Verb similarity on the taxonomy of WordNet

Dongqiang Yang and 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. In this paper, we introduce two kinds of word similarity algorithms to investigate the capability of WordNet in measuring verb similarity. Both are tested on two noun and two verb data sets. The noun set is a standard set but in the absence of a standard verb set we have proposed and tested human and computer results on a similar verb set.

Introduction

Many researchers have explored the similarity of nouns using a variety of methods including methods based on WordNet. However, little attention has been paid to verbs [1], there is no standard evaluation set, and it is not clear that the WordNet verb hierarchy is rich enough to support verb similarity assessment. In this study we introduce an verb evaluation set with both tuning and evaluation partitions, we present and adapt a successful noun similarity method based on WordNet to the verb similarity task, and we present a hybrid technique that seeks to increase accuracy by cross mapping into the noun hierarchy and back.

Measuring word similarity can be classified into knowledge-rich and knowledge-poor methods [2, 3]. We introduce both approaches before presenting our own results using knowledge-rich methods.

Knowledge-poor methods

Knowledge-poor methods mainly depend on information or probability information derived from a corpus rather than a knowledge base. Such methods may be further categorized according to how co-occurrence frequency data is handled:

Vector space.
These approaches assume that semantically related words are more likely to co-occur in the corpus. A matrix is constructed in word-by-word or word-by-document order with a cell value such as term frequency (TF) or TF*IDF (inverse document frequency, but more accurately the information conveyed by the fact of occurrence in

a document). Word similarity is established by comparing distance measures such as the cosine coefficient or Euclidean distance.

Syntactic parsing.
These approaches assume that the semantic relatedness of words leads to their use in similar grammatical structures. Judging word similarity is achieved by tagging parts-of-speech in the corpus, shallow parsing of sentences, specifying the relationship between chunks and comparing the syntactic components along with their dependency relations [2].

Knowledge-rich approaches

Knowledge-rich methods require semantic networks or a semantically tagged corpus to define the concept of word in the relation with other concepts or in the surrounding context. Most methods that calculate semantic distance using ontology or thesaurus knowledge such as WordNet [4] or Roget fall into this category. The popular methodologies for measuring semantic relatedness with the help of a thesaurus can be classified into two categories: one uses the solely semantic links (i.e. edge-counting), the other combines corpus statistics with the taxonomic distance.

Edge-counting

The edge-counting or shortest path method derives from the geometric model in Cognitive Psychology, where the shorter distance entails the stronger association between stimuli and response. It can be traced back to Quillian's semantic memory model [5, 6] where concept nodes are planted within the hierarchical network and the number of hops between the nodes specifies the similarity of the concepts. Generally the similarity of words in the thesaurus space can be described as

$$Sim(i, j) = 2D - Dist(i, j) \ . \qquad (1)$$

where D is a constant (e.g. the maximum depth in the taxonomy of WordNet, viz. 16 if we presume all the hierarchies have a common node), Dist(I,j) is links between two concept nodes I and j. In the edge-counting methods distance is typically assessed by counting the edges traversed from $c1$ to $c2$ via *ncn*, $Dist(c1, c2)$ – we will introduce a few popular edge-counting models working in the semantic hierarchy [7].

Wu and Palmer [8, 9] proposed to measure the verbal concept similarity in the projected domain hierarchy when translating from English verbs to Chinese. According to Wu and Palmer, the relatedness of two words is the weighted sum of all their senses comparison,

$$Sim(v_i, v_j) = \sum_k w_k * \frac{2 * dep(ncn(c_{i,k}, c_{j,k}))}{dep(c_{i,k}) + dep(c_{j,k})} \ , \qquad (2)$$

where $ncn(c_{i,k},c_{j,k})$ is the nearest common node (*ncn*) for the conceptual nodes $c_{i,k}$, $c_{j,k}$ of verbs v_i and v_j, *dep* is the depth of node relative to the root, w_k is the weight of each pair of concepts in each domain. The sum of w_k is 1. This model is appropriate for measuring both verbs and nouns in the "IS-A" hierarchical concept net.

Leacock and Chodorow [10] adapted the concept of information content [11] to evaluate the relatedness of two words using the following model:

$$Sim(W_i, W_j) = Max\left[-\log\frac{Dist(c_{i,}c_j)}{2*D}\right] \tag{3}$$

$$= Max\left[\log 2D - \log Dist(c_i, c_j)\right] \ ,$$

where *Dist(c_i,c_j)* is the shortest distance between concepts c_i and c_j. In addition, they defined the similarity of two words as the maximized value of all the pairwise similarities. Note that in Equation (3)

$$Dist(c_i,c_j)=dep(c_i)+ dep(c_j)-2*dep(ncn(c_i,c_j)) \ , \tag{4}$$

$$Sim(W_i, W_j) = Max\left[\log\frac{2D}{Dist(c_i, c_j)}\right] . \tag{5}$$

Hence, the concept model is similar to Wu and Palmer's apart from the *log* normalization.

