Determining the specificity of nouns from text

Sharon A. Caraballo and **Eugene Charniak**

Dept. of Computer Science
Brown University
Providence, RI 02912
{sc,ec}@cs.brown.edu

Abstract

In this work, we use a large text corpus to order nouns by their level of specificity. This semantic information can for most nouns be determined with over 80% accuracy using simple statistics from a text corpus without using any additional sources of semantic knowledge. This kind of semantic information can be used to help in automatically constructing or augmenting a lexical database such as WordNet.

1 Introduction

Large lexical databases such as WordNet (see Fellbaum (1998)) are in common research use. However, there are circumstances, particularly involving domain-specific text, where WordNet does not have sufficient coverage. Various automatic methods have been proposed to automatically build lexical resources or augment existing resources. (See, e.g., Riloff and Shepherd (1997), Roark and Charniak (1998), Caraballo (1999), and Berland and Charniak (1999).) In this paper, we describe a method which can be used to assist in this problem.

We present here a way to determine the relative specificity of nouns; that is, which nouns are more specific (or more general) than others, using only a large text corpus and no additional sources of semantic knowledge. By gathering simple statistics, we are able to decide which of two nouns is more specific to over 80% accuracy for nouns at "basic level" or below (see, e.g., Lakoff (1987)), and about 59% accuracy for nouns above basic level.

It should be noted that specificity by itself is not enough information from which to construct a noun hierarchy. This project is meant to provide a tool to support other methods. See Caraballo (1999) for a detailed description of a method to construct such a hierarchy.

2 Previous work

To the best of our knowledge, this is the first attempt to automatically rank nouns based on specificity.

Hearst (1992) found individual pairs of hypernyms and hyponyms from text using pattern-matching techniques. The sparseness of these patterns prevents this from being an effective approach to the problem we address here.

In Caraballo (1999), we construct a hierarchy of nouns, including hypernym relations. However, there are several areas where that work could benefit from the research presented here. The hypernyms used to label the internal nodes of that hierarchy are chosen in a simple fashion; pattern-matching as in Hearst (1992) is used to identify candidate hypernyms of the words dominated by a particular node, and a simple voting scheme selects the hypernyms to be used. The hypernyms tend to lose quality as one looks closer to the root of the tree, generally by being too specific. This work could help to choose more general hypernyms from among the candidate words. In addition, it could be used to correct places where a more specific word is listed as a hypernym of a more general word, and to select between hypernyms which seem equally good by the voting scheme (which is currently done arbitrarily).

3 Methods for determining specificity

We tested several methods for ordering nouns by their specificity. Each of these methods was trained on the text of the 1987 Wall Street Journal corpus, about 15 million words. When parsed text was needed, it was obtained using a parser recently developed at Brown which performs at about the 88% level on the standard precision and recall measures.

One possible indicator of specificity is how often the noun is modified. It seems reasonable to suppose that very specific nouns are rarely modified, while very general nouns would usually be modified. Using the parsed text, we collected statistics on the probability that a noun is modified by a prenominal adjective, verb, or other noun. (In all of these measures, when we say "noun" we are referring only to common nouns, tagged NN or NNS, not proper nouns tagged NNP or NNPS. Our results were consistently better when proper nouns were eliminated, probably since the proper nouns may conflict with identically-spelled common nouns.) We looked at both the probability that the noun is modified by any of these modifiers and the probability that the noun is modified by each specific category. The nouns, adjectives, and verbs are all stemmed before computing these statistics.

$$P_{adj}(noun) = \frac{count(noun \text{ with a prenominal adjective})}{count(noun)}$$

$$P_{vb}(noun) = \frac{count(noun \text{ with a prenominal verb})}{count(noun)}$$

$$P_{nn}(noun) = \frac{count(noun \text{ with a prenominal noun})}{count(noun)}$$

$$P_{mod}(noun) = \frac{count(noun \text{ with prenom. adj, vb, or nn})}{count(noun)}$$

However, if a noun almost always appears with exactly the same modifiers, this may be an indication of an expression (e.g., "ugly duckling"), rather than a very general noun. For this reason, we also collected entropy-based statistics. For each noun, we computed the entropy of the rightmost prenominal modifier.

$$H_{mod}(noun) = -\sum_{modifier} [P(modifier|noun) * \log_2 P(modifier|noun)]$$

where $P(modifier|noun)$ is the probability that a (possibly null) *modifier* is the rightmost modifier of *noun*. The higher the entropy, the more general we believe the noun to be. In other words, we are considering not just how often the noun is modified, but how much these modifiers vary. A great variety of modifiers suggests that the noun is quite general, while a noun that is rarely modified or modified in only a few different ways is probably fairly specific.

