Refining Automatically-Discovered Lexical Relations: Combining Weak Techniques for Stronger Results

Marti A. Hearst
Computer Science Division
571 Evans Hall
University of California, Berkeley
Berkeley, CA 94720
marti@cs.berkeley.edu

Gregory Grefenstette
Department of Computer Science
210 MIB
University of Pittsburgh
Pittsburgh, PA 15260
grefen@cs.pitt.edu

Abstract

Knowledge-poor corpus-based approaches to natural language processing are attractive in that they do not incur the difficulties associated with complex knowledge bases and real-world inferences. However, these kinds of language processing techniques in isolation often do not suffice for a particular task; for this reason we are interested in finding ways to combine various techniques and improve their results.

Accordingly, we conducted experiments to refine the results of an automatic lexical discovery technique by making use of a statistically-based syntactic similarity measure. The discovery program uses lexico-syntactic patterns to find instances of the hyponymy relation in large text bases. Once relations of this sort are found, they should be inserted into an existing lexicon or thesaurus. However, the terms in the relation may have multiple senses, thus hampering automatic placement. In order to address this problem we applied a term-similarity determination technique to the problem of choosing where, in an existing lexical hierarchy, to install a lexical relation. The union of these two corpus-based methods is promising, although only partially successful in the experiments run so far. Here we report some preliminary results, and make suggestions for how to improve the technique in future.

Introduction

Knowledge-poor corpus-based approaches to natural language processing are attractive in that they do not incur the difficulties associated with complex knowledge bases and real-world inferences, while they promise to offer efficient means of exploiting ever-growing quantities of on-line text. However, coarse-level language processing techniques in isolation often do not suffice for a particular task; for this reason we are interested in finding ways to combine various approaches and improve their results.

Accordingly, we conducted experiments to refine the results of an automatic lexical discovery technique by making use of a statistically-based syntactic similarity measure, and integrating them with an existing knowledge structure. The discovery program uses lexico-syntactic patterns to find instances of the hyponymy (i.e., ISA) relation in large text bases. Once relations of this sort are found, they should be inserted into an existing lexicon or thesaurus. However, the terms in the relation may have multiple senses, thus hampering automatic placement. In order to address this problem we applied a term-similarity determination technique to the problem of choosing where, in an existing lexical hierarchy, to install a lexical relation. The union of these two corpus-based methods is promising, although only partially successful in the experiments run so far.

These ideas are related to other recent work in several ways. We make use of restricted syntactic information as do Brent's (Brent 1991) verb subcategorization frame recognition technique and Smadja's (Smadja & McKeown 1990) collocation acquisition algorithm. The work reported here attempts to find semantic similarity among terms based on the contexts they tend to occur in; (Church & Hanks 1990) uses frequency of co-occurrence of content words to create clusters of semantically similar words, (Hindle 1990) uses both simple syntactic subject-verb-object frames and frequency of occurrence of content words to determine similarity among nouns, and (Calzolari & Bindi 1990) use corpus-based statistical association ratios to determine lexical information such as prepositional complementation relations, modification relations, and significant compounds. This paper presents an attempt to combine knowledge-poor techniques; (Wilks

et al. 1992) discusses the potential power behind combining weak methods and describes advances achieved using this paradigm.

The next section describes in detail the problem being addressed and the two existing coarse-level language processing techniques that are to be combined. This is followed by a description of how the similarity calculations are done, the results of applying these calculations to several examples, the difficulties that arise in each case, and a sketch of some solutions for these difficulties. We then illustrate a side-effect of the integration of statistical techniques with a lexical knowledge source, followed by a brief conclusion.

The Problem: Integration of Lexical Relations

(Hearst 1992) reports a method for the automatic acquisition of the hyponymy lexical relation from unrestricted text. In this method, the text is scanned for instances of distinguished lexico-syntactic patterns that indicate the relation of interest.

For example, consider the lexico-syntactic pattern

... NP {, NP} {,} or other NP ...*

When a sentence containing this pattern is found (with some restrictions on the syntax to the left and the right of the pattern) it can be inferred that the NP's on the left of *or other* are hyponyms of the NP on the right (where NP indicates a simple noun phrase). From the sentence

Bruises, wounds, broken bones or other injuries are common.

we can infer:

hyponym(bruise, injury)
hyponym(wound, injury)
hyponym(broken bone, injury)

This pattern is one of several that have been identified as indicating the hyponymy relation.