Resnik's information content

Resnik [11] argues that the links in the hierarchy of WordNet representing a uniform distance in the edge-counting measurement can not account for the semantic variability of a single link. He defines information content of *ncn* to explain the similarity of two words through frequency statistics retrieved from a corpus, not through the distance of edge-counting. Here the frequency of *ncn* subsumes all the frequency data of subordinate concept nodes. The information content can be quantified as the negative of the log likelihood, *-logP(c).*

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.

Jiang and Conrath's model

Building on Resnik's work, Jiang and Conrath [12] further assumed that a combination of information content and edge-counting will improve the correlation co-efficient (compared with human judgment). They also considered the link type, depth, conceptual density, and information content of concepts. Their simplified formula can be expressed as follows:

$$Dist(c_i,c_j) = IC(c_i) + IC(c_j)-2*IC(ncn(c_i, c_j)) \ . \tag{6}$$

$$Sim(c_i, c_j) = -Dist(c_i, c_j) \ . \tag{7}$$

Lin's model

Lin [13] introduced another way of in computing the similarity to disambiguate word sense,

$$Sim(c_i, c_j) = \frac{2 * IC(ncn(c_i, c_j))}{IC(c_i) + IC(c_j)} \ , \tag{8}$$

which is essentially another normalized form of Jing and Conrad's model.

Multiplicative Models

The noun model

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

$$Sim(c1, c2) = 2\gamma \div (\alpha + \beta) \ , \tag{9}$$

$$Sim(c1, c2) = 2\gamma - (\alpha + \beta) \ . \tag{10}$$

where *α, β, γ,* respectively denote attributes of concepts *c*1, *c*2, and the *ncn* of *c*1,*c*2 in the "IS-A" hierarchy. The attribute can be viewed as some function of the depth in the taxonomy or the information content extracted from the outer corpus.

Yang and Powers [14] proposed a new model to measure semantic similarity in the taxonomy of WordNet, based on a variation of edge-counting. In contrast with the above methods they also take into account the part-whole (hol/meronym) relationships in WordNet and compare two searching algorithms, a bidirectional depth-limit search (BDLS) and unidirectional breadth-first search (UBFS).

On the assumption 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, λ, from one node to another, they define the similarity of two concepts multiplicatively as,

$$Sim(c1, c2) = \alpha_t \beta^{\lambda} \ . \tag{11}$$

Partially inspired by Hirst and St. Onge's algorithm for the detection and correction of malapropisms [15] which different weights for identical words, synonyms or antonyms, and hyper/hyponym, Yang and Powers deal with the identity case where *c*1 and *c*2 are identical as α_{id} = 1, *γ* = 0, the syn/antonym as an intermediate weight, α_{sa} = 0.9, *γ* = 0, assigning the lowest weight (e.g. $\alpha = \alpha_{hh} = \alpha_{hm}$ = 0.85, $\beta = \beta_{hh} = \beta_{hm}$ = 0.7) for the hyper/hyponym, hol/meronym where searching depth *γ* is more than one – these weights being the result of tuning noun similarity.

These models are evaluated 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 noun pair dataset [11] 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, they also validate on an independent 37 noun pair test set (0.876) and present cross-validated results for the full 65 noun-pair superset (0.897) [16]. Note that their best performance on these data sets is achieved for the maximum score across distinct sense in relation to the common case of words that are polysemous.

A multistrategy verb model

To investigate the appropriateness of such a model for judging word similarity we have sought to adapt it to apply to verbs, which are another significant hierarchy in WordNet. Unlike the noun taxonomy, which is rich in complexity and links, the verbs are organized into a relatively shallow hierarchy according to their hyper/troponymy relations and WordNet does not represent holo/metonymy relations. The maximum distance between contentive verbs (excluding stopwords like ‘be’, ‘make’ and ‘do’) is around 4 nodes, which make it more difficult to find relationships between verbs [17]. Based on the Yang and Powers noun model and approach, we designed and tuned a new algorithm to account for the similarity of verbs in the face of the sparseness and limitations of the WordNet verb hierarchy. To supplement the verb hierarchy, we also consider derivational mapping into the noun hierarchy, the use definitions (glosses), and effect of stemming. Thus we consider the following factors in constructing this model of verb similarity, where at this stage stemming refers only to the simple suffix removal functions provided with WordNet 2.

