breadth-first search (UBFS). Both employ the noun taxonomy of WordNet but take into account hol/meronym links and hyper/hyponym links as well as syn/antonym links, with a weight dependent on the link type t for the link (β_t) or that characterizes a path α_t.

These models are all based on the geometric model from human cognitive psychology because it performs more powerfully in a well-organized hierarchy than other psychological models (Tversky 1977). We also make the standard assumptions that a single link in the taxonomy always stands for the same depth-independent distance and that the distance between two conceptual nodes is the least number of links from one node to another. We therefore define the similarity of two concepts as:

$$Sim(c1,c2) = \alpha_t \prod_{i=1}^{dist(c1,c2)} \beta_{t_i}, \quad dist(c1,c2) < \gamma \qquad (3)$$

or

$$Sim(c1,c2) = 0, \quad dist(c1,c2) \geq \gamma \qquad (4)$$

where $0 \leq Sim(c1,c2) \leq 1$ and

- t = *hh* (hyper/hyponym), *hm* (hol/meronym), *sa* (syn/antonym)
- α_t: a link type factor applied to a sequence of links of type t. ($0 < \alpha_t \leq 1$).
- β_t: the depth factor, which also depends on the link type.
- γ: an arbitrary threshold on the distance introduced for efficiency, representing human cognitive limitations.
- $c1$, $c2$: concept node 1 and concept node 2.
- *dist*($c1$, $c2$): the distance (the shortest path) of $c1$ and $c2$.

There are three sorts of cases when we process the relationships existing within the taxonomy of WordNet. The most strongly related concepts are the identity case where $c1$ and $c2$ are identical, $\alpha_{id} = 1$, *dist*($c1$, $c2$) = 0. For the link type of syn/antonym, we assign an intermediate weight, $\alpha_{sa} = 0.9$. Similarly, we assign a lower weight (e.g. $\alpha = \alpha_{hh} = \alpha_{hm} = 0.85$, $\beta = \beta_{hh} = \beta_{hm} = 0.7$) for the hyper/hyponym, hol/meronym. Note that syn/antonym and identity links constitute entire paths and cannot be part of a multilink path. Also we give equal weight to hyper/hyponym and hol/meronym path and link types to avoid deriving different similarity values for equivalent paths.

WordNet connects concepts or senses, but most words have multiple senses. So having defined conceptual similarity, we now define three kinds of functions to evaluate word relatedness, that is the similarity among all the n_i senses $c_{i,j}$ of word w_i. This also relates to one of the questions the experiments explore, namely how do the different senses affect a person's similarity judgments? Is it the sum of all the matching scores in an arbitrary comparison of different senses, the mean value, or the maximum value?

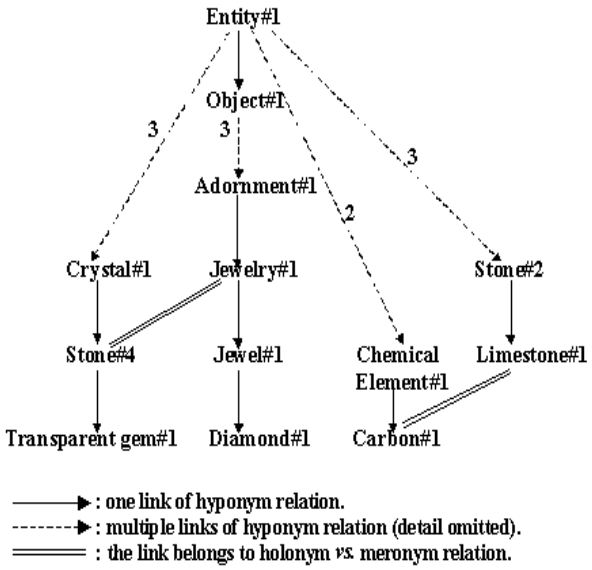

Figure 1: a part of WordNet-style hierarchy

To address this we also compare three functions of word similarity, namely,

- The maximum of the sense distances.

$$Sim_{max}(w1,w2) = \max_{(i,j)}[Sim(c_{1,i},c_{2,j})] \qquad (5)$$

- The sum of the sense distances.

$$Sim_{sum}(w1,w2) = \sum_{(i,j)}[Sim(c_{1,i},c_{2,j})] \qquad (6)$$

- The unweighted mean of the sense distances.

$$Sim_{mean}(w1,w2) = \sum_{(i,j)}[Sim(c_{1,i},c_{2,j})] \Big/ n_1 n_2 \qquad (7)$$

Weighted means could also be used given statistics on the relative frequency of the senses.

