Using Wordnet to Supplement Corpus Statistics

Rose Hoberman
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
roseh@cs.cmu.edu

Roni Rosenfeld
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
roni@cs.cmu.edu

Abstract

Data-driven techniques, although commonly used for many natural language processing tasks, require large amounts of data to perform well. Even with significant amounts of data there is always a long tail of infrequent linguistic events, which results in poor statistical estimation. To lessen the effect of these unreliable estimates, we propose augmenting corpus statistics with linguistic knowledge encoded in existing resources. This paper evaluates the usefulness of the information encoded in WordNet for two tasks: improving perplexity of a bigram language model trained on very little data, and finding longer-distance correlations in text. Word similarities derived from WordNet are evaluated by comparing them to association statistics derived from large amounts of data. Although we see the trends we were hoping for, the overall effect is small. We have found that WordNet does not currently have the breadth or quantity of relations necessary to make substantial improvements over purely data-driven approaches for these two tasks.

1 Motivation and Outline

Data-driven techniques are commonly used for many natural language processing tasks. However, these techniques require large amounts of data to perform well, and even with significant amounts of data there is always a long tail of infrequent linguistic events. The majority of words, for example, occur only a few times even in a very large corpus. Poor statistical estimation of these rare events will always be a problem when relying on data-driven techniques, especially when only small amounts of data are available.

One proposed solution is to augment corpus-derived statistics with linguistic knowledge, available in the form of existing lexical and semantic resources. Such resources include lexical databases like WordNet (Fellbaum, 1998), knowledge bases like Cyc (Lenat, 1995), thesauri like Roget's (Chapman, 1977), and machine readable dictionaries like the Longman Dictionary of Contemporary English (Proctor, 1978). These linguistic resources have been used for many natural language processing tasks, such as resolving syntactic ambiguity (Resnik, 1999), identifying spelling errors (Budanitsky and Hirst, 2001), and disambiguating word senses (Agirre and Rigau, 1996). However, as they are not frequency-based, it is not clear in general how to use them within a statistical framework.

In this paper, we consider the use of linguistic knowledge derived from WordNet for combating data sparseness in two language modeling tasks. WordNet is a large, widely-used, general-English semantic network that groups words together into synonym sets, and links these sets with a variety of linguistic and semantic relations. The taxonomic structure of WordNet enables us to automatically derive word similarities, based on variants of a method proposed in (Resnik, 1995). The goal of this paper is to examine the usefulness of these WordNet-derived word similarities for supplementing corpus statistics in the following two tasks.

The first task is to improve the perplexity of a bigram language model trained on very little data. If

semantically similar words have similar bigram distributions, then for rare words we can use WordNet to find similar, more common *proxy* words. By combining the bigram data of rare words with their nearest neighbors, we hope to reduce data sparseness and thus better approximate each word's true bigram distribution.

The second task is to find long-distance correlations in text; specifically, we would like to find words that tend to co-occur within a sentence. By identifying these "sticky pairs" we can build language models that better reflect the semantic coherence evident within real sentences. Association statistics collected from data are only reliable, however, for high frequency words. Long-distance associations are generally semantic, and thus WordNet seems an appropriate resource for augmenting the data.

In section 2 we describe three measures proposed in the literature for measuring the semantic similarity between concepts in a taxonomy. In addition, we introduce a novel measure of word similarity that takes into account sense frequency. Sections 3 and 4 present a more detailed motivation, methodology, results and error analysis for the bigram and semantic coherence tasks, respectively. Finally, in Section 5 we conclude and discuss future work.

2 Measuring Similarity in a Taxonomy

2.1 Measuring Concept Similarity

WordNet is a large, general-English semantic network that represents concepts as groups of synonymous words, called *synsets*. WordNet is comprised of approximately 110K synsets, which are connected by about 150K edges representing a variety of linguistic and semantic relations. The largest component of WordNet consists of noun synsets connected by hypernym (IS-A) relations, effectively forming a noun taxonomy.