We also looked at a simpler measure which can be computed from raw text rather than parsed text. (For this experiment we used the part-of-speech tags determined by the parser, but that was only to produce the set of nouns for testing. If one wanted to compute this measure for all words, or for a specific list of words, tagging would be unnecessary.) We simply looked at all words appearing within an n-word window of any instance of the word being evaluated, and then computed the entropy measure:

$$H_n(noun) = -\sum_{word} [P(word|noun) * \log_2 P(word|noun)]$$

where $P(word|noun)$ is the probability that a word appearing within an n-word window of *noun* is *word*. Again, a higher entropy indicates a more general noun. In this measure, the nouns being evaluated are stemmed, but the words in its n-word window are not.

Finally, we computed the very simple measure of frequency ($freq(noun)$). The higher

the frequency, the more general we expect the noun to be. (Recall that we are using tagged text, so it is not merely the frequency of the word that is being measured, but the frequency of the word or its plural tagged as a common noun.)

This assumed inverse relationship between frequency and the semantic content of a word is used, for example, to weight the importance of terms in the standard IDF measure used in information retrieval (see, e.g., Sparck Jones (1972)), and to weight the importance of context words to compare the semantic similarity of nouns in Grefenstette (1993).

4 Evaluation

To evaluate the performance of these measures, we used the hypernym data in WordNet (1998) as our gold standard. (A word X is considered a hypernym of a word Y if native speakers accept the sentence "Y is a (kind of) X.") We constructed three small hierarchies of nouns and looked at how often our measures found the proper relationships between the hypernym/hyponym pairs in these hierarchies.

To select the words for our three hierarchies, we wanted to use sets of words for which there would be enough information in the Wall Street Journal corpus. We chose three clusters produced by a program similar to Roark and Charniak (1998) except that it is based on a generative probability model and tries to classify all nouns rather than just those in pre-selected clusters. (All data sets are given in the Appendix.) The clusters we selected represented vehicles (car, truck, boat, ...), food (bread, pizza, wine, ...), and occupations (journalist, engineer, biochemist, ...). From the clustered data we removed proper nouns and words that were not really in our target categories. We then looked up the remaining words in WordNet, and added their single-word hypernyms to the categories in the correct hierarchical structure. (Many WordNet categories are described by multiple words, e.g., "motorized vehicle", and these were omitted for obvious reasons.)

For each of these three hierarchies, we looked at each hypernym/hyponym pair within the hierarchy and determined whether each specificity measure placed the words in the proper order. The percentage each specificity measure placed correctly are presented in Table 1.

Clearly the better measures are performing much better than a random-guess algorithm which would give 50% performance.

Among the measures based on the parsed text (P_{mod} and its components and H_{mod}), the entropy-based measure H_{mod} is clearly the best performer, as would be expected. However, it is interesting to note that the statistics based on adjectives alone (P_{adj}) somewhat outperform those based on all of our prenominal modifiers (P_{mod}). The reasons for this are not entirely clear.

Although H_{mod} is the best performer on the vehicles data, $freq$ and H_{50} do marginally better overall, with each having the best results on one of the data sets. All three of these measures, as well as H_2 and H_{10}, get above 80% correct on average.

In these evaluations, it became clear that a single bad node high in the hierarchy could have a large effect on the results. For example, in the "occupations" hierarchy, the root node is "person," however, this is not a very frequent word in the Wall Street Journal corpus and rates as fairly specific across all of our measures. Since this node has eight children, a single bad value at this node can cause eight errors. We therefore considered another evaluation measure: for each internal node in the tree, we evaluated whether each specificity measure rated this word as more general than *all* of its descendants. (This is somewhat akin to the idea of edit distance. If we sufficiently increased the generality measure for each node marked incorrect in this system, the hierarchy would match WordNet's exactly.) The results for this evaluation are presented in Table 2. Although this is a harsher measure, it isolates the effect of individual difficult internal nodes.

Although the numbers are lower in Table

Specificity measure	Vehicles	Food	Occupations	Average
P_{mod}	65.2	63.3	66.7	65.0
P_{adj}	65.2	67.3	69.7	67.4
P_{vb}	73.9	42.9	51.5	56.1
P_{nn}	65.2	57.1	51.5	58.0
H_{mod}	91.3	79.6	72.7	81.2
H_2	87.0	79.6	75.8	80.8
H_{10}	87.0	79.6	75.8	80.8
H_{50}	87.0	85.7	75.8	82.8
$Freq$	87.0	83.7	78.8	83.1

Table 1: Percentage of parent-child relationships which are ordered correctly by each measure.

