Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction

Appeared in Conference on Natural Language Learning (CoNLL-2003), Edmonton, Canada, pages 111-118

Scott Cederberg **Dominic Widdows**
Center for the Study of Language and Information
210 Panama Street
Stanford University
Stanford CA 94305
{cederber,dwiddows}@csli.stanford.edu

Abstract

In this paper we demonstrate methods of improving both the recall and the precision of automatic methods for extraction of hyponymy (IS_A) relations from free text. By applying latent semantic analysis (LSA) to filter extracted hyponymy relations we reduce the rate of error of our initial pattern-based hyponymy extraction by 30%, achieving precision of 58%. Applying a graph-based model of noun-noun similarity learned automatically from coordination patterns to previously extracted correct hyponymy relations, we achieve roughly a five-fold increase in the number of correct hyponymy relations extracted.

1 Introduction

This paper demonstrates that mathematical models for measuring semantic similarity between concepts can be used to improve the learning of hyponymy relationships between concepts from free text. In particular, we show that latent semantic analysis can be used to filter results, giving an increase in precision, and that neighbors in a graph built from coordination information can be used to improve recall.

The goal of extracting semantic information from text is well-established, and has encouraged work on lexical acquisition (Roark and Charniak, 1998), information extraction (Cardie, 1997), and ontology engineering (Hahn and Schnattinger, 1998). The purpose of this kind of work is to collect information about the meanings of lexical items or phrases, and the relationships between them, so that the process of building semantic resources (such as ontologies and dictionaries) by hand can be automated or at least helped.

One of the standard ways of arranging concepts is in a concept hierarchy or *taxonomy* such as the WordNet noun taxonomy (Fellbaum, 1998). The fundamental relationship between objects in a taxonomy is called *hyponymy*, where y is a hyponym of x if every y is also an x. For example, every *trout* is also a *fish*, so we say that *trout* is a hyponym ("below name") of *fish* and conversely, *fish* is a hypernym ("above name") of *trout*. Other names exist for variants of the hyponymy relationship, such as an IS_A relationship, a parent-node / child-node relationship, and a broader term / narrower term relationship. It is also noted that the *genus* of an object, in traditional lexicographic terms, is often a hypernym of that object (Guthrie et al., 1996). Throughout this paper we will write $y \sqsubset x$ for the relationship "y is a hyponym of x". In this paper, we use the hyponymy relationship to describe subset relationships, so we regard $y \sqsubset x$ to be true if the set of y's can reasonably be said to be a subset of the set of x's.[1]

Because hyponymy relationships are so central to knowledge engineering, there have been numerous attempts to learn them from text, beginning with those of Hearst (1992). We review this work in Section 2, where we reproduce similar experiments as a baseline from which to expand. The rest of the paper demonstrates ways in which other mathematical models built from text corpora can be used to improve hyponymy extraction. In Section 3, we show how latent semantic analysis can be used to filter potential relationships according to their "semantic plausibility". In Section 4, we show how correctly extracted relationships can be used as "seed-cases" to extract several more relationships, thus improving recall; this work shares some similarities with that of Caraballo (1999). In Section 5 we show that combining the techniques of Section 3 and Section 4 improves both precision and recall. Section 6 demonstrates that

[1] Another possible view is that "hyponymy" should only refer to core relationships, not contingent ones (so *pheasant* $\sqsubset$ *bird* might be accepted but *pheasant* $\sqsubset$ *food* might not be, because it depends on context and culture). We use the broader "subset" definition because contingent relationships are an important part of world-knowledge (and are therefore worth learning), and because in practice we found the distinction difficult to enforce. Another definition is given by Caraballo (1999): "...a word A is said to be a *hypernym* of a word B if native speakers of English accept the sentence 'B is a (kind of) A.' "

linguistic tools such as lemmatization can be used to reliably put the extracted relationships into a normalized or "canonical" form for addition to a semantic resource.