1. Similarity on the verb taxonomy is evaluated in the same basic way as for the noun hierarchy, viz. equation (11) and (12), except that we there is no correlate of the holo/meronym relationships (viz. no metonymy by which a part of an action/scene may be related to the whole). We thus need to set up and tune parameters for the syno/antonyms and hyper/troponyms in the same way as with the noun model.
2. Some verbs have the *noun* forms as a *stem*, or vice versa, as they are *derivationally* related. Thus we can project to the noun hierarchy from the verb hierarchy to enrich the relationships among verbs, introducing α_{der} as discount factor or weight.
3. The definition of a verb, its *gloss*, can give a hint to the relation with other verbs *when there are no apparent linkages in the verb and noun hierarchies*. Lesk [18] proposed to calculate the overlaps of target word and other words in the context in the definitions to select an appropriate sense. Pedersen et al. [7] treat the definitions in WordNet as a million word corpus, and build a co-occurrence matrix to specify how many times the two concepts turn up together in the gloss of WordNet. In this paper we assume verbs in the definition of WordNet, which are not in the frequent word list like “make”, “do”, etc., bring about a strong semantic relation with its target word. This thus introduces α_{gls}.
4. The *stemming* effect seen above can also connect related verbs in the verb hierarchy without considering their individual senses, but rather allows us to capture a wider class of relationship that relate to the etymology of the word and its root meeting, but should not represent as strong a relationship as those that are represented directly by links. This gives us weight α_{stm}.

Comprehensively considering these new factors and the existing link type and depth factors that we need to tune for the WordNet verb taxonomy, noting that Yang and Powers have already well tuned for noun similarity and needs no adjustments for links within the noun hierarchy, the new model is

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

$$Sim(c1,c2) = 0, \qquad dist(c1, c2) \geq \gamma, \tag{12}$$

$$Sim_{max}(v1,v2) = \underset{(i,j)}{Max}\left[Sim(c_{1,i}, c_{2,j})\right], \tag{13}$$

- where $0 \leq Sim(c1, c2) \leq 1$,
- $t = ht$ (hyper/troponym), *sa* (syn/antonym), *der* (derived nouns) or *gls* (definition),
- α_t is a link type factor applied to a sequence of links of type *t*. ($0 < \alpha_t \leq 1$),
- α_{stm} is the stemming factor, if c1 is linking c2 without stemming, $\alpha_{stm}= 1$
- β_t is the depth factor depending on the link type
- γ is an arbitrary threshold on the distance, which will no more than five in the verb taxonomy
- *dist*(*c*1, *c*2) is the distance (the shortest path) of *c*1 and *c*2
- *c*1, *c*2 represent concept nodes

The most strongly related concepts are the identity case where *c*1 and *c*2 are identical, $\alpha_{id} = 1$ and *Dist(c*1*, c*2*)* = 0. For the link type of syn/antonym, we again assign an intermediate weight (e.g. $\alpha_{sa} = 0.9$, *Dist(c*1*, c*2*)* = 0), and we again tune to assign the lowest weight (e.g. $\alpha_{ht} = 0.85$) for hyper/troponymy. Note that any syn/antonym and identity links constitute entire paths and cannot be part of a multilink path.

Given the fact that most verbs are polysemous we will again assign the maximum value of the similarity among all the n_i senses $c_{i,j}$ of any polysemous word v_i. To make clear the final model of verb similarity in the WordNet we present it succinctly but informally as the following algorithm. The bidirectional search is as described in the original Yang and Powers algorithm, deciding first if it is a direct identity or synonym path, or otherwise discounting as a hyper/tropo path and calculating the additional distance d required to connect them, except that if unsuccessful is redone with a further discount allowing connection through any derivationally related stem, not just through specific senses.

The basic algorithm is as follows where the noun similarity and maximum similarity steps are exactly as described by Yang and Powers:

```
for each sense c1 and c2 of v1 and v2 resp.
  if c1 and c2 are synonymous or antonymous
    assign sim_sa(c1,c2)= $\alpha_{sa}$; Goto next loop
  elsif c1 and c2 are hyper- tropo- and/or antonym connected
            with depth d less than $\gamma$
      sim(c1,c2) = sim_hta(c1,c2)= $\alpha_{ht} * \beta_{ht}^{d}$
      if=0 & c1 and c2 are stem hyper/tropo/antonym connected
            with depth d less than $\gamma$
        sim(c1,c2) = sim_stm(c1,c2)= $\alpha_{stm} * \alpha_{ht} * \beta_{ht}^{d}$
  endif
endfor
calculate the maximum similarity score,
sim$_{max}$(c1$\in$v1,c2$\in$v2)
if≠0
        sim(v1,v2) = sim$_{max}$(c1$\in$v1,c2$\in$v2)
elsif v1 can find v2 in its definition or vice versa
        sim(v1,v2) = sim_gls(v1,v2)= $\alpha_{gls}$
else
  if both v1 and v2 have derived noun form
    go into noun taxonomy and perform BDLS search:
            sim(v1,v2) = sim_der(c1,c2)= $\alpha_{der}$ * sim_noun(c1,c2)
endif
endif
```