Specificity measure	Vehicles	Food	Occupations	Average
P_{mod}	44.4	57.9	53.3	51.9
P_{adj}	33.3	52.6	60.0	48.7
P_{vb}	33.3	21.1	40.0	31.5
P_{nn}	55.6	21.1	33.3	36.6
H_{mod}	77.8	63.2	66.7	69.2
H_2	66.7	57.9	60.0	61.5
H_{10}	66.7	63.2	60.0	63.3
H_{50}	66.7	73.7	60.0	66.8
$Freq$	66.7	63.2	60.0	63.3

Table 2: Percentage of internal nodes having the correct relationship to all of their descendants.

2, the same measures as in Table 1 perform relatively well. However, here H_{mod} has the best performance both on average and on two of three data sets, while the $freq$ measure does a bit less well, now performing at about the level of H_{10} rather than H_{50}. The fact that some of the numbers in Table 2 are below 50% should not be alarming, as the average number of descendants of an internal node is over 5, implying that random chance would give performance well below the 50% level on this measure.

Some of these results are negatively affected by word-sense problems. Some of the words added from the WordNet data are much more common in the Wall Street Journal data for a different word sense than the one we are trying to evaluate. For example, the word "performer" is in the occupations hierarchy, but in the Wall Street Journal this word generally refers to stocks or funds (as "good performers", for example) rather than to people. Since it was our goal not to use any outside sources of semantic knowledge these words were included in the evaluations. However, if we eliminate those words, the results are as shown in Tables 3 and 4.

It is possible that using some kind of automatic word-sense disambiguation while gathering the statistics would help reduce this problem. This is also an area for future work. However, it should be noted that on the evaluation measures in Tables 3 and 4, as in the first two tables, the best results are obtained with H_{mod}, H_{50} and $freq$.

The above results are primarily for nouns at "basic level" and below, which includes the vast majority of nouns. We also considered a data set at basic level and above, with "entity" at its root. Table 5 presents the

Specificity measure	Vehicles	Food	Occupations	Average
P_{mod}	65.0	62.5	67.7	65.1
P_{adj}	70.0	66.7	71.0	69.2
P_{vb}	80.0	43.8	48.4	57.4
P_{nn}	70.0	56.3	51.6	59.3
H_{mod}	100.0	81.3	74.2	85.1
H_2	95.0	79.2	77.4	83.9
H_{10}	95.0	79.2	77.4	83.9
H_{50}	95.0	85.4	77.4	85.9
$Freq$	95.0	83.3	80.6	86.3

Table 3: Percentage of correct parent-child relationships when words with the wrong predominant sense are removed.

Specificity measure	Vehicles	Food	Occupations	Average
P_{mod}	50.0	55.6	61.5	55.7
P_{adj}	33.3	50.0	61.5	48.3
P_{vb}	33.3	16.7	38.5	29.5
P_{nn}	66.7	22.2	30.8	39.9
H_{mod}	100.0	55.6	76.9	77.5
H_2	83.3	55.6	69.2	69.4
H_{10}	83.3	61.1	61.5	68.7
H_{50}	83.3	72.2	69.2	74.9
$Freq$	83.3	61.1	69.2	71.2

Table 4: Percentage of internal nodes with the correct relationship to all descendants when words with the wrong predominant sense are removed.

results of testing on this data set and each measure, for the evaluation measures described above, percentage of correct parent-child relationships and percentage of nodes in the correct relationship to all of their descendants.

Note that on these nouns, $freq$ and H_{50} are among the worst performers; in fact, by looking at the parent-child results, we can see that these measures actually do worse than chance. As nouns start to get extremely general, their frequency appears to actually decrease, so these are no longer useful measures. On the other hand, H_{mod} is still one of the best performers; although it does perform worse here than on very specific nouns, it still assigns the correct relationship to a pair of nouns about 59% of the time.

5 Conclusion

Determining the relative specificity of nouns is a task which can help in automatically building or augmenting a semantic lexicon. We have identified various measures which can identify which of two nouns is more specific with over 80% accuracy on basic-level or more specific nouns. The best among these measures seem to be H_{mod}, the entropy of the rightmost modifier of a noun, H_{50}, the entropy of words appearing within a 50-word window of a target noun, and $freq$, the frequency of the noun. These three measures perform approximately equivalently.

If the task requires handling very general nouns as well as those at or below the basic level, we recommend using the H_{mod} measure. This measure performs nearly as well as the other two on specific nouns, and much better on general nouns. However,

Specificity measure	Parent-child	All descendants
P_{mod}	59.1	46.4
P_{adj}	60.2	46.4
P_{vb}	50.0	35.7
P_{nn}	50.0	28.6
H_{mod}	59.1	39.3
H_2	53.4	25.0
H_{10}	45.5	32.1
H_{50}	46.6	32.1
$Freq$	45.5	32.1

Table 5: Evaluation of the various specificity measures on a test set of more general nouns.

if it is known that the task will only involve fairly specific nouns, such as adding domain-specific terms to an existing hierarchy which already has the more general nouns arranged appropriately, the easily-computed $freq$ measure can be used instead.