3 Search order in WordNet

The previous work (Wu and Palmer 1994; Hirst and St.Onge 1995; Resnik 1995; Jiang and Conrath 1997; Lin 1997; Leacock and Chodorow 1998) mainly focuses on the 'IS-A' hierarchy of WordNet. One of the reasons is that the hyper/ hyponym relationship accounts for nearly 80 percent of all link types. Nevertheless, the 'PART-OF' hierarchy also plays an important role in weaving interconnections into WorldNet's 11 'IS-A' noun hierarchies. The role is however very shallow—few hol/meronym links are needed.

Notation: We first define that $c2 = c1.hype$ symbolizes that the concept $c2$ is the hypernym of concept $c1$, which is identical to $c1 = c2.hypo$ (i.e. the hyponym of $c2$ is $c1$). Similarly if $c2 = c1.holo$, the concept $c2$ is the holonym of concept $c1$, which is equivalent to $c1 = c2.mero$ (i.e. the meronym of $c2$ is $c1$). We use *word* to refer to the token itself and <*word*#*n*> to refer to the *n*th sense of a polysemous word.

In Figure 1, *<entity#1>* is the nearest common node of *<jewel#1> ("a precious or semiprecious stone incorporated into a piece of jewelry")* together with *<stone#4> ("a crystalline rock that can be cut and polished for jewelry"),* if we ignore the hol/meronym relationship. However, the word similarity search still takes place in the 'IS-A' hierarchy.

In the 'IS-A' hierarchy,
dist(jewel#1, stone#4) = 10,
which is derived from
dist(jewel#1, entity#1) + dist(stone#4, entity#1),
if we use edge-counting. Similarly
dist(diamond#1, stone#4) = 11,
and
dist(carbon#1, limestone#1) = 7.

Because the hol/meronym link for *<stone#4>* to *<jewel#1>* is ignored, the similarity of the word pair *(jewel, stone)* fails to reflect our human intuition of their semantic relatedness.

However we can see that
<stone#4>.holo = <jewelry#1>,
and
<limestone#1>.mero = <carbon#1>,
when we augment the 'IS-A' hierarchy of hyper/hyponym links with a 'PART-OF' hierarchy of hol/meronym links.

Hence we now have
dist(jewel#1, stone#4) = 2,
dist(carbon#1, limestone#1) = 1,
and for $\alpha = 0.85$, $\beta = 0.7$,
Sim(jewel#1, stone#2) = 0.049,
Sim(jewel#1, stone#4) = 0.595,
and
Sim(carbon#1, limestone#1) = 0.85.

So
Sim_{max}*(jewel, stone)* = 0.595,
and
Sim_{max}*(carbon, limestone)* = 0.85.

This interconnectivity of 'IS-A' and 'PART-OF' hierarchies produce benefits like evaluating the relatedness of a word pair like *(stone, jewel)* faster and more accurately. There is a trade-off if the shortest path in the 'IS-A' hierarchy has exceeded the threshold before reaching the root node in the hierarchy. Such searching order sometimes avoids returning to the top of the hierarchy, which could make the searching process pointless.

Hence, we introduce the 'PART-OF' hierarchy into our search order, in which the node is expanded into four directions in the searching of WordNet viz. the hypernyms, hyponym, holonym, meronym of each sense.

4 Search in WordNet

4.1 Bidirectional depth-limit search (BDLS)

As we are investigating the similarity of two concepts by looking at the distance between them, we define the bidirectional depth-limit search (BDLS) as a concurrent expansion of a subpath from each node, but limit each subpath to a single link type and direction. The sum of these two distances is limited to γ and in this model we limit each of the two subpaths to $\gamma/2$. For example, for the concept pair *(word1#n, word2#m),* where *<word1#n>* is the *n*th sense of *word1* and *<word2#m>* is the *m*th sense of *word2* in WordNet, we can treat *<word1#n>* as the stimulus node, *<word2#m>* as the response node. We develop partial paths (subpaths) from both nodes to find their common node in the hierarchy instead of developing a path from the stimulus node until it encounters the response node.