Evaluation

Task

Unfortunately, there is no benchmark data set for verbs in the literature. We have thus had to make our own data set and offer it as a standard for testing verb similarity. We selected 20 verb synonym tests from the 80 TOEFL (Test of English as a Foreign Language)[12] questions used by [19], and 16 from a set of 50 ESL (English as a second language) questions [20] – these are widely used to assess non-native eligibility for university entry or employment in English speaking countries and we judged them as representing different levels of difficulty for non-native speakers, but as all well within the competence of a native speaker or university graduate in an English speaking country. Each these 36 multiple choice questions consists of a question or target word and four other words or phrases to choose from. We tried to select examples with words rather than phrases, and then used each target word together with one of the four choices to construct a pair of verbs in the questionnaire, giving a total 144 pairs verbs. We randomly arrange these word pairs and randomly reverse the order of target verb and choice verb. Six colleagues (2 academic staffs and 4

[1] Test of English as a Foreign Language (TOEFL), Educational Testing Service, Princeton, New Jersey, http://www.ets.org/.

postgraduates) voluntarily rated these pairs for similarity. Four of them are native English speakers; the other two have used English as a second language and a main communication tool in the academic and ordinary life for over ten years. We gave them the following instructions:

Indicate how strongly these words are related in meaning using integers from 0 to 4. The following are given as examples of kinds of descriptions that might apply to each number, but you must give your own judgement and if you think something falls in between two of these categories you must push it up or down (no halves or decimals).

0: not at all related

1: vaguely related

2: indirectly related

3: strongly related

4: inseparably related

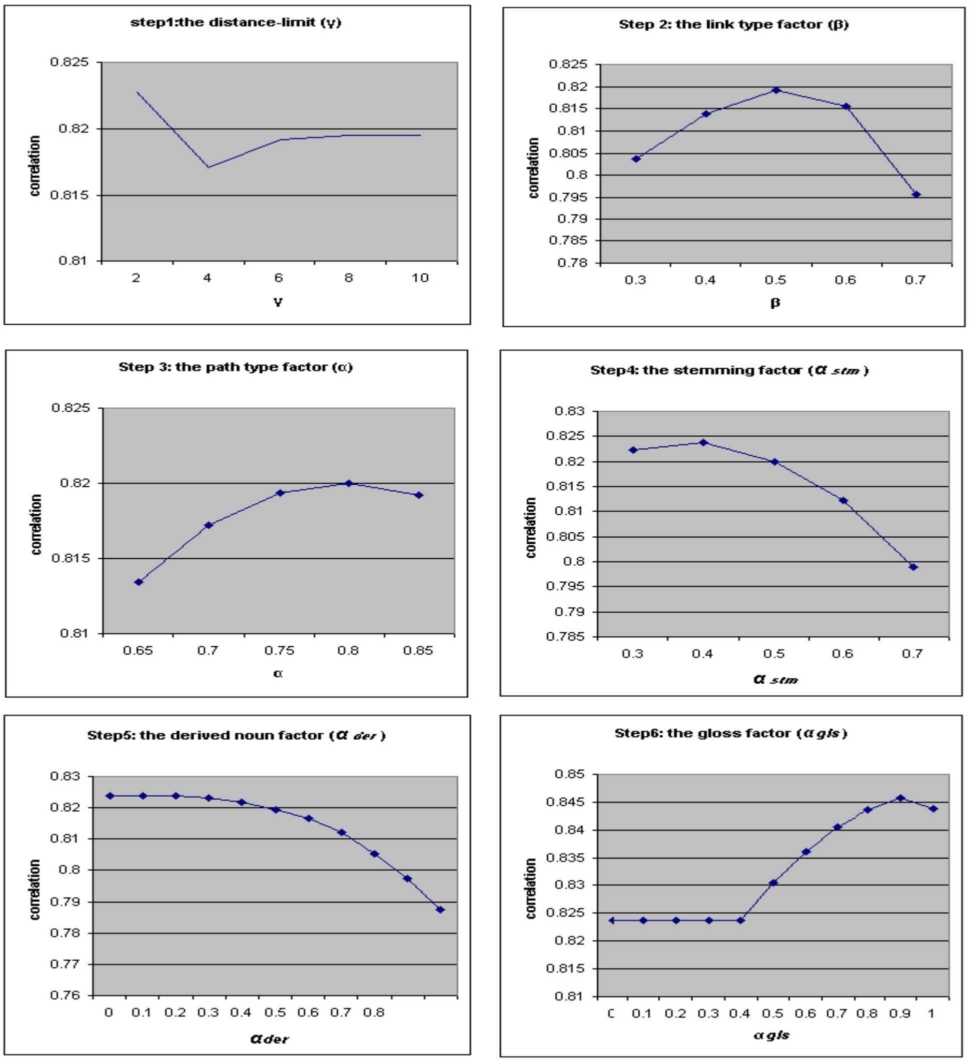
The word pairs were sorted in descending order of average score, and divided up to achieve a balanced set with 26 words in each category (eliminating some words with averages below 2 to eliminate an expected imbalance due to the questions being designed to have exactly one best answer and being biased to include more dissimilar words). We then randomly assigned 13 words from each category to one of two data sets, data1 and data2. The average correlation among these six subjects was r = 0.866.