The simplest measure of semantic similarity between two concepts in a taxonomy (synsets in WordNet) is the length of the shortest path between them. The shorter the path, the more similar the concepts should be. However, this simple correspondence between path length and similarity is not always valid because edges in a taxonomy often span significantly different semantic distances (Resnik, 1999). For example, in WordNet, eight edges separate *rabbit* from *organism* (*rabbit* IS-A *leporid* ... IS-A *mammal* ... IS-A *organism*), whereas only one edge separates *plankton* from *organism* (*plankton* IS-A *organism*).

(Resnik, 1995) has proposed an alternative measure of semantic distance in a taxonomy, based on the intuition that the similarity between two concepts is equal to the amount of information they share. He collects counts from a corpus to estimate the probability of each concept in the taxonomy, then uses these probabilities to obtain the *information content* (negative log likelihood) of a concept. The Resnik similarity between two concepts c_1 and c_2 is defined as the information content of their least common ancestor (lca):

$$sim_{res}(c_1, c_2) = -\log(p(lca(c_1, c_2))$$

The Resnik similarity measure has some properties that may not be desirable. For instance, the extent of self-similarity depends on how specific a concept is; two items x and y can be more similar to each other than another item z is to itself. In an attempt to address this and other issues, many other similarity measures have subsequently been proposed. We selected two additional measures suggested in the literature which have shown to be highly correlated with human judgments of similarity.

(Jiang and Conrath, 1997) calculate the inverse of semantic similarity – semantic distance. However, since any distance measure can easily be converted into a measure of similarity through simple algebraic manipulation, we still refer to it as a measure of similarity. Although Jiang and Conrath motivate their measure differently, it essentially quantifies the distance between two concepts as the amount of information that is not shared between them. This is just the total amount of information in the two synsets minus their shared information:

$$sim_{jc}(c_1, c_2) = \\ 2 \cdot \log(p(lca(c_1, c_2))) - (\log(p(c_1)) + \log(p(c_2)))$$

(Lin, 1998) takes into account both similarities and differences between the two concepts. He normalizes the amount of shared information by the total amount of information, essentially calculating the

percentage of information which is shared:

$$sim_{lin}(c_1, c_2) = \frac{2 \cdot \log(p(lca(c_1, c_2)))}{\log(p(c_1)) + \log(p(c_2))}$$

2.2 Measuring Word Similarity

The three similarity measures just described are all designed for determining the similarity between concepts in a taxonomy. However, most words have multiple possible senses corresponding to distinct concepts in the taxonomy. Given a concept similarity measure $sim(c_1, c_2)$, Resnik defines the similarity between *words* as the maximum similarity between any of their senses. He searches over all possible combinations of senses and takes the most informative ancestor:

$$wsim_{max}(w_1, w_2) = \max_{c_1, c_2}[sim(c_1, c_2)]$$

where c_1 ranges over the senses of w_1 and c_2 ranges over the senses of w_2.

However, as (Resnik, 1999) notes, this method "sometimes produces spuriously high similarity measures for words on the basis of inappropriate word senses." In particular, it measures words as highly similar even if it is only their rare senses that are similar in meaning. For instance, $wsim_{max}$ will find that *brazil* and *pecan* are very similar because they are both *nuts*, which is a rare concept and thus has high information content. However, in broadcast news *brazil* almost always refers to a country and so on average it is not at all similar to *pecan*.

To address this issue, we devised another word similarity measure, $wsim_{wgt}$, which is the weighted sum of $sim(c_1, c_2)$, over *all* pairs of senses c_1 of w_1 and c_2 of w_2, where more frequent senses are weighted more heavily:

$$wsim_{wgt}(w_1, w_2) = \sum_{c_1, c_2} p(c_1|w_1) \cdot p(c_2|w_2) \cdot sim(c_1, c_2))$$

Here, $p(c_j|w_i)$ is the probability of word i mapping to sense j and is derived from a sense tagged corpus (Miller et al., 1993).

3 Improving Bigram Estimation

A general technique for combating data sparseness for n-gram language models is to derive clusters of similar words or phrases. The hope is that pooling words in the same equivalence class will result in more reliable estimation of model parameters and better generalization to unseen sequences. Several algorithms have been suggested for automatically clustering the vocabulary using information theoretic criteria (Brown et al., 1992; Kneser and Ney, 1993). Another method (Dagan et al., 1999) groups words according to their distributional similarity. All data-driven vocabulary clustering algorithms, however, suffer from the same limitation: when a word is rare there is usually not enough data to identify an appropriate equivalence class. Rare words, which could most benefit from being assigned to an appropriate cluster, cannot be reliably clustered. In this section we test whether the bigram distributions of semantically similar words (according to WordNet) can be combined to reduce the bigram perplexity of rare words.