This approach differs from statistical techniques in an interesting way. Both require as their primary resource a large text collection, but whereas statistical techniques try to find correlations and make generalizations based on the data of the entire corpus, only a single instance of a lexico-syntactic pattern need be found in order to have made a discovery.

Once the relation is found, it is desirable to integrate it as a part of an existing network of lexical relations. We want to develop a means to correctly insert an instance of the hyponymy relation into an existing hyponymically-structured network (hyponymy is reflexive, and transitive, but not symmetric.).

For our experiments we use the manually constructed thesaurus WordNet (Miller *et al.* 1990). In WordNet, word forms with synonymous meanings are grouped into sets, called synsets. This allows a distinction to be made between senses of homographs. For example, the noun *"board"* appears in the synsets {*board, plank*} and {*board, committee*}, and this grouping serves for the most part as the word's definition. In version 1.1, WordNet contains about 34,000 noun word forms, including some compounds and proper nouns, organized into about 26,000 synsets. Noun synsets are organized hierarchically[1] according to the hyponymy relation with implied inheritance and are further distinguished by values of features such as meronymy. WordNet's coverage is impressive and provides a good base for an automatic acquisition algorithm to build on.

Now, assuming we have discovered the relation *hyponym(X, Y)*, indicating that X is a kind of Y, we wish to enter this relation into the WordNet network.

If the network is sufficiently mature, as is WordNet, we can assume that most of the highly ambiguous words are already present and appear in higher levels of the network. Therefore, most of the time we will be trying to insert a rather specific term that itself does not need to be disambiguated (i.e., it has only one main sense) as a hyponym of a term that can have one or more senses. If we assume this is indeed the case, then there are two scenarios to consider:

(Scenario 1) Each sense of Y has several child subtrees and the task is to determine which subtree X shares context with. This in turn implies which sense of Y the hyponym relation refers to.

(Scenario 2) One or more of the senses of Y have no children. Thus there are no subtrees to compare X against.

There are two considerations associated with Scenario 1:

(1a) X is not a direct descendent of Y, but belongs two or more levels down.

(1b) X belongs in a new subtree of its own, even though the correct sense of Y has one or more child subtrees.

In the work described here we address only the situation associated with Scenario 1, since our technique uses the child subtrees to determine which sense of Y to associate X with.

It has been observed (e.g., (Kelly & Stone 1975)), that the sense of a word can be inferred from the lexical contexts in which the word is found. As a (simplified) example, when 'bank' is used in its riverbank sense, it is often surrounded by words having to do with bodies of water, while when used in its financial institution sense, it appears with appropriate financial terms. The strategy we present here makes use of an extension of this idea; namely, we will look at the contexts of each subtree of the hypernym of interest, and see which subtrees' contexts coincide most closely with the contexts that the target hyponym tends to occur in.

[1] Although WordNet's hyponyms are structured as a directed network, as opposed to as a tree, for the purposes of this paper, we treat it as if it were a tree.

To restate: in order to place the hyponym relation into the network, we propose the following:

> Similarity Hypothesis: when comparing the contexts that hyponym X occurs in with the contexts of the subtrees (e.g., senses) of hypernym Y, X's contexts will be found to be most similar to those of the subtree of Y in which X belongs.

How will the context comparison be done? (Grefenstette 1992b) has developed a weak technique, embodied in a program called SEXTANT, which, given a target word, determines which other words in a corpus are most similar to it based on their syntactic usage patterns. For Scenario (1) we will employ this pairwise-similarity determination method on all the children of each sense of Y, in an effort to determine which subtree X is most similar to. Bearing consideration (1a) in mind, we may be matching against sub-sub-trees, down several levels, although presumably the more detailed the divisions, the more difficult it will be to distinguish among them using context alone.

Similarity Determination

This section describes the mechanism by which the syntactic similarity determination is done, using one running example. The results of determining syntactic similarity for several other examples appear in the following section.