The heuristic behind the algorithm is that different concepts and link types have different branching factors in WordNet. Typically a node has only one hypernym and one holonym but has many, *n* say, hyponyms or meronyms. In this case we have to compare *n* times with a response node in order to find if it is one of *n* hyponyms or meronyms of the stimulus node. It needs however only a single comparison to match the single hypernym or holonym.

Notice that two types of redundant path in WordNet should be weeded out due to the reversible relation between hypernym *vs.* hyponym and holonym *vs.* meronym. Clearly we need to avoid cycles, but also we need to avoid various kinds of equivalent path. Suppose each concept node has one single child in each direction, and
*c*1.*hype.hype.hype* = *c*2,
then three redundant paths occur in the hierarchy, viz:
*c*1 = *c*2.*hypo.hypo.hypo,*
*c*1.*hype* = *c*2.*hypo.hypo,*
*c*1.*hype.hype* = *c*2.*hypo*,
leading to overcounting for the values of Sim_{sum} and Sim_{mean}. This can be handled by marking nodes or links as used.

In addition we are using some savage pruning heuristics where after following a link of type *t* all other link types are excluded rather than just the specific inverse of the link used, viz. for each of the two halves of the BDLS, the stimulus subpath *s* and the response subpath *r*, we only follow links of type t_s and t_r respectively.

This achieves time and cost efficiency at the expense of severely pruning the space searched.

The motivation for this is that a part of a part of a part is still a part. A tooth is still part of an animal. Similarly a poodle is still an animal even though *dog* and *mammal* are intervening hypo/hypernyms. But complex mixed paths should normally admit less complex alternatives in which only one change of type is allowed. Thus we don't relate head to tooth and tooth to animal and animal to dog, but head to animal to dog.

4.2 Unidirectional breadth-first search (UBFS)

In UBFS, we discard these limitations and use a standard unidirectional search that expands each node in each direction starting from *s*, while again avoiding loops and redundant paths by marking nodes and links as 'used'.

5 Evaluation

5.1 Task

There is a distinct lack of standards for evaluation of lexical similarity. It is appropriate to compare our computational model with human judgment, and Rubenstein and Goodenough (1965-RG) for their experiment on judging synonymy of word pairs, hired 51 subjects, two groups of college undergraduates, to evaluate 65 pairs of nouns by assigning a similarity from 0 to 4. Miller and Charles (1991-MC) extracted 30 pairs of nouns from RG dataset and repeated their experiment with 38 subjects, achieving correlation of 0.968 with the original experiment. In measuring word similarity using information content, Resnik replicated evaluation on the 28 pairs of nouns from MC dataset that occurred in WordNet with 10 subjects, achieving a correlation between his mean ratings and the MC means of 0.96.

To compare research approaches we can use Resnik's 28 pairs of nouns as tuning data, use the other 37 pairs from '65' dataset as test data, and compare the correlation coefficient with the human judgment results of the MC experiment. However in order to present valid results for the full 65 pairs we use cross validation techniques, viz. we also train on the 35 pairs not included in the '28' or '30' datasets so that we can evaluate on the '28' and '30' datasets as well. In separate experiments using the taxonomy of WordNet we employ BDLS and UBFS with three functions for evaluating word pairs: Sim_{max}, Sim_{sum} and Sim_{mean} (we also append the letter B or U to each of these to indicate whether we used the BDLS or UBFS search methods).

5.2 Tuning

In order to evaluate the two algorithms, we must first set the α, β and γ parameters. To do this we use a tuning set to explore the role of α, β and γ in assigning weights to paths of length 2 or more (with intermediate nodes). To simplify the process, we use the same α and β for both hyper/hyponym and hol/meronym, and tune these relative to the weighting of $\alpha_{id} = 1$ for the length 0 identity path and $\alpha_{sa} = 0.9$ for the length 1 syn/antonym path. This process is presented in Fig. 2 for the BDLS and we now explain the steps used. In fact we optimize the concept similarity model separately for each measure and each algorithm, but as Sim_{maxB} is best we use that as our primary example.