2 Pattern-Based Hyponymy Extraction

The first major attempt to extract hyponyms from text was that of Hearst (1992), described in more detail in (Hearst, 1998), who extracted relationships from the text of *Grolier's Encyclopedia*. The method is illustrated by the following example. The sentence excerpt

> Even then, we would trail behind other European Community members, such as Germany, France and Italy... (BNC)[2]

indicates that Germany, France, and Italy are all European Community members. More generally, phrases of the form

x such as y_1 (y_2, ..., and/or y_n)

frequently indicate that the y_i are all hyponyms of the hypernym x. Hearst identifies several other constructions that have a tendency to indicate hyponymy, calling these constructions *lexicosyntactic patterns*, and analyses the results. She reports that 52% of the relations extracted by the "*or other*" pattern (see Table 1) were judged to be "pretty good relations". A more recent variant of this technique was implemented by Alfonseca and Manandhar (2001), who compare the collocational patterns of words from *The Lord of the Rings* with those of words in the WordNet taxonomy, adding new nouns to WordNet with an accuracy of 28%. Using a much more knowledge-intensive approach, Hahn and Schnattinger (1998) improve "learning accuracy" from around 50% to over 80% by forming a number of hypotheses and accepting only those which are most consistent with their current ontology. Their methods are like ours in that the "concept learning" combines information from several occurrences, but differ in that they rely on a detailed existing ontology into which to fit the new relationships between concepts.

Our initial experiment was to construct a hyponymy extraction system based on the six lexicosyntactic patterns identified in (Hearst, 1998), which are listed in Table 1. We first used a chunker to mark noun groups, and then recognized and extracted noun groups occurring as part of one of the extraction patterns.[3]

We applied these extraction patterns to an approximately 430,000-word extract from the beginning of the British National Corpus (BNC). The patterns extracted 513 relations. We selected 100 of the extracted relations at random and each author evaluated them by hand, scoring each relation on a scale from 4 (correct) to 0 (incorrect), defined as follows:

x such as y_1 (, y_2, ..., and/or y_n)
such x as y_1 (, y_2, ..., and/or y_n)
y_1 (, y_2, ..., y_n,) or other x
y_1 (, y_2, ..., y_n,) and other x
x, including y_1 (, y_2, ..., and/or y_n)
x, especially y_1 (, y_2, ..., and/or y_n)

Table 1: The lexicosyntactic patterns described by Hearst (1998), which we used in the work described in this paper. Each of these patterns is taken to indicate the hyponymy relation(s) $y_i \sqsubset x$.

4. Extracted hypernym and hyponym exactly correct as extracted.

3. Extracted hypernym and hyponym are correct after a slight modification, such as depluralization or the removal of an article (*e.g.* a, the) or other preceding word.

2. Extracted hypernym and hyponym have something correct, *e.g.* a correct noun without a necessary prepositional phrase, a correct noun with a superfluous prepositional phrase, or a noun + prepositional phrase where the object of the preposition is correct but the preposition itself and the noun to which it attaches are superfluous. Thus these hyponymy relations are potentially correct but will require potentially difficult processing to extract an exactly correct relation. Some of the errors which would need to be corrected were in preprocessing (*e.g.* on the part of the noun-group chunker) and others were errors caused by our hyponymy extractor (*e.g.* tacking on too many or too few prepositional phrases).

1. The relation extracted is correct in some sense, but is too general or too context specific to be useful. This category includes relations that could be made useful by anaphora resolution (*e.g.* replacing "this" with its referent).

0. The relation extracted is incorrect. This results when the constructions we recognize are used for a purpose other than indicating the hyponymy relation.

The results of each of the authors' evaluations of the 100-relation random sample are show in Table 2.[4] For

[2]This excerpt and others in this paper are from the British National Corpus.

[3]The chunker used was LT CHUNK, from the University of Edinburgh's Language Technology Group. It can be downloaded from `http://www.ltg.ed.ac.uk/software/chunk/`.