We then optimized the verb model for each data set through calculating the correlation with average human scores, using a greedy approach to optimizing the parameters (choosing the mid-value when there was no significant difference). Here we show how we regulated the verb model on data1.

To distinguish the different effect of each factor we proposed, we assumed the contribution of the verb hierarchy similarity, derived noun hierarchy similarity and gloss similarity are independent. Thus we first sought the optimal parameterization for the verb hierarchy, and then to set without any interaction with α_{der} and α_{gls} considered how helpful the derived noun similarity was and then how helpful the gloss similarity was.

Tuning

There were three parameters we needed to adjust in relation to the application of the Powers and Yang algorithm to the verb hierarchy, the path type factor α_{t}, the link type factor β and the depth factor γ (optional, noting that this last factor was originally and primarily conceived to minimize cpu time, but also may also serve as a threshold to stop relationships that are too strained being discovered). Then in order to factor in the alternative source of information we needed to set the stem similarity weighting α_{stm}, the derived noun similarity weighting α_{der}, and the gloss similarity weighting α_{gls}. In this case the three values are fallback weights: given the algorithm for the verb hierarchy hasn't given us a non-zero value, we retry using ignore sense and inflectional variations of verbs (discounted using α_{stm}), and if it is still non-zero, we use the noun version algorithm to seek a value for derivationally related nouns (discounted by α_{der}), or failing that we try to find a connection via the glosses (α_{gls}).



step1:the distance-limit (γ)

Step 2: the link type factor (β)

Step 3: the path type factor (α)

Step4: the stemming factor (α stm)

Step5: the derived noun factor (α der)

Step6: the gloss factor (α gls)

Fig. 1. The tuning process on the RHE

Step 1: the distance-limit (γ)

Once the values of α, α_{stm} and β had been assigned initially, i.e. respectively 0.85, 0.5 and 0.5, we varied the distance-limit γ (for the combined path length), enlarging the search distance of each node from 1 to 5 (essentially the maximum distance is no more than 5 in the WordNet), viz. the total distance of two node in the BDLS is from 2 to 10, to investigate if by expanding the distance-limit, the model could produce a judgment that is more accurate. We can see in Fig. 1(*γ*) that there is a drop in the correlation when we increase the searching scope from 1 level to 2 level, after that the curve approached level. Our purpose in the paper is to investigate the function of verb hierarchy, so we use $\gamma = 6$ for a rich hierarchy exploration (RHE) but also use $\gamma = 2$ as a reference point for shallow hierarchy exploration (SHE). In the following part we just illustrate how to calibrate the model using the RHE variant.

Step 2: the link type factor (*β*)

We tested *β* over the range 0.3 to 0.7 tuning with increments of 0.1, to see if it affects the correlation with human judgment. Note that each link in the taxonomy is of uniform distance if we give $\beta = 1$. In fact, we see from Fig. 1(*β*) that the performance of the system begins to deteriorate significantly for *β* bigger than 0.6 with the maximum at 0.5.

Step 3: the path type factor (*α*)

We varied the value of *α*, by increments of 0.05 from 0.5 to 0.95. The optimal value for *α* is around 0.8 but there is very little sensitivity to its precise value as seen in in Fig. 1(*α*).

Step 4: the stemming factor (α_{stm})

After the optimal value, 0.4, Fig. 1(α_{stm}) shows that the correlation begins to drop quickly but prior to that there is little change.