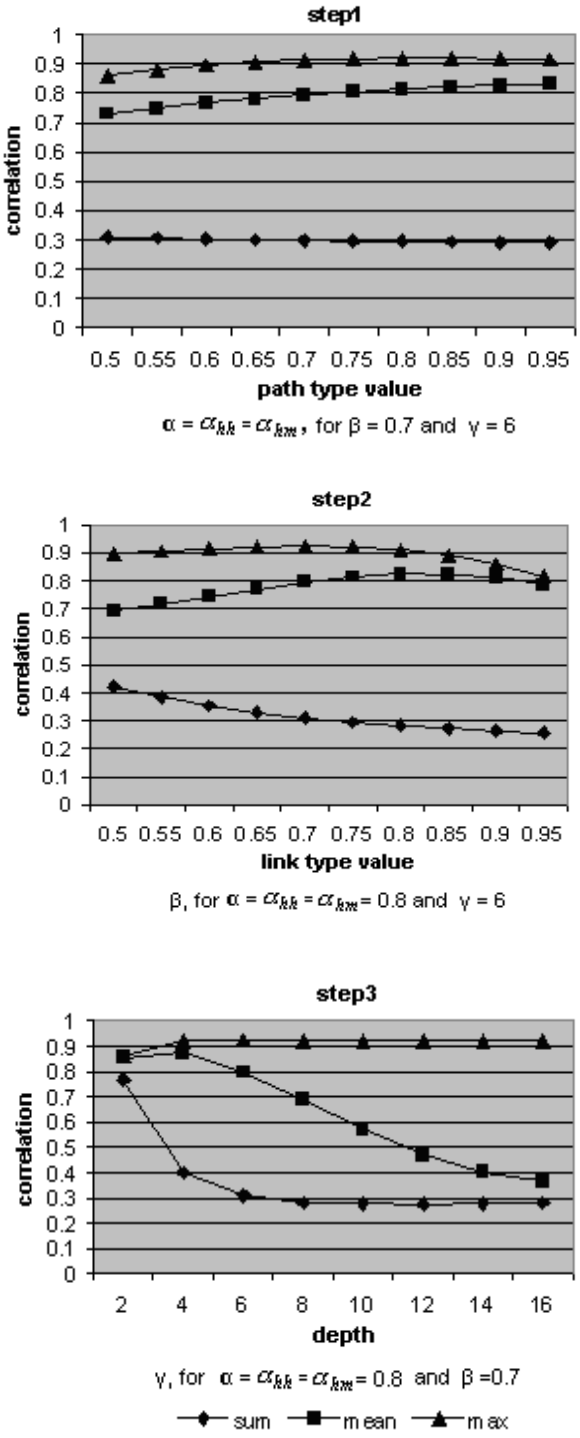

Figure 2: the process for the tuning concept similarity model.

Step 1: the path type factor (α)

To set $\alpha = \alpha_{hh} = \alpha_{hm}$, we first set up the maximum search depth for each node to be no more than 3. Hence, the maximum distance in the hierarchy of WordNet between two concepts should be 6 (i.e. $\gamma = 6$). Then we fixed the value of β_t as 0.7. We varied the value of α, by increments of 0.05 from 0.5 to 0.95. After collecting the similarity score for each pair of words, we calculated the correlation with the benchmark of MC mean human judgment. The optimal value for α around 0.8 for both the '28' and '35' subsets, but there is very little sensitivity to its precise value for either training set.

Step 2: the link type factor (β)

We tested $\beta = \beta_{hh} = \beta_{hm}$ over the range 0.5 to 0.95 tuning with increments of 0.05, to see if it affects the correlation with human judgment. Note that we want the weight of a path to decrease monotonically with length, for length 0 we have $\alpha_{id} = 1$, and for length 1 $\alpha_{sa} = 0.9$. In fact, we found that the performance of the system began to deteriorate as β exceeded 0.8. The best value for β is 0.7 for both the '28' and '35' subsets.