[4]Table 2 suggests that although there is significant disagreement about how to assign scores of 1 and 0, inter-annotator

score	Author 1	Author 2
4	4	2
3	34	35
2	14	13
1	35	22
0	13	28

Table 2: Number of the 100 randomly selected hyponymy relations (of 513 extracted) to which each of the authors assigned the five available scores.

purposes of calculating precision, we consider those relations with a score of 4 or 3 to be correct and those with a lower score to be incorrect. After discussion between the authors on disputed annotations to create "gold standard" annotations, we found that 40 of the 100 relations in our random sample were correct according to this criterion. In other words, 40% of the relations extracted were exactly correct or would be correct with the use of minor post-processing consisting of lemmatization and removal of common types of qualifying words. (We describe our application of such post-processing in Section 6.)

Thus our initial implementation of Hearst-style hyponymy extraction achieved 40% precision. This is less than the 52% precision reported in (Hearst, 1998). We believe this discrepancy to be mainly due to the difference between working with the BNC and Grolier's encyclopedia—as noted by Hearst, the encyclopedia is designed to be especially rich in conceptual relationships presented in an accessible format.

Various problems with the pattern-based extraction method explain the 60% of extracted relations that were incorrect and/or useless. One problem is that the constructions that we assume to indicate hyponymy are often used for other purposes. For instance, the pattern

x including y_1, y_2, ..., and y_n

which indicates hyponymy in sentences such as

> Illnesses, including chronic muscle debility, herpes, tremors and eye infections, have come and gone. (BNC)

and is a quite productive source of hyponymy relations, can be used instead to indicate group membership:

> Often entire families including young children need practical home care ... (BNC)

While all children are members of families, the hyponymy relationship *child* $\sqsubset$ *family* does not hold, since it is not true that all children *are* families.

Another source of errors in lexicosyntactic hyponymy extraction is illustrated by the sentence

> A kit such as Edme Best Bitter, Tom Caxton Best Bitter, or John Bull Best Bitter will be a good starting kit. (BNC)

which indicates the (potentially useful) relations *Edme Best Bitter* $\sqsubset$ *beer-brewing kit*, *Tom Caxton Best Bitter* $\sqsubset$ *beer-brewing kit*, and *John Bull Best Bitter* $\sqsubset$ *beer-brewing kit*, but only when we use the context to infer that the type of "kit" referred to is a beer-brewing kit, a process that is difficult by automatic means. Without this inference, the extracted relations *Edme Best Bitter* $\sqsubset$ *kit*, etc., while correct in a certain sense, are not helpful. One frequent source of such problems is anaphora that require resolution.

There are also problems related to prepositional phrase attachment.

3 Improving Precision Using Latent Semantic Analysis

Solving all of the problems with pattern-based hyponymy extraction that we describe above would require near-human-level language understanding, but we have applied a far simpler technique for filtering out many of the incorrect and spurious extracted relations with good results, using a variant of latent semantic analysis (LSA) (Deerwester et al., 1990; Baeza-Yates and Ribiero-Neto, 1999, p. 44). LSA is a method for representing words as points in a vector space, whereby words which are related in meaning should be represented by points which are near to one another. The LSA model we built is similar to that described in (Schütze, 1998). First 1000 frequent content words (*i.e.* not on the stoplist)[5] were chosen as "content-bearing words". Using these content-bearing words as column labels, the other words in the corpus were assigned row vectors by counting the number of times they occured within a 15-word context window of a content-bearing word. Singular-value decomposition (Deerwester et al., 1990) was then used to reduce the number of dimensions from 1000 to 100. Similarity between two vectors (points) was measured using the cosine of the angle between them, in the same way as the similarity between a query and a document is often measured

agreement regarding the assignment of scores of 4, 3, and 2 is quite high. Indeed, considering the rougher distinction we use for reporting precision, in which scores of 4 and 3 are deemed correct and scores of 2, 1, and 0 are deemed incorrect, we found that inter-annotator agreement across all relations annotated (including those from this random sample and those from the sample described in Section 3) was 86%. We discussed each of the relations in the 14% of cases where we disagreed until we reached agreement; this produced the "gold standard" annotations to which we refer.