Generating the corpus

One of the relations extracted from an on-line encyclopedia (Grolier 1990) using the technique described in (Hearst 1992), is *hyponym(Harvard, institution)*. As described above, if we wish to insert this relation into a hierarchical structure such as WordNet, we have to decide which sense of 'institution' is appropriate. Below we list an abbreviated version of the WordNet synsets associated with the various senses of 'institution'. Each sense is followed by the hyponymic subtrees associated with it, indicated by an arrow.

```
Institution: (hyponyms)
institution, establishment
          => charity
          => religion, faith, church
          => vicariate, vicarship
          => school, educational institution
          => academy, honorary society
          => foundation
          => bank, commercial bank

institution
          => orphanage, orphans' asylum
          => penal institution

constitution, establishment, formation,
initiation, founding, foundation,
institution, origination, setting up,
creation, instauration
          => colonization, settlement
```

Our goal is to see if, by examining the syntactic contexts of these terms in a corpus of text, we can decide under which synset to place 'Harvard'.

Given a large enough text sample, SEXTANT can tell us what words are used in the most similar ways to 'Harvard'. In order to generate this text, we took all the individual words from the above list, giving the list of words

> institution, establishment, charity, religion, faith, church, vicariate, vicarship, school, educational, academy, honorary society, foundation, bank, commercial bank, orphanage, orphans' asylum, penal institution, constitution, establishment, formation, initiation, founding, foundation, institution, origination, setting up, creation, instauration, colonization, settlement

and extracted all the sentences from (Grolier 1990) that contained those terms. This generated a corpus of 3.7 Million characters (630,000 words, 22,000 sentences). Sample sentences are:

```
Aachen is the site of an important
engineering school ( Technische Hochschule
Aachen ) and is a rail center of a large
coal-mining region


Zwingli's Sixty-seven Articles (1523)
for disputation became a basic doctrinal
document for the Swiss reformed church.
```

Syntactic Analysis of the Corpus

These sentences were grammatically analyzed using a robust system developed for CLARIT (Evans *et al.* 1991). At this stage, the following type of information is output, associating one grammatical category and a normalized word form for each word:

```
"aachen" ukw? aachen
"is" auxb be
"the" d the
"site" sn site
"of" prep of
"an" d an
"important" adj important
"engineering" vb-prog engineer
"school" sn school
"\(" *left-paren* \(
"technische" ukw? technische
"hochschule" ukw? hochschule
"aachen" ukw? aachen
"\)" *right-paren* \)
"and" cnj and
```

```
"is" auxb be
"a" d a
"rail" sn rail
"center" sn center
"of" prep of
"a" d a
"large" adj large
"coal" sn coal
"mining" vb-prog mine
"region" sn region
"\." *period* \.
...
```

The SEXTANT system takes this output and uses a number of simple robust algorithms (Grefenstette 1992a) to divide the sentences of the corpus into complex noun phrases (NP) and verb phrases (VP) as shown below:

```
NP      aachen
VP      be
NP      the site of an important engineering
        school
--      (
NP      technische hochschule aachen
--      ) and
VP      be
NP      a rail center of a large coal
        mine region
--      .
```

Then, the modifiers are connected to head nouns and the verbs are connected to their subjects and objects. For the above sample setence, SEXTANT produces the following list of connections, where **NN** signifies a noun-noun relation; **NNPREP** a noun-noun relation with an interposed preposition; **ADJ** an adjective; **SUBJ**, **DOBJ**, and **IOBJ** signify subject and direct and indirect object relations; and **MOD** is the inverse of **NN**, **NNPREP**, and **ADJ**:

These relations are not syntactically perfect, for example the system does not decide if 'coal' modifies 'mining' or 'region' but keeps both possibilities.