Word Pair		Miller & Charles	Bidirectional Depth-Limit search			Unidirectional Breadth-First Search		
			Sim_{maxB}	Sim_{sumB}	Sim_{meanB}	Sim_{maxU}	Sim_{sumU}	Sim_{meanU}
car	automobile	3.9200	0.9000	1.8799	0.1343	0.9000	2.4539	0.4908
gem	jewel	3.8400	0.9000	3.5187	0.1257	0.9000	2.4375	0.8125
journey	voyage	3.8400	0.8500	3.2702	0.2516	0.8500	1.3281	0.6641
boy	lad	3.7600	0.8500	9.4807	0.1031	0.8500	3.7885	0.4736
coast	shore	3.7000	0.8500	3.6387	0.3032	0.8500	1.3281	0.6641
asylum	madhouse	3.6100	0.8500	1.7926	0.2241	0.8500	0.8500	0.8500
magician	wizard	3.5000	0.9000	2.8372	0.1182	0.9000	1.7965	0.4491
midday	noon	3.4200	0.9000	0.9000	0.9000	0.9000	0.9000	0.9000
furnace	stove	3.1100	0.5950	0.9264	0.1029	0.6375	0.6375	0.6375
food	fruit	3.0800	0.4165	26.8375	0.1335	0.4781	0.8367	0.4184
bird	cock	3.0500	0.8500	3.6649	0.2036	0.8500	2.2047	0.5512
bird	crane	2.9700	0.4165	2.2482	0.0681	0.4781	0.8367	0.4184
tool	implement	2.9500	0.8500	51.9400	0.1234	0.8500	0.8500	0.8500
brother	monk	2.8200	0.8500	3.5560	0.0850	0.8500	0.8500	0.8500
crane	implement	1.6800	0.2916	0.9899	0.0660	0.3586	0.3586	0.3586
lad	brother	1.6600	0.2916	4.6171	0.0710	0.3586	0.7172	0.3586
journey	car	1.1600	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
monk	oracle	1.1000	0.1000	0.3599	0.0400	0.0000	0.0000	0.0000
food	rooster	0.8900	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
coast	hill	0.8700	0.2916	1.7487	0.0795	0.4781	0.8367	0.4184
forest	graveyard	0.8400	0.0490	0.0490	0.0490	0.0000	0.0000	0.0000
monk	slave	0.5500	0.2916	2.8856	0.0962	0.3586	0.7172	0.3586
coast	forest	0.4200	0.1429	0.4257	0.1046	0.0000	0.0000	0.0000
lad	wizard	0.4200	0.2916	2.7742	0.0867	0.3586	0.7172	0.3586
chord	smile	0.1300	0.0343	0.0343	0.0343	0.0000	0.0000	0.0000
glass	magician	0.1100	0.1000	1.2902	0.0403	0.0000	0.0000	0.0000
noon	string	0.0800	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
rooster	voyage	0.0800	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Correlation with Miller & Charles (r)		1.0000	0.9210	0.2730	0.4820	0.9123	0.7465	0.8248
Time cost			12,348.98			306,878.29		

Table 1: Results of comparison between BDLS and UBFS on the 28 pair dataset.

Step 3: the depth-limit (γ)

Once α and β have been adjusted, we varied the depth-limit: γ. We enlarged the search scope from 2 to 16 (nearly the maximum depth in WordNet) to investigate if by expanding of the depth-limit, the model can produce a judgment that is more accurate. For the maximizing sense measure of two words: Sim_{maxB}, the modification of γ has a very tiny effect on the correlation. Many word pairs, which have mostly intermediate or low similarity according to human judgment, get an empty association when we shrink the depth-limit to $\gamma = 2$, increasing the volatility (standard deviation) of the ratings. We found $\gamma = 12$ as the most favourable depth-limit for both the '28' and '35' subsets. Note that if γ is set to 12 for BDLS we halve this so that the maximum depth for each node is 6. As is clear from Fig. 2 we note that Sim_{maxB} is clearly superior throughout the tuning process. After the three steps, we repeated step 1 and step 2 to confirm if the parameters α and β in the concept similarity model remained optimal and adjusted α.

Finally, we select $\alpha = \alpha_{hh} = \alpha_{hm} = 0.85$, $\beta = 0.7$, $\gamma = 12$ as the fixed point of the model for Sim_{maxB}, and confirmed that $\alpha_{id} > \alpha_{sa} > \alpha_{hh} > \alpha_{hm}$, $\alpha_{sa} = 0.9$ was optimal. And again this was the case for both the '28' and '35' subsets.