[5] A "stoplist" is a list of frequent words which have little semantic content in themselves, such as prepositions and pronouns (Baeza-Yates and Ribiero-Neto, 1999, p. 167).

score	Author 1	Author 2
4	4	5
3	57	52
2	18	14
1	12	19
0	9	10

Table 3: Number of the 100 top-ranked hyponymy relations (of 513 extracted) to which each of the authors assigned the five available scores.

in information retrieval (Baeza-Yates and Ribiero-Neto, 1999, p. 28). Effectively, we could use LSA to measure the extent to which two words x and y usually occur in similar contexts. This LSA similarity score will be called $\mathrm{sim}(x, y)$.

Since we expect a hyponym and its hypernym to be semantically similar, we can use the LSA similarity between two terms as a test of the plausibility of a putative hyponymy relation between those terms. If their similarity is low, it is likely that they do not have a true and useful hyponymy relationship; the relation was probably extracted erroneously for one or more of the reasons listed above. If the similarity between two terms is high, we have increased confidence that a hyponymy relationship exists between them, because we know that they are at least in similar "semantic regions".

We ranked the 513 putative hyponym/hypernym pairs that we extracted from our trial excerpt of the BNC according to the similarity between the putative hypernym and the putative hyponym in each pair; *i.e.* for each pair x and y where the relationship $y \sqsubset x$ had been suggested, we calculated the cosine similarity $\mathrm{sim}(x, y)$, then we ranked the extracted relations from highest to lowest similarity. We then manually evaluated the accuracy of the top 100 extracted relations according to this ranking using the 5-point scale described in Section 2. We found that 58 of these 100 top-ranked relations received scores of 4 or 3 according to our "gold standard" annotations. Comparing this 58% precision with the 40% precision obtained on a random sample in Section 2, we determine that LSA achieved a 30% reduction in error (see Table 3 for a breakdown of annotation results by author).[6]

Thus LSA proved quite an effective filter. LSA provides broad-based semantic information learned statistically over many occurences of words; lexicosyntactic hyponymy extraction learns semantic information from specific phrases within a corpus. Thus we have benefitted from combining local patterns with statistical information. Considered in analogy with the process by which humans learn from reading, we might think of the semantic information learned by LSA as background knowledge that is applied by the reader when determining what can accurately be gleaned from a particular sentence when it is read.

[6]It should be noted that 24 of the top 100 hyponymy relations evaluated in this section were also in the randomly-chosen sample of 100 relations described in Section 2. Thus there were a total of 176 distinct hyponymy relations across both test sets.

4 Improving Recall Using Coordination Information

One of the main challenges facing hyponymy extraction is that comparatively few of the correct relations that might be found in text are expressed overtly by the simple lexicosyntactic patterns used in Section 2, as was apparent in the results presented in that section.

This problem has been addressed by Caraballo (1999), who describes a system that first builds an unlabelled hierarchy of noun clusters using agglomerative bottom-up clustering of vectors of noun coordination information. The leaves of this hierarchy (corresponding to nouns) are assigned hypernyms using Hearst-style lexicosyntactic patterns. Internal nodes in the hierarchy are then labelled with hypernyms of the leaves they subsume according to a vote of these subsumed leaves.

We proceed along similar lines, using noun coordination information and an alternative graph-based clustering method. We do not build a complete hierarchy, but our method nonetheless obtains additional hypernym-hyponym pairs not extracted by lexicosyntactic patterns.

Our method is based on the following sort of inference. Consider the sentence

> This is not the case with sugar, honey, grape must, cloves and other spices which increase its merit. (BNC)

which provides evidence that *clove* is a kind of *spice*. Given this, the sentence

> Ships laden with nutmeg or cinnamon, cloves or coriander once battled the Seven Seas to bring home their precious cargo. (BNC)

might suggest that *nutmeg*, *cinnamon*, and *coriander* are also spices, because they appear to be similar to *cloves*. Thus we can learn the hyponymy relations *nutmeg* $\sqsubset$ *spice*, *cinnamon* $\sqsubset$ *spice*, and *coriander* $\sqsubset$ *spice* that are not directly attested by lexicosyntactic patterns in our training corpus.