6 Acknowledgments

Thanks to Mark Johnson and to the anonymous reviewers for many helpful suggestions. This research is supported in part by NSF grant IRI-9319516 and by ONR grant N0014-96-1-0549.

References

Matthew Berland and Eugene Charniak. 1999. Finding parts in very large corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics.*

Sharon A. Caraballo. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics.*

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

Gregory Grefenstette. 1993. SEXTANT: Extracting semantics from raw text implementation details. *Heuristics: The Journal of Knowledge Engineering.*

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics.*

George Lakoff. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind.* University of Chicago Press.

Ellen Riloff and Jessica Shepherd. 1997. A corpus-based approach for building semantic lexicons. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 117–124.

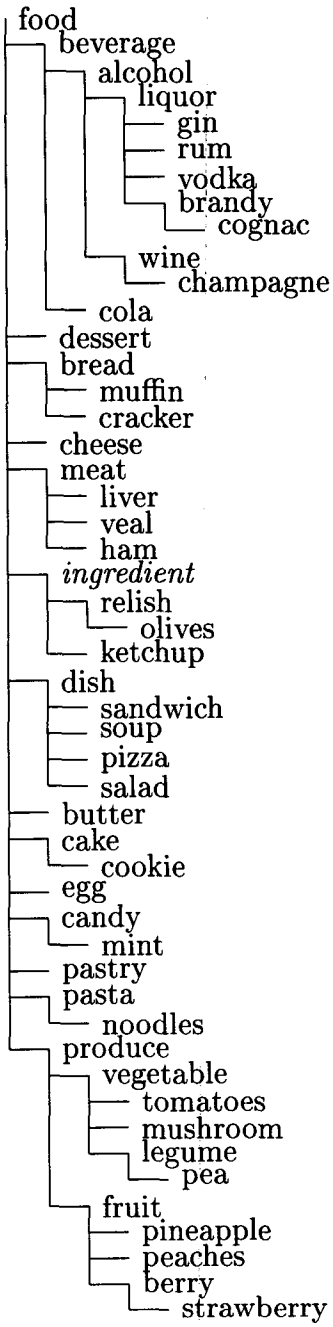
Brian Roark and Eugene Charniak. 1998. Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. In *COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics: Proceedings of the Conference*, pages 1110–1116.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

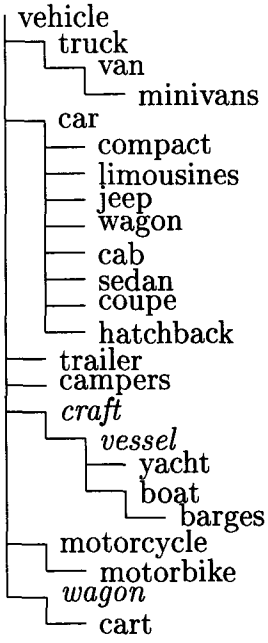
Appendix

Below are the data sets used in these experiments. Words shown in *italics* are omitted from the results in Tables 3 and 4 because the predominant sense in the Wall Street Journal text is not the one represented by the word's position in this hierarchy.

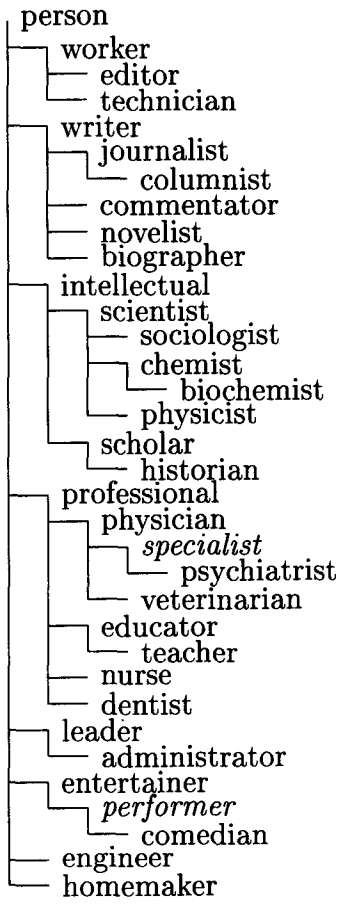
Food:



Vehicles:



Occupations:



Entities (data used for Table 5):