Although we specified a 2-fold cross validation (2CV) approach to give tuning independent results for all 65

Method	r	r^2
Resnik	0.791	0.626
Jiang & Conrath	0.828	0.686
Lin	0.834	0.696
Average human	0.902	0.814
Sim_{maxB}	**0.921**	**0.848**

Table 2: Results for benchmark methods on the 28 pairs of nouns from the MC data list that were available in WordNet 1.5. This table presents correlations with respect to the results published by the original authors.

pairs, in the end we found the same parameterization was optimal for both training sets. Thus as our results for 28, 30 and 35 pair subsets are independent of their complementary tuning sets, our results on the full 65 pairs correspond to the 30:35 2CV thus the results presented in Table 1 on the 28 pair dataset can be regarded as tuning independent (as the model was tuned for the complementary 35 pair dataset). Furthermore our results can be validly compared with results in the literature for the 28, 30 and 65 pair subsets. (Note that the results reported in the literature *are* tuning dependent and are likely to be overstated.)

By way of comparison, we also investigated the naive unidirectional version of the algorithm, UBFS, and selected $\alpha = \alpha_{hh} = \alpha_{hm} = 0.85$, $\beta = 0.75$, $\gamma = 5$ based on a second series of tuning experiments, and we again found that performance was not overly sensitive to the precise parameters.

5.3 Results

We have found for our proposed algorithm, Sim_{maxB}, the maximum semantically meaningful distance between two nodes in the taxonomy is $\gamma = 12$, and so set $Sim(c1, c2)$ to 0 if the shortest path between nodes is more than 12 edges. The discount factor for multiple link paths involving hyper/hyponym and hol/meronym has been set to $\alpha = 0.85$ with respect to the $\alpha_{id} = 1$ weight allocated for identity and the $\alpha_{sa} = 0.9$ given to syn/antonym links; each additional link in the path shrinks the path's weight by $\beta = 0.7$.

The full results for the '28' set are presented in Table 1, achieving a correlation of 0.921 for this parameterization. We also note that Sim_{max} performs best of the three functions Sim_{max}, Sim_{sum} and Sim_{mean} for both BDLS and UBFS, and is competitive compared with existing algorithms as shown in Table 2. In general the use of Sim_{sum} dramatically overstates the similarity of words with multiple senses whilst Sim_{mean} tends to understate it. There is no evidence of degradation of performance due to pruning in BDLS provided we use the best function, Sim_{max}, but our increase in efficiency is immense because of the exponential explosion of partial path and the huge branching factor caused by some nodes' meronym children. The only advantage of UBFS is that much more complete paths can be found.

Method	r			65		
	30	35	65	σ/μ	z-score	Signifi-cance
HSO +	0.689	0.752	0.732	1.371	-0.793	0.215
JC +	0.695	0.755	0.731	1.292	1.027	0.108
R *	0.775	0.840	0.800	0.885	2.182	0.000
LC *	0.821	0.859	0.852	0.470	2.697	0.000
L *	0.823	0.841	0.834	0.842	1.805	0.031
Human #	0.846	N/A	N/A	0.669	-5.878	<0.001
JS o	0.878	0.784	0.818	0.644	-0.321	0.400
Sim_{maxB}	**0.921**	**0.877**	**0.897**	0.957	0.000	1.000

Table 3: Correlation and significance results on the standard subset of 65 noun pairs of RG[1]. (+: moot significance due to low correlation r or high standard error σ relative to mean μ, *: significant to 0.05 level, #: rating scores of 13 human subjects[2], o: JS used Roget not WordNet and so it not directly comparable, N/A: not available). Significance is calculated for the 65 pair set except in the case of the human subjects where data for only the 30 pair subset was available. Correlations on the 30 and 35 pair subsets are also shown for completeness. Sim_{maxB} results are independent of tuning due to use of 2CV, but as other authors' algorithms are tuned using the entire dataset their performance may be overstated.

On the less commonly tested 37 pairs of nouns using a model (tuned using the complementary 28 pairs), we generated a correlation of 0.876 with the mean human ratings from the experiment of RG. For the total set of 65 we have r = 0.897, which can be compared with published figures for the methods proposed by Resnik (1995-R), Lin (1997-L), Leacock & Chodorow (1998-LC), Jiang & Conrath (1997-JC), Hirst & St. Onge (1995-HSO), which targeted the full 65 as shown in Table 3.