This kind of information from coordination patterns has been used for work in automatic lexical acquisition (Riloff and Shepherd, 1997; Roark and Charniak, 1998; Widdows and Dorow, 2002). The basic rationale behind these methods is that words that occur together in lists are usually semantically similar in some way: for example, the phrase

y_1, y_2, and y_3

suggests that there is some link between y_1 and y_2, etc. Performing this analysis on a whole corpus results in a data structure which holds a collection of nouns and observed noun-noun relationships. If we think of the nouns as *nodes* and the noun-noun relationships as *edges*, this data structure is a graph (Bollobás, 1998), and combinatoric methods can be used to analyze its structure.

Work using such techniques for lexical acquisition has proceeded by building classes of related words from a single "seed-word" with some desired property (such as being a representative of a paticular semantic class). For example, in order to extract a class of words referring to kinds of *disease* from a corpus, you start with a single seed-word such as *typhoid*, and then find other nouns that occur in lists with *typhoid*. Using the graph model described above, Widdows and Dorow (2002) developed a combinatoric algorithm for growing clusters from a single seed-word, and used these methods to find correct new members for chosen categories with an accuracy of over 80%.

The idea that certain patterns can be identified using finite-state techniques and used as evidence for semantic relationships is the same as Hearst's (1992), but appears to be more effective for finding just similar words rather than hypernyms because there are many more instances of simple coordination patterns than of hypernymy patterns—in the lists we used to extract these relationships, we see much more cooccurence of words on the same ontological level than between words from different ontological levels. For example, in the BNC there are 211 instances of the phrase "fruit and vegetables" and 9 instances of "carrots and potatoes", but no instances of "fruit and potatoes", only 1 instance of "apples and vegetables", and so on.

This sort of approach should be ideal for improving the recall of automatic hyponymy extraction, by using the hyponym from each of the correct hypernym/hyponym pairs as a seed-word for the category represented by the hypernym—for example, from the relationship *clove* $\sqsubset$ *spice*, the word *clove* could be taken as a seed-word, with the assumption that words which frequently occur in coordination with *clove* are also names of spices.

We used the algorithm of (Widdows and Dorow, 2002) on the British National Corpus to see if many more hyponymy relations would be extracted in this way. For each correct pair $y \sqsubset x$ where y was a single-word hyponym of x discovered by the lexicosyntactic patterns of Section 2, we collected the 10 words most similar to y according to this algorithm and tested to see if these neighbors were also hyponyms of x.

Of the 176 extracted hyponyms that we evaluated by hand in the overlapping test sets described in Section 2 and Section 3, 95 were rated 4 or 3 on our 5-point scoring system (Section 2) by at least one of the authors. Considering these correct or nearly-correct relations in their hand-corrected form, we found that 45 of these 95 relations involved single-word hyponyms. (We restricted our attention to these 45 relations because the graph model was built using only single words as nodes in the graph.)

This set of 45 correct hypernym "seed-pairs" was extended by another potential 459 pairs (slightly more than 10 for each seed-pair because if there was a tie for 10th place both neighbors were used). Of these, 211 (46%) were judged to be correct hypernym pairs and 248 (54%) were not.[7] This accuracy compares favorably with the accuracy of 40% obtained for the raw hyponymy extraction experiments in Section 2, suggesting that inferring new relations by using corpus-based similarities to previously known relations is more reliable than trying to learn completely new relations even if they are directly attested in the corpus. However, our accuracy falls way short of the figure of 82% reported by Widdows and Dorow (2002). We believe this is because the classes in (Widdows and Dorow, 2002) are built from carefully selected seed-examples: ours are built from an uncontrolled sample of seed-examples extracted automatically from a corpus. We outline three cases where this causes a critical difference.

The ambiguity of "mass"

One of the correct hyponymy relations extracted in our experiments in Section 2 was *mass* $\sqsubset$ *religious service*. Using *mass* as a seed suggested the following candidates as potential hyponyms of *religious service*:

Seed	Semantically Similar Words
mass	length weight angle shape depth height range charge size momentum

All these neighbors are related to the "measurement of physical property" sense of the word *mass* rather than the "religious service" sense. The inferred hyponymy relations are all incorrect because of this mismatch.