Step 5: the derived noun factor (α_{der})

Similarly, there is little difference as α_{der} increase from 0 to 0.5, but after that the correlation deteriorated slowly – see Fig. 1(α_{der}). We chose 0.4 as a compromise value, as with the shallower verb hierarchy we did expect to see smaller values, but a larger value will maximize utilization of the information in the network.

Step 6: the gloss factor (α_{gls})

There is an initial jump at 0.4, rising to a clear optimum at 0.9, as seen in Fig. 1(α_{gls}).

Results

After we had tuned the verb model on each data set we found the selected values did not correspond very well with each other, reducing the score for the 2-fold cross validation. This was not unexpected due to the relative flatness (lack of significant difference) for much of the curves, which forced an arbitrary selection within a range. Unfortunately the tuning is a time intensive process, so we have not yet been able to perform a higher order cross validation. Owing to the sensitivity of each data set as measured by the correlation, r, to tuning on the other, we adopted a compromise tuning based on both subsets for future comparison against human performance, noting that apart from the Yang and Powers paper where identical results were achieved for each mode of the cross-validation, results for work on noun similarity do *not* do tuning and validation on separate subsets of the data. Table 1 shows the final parameters and correlations with the average human scores for both RHE and SHE. There is no big difference on the final verb model due to the choice of RHE or SHE.

Table 1. the final result on the each 65 data sets and the total dataset. (r_t: the correlation on the tuning set, r_e: the correlation on the evaluation set, where data1 is the evaluation set for data2, and vice versa.)

		γ	β	α	α_{stm}	α_{der}	α_{gls}	r_t	r_e
RHE	Data1(65)	2	0.5	0.8	0.4	0.1	0.9	0.846	0.775
	Data2(65)	2	0.2	0.85	0.7	0.8	0.5	0.864	0.823
	Total (130)	2	0.5	0.8	0.5	0.75	0.6	0.808	
SHE	Data1 (65)	0	0.6	0.75	0.4	0.7	0.9	0.838	0.824
	Data2 (65)	0	0.4	0.8	0.6	0.7	0.5	0.846	0.835
	Total (130)	0	0.5	0.8	0. 5	0.75	0.6	0.833	

Discussion

The Yang and Powers noun similarity study advocated the Wilcoxon Signed Rank Test as a principled non-parametric modification to the two-sample t test for comparing their results against human judgment. We similarly performed this test for the present verb similarity study, achieving the results listed in the Table 2. The choice of RHE versus SHE makes no significant difference in the ability of judging verb similarity, and they are only significantly better than one subject (a non-native speaker). However, three other subjects fail to do significantly better than SHE (shallow), whilst just one just misses out on being significantly better than RHE (rich), although all their judgments retain a high correlation with the average human. Thus while there is no significant difference between the rich and shallow variants themselves with respect to the group, the richer variant doesn't keep step with individual human subjects as well as the shallower variant, implying that the additional levels of the verb hierarchy are less useful in modeling human behavior than the gloss derived noun fallbacks we have introduced.

Table 2. significance test on both RHE and SHE, r_a: the correlation with average human, σ: standard deviation, μ: mean

			RHE		SHE	
	r_a	σ/μ	z-score	Significance	z-score	Significance
Subject1	0.88	0.292	-3.25	0.001	-2.113	0.035
Subject2	0.733	0.45	0	1	-0.802	0.423
Subject3	0.878	0.488	-3.07	0.002	-3.421	<0.001
Subject4	0.926	0.485	-3.52	<0.001	-1.14	0.254
Subject5	0.913	0.397	-4.47	<0.001	-3.596	<0.001
Subject6	0.868	0.402	-1.89	0.059	-1.61	0.107
RHE	0.808	0.308	0	1	-1.484	0.138

SHE	0.833	0.561	-1.484	0.138	0	1

Conclusions and Future Work

The maximum links each node can reach in the verb model are much less than the γ in the noun model. Moreover the link type factor in the verb model also more quickly reduce the similarity of node in the next level with the target node. So do the path type factor. All of these facts partly tell us that verb hierarchy exists in a very shallow way in human, or the hierarchy does a limit help in assessing the similarity of verbs.

Thus the Yang and Powers noun similarity model does not adapt so directly to verbs in the WordNet hierarchy. This is clearly connected to this observation that the verb taxonomy is shallower but another factor is that the verb hierarchy does not include a second part-whole analog to the holo/meronym links of the noun hierarchy.