6 Discussion

6.1 Comparison with human judgement

Our preferred model, Sim_{maxB}, correlates fairly well with human judgment for the tuning set (i.e. 28 pairs of words in the MC list), and the correlation, r, is 0.921 which means the strength of the correlation is 84.8% as given by the coefficient of determination r^2. Lin is only 69.6% as strong as it possibly could be, and Jiang & Conrath only 68.6%, as shown in Table 2.

Note that Resnik suggested an upper bound for any similarity measures is r = 0.9015 for his 28 pair dataset, being the *average correlation* achieved by his 10 human subjects against MC. However our model performs

[1] The correlation data is taken from Jarmasz and Szpakowicz's results for independent implementations and tests.

[2] Data from 13 subjects from the wordsimilarity-353 test collection (Finkelstein et al., 2002).

significantly *better* than typical human subjects (see Table 3), with r = 0.921 on the 28 pair benchmark. This does not mean our computational method (i.e. Sim_{maxB}) has gone beyond human judgment, because Resnik's upper bound represents only the judgment of the average individual compared with the group results. The correct interpretation is that our model is more accurate than most of the individual subjects and that a more appropriate upper bound for the 28 pairs is the r = 0.96 obtained from comparison of the human group judgment in Resnik's experiment and that of the MC experiment or the r = 0.968 for MC versus RG.

Unfortunately the raw rankings are not available for Resnik's human subjects, so we had to demonstrate the significance of our results versus human performance by using the corresponding results from Finkelstein et al. (2002).

6.2 Comparison with Roget's thesaurus

The measures in the two types of searching algorithms are not without problems. None of them can detect the relatedness of word pairs such as (*car, journey*) because there is no direct connection in the taxonomy of WordNet. However, in Roget's Thesaurus, we find the following item,

Entry: *lift*
Function: *noun*
Definition: *transportation*
Synonyms: *car ride, drive, journey, passage, ride, run, transport*
Concept: *transportation action* [3]

So that *car* and *journey* have a strong association in the concept of transportation action.

Jarmasz and Szpakowicz (2003-JS) have employed a simple edge-counting model to measuring the semantic similarity in Roget's Thesaurus in which *(car, journey)* has intermediate similarity. They catch fine correlations with MC's 30 pairs and RG's 65 pairs data, viz. 0.878 and 0.818. However our model Sim_{maxB} again performs better in both the 28 pair tuning dataset and the 37 pair *held out* test set and hence across the full 65 pair dataset, with respectively r = 0.921 (28 or 30 pairs), r = 0.876 (37 pairs) and r = 0.897 (65 pairs).

6.3 Comparison of WordNet algorithms

We also performed a statistical significant test on the full dataset of 65 pairs of RG data to check if the outstanding performance of Sim_{maxB} compared with other WordNet algorithms is attributable to tuning or chance. Unfortunately previous authors have not been careful to use separate training/tuning and test sets and have not reported significance. Moreover care needs to be taken with the development of a suitable significance test for the comparison of rankings.

The significance of our results on the 65 pair dataset were shown in the Table 3 along with correlations for the 30, 35 and 65 subsets and a significance test versus 13 human subjects on the 30 pair subset.

Note that the two-sample t-test for the significance of the difference between means is inappropriate for straight-forward calculating on the scores of word pairs in two different measuring methods since we cannot assume the scales are comparable or equal-interval. As a non-parametric alternative to the t-test the Wilcoxon Signed Rank Test was applied to the rank differences for each method compared with RG's human judgements.

The procedure we used was to calculate the absolute value of rank-difference between the human judgment and our measure (i.e. Sim_{maxB}), denoted the random variable *a*. We then obtain the variable *b* for each of other methods in the same way. The alternative hypothesis for the variables *a, b* is *HA:* $a > b$ and we set the confidence level at 95%. This directional hypothesis specifies that our method makes better judgments compared with others across the 65 word pairs. The Wilcoxon Signed Ranks Test then reranks the rank-differences and evaluates significance with respect to a z-distribution.