The specific properties of "nitrogen"

Another true relation we extracted was *nitrogen* $\sqsubset$ *nutrient*. Using the same process as above gave the following neighbors of nitrogen:

Seed	Semantically Similar Words
nitrogen	methane dioxide carbon hydrogen methanol vapour ammonia oxide oxygen monoxide water

These neighboring terms are not in general nutrients, and the attempt to infer new hyponymy relations is a fail-

[7] As before, we consider scores of 4 and 3 on our 5-point scale to be correct and lower scores to be incorrect. The precision of graph-model results (reported in this section and in Section 5), unlike those reported elsewhere, are based on the annotations of a single author.

ure in this case. While the relationship *nitrogen* $\sqsubset$ *nutrient* is one of the many facts which go to make up the vast store of world-knowledge that an educated adult uses for reasoning, it is not a necessary property of nitrogen itself, and one could arguably "know" the meaning of nitrogen without being aware of this fact. In traditional lexicographic terms, the fact that nitrogen is a nutrient might be regarded as part of the differentiae rather than the genus of nitrogen. Had our seed-pair instead been *nitrogen* $\sqsubset$ *gas* or *nitrogen* $\sqsubset$ *chemical element*, many correct hyponymy relations would have been inferred by our method, and both of these classifications are central to the meaning of *nitrogen*.

Accurate levels of abstraction for "dill"

Finally, even when the hyponymy relationship $y \sqsubset x$ used as a seed-case was central to the meaning of y and all of the neighbors of y were related to this meaning, they were still not always hyponyms of x but sometimes members of a more general category. For example, using the correct seed-pair *dill* $\sqsubset$ *herb* we retrieved the following suggested hyponyms for *herb*:

Seed	Semantically Similar Words
dill	rind fennel seasoning juice sauce pepper parsley vinegar oil pur

All of these items are related to dill, but only some of them are herbs. The other items should also be placed in the same general area of a taxonomy as *dill*, but as cooking ingredients rather than specifically herbs.

In spite of these problems, the algorithm for improving recall by adding neighbors of the correct hyponyms worked reasonably well, obtaining 211 correct relationships from 45 seeds, an almost fivefold increase in recall, with an accuracy of 46%, which is better than that of our baseline pattern-matching hyponymy extractor.

It is possible that using coordination (such as cooccurence in lists) as a measure of noun-noun similarity is well-adapted for this sort of work, because it mainly extracts "horizontal" relationships between items of similar specificity or similar generality. Continuing the geometric analogy, these mainly "horizontal" relationships might be expected to combine particularly well with seed examples of "vertical" relationships, *i.e.* hyponymy relationships.

5 Combining LSA and Coordination to Improve Precision and Recall

Having used two separate techniques to improve precision and recall in isolation, it made sense to combine our methods to improve performance overall. This was accomplished by applying LSA filtering as described in Section 3 to the results obtained by extending our initial hypernym pairs with coordination patterns in Section 4.

LSA filtering of extended results: phase I

The first application of filtering to the additional hyponymy relations obtained using noun-cooccurrence was straightforward. We took the 459 potential hyponymy relationships obtained in Section 4. For each of the prospective hyponyms y of a given hypernym x, we computed the LSA similarity $\mathrm{sim}(x, y)$. We then considered only those potential hyponyms whose LSA similarity to the hypernym surpassed a certain threshhold. Using this technique with an experimentally determined threshhold of 0.15, we obtained a set of 260 hyponymy relations of which 166 were correct (64%, as opposed to the 46% correct in the unfiltered results). The LSA filtering had removed 154 incorrect relationships and only 45 correct ones, reducing the overall error rate by 33%.

In particular, this technique removed all but one of the spurious *religious service* hyponyms which were obtained through inappropriate similarities with *mass* in the example in Section 4, though it was much less effective in filtering the neighbors of *nitrogen* and *dill*, as might be expected.

LSA filtering of extended results: phase II

For some of the hyponymy relations to which we applied our extension technique, the hypernym had multiple words.[8] In some of these cases, it was clear that one of the words in the hypernym had a meaning more closely related to the original (correct) hyponym. For instance, in the *mass* $\sqsubset$ *religious service* relation, the word *religious* tells us more about the appropriate meaning of *mass* than does the word *service*. It thus seemed that, at least in certain cases, we might be able to get more traction in LSA filtering of potential additional hyponyms by first selecting a particular word from the hypernym as the "most important" and using that word rather than the entire hypernym for filtering.[9]

We thus applied a simple two-step algorithm to refine the filtering technique presented above:

1. The LSA similarity between the original (correct) hyponym and each word in the hypernym is computed. The words of the hypernym are ranked according to these similarities.