Such relationships do exist and correspond to the concept of metonymy where there is a relationship between a word that describe a complex action or scene and one that describes a more specific aspect of that activity. For example, one of the poorly handled pairs in our data set is ‘market’ versus ‘sell’. If we could compare the noun sense of ‘market’ with ‘sell’ or ‘sale’ we would do much better. Similarly if we could recognize that marketing is a complex activity which involves price-setting, product packaging, advertising, and selling, as metonymously related activities, we could again do better. The first improvement can be made by connecting the two hierarchies into one and using a single bidirectional search to evaluate similarity of any noun or verb against any other noun or verb – this is straightforward and is planned as part of our refinement of these techniques. The second improvement is not so straightforward as it would seem to require manual augmentation of WordNet with the additional hierarchy, although of course there is always the possibility that WordNet-like hierarchies and variations could be self-organized based on corpus data.

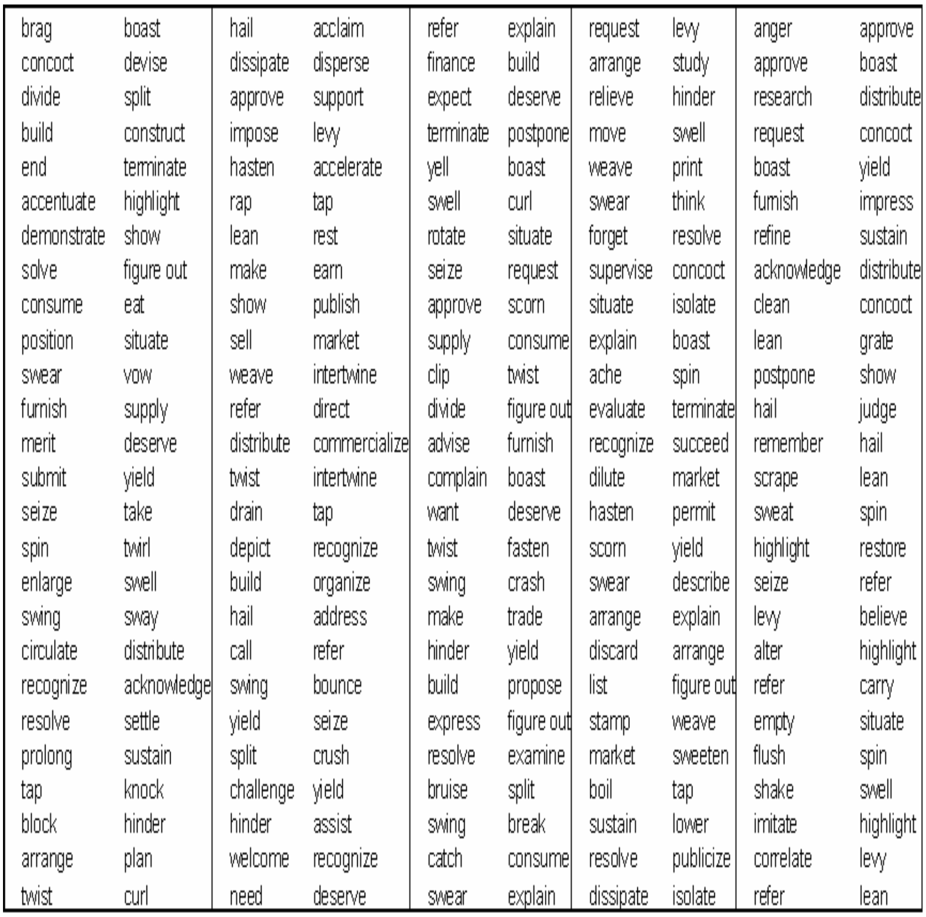
The fallback into the use of glosses, stems, or noun similarity, do improve the situation but this increases the set of parameters to nine – three for the noun similarity, three for the basic verb similarity, and three for the three fallback options. However, this increase in the number of parameters does not seem to make the system brittle, as the tuning curves have fairly flat peaks and the tuning effects are relatively minor compared with the improvement due to the fallback mechanisms.

Reference

1. Resnik, P., Diab, M.: Measuring verb similarity. In: Proc. the 22nd Annual Meeting of the Cognitive Science Society (CogSci 2000) (2000) 399--404
2. Grefenstette, G.: Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and Window Based Approaches. In: Proc. Workshop on acquisition of lexical knowledge from text columbus (1993)
3. Gasperin, C., Gamallo, P., Agustini, A., Lopes, G., Lima, V. d.: Using Syntactic Contexts for Measuring Word Similarity. In: Proc. Workshop on Semantic Knowledge Acquisition & Categorisation (ESSLLI 2001). (2001)