3.1 Methodology

We use simple linear interpolation to combine a target word's bigram estimate with that of a similar, but more common *proxy* word. This use of *proxy* is similar to the notion of *synonym* in (Jelinek et al., 1990).

Formally, let $p_{ml}(\cdot|w)$ be the unsmoothed (maximum likelihood) bigram distribution following word w as derived from the training corpus and $p_{gt}(\cdot|w)$ be the corresponding Good-Turing smoothed distribution. Then $p_{gt}(\cdot|t)$ is the baseline word predictor following target word t and

$$p^s(\cdot|t) = (1 - \lambda)p_{gt}(\cdot|t) + \lambda p_{ml}(\cdot|s)$$

is the hopefully improved prediction using proxy s, where λ is optimized using 10-way cross-validation on the training set. When the amount of training data is small, the 10-way cross-validation leads to highly erratic weights because counts are often small, and sometimes even zero. To avoid extreme values, we smooth (shrink) λ towards 0.20, with less shrinkage as the amount of training data grew.

For each word t we use WordNet to choose a proxy s from the set of candidate proxies; we choose the proxy which is most similar to t according to the WordNet similarity measure. To evaluate WordNet's proxy selection, we compare to the proxies selected by two measures of similarity derived solely

from the training data. The first is the Kullback-Leibler distance between the target distribution and the proxy distribution: $D(p_{gt}(\cdot|t) \mid\mid p_{ml}(\cdot|s))$. The second measure is the training set perplexity reduction of word s, i.e. the improvement in perplexity of the interpolated model $p^s(\cdot|t)$ compared to the 10-way cross-validated model.

To evaluate each proxy s we calculate its percent reduction in perplexity on a test set, comparing the perplexity of $p^s(\cdot|t)$ with the baseline model $p_{gt}(\cdot|t)$.

3.2 Experiments

Our corpus consists of 140 Million Words (MW) of broadcast news (Graff, 1997). Of this data, 40MW was reserved for testing, and various subsets of the remainder were used for training.

The set of target words and their candidate proxies were selected from the nouns included in WordNet. We randomly selected a sample of 150 target words that occurred only once or twice in the first one million words of our training set but at least 50 times in the test set (so that we can accurately estimate perplexity reduction). For the proxies we selected all the nouns occurring at least 50 times in the first one million words. This gave us a set of 1862 candidate proxies.

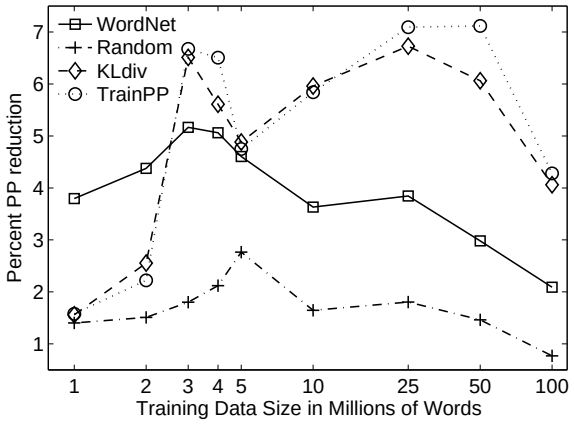

Figure 1: Perplexity reduction as a function of training data size for four similarity measures: WordNet, random, KL divergence, and training set PP reduction.