The one-tailed Wilcoxon Signed Ranks Test on the rank-differences indicated that Sim_{maxB} is significantly better than the three algorithms with $r \geq 0.8$ and $\sigma/\mu < 1$ on the full 65 pair dataset. However, the observed values are not significant for the two cases with $r < 0.8$ or $\sigma/\mu > 1.25$, due to their low correlations and high variance (σ^2). For results with low correlations or high variance, there is a much greater likelihood of our results being better by chance, so a large volume of data is needed to achieve significance. But these attributes themselves indicate that the methods are relatively unreliable.

For completeness the Jarmasz and Szpakowicz (2003) application of an algorithm similar to HSO to Roget (rather than WordNet) is included. It is noted that it performs relatively well on the tuning set ('30') and relatively poorly on the test set ('35') which includes more pairs of dissimilar words.

Finkelstein et al. (2002) created a large data set that included the original MC '30' dataset for their study of human association performance. Unfortunately they did not include the rest of the RG dataset so we are unable to present significance results for Sim_{maxB} versus human subjects for the '35' and '65' subsets because we have no human ranking data for the other 35 pairs. As this study used 13 subjects and our significance test on the '30' subset compared the rankings across the whole group of 390 tests we were able to establish a very high degree of significance for our word similarity results versus human performance ($p < 0.001$).

7 Conclusion

This paper has presented a new path-weighting model to measure semantic similarity in the taxonomy of WordNet. We assess our model on traditional and widely used datasets, but this is complicated by the lack of segmentation of tuning and test sets in the literature, as well as inconsistent use of 28, 30 and 65 pair subsets of the data according to the source of the data and the coverage of the version of WordNet used. The correlation

[3] Roget's New Millennium Thesaurus, 1st ED (v 1.0.5)

with human judgment is r = 0.921 on the standard Resnik dataset, which is better than present findings in the literature. It is r = 0.897 on the full Rubenstein and Goodenough dataset for which it is also a better fit to human data than any other algorithms we have found (although in general they are optimized for this full set of 65, whereas we trained using a 2CV paradigm).

Our results also show that the geometric model can fit particularly well in simulating human judgments on semantic similarity.

In future research, we will emphasize the analysis of the textual WordNet definitions to investigate latent features of concepts. Moreover, we will attempt to evaluate our model on a large dataset and in specific applications.

8 References

Collins, A. M. and M. R. Quillian (1969). Retrieval Time from Semantic Memory. Journal of Verbal Learning and Verbal Behavior 8: 240-47.

Finkelstein, Lev et al (2002). Placing Search in Context: The Concept Revisited, ACM Transactions on Information Systems, 20(1): 116-131

Hirst, G. and D. St.Onge (1995). Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. WordNet. C. Fellbaum. Cambridge, MA, The Mit Press.

Jarmasz, M. and S. Szpakowicz (2003). Roget's Thesuarus and Semantic Similarity. http://www.site.uottawa.ca/~szpak/recent_papers/TR-2003-01.pdf

Jiang, J. and D. Conrath (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. The International Conference Research on Computational Linguistics (ROCLING X), Taiwan.

Leacock, C. and M. Chodorow (1998). Combining Local Context and WordNet Similarity for Word Sense Identification. WordNet: An Electronic Lexical Database. C. Fellbaum, MIT Press: 265-283.

Lin, D. (1997). Using Syntactic Dependency as a Local Context to Resolve Word Sense Ambiguity. The 35th Annual Meeting of the Association for Computational Linguistics, Madrid.

Miller, G. (1995). A Lexical Database for English. Communications of the ACM 38,11: 39-41.

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

Moldovan, D. I. and R. Mihalcea (2000). "Using WordNet and Lexical Operators to Improve Internet Searches." IEEE Internet Computing 4(1): 34-43.

Quillian, M. R. (1967). Word concepts: A theory and Simulation of Some Basic Semantic Capabilities. Behavioral Science 12: 410-30.

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. IJCAI-95.

Rubenstein, H. and J. B. Goodenough (1965). Contextual Correlates of Synonymy. Communications of the ACM 8(10): 627-633.

Tversky, A. (1977). Features of Similarity. Psychological Review, 84, 327-352.

Wu, Z. and M. Palmer (1994). Verb Semantics and Lexical Selection. The 32nd. Annual Meeting of the Association for Computational Linguistics.

Zipf, G. K. (1965). Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. N.Y., Hafner Pub. Co.