2. The word in the hypernym that has the highest LSA similarity to the original (correct) hyponym is used instead of the entire hypernym for phase-I-style filtering.

[8] The graph model used to obtain new candidate hyponyms was built using single words, which is why our extended results include some multiword expressions among the hypernyms but only single word hyponyms.

[9] When using an entire multiword hypernym for filtering, a term-vector was produced for the multiword hypernym by averaging the LSA vectors for the constituent words.

This filtering technique, with an LSA-similarity threshhold of 0.15, resulted in the extraction of 35 correct and 25 incorrect relationships. In contrast, using LSA similarity with the whole expression rather than the most important word resulted in the extraction of 32 correct and 30 incorrect relationships for those hypernyms with multiple words. On the face of it, selecting only the most important part of the hypernym for comparison enabled us to obtain more correct and fewer incorrect relations, but it is also clear that by this stage in our experiments our sample of seed-relationships had become too small for these results to be statistically significant.

However, the examples we considered did demonstrate another point—that LSA could help to determine which parts of a multiword expression were semantically relevant. For example, one of the seed-relationships was *France* $\sqsubset$ *European Community member*. Finding that

$$\mathrm{sim}(\textit{france}, \textit{european}) > \mathrm{sim}(\textit{france}, \textit{community}),$$

we could infer that the adjective *European* was central to the meaning of the hyponym, whereas for the example *wallflowers* $\sqsubset$ *hardy biennials* the opposite conclusion, that *hardy* is an adjectival modifier which isn't central to the relationship, could be drawn. However, these conclusions could also be drawn by using established collocation extraction techniques (Manning and Schütze, 1999, Ch. 5) to find semantically significant multiword expressions.

6 Obtaining Canonical Forms for Relations

An important part of extracting semantic relations like those discussed in this paper is converting the terms in the extracted relations to a canonical form. In the case of our extracted hyponymy relations, such normalization consists of two steps:

1. Removing extraneous articles and qualifiers. Our extracted hyponyms and hypernyms were often in the form "another x", "some x", and so forth, where x is the hypernym or hyponym that we actually want to consider.

2. Converting nouns to their singular form. This is elementary morphological analysis, or a limited form of lemmatization.

We performed the second of these steps using the `morph` morphological analysis software (Minnen et al., 2001).[10] To perform the first step of removing modifiers, we implemented a Perl script to do the following:

- Remove leading determiners from the beginning of the hypernym and from the beginning of the hyponym.
- Remove leading prepositions from the beginning of the hypernym. Doing this after removing leading determiners eliminates the common "those of" construction.
- Remove cardinal numbers from the hypernym and the hyponym.
- Remove possessive prefixes from the hypernym and the hyponym.
- Remove "set of" and "number of" from the hypernym and the hyponym. This *ad hoc* but reasonable procedure eliminates common troublesome constructions not covered by the above rules.
- Remove leading adjectives from hypernyms, but not from hyponyms. In addition to removing "other", this amounts to playing it safe. By removing leading adjectives we make potential hypernyms more general, and thus more likely to be a superset of their potential hyponym. While this removal sometimes makes the learned relationship less useful, it seldom makes it incorrect. We leave adjectives on hyponyms to make them more specific, and thus more likely to be a subset of their purported hypernym.

Using these simple rules, we were able to convert 73 of the 78 relations orginally scored as 3 (see Section 2) to relations receiving a score of 4. This demonstrates as a "proof of concept" that comparatively simple language processing techniques can be used to map relationships from the surface forms in which they were observed in text to a canonical form which could be included in a semantic resource.