In order to evaluate the usefulness of WordNet for perplexity reduction on varying training set sizes, we created nine random subsets of the training corpus,

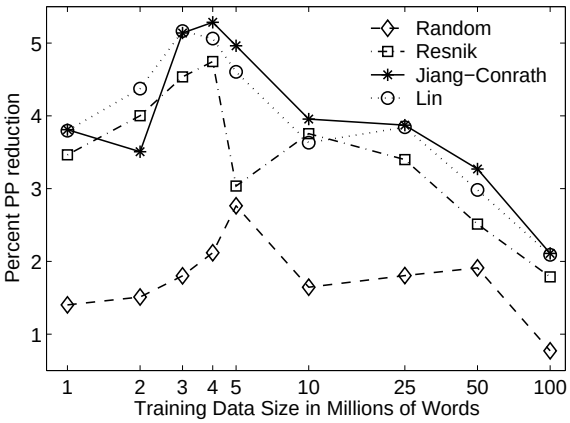

Figure 2: Perplexity reduction as a function of training data size for a random baseline and three WordNet similarity measures: Resnik, Jiang-Conrath, and Lin.

ranging in size from 1MW to the entire 100MW.

For each training set size we chose the highest scoring proxy word for each target word, according to each of the similarity measures. For each proxy selected we then calculated the resulting reduction in perplexity on words that follow the target word in the test set. The average perplexity reduction is defined as the weighted average of the perplexity reduction achieved for each target word, where each target word is weighted by its frequency in the test set.

Figure 1 shows the average perplexity reduction on the test set as a function of training data size. With large amounts of training data the proxies selected by WordNet reduce perplexity more than the randomly selected proxies, but less than the proxies selected from the data. On the other hand, when the amount of training data is small (1-2MW), the proxies selected by the data alone have close to random performance and WordNet achieves a larger reduction in perplexity. Although the expected trends are present, the magnitude of WordNet's perplexity reduction is small, with a maximum at around 5%.

The WordNet perplexity reduction remains relatively consistent regardless of which measure of word or concept similarity is selected. Figure 2 compares the performance of the three different WordNet concept similarity measures and a random baseline. Although all three WordNet measures perform

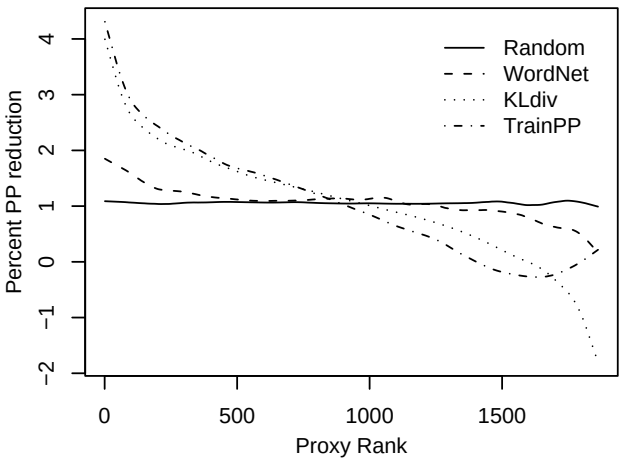

Figure 3: Perplexity reduction as a function of proxy rank for four similarity measures (with 100MW training data)

%	Classification	Example
15	Closely related in WN	blizzard-storm
40	Not closely related in WN	aluminum-steel
45	Not IS-A related	rug-arm

Table 1: Classification of best proxies for 150 target words.

significantly better than the baseline, the differences between them is small, and none of the variants performs consistently better over all data sizes.

Another interesting question is to what extent the average perplexity reduction changes if we choose a lower-ranked proxy instead of the highest-ranked. For each of the three similarity measures, as well as a random proxy ranking, Figure 3 shows the reduction in perplexity achieved by always using the n^{th} ranked proxy, where n ranges from 1 to 1862. For the three similarity measures, we can achieve a reduction in perplexity by using any proxy ranked among the first few hundred. The random performance, as expected, remains the same at every rank.

3.3 Error Analysis

In order to understand why the optimal proxies aren't often chosen by WordNet, we selected a small number of proxies to analyze manually. For each target word we found the proxy which resulted in the largest test set perplexity reduction, and classified it into one of three categories, shown in Table 1.

Our analysis shows that about 45% of the best proxies are not related by an IS-A relation; in most cases we could identify no close semantic relationship. For instance, the best proxy for *rug* is *arm*, because they both tend to be followed by words like *draped, pulled, behind, under*. Likewise, the best proxy for *name* is *spelling*, because in broadcast news they tend to be followed by words like *please* and *correctly*. In a small number of cases (e.g. *testament-religion*) the words are related topically, but not via an IS-A relation. A few more cases consist of words whose usage is domain specific. For instance, in the broadcast news corpus, *glove* refers almost exclusively to a piece of evidence in the OJ Simpson trial, and *Beard* refers to a man.