Syntactically-Derived Similarity

From these relations we retain a simplified version of head-modifier pairs and the modifier-head pairs which becomes the input to a similarity calculation. Using measures derived in the social sciences for comparing two individuals, each described by a number of attributes (Romesburg 1984), we compare each of the terms in the corpus derived from the words in the list above. The attributes of each word are those other words found in syntactic relation to it by the simple syntactic processing of SEXTANT. The data at this stage are a list of words and modifying attributes:

head	*modifier*	*type*
aachen	site	NN
engineering	important	ADJ
school	important	ADJ
school	engineer	SUBJ
site	school	NNPREP
hochschule	technische	NN
aachen	technische	NN
aachen	hochschule	NN
engineer	aachen	DOBJ
center	rail	NN
aachen	center	NN
coal	large	ADJ
region	large	ADJ
mining	coal	NN
region	coal	NN
coal	mine	SUBJ
center	region	NNPREP

Table 1: SEXTANT's Syntactic Analysis

```
aachen site
engineering important
school important
school engineer-SUBJ
site school
hochschule technische
aachen technische
aachen hochschule
aachen engineer-DOBJ
center rail
aachen center
coal large
region large
mining coal
region coal
coal mine-SUBJ
center region
site aachen-MOD
important engineering-MOD
important school-MOD
school site-MOD
technische hochschule-MOD
technische aachen-MOD
hochschule aachen-MOD
rail center-MOD
center aachen-MOD
large coal-MOD
large region-MOD
coal mining-MOD
coal region-MOD
region center-MOD
```

We use, as the similarity measure, the Jaccard coefficient, which is the sum of shared attributes divided by the sum of unique attributes in the two objects being compared. We use a weighted version of the Jaccard coefficient; each attribute is weighted between 0 and 1

term	*freq*	*closest term*
establishment	1140	creation
charity	76	devotion
religion	2347	religious
faith	835	religion
church	8308	school
vicariate	3	prothonotary
school	10012	institution
academy	2254	university
foundation	1697	institution
bank	2612	institution
orphanage	11	yverdon
asylum	11	promulgation
orphan	4	mottel
penal	16	roanoke
constitution	2062	state
establishment	1140	creation
formation	1764	creation
initiation	133	rite
founding	396	creation
foundation	1697	institution
origination	6	coorigination
creation	1180	establishment
colonization	228	colony
settlement	2649	institution

Table 2: Results of Harvard Experiment

as a function of how many different words it modifies. See (Grefenstette 1992a) for details.

Several Examples

Example 1

As described in the preceeding section, we processed 3.7 megabytes of text for the 'Harvard' example, performing morphological analysis, dictionary look-up, rough grammatical disambiguation, division into noun and verb phrases, parsing, and extraction of lexical attributes for each noun in the corpus.

Many of the terms in the institution synsets were found to have reasonable associations using these rough techniques. Table 2 below shows, for each word listed in WordNet as an immediate hyponym of 'institution', the word whose lexico-syntactic context was most similar to it among all the 22,000 unique words examined. Terms associated with words with low frequency (such as 'orphanage' and 'asylum') tend to be less plausible than higher frequency words.

As for our original concern as to where to place 'Harvard' as an 'institution', SEXTANT finds that 'Harvard' is used most similarly to 'Yale, Cambridge, Columbia, Chicago, Oxford, and Juilliard'. For example, 'Harvard' and 'Yale' are found to be similar because they both modify or are both modified by the following collection of words:

```
school study-SUBJ school-MOD
university-MOD law-MOD faculty-MOD
college-MOD graduate-MOD
review-MOD divinity-MOD
```

The fact that SEXTANT places 'Harvard' as being most similar to other university names, but not the term 'university' itself, points out one difficulty with our task. This particular portion of WordNet does not contain listings of specific instances of university names (whereas the unabbreviated version of the network beneath 'institution' does contain names of specific religious groups). If it had, then we could reasonably assign 'Harvard' to that same subtree. The difficulty lies in where to place 'Harvard' in the absence of knowledge of the terms it is found closest to. If we ask SEXTANT to compare 'Harvard' only to the words which were used to generate the corpus, we find that 'Harvard' is closest to 'academy'. 'Academy' is in the correct subtree, but now we must ask, should 'Harvard' be placed as a child of 'academy', on the same level as 'academy' or somewhere else in the subtree? To complicate matters 'academy' reappears as a child of the synset 'school, educational institution'. We do not yet know how to answer this placement problem.