4. Miller, G.: A lexical database for English. Communications of the ACM 38,11 (1995) 39-41
5. Collins, A. M., Quillian, M. R.: Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior 8 (1969) 240-47
6. Quillian, M. R.: Word concepts: A theory and simulation of some basic semantic capabilities. Behavioral Science 12 (1967) 410-30
7. Pedersen, T., Banerjee, S., Patwardhan, S.: Maximizing Semantic Relatedness to Perform Word Sense Disambiguation. (2003)
8. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: Proc. 32nd. Annual Meeting of the Association for Computational Linguistics (1994) 133 --138
9. Rada, R., Mili, H., Bicknell, E., M.Blettner: Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics 19 (1989) 17-30
10. Leacock, C., Chodorow, M.: Combining local context and WordNet similarity for word sense identification. In: C. Fellbaum, (ed.) WordNet: An electronic lexical database. MIT Press (1998) 265-283
11. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proc. Proceedings of IJCAI-95 (1995)
12. Jiang, J., Conrath, D.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proc. proceedings on international conference on research in computational linguistics (1997) 19-33
13. Lin, D.: Using syntactic dependency as a local context to resolve word sense ambiguity. In: Proc. Proceedings of the 35th annual meeting of the association for computational linguistics (1997) 64-71
14. Yang, D., Powers, D. M. W.: Measuring Semantic Similarity in the Taxonomy of WordNet. In: Proc. Twenty-Eighth Australasian Computer Science Conference (ACSC2005) (2005) 315-322
15. Hirst, G., St.Onge, D.: Lexical chains as representations of context for the detection and correction of malapropisms. In: C. Fellbaum, (ed.) WordNet. The Mit Press (1995)
16. Rubenstein, H., Goodenough, J. B.: Contextual correlates of synonymy. communications of the ACM 8 (1965) 627-633
17. Fellbaum, C.: An Electronic Lexical Database. The MIT Press, London, England (1998)
18. Lesk, M.: Automatic sense diaambiguation using machine readable dictionaries: how to tell a pine code from an ice cream cone. In: Proc. the 5th annual international conference on systems documentation (1986) 22-26
19. Landauer, T. K., Dumais, S. T.: A solution to Plato's problem: The latent
semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review 104 (1997) 211-240
20. Tatsuki, D. Basic 2000 Words - Synonym Match 1. In: Interactive JavaScript Quizzes for ESL Students. http://www.aitech.ac.jp/~iteslj/quizzes/js/dt/mc-2000-01syn.html.[Online]. Available:

Appendix: The 130 pairs of verbs



brag	boast	hail	acclaim	refer	explain	request	levy	anger	approve
concoct	devise	dissipate	disperse	finance	build	arrange	study	approve	boast
divide	split	approve	support	expect	deserve	relieve	hinder	research	distribute
build	construct	impose	levy	terminate	postpone	move	swell	request	concoct
end	terminate	hasten	accelerate	yell	boast	weave	print	boast	yield
accentuate	highlight	rap	tap	swell	curl	swear	think	furnish	impress
demonstrate	show	lean	rest	rotate	situate	forget	resolve	refine	sustain
solve	figure out	make	earn	seize	request	supervise	concoct	acknowledge	distribute
consume	eat	show	publish	approve	scorn	situate	isolate	clean	concoct
position	situate	sell	market	supply	consume	explain	boast	lean	grate
swear	vow	weave	intertwine	clip	twist	ache	spin	postpone	show
furnish	supply	refer	direct	divide	figure out	evaluate	terminate	hail	judge
merit	deserve	distribute	commercialize	advise	furnish	recognize	succeed	remember	hail
submit	yield	twist	intertwine	complain	boast	dilute	market	scrape	lean
seize	take	drain	tap	want	deserve	hasten	permit	sweat	spin
spin	twirl	depict	recognize	twist	fasten	scorn	yield	highlight	restore
enlarge	swell	build	organize	swing	crash	swear	describe	seize	refer
swing	sway	hail	address	make	trade	arrange	explain	levy	believe
circulate	distribute	call	refer	hinder	yield	discard	arrange	alter	highlight
recognize	acknowledge	swing	bounce	build	propose	list	figure out	refer	carry
resolve	settle	yield	seize	express	figure out	stamp	weave	empty	situate
prolong	sustain	split	crush	resolve	examine	market	sweeten	flush	spin
tap	knock	challenge	yield	bruise	split	boil	tap	shake	swell
block	hinder	hinder	assist	swing	break	sustain	lower	imitate	highlight
arrange	plan	welcome	recognize	catch	consume	resolve	publicize	correlate	levy
twist	curl	need	deserve	swear	explain	dissipate	isolate	refer	lean