Another large category of misses is due to idiosyncrasies, errors, or incompleteness in the WordNet taxonomy. For instance, *aluminum* is classified as a chemical element whereas *steel* is classified as an alloy/mixture; a *bomb* is classified as an explosive device while a *shell* is listed under weaponry; *commentary* and *testimony* are both messages but have no closer common ancestor.

In summary, about half of WordNet's misses are due to inherent limitations in the types of relationships it seeks to encode, and the other half are due to idiosyncrasies or errors in the current database.

4 Semantic Coherence

Perhaps the most salient deficiency of conventional n-gram language models is their complete failure at modeling semantic coherence. These models capture short distance correlations among words in a sentence, yet are unable to distinguish meaningful sentences (where the content words come from the same semantic domain) from "fake" sentences (where the content words are drawn randomly). If we could devise an accurate measure of semantic similarity we could incorporate it as a feature into a language model. An exponential language model (Rosenfeld, 1997), for instance, could include a constraint on the expected similarity between all content words in a sentence.

One way of extracting similarity statistics from data is to collect a large corpus of utterances and then statistically analyze the sentences to determine groups of words that tend to co-occur (Cai et al.,

2000; Eneva et al., 2001).

Many measures of association have been proposed which work well when calculated from large amounts of data. For words that occur only a few times, however, it is usually not possible to identify with confidence which other words are likely to co-occur with them in the future. If semantically similar words tend to co-occur, however, then we can use WordNet to find associated words even for rare words.

4.1 Methodology and Experiments

To assess the potential effectiveness of WordNet for finding long distance word correlations, we compare the similarities derived from WordNet to a statistical measure of association calculated from *large* amounts of data. These statistical associations will serve as a "ground truth" with which to evaluate WordNet. If frequent words that co-occur in our corpus tend to have high WordNet scores then we will feel confident in relying on WordNet's similarity judgments in situations where only little data is available.

For these experiments we selected a set of about 500,000 noun pairs, where the expected number of chance co-occurrences of each pair (assuming pairwise independence) in our 100MW corpus was required to be greater than five. This constraint on the expectation ensures that the statistical associations are based on enough data to be credible, and can thus serve as an effective "ground truth."

	WORD 1 YES	WORD 1 NO
WORD 2 YES	C_{11}	C_{12}
WORD 2 NO	C_{21}	C_{22}

Figure 4: Contingency Table

To evaluate word association we chose to use the Q measure of association (also known as Yule's statistic (Cai et al., 2000)). It is based on a 2×2 contingency table, shown in Figure 4.1. C_{11} is the number of sentences in the training corpus which contain both words, C_{21} is the number of sentences which contain only word 1, etc.

From the counts in the contingency table we can compute the Q statistic for each pair of words:

$$Q = \frac{C_{11} \cdot C_{22} - C_{12} \cdot C_{21}}{C_{11} \cdot C_{22} + C_{12} \cdot C_{21}}$$

The values of Q range from -1 to 1; Q is -1 when two words have never occurred in the same sentence, and is 1 when they always occur together.

For each wordpair we calculated its Q score and its WordNet similarity score. The overall correlation between these two scores can be seen in Figure 5. The graph shows a sample of 10,000 noun pairs, with their $wsim_{wgt}$ scores plotted against their Q score. In general, words with high WordNet similarity scores are likely to co-occur. Figure 6 illustrates this even more clearly. When the WordNet scores are high ($wsim_{wgt} > 6$), the Q scores are generally positive. Unfortunately, however, this happens very rarely. Only 0.1% of the wordpairs have Word-