7 Conclusion and Further Work

The results presented in this paper demonstrate that the application of linguistic information from automatically-learned mathematical models can significantly enhance both the precision and the recall of pattern-based hyponymy extraction techniques. Using a graph model of noun similarity we were able to obtain an almost five-fold improvement in recall, though the precision of this technique is clearly affected by the correctness of the "seed-relationships" used. Using LSA filtering we eliminated spurious relations extracted by the original pattern method, reducing errors by 30%. Such filtering also eliminated spurious relations learned using the graph model that were the result of lexical ambiguity and of seed hyponymy relations inappropriate for the technique, reducing errors by 33%.

[10] This software is freely available from `http://www.cogs.susx.ac.uk/lab/nlp/carroll/morph.html`.

This paper suggests many possibilities for future work. First of all, it would be interesting to apply LSA to a system for building an entire hypernym-labelled ontology in roughly the way described in (Caraballo, 1999), perhaps by using an LSA-weighted voting method to determine which hypernym would be used to label each node. We are considering how to extend our techniques to such a task.

Also, systematic comparison of the lexicosyntactic patterns used for extraction to determine the relative productiveness and accuracy of each pattern might prove illuminating, as would comparison across different corpora to determine the impact of the topic area and medium/format of documents on the effectiveness of hyponymy extraction. Ultimately, the ability to predict *a priori* how well a knowledge-extraction system will work on a previously unseen corpus will be crucial to its usefulness.

Applying the techniques of this paper to a system that used mutual bootstrapping (Riloff and Jones, 1999) to find additional extraction patterns would also be interesting (such an approach is suggested in (Hearst, 1998)). And of course, further refinement of the mathematical models we use and our methods of learning them, including more sophisticated use of available tools for linguistic pre-processing, such as the identification and indexing of multiword expressions, could further improve the precision and recall of hyponymy extraction techniques.

Acknowledgements

This research was supported in part by the Research Collaboration between the NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and CSLI, Stanford University, and by EC/NSF grant wIST-1999-11438 for the MUCHMORE project. Thanks also to Stanley Peters for his helpful comments on an earlier draft.

References

Enrique Alfonseca and Suresh Manandhar. 2001. Improving an ontology refinement method with hyponymy patterns. In *Third International Conference on Language Resources and Evaluation*, pages 235–239, Las Palmas, Spain.

Ricardo Baeza-Yates and Berthier Ribiero-Neto. 1999. *Modern Information Retrieval*. Addison Wesley / ACM Press.

Béla Bollobás. 1998. *Modern Graph Theory*. Number 184 in Graduate Texts in Mathematics. Springer-Verlag.

Sharon Caraballo. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. In *37th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference*, pages 120–126.

Claire Cardie. 1997. Empirical methods in information extraction. *AI Magazine*, 18:65–79.

Scott Deerwester, Susan Dumais, George Furnas, Thomas Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

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

L Guthrie, J Pustejovsky, Y Wilks, and B Slator. 1996. The role of lexicons in natural language processing. *Communications of the ACM*, 39(1):63–72.

Udo Hahn and Klemens Schnattinger. 1998. Towards text knowledge engineering. In *AAAI/IAAI*, pages 524–531.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING*, Nantes, France.

Marti A. Hearst, 1998. *WordNet: An Electronic Lexical Database*, chapter 5, Automated discovery of WordNet relations, pages 131–152. In Fellbaum (Fellbaum, 1998).

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of english. *Natural Language Engineering*, 7(3):207–223.

Ellen Riloff and Rosie Jones. 1999. Learning dictionaries for infomation extraction by multi-level bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, pages 472–479. AAAI.

Ellen Riloff and Jessica Shepherd. 1997. A corpus-based approach for building semantic lexicons. In Claire Cardie and Ralph Weischedel, editors, *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 117–124. Association for Computational Linguistics, Somerset, New Jersey.

Brian Roark and Eugene Charniak. 1998. Noun-phrase co-occurence statistics for semi-automatic semantic lexicon construction. In *COLING-ACL*, pages 1110–1116.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–124.

Dominic Widdows and Beate Dorow. 2002. A graph model for unsupervised lexical acquisition. In *19th International Conference on Computational Linguistics*, pages 1093–1099, Taipei, Taiwan, August.