Supposing that there did exist a subtree containing 'Yale, Cambridge, Columbia, ...', we would most probably find 'Harvard' already there. However, this ability to place a new term near its closest neighbors might be useful in non-static domains in which new terms are introduced over time (e.g., in the medical domain). If a knowledge structure exists for the domain, new text will produce new terms that must be integrated into this structure. This experiment provides an indication that such integration may be automated using existing lexicons, grammars, and text corpora that cover the relevant phenomena well.

Example 2

As another example, the acquisition algorithm discovered the lexical relation between 'rice' and 'cereal'. WordNet has two senses of 'cereal': one as a grain, and one as a breakfast product. We extracted 260,000 characters of text involving the following strings:

```
frumenty kasha grits hominy
grits farina millet
oatmeal hasty pudding mush
burgoo flummery gruel loblolly
porridge rice oat cornmeal meal
corn wheat buckwheat barley
cold cereal puffed wheat
puffed rice wheatflakes cornflakes
granola
```

In Table 3, we see that 'rice' is found to be closest to 'wheat', and vice versa. This table also shows the

word	*freq*	*closest terms*
wheat	234	rice, corn
rice	217	wheat, corn
corn	182	rice, wheat
crop	179	wheat, cultivation, rice, grain, production
meal	129	corn grain food, export, product, rice, hay,
product	102	export, production, industry, cattle
grain	101	export, growing, plant, cereal, producer
production	93	cultivation, growing
area	90	region, farm, land
food	89	export, grain, crop
plant	84	variety grain, production, seed
center	79	distribution, production, product, export, farmer
cultivation	71	growing, production,
farming	65	field, farm, export
	...	
millet	59	francois, sorghum, barley
barley	46	potatoe, cotton, rye
oat	44	barley, vegetable, potatoe, cotton
cereal	41	producer, starch, flour,
buckwheat	10	bread, sugarcane, pineapple, sorgum
grit	10	wing, farina, estrilidae
loblolly	8	grade drought, pine
gruel	3	second, conglomerate
mush	3	cornmeal
oatmeal	3	quaker, gascony
porridge	3	symbol, ale, roll
cornmeal	1	counterpart, sports-writer, literature
farina	1	embayment, peel
pudding	1	NO RELATIONS

Table 3: Results of Cereal Experiment

results involving the breakfast product terms (some did not occur in the corpus); note that the frequency of these terms is too low for valid assessments to be made. For example, the cold cereal terms that were unambiguous, such as 'wheatflakes', 'cornflakes', and 'granola', as well as all of the hot cereal items are underrepresented in the corpus. This fact reduces the possibility that 'rice' could be found similar to the breakfast product terms, although the fact that it is strongly related to the 'wheat' sense does lend some validity to the result.

In summary, this example highlights another difficulty; underrepresentation of data in the corpus can make the position assignment technique unapplicable.

Example 3

A third relation that we considered is *hyponym(logarithm, calculation)*. WordNet records 'logarithm' as a kind of 'exponent' which in turn is a kind of 'mathematical notation'. However, the sense of logarithm as a calculation should also be considered. The WordNet structure for 'calculation' is:

```
calculation, computation, figuring, reckoning
       => extrapolation
       => interpolation
       => estimate, estimation
            => guess, guesswork, shot, dead
               reckoning
            => approximation
       => integral
            => indefinite integral
            => definite integral

calculation, computation
       => mathematical process, operation
            => differentiation
            => division
            => integration
            => multiplication
            => subtraction
            => summation, addition
            => exponentiation, involution
```

Although there is no common one-word term for computing a logarithm, there should be an entry for 'logarithm' alongside 'exponentiation' in the second subtree (or perhaps 'taking a logarithm'). Our results found 'logarithm' to be closest to 'extrapolation' and 'multiplication', one term from each subtree, and thus eluded correct classification. (It isn't clear, however, that the first subtree is entirely well-defined. Why is 'integral' associated with figuring, as opposed to listing 'integration' in the second subtree?)

This example shows that difficulties arise when the shades of differences among the terms are subtle, fine-grained, or somewhat arbitrary. The original goal was

to make coarser-level distinctions, but because WordNet is so well-developed, it turns out that many decisions will require choosing between finely divided subtrees.