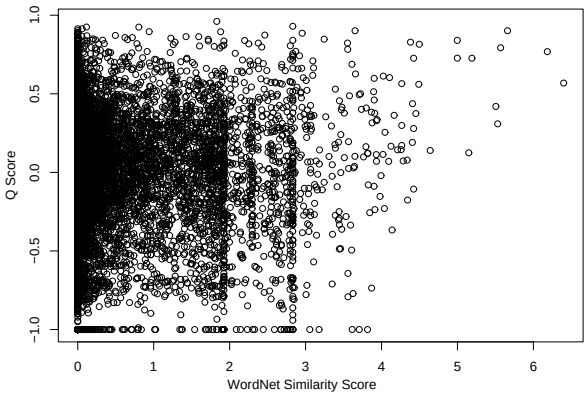

Figure 5: WordNet similarity scores versus Q scores for 10,000 noun pairs

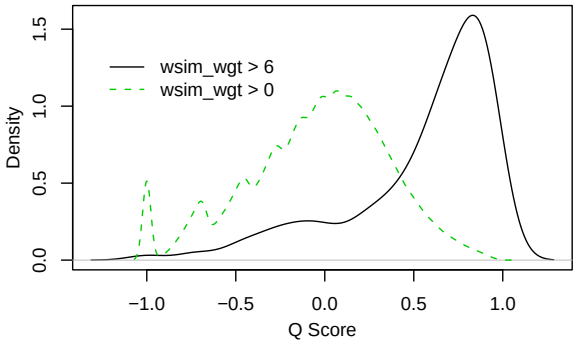

Figure 6: Distribution of Q scores for wordpairs with high WordNet similarity ($wsim_{wgt} > 6$) vs. distribution of Q scores for all wordpairs.

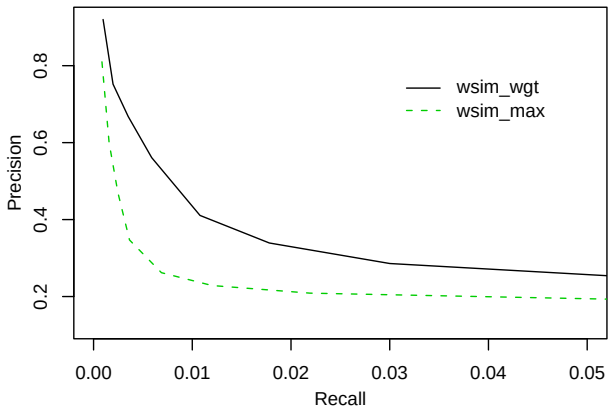

Figure 7: Precision/recall curve for $wsim_{wgt}$ vs. $wsim_{max}$, for wordpairs with $Q > 0.5$.

Net similarity scores above 5 and only 0.03% are above 6. Thus, over the whole range of possible values, the two scores have essentially no correlation ($\rho = 0.01$).

Although there was no significant difference in performance among the three information based concept similarity measures (discussed in Section 2.1), they all performed significantly better than the simple edge counting technique.

The two *word* similarity measures (discussed in Section 2.2), however, did yield significantly different results. Figure 7 compares $wsim_{wgt}$ to $wsim_{max}$. This graph is a precision/recall curve where we assume there is some reference standard (some set of truly similar words), then plot the tradeoff between precision and recall as the WordNet similarity threshold is lowered. For this graph we defined "truly associated" to be a Q score greater than 0.5. This is a somewhat arbitrary choice, but when we vary this cutoff the graph does not change much. The weighted similarity ($wsim_{wgt}$) performs better; it consistently has higher precision for equivalent recall values. However, for both measures, the recall is disappointingly low.

4.2 Error Analysis

We manually conducted a small analysis to try to understand WordNet's poor recall on the semantic coherence task. In particular, we tried to determine the cause of misses (when nouns that co-occur are given low WordNet similarity scores).

From the original set of 500K noun pairs we selected those pairs which were most highly associated ($Q > 0.9$). These pairs were then manually classified into categories based on the relationship between the words (e.g. part-whole, antonyms,...) The final categories and counts are listed in Table 2. The majority of co-occurring word pairs were related only topically. WordNet does not currently encode these topical relationships, explaining why only a third of the pairs were found to be connected by a relationship that WordNet was designed to encode. Furthermore, only 59% of that third were actually found to be related in the current version of WordNet. This lack of coverage considerably limits the usefulness of WordNet for predicting which words will co-occur together in the news.

Relation	**Num**	**Examples**
WN	**277(163)**	
part/member	87 (15)	student-school
phrase is-a	65 (47)	*death* tax IS-A *tax*
coordinates	41 (31)	house-senate, gas-oil
morphology	30 (28)	hospital-hospitals
is-a	28 (23)	gun-weapon
antonyms	18 (13)	majority-minority
other	8 (6)	doctor-patient
non-WN	**461**	
topical	336	evidence-guilt
news/events	102	iraq-weapons
other	23	*end* of the *spectrum*

Table 2: Relation types of wordpairs that tend to co-occur in broadcast news. Counts in parentheses are the number of pairs that were actually related in the current version of WordNet.

5 Conclusions and Future Work

We find that word similarities derived from WordNet are a very weak source of knowledge for these two language modeling tasks. Interpolating the bigram distribution of rare words with proxies selected from WordNet results in a small perplexity improvement compared to proxies selected only from the data. Although words with very high WordNet similarity do tend to co-occur within sentences, recall is poor because most words with strong long-distance correlations are topically related, and WordNet currently does not include topical links. However, top-

ical clusters are being added to the next version of WordNet, so performance on this task might improve. Nonetheless, we have found that the limited types and quantities of relationships in WordNet compared to the spectrum of relationships found in real data make it difficult to make substantial improvements for these tasks by using a linguistic knowledge source even as large as WordNet.

Despite the disappointing performance of the WordNet-derived similarity measures for these two tasks, we see some possible directions for future work. While the experiments in this paper interpolate each target word with only a single proxy, Figure 3 shows that *many* of the top-ranked proxies reduce perplexity. It seems likely, therefore, that interpolating each target word with multiple proxies would result in further perplexity reduction. In addition, while our present work merely compares the usefulness of the proxies selected by the data and WordNet individually, a straightforward extension would be to *combine* the two knowledge sources. Incorporating WordNet's word similarities as a prior would allow them to dominate when there is little data, but eventually be washed out by the more reliable statistical measures.

Another possibility is to try similar techniques on a slightly different task, like trigram prediction, where data sparsity is a more significant problem. Prediction of distance two bigrams might also be more successful, because at larger distances linguistic behavior tends to be governed more by semantic rather than syntactic relations. Another more long term direction is to automatically learn which subsets of WordNet are most reliable. For example, anecdotal evidence indicates that proxies derived from WordNet's *organism* hierarchy are consistently useful, whereas proxies from the *communication* hierarchy are less reliable.

References

E. Agirre and G. Rigau. 1996. Word sense disambiguation using conceptual density. In *COLING*.

P. Brown, V. DellaPietra, P. deSouza, J. Lai, and R. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

A. Budanitsky and G. Hirst. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and other lexical resources*.

C. Cai, R. Rosenfeld, and L. Wasserman. 2000. Exponential language models, logistic regression, and semantic coherence. In *NIST/DARPA Speech Transcription Workshop*.

R. Chapman, editor. 1977. *Roget's International Thesaurus*. Harper and Row, fourth edition.

I. Dagan, L. Lee, and F. Pereira. 1999. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1-3):43–69. Special issue on natural language learning.

E. Eneva, R. Hoberman, and L. Lita. 2001. Within-sentence semantic coherence. In *EMNLP*.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

D. Graff. 1997. The 1996 broadcast news speech and language model corpus. In *DARPA Workshop on Spoken Language Technology*.

F. Jelinek, R. Mercer, and S. Roukos. 1990. Classifying words for improved statistical language models. In *ICASSP*.

J. Jiang and D. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *ROCLING*.

R. Kneser and H. Ney. 1993. Improved clustering techniques for class-based statistical language modelling. In *EUROSPEECH*.

D. B. Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11), November.

D. Lin. 1998. An information-theoretic definition of similarity. In *ICML*.

G. Miller, C. Leacock, R. Tengi, and R. Bunker. 1993. A semantic concordance. *ARPA WorkShop on Human Language Technology*.

P. Proctor, editor. 1978. *The Longman Dictionary of Contemporary English (LDOCE)*. Longman Group.

P. Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*.

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.

R. Rosenfeld. 1997. A whole sentence maximum entropy language model. In *IEEE Workshop on Automatic Speech Recognition and Understanding*.