Other Examples

The previous three examples have shown various difficulties associated with this approach. Foremost among them is the fact that WordNet senses have been manually created, and correspond to a human conception of what attributes of two words or concepts are saliently similar, whereas SEXTANT finds similarity based on frequency of usage in a specific corpus. It may well be that two concepts considered to be similar in the WordNet hierarchy do not display the kinds of regular contextual similarities that our methods can recognize in a particular corpus.

There are other difficulties as well. For example, in one instance we found *hyponym(granite,rock)*. In WordNet there are very fine differences among the senses of 'rock' and in fact 'granite' appears in more than one of its subtrees. Other difficulties were: the hypernym senses have no child subtrees to compare against, the hypernym is a very general term and thus has hundreds of children, and the hyponym does not appear frequently enough in the corpus to take statistics on.

Suggestions for Improvements

Although none of the examples here yield perfect results, the implications, especially in the 'Harvard' case, are promising. In this section we discuss some paths for improvement.

Improving the Similarity Calculation

There are several drawbacks to the statistical similarity calculation as currently formulated. These drawbacks derive from a number of limitations in the SEXTANT system, the principal drawback begin that SEXTANT was designed as a word-based system. Here we would like to compare an unknown word to a group of known senses. This task mismatch leads to three main drawbacks, which we elucidate below, in hopes of pointing toward a more successful approach.

(1) In the test runs reported here, only pairwise comparisons were made; thus we compared the contexts of 'harvard' to the contexts of the individual terms in each subtree of 'institution', instead of grouping the terms in each subtree and comparing against the group as a whole. This might be remedied by taking each word in a subtree and replacing it with a dummy word that represents the subtree as a whole. This solution is however complicated by the fact that the same string can appear in different subtrees as seen above with 'foundation', 'establishment', and 'academy'. The fact that these words appear in different subtrees means, of course, that different nuances of meanings are associated with the word in different contexts. When SEXTANT encounters a string it has no way of knowing with which nuance it is dealing.

(2) As in any comparison of objects using their attributes, confidence in the results varies with the degree of similarity found between the objects. A relative comparison is not reliable if the similarity score between any pair of terms is not high enough. For this reason, although we can compare the similarity score between, say, 'school' and 'harvard' with the score between 'church' and 'harvard' ('school' is more similar), to see which is higher, we have no automatic way of determining a threshold below which relative comparisons become meaningless. Thus, if 'school' is found to be mildly more similar to 'harvard' there is no way to make a decision about how confident we can be.

(3) Using syntactic context (as opposed to strict lexical context) for similarity comparison may not be the most appropriate context for the discovered hypernym placement task. One reason using solely local syntactic context may be inappropriate is that proper nouns tend to be modified differently than common nouns, e.g., authors don't apply scalar adjectives such as 'large' to proper nouns like 'harvard' (at least not when it acts as the head of the noun phrase).

To get around these difficulties we would like to employ a statistical disambiguation algorithm that allows comparison among terms and doesn't rely on modificational syntactic context. (Yarowsky 1992) describes a very promising lexical disambiguation algorithm that determines the similarity of words to categories of an on-line version of Roget's Thesaurus. Roget categories can be though of as indicating word senses, e.g., the word 'crane' can be seen to fall into either the *TOOL* or the *ANIMAL* Roget category. We would like to try the technique that Yarowsky describes, using WordNet subtrees instead of Roget categories. There are three main differences between the techniques: first, Yarowsky's method uses a larger surrounding context than that which we were considering ($\pm$ 50 words around the problematic term, as opposed to only the syntactic relations in which it enters), second, it uses lexical context only, without the need of syntactic information, and finally, it uses a Bayesian framework for combining the evidence for each sense category.

Improving the Problem Statement

As seen in the examples above, in many cases WordNet's hierarchical distinctions are so fine that their arrangement is not necessarily incontrovertible. We are considering reworking the structure of the network in order to make it more amenable to use in coarse-level techniques, perhaps by "collapsing" subtrees with several levels into one.

word	*freq*	*closest terms*
state	2281	government, university, constitution
art	1849	architecture, music, education
system	1803	institution, education, government
education	1647	program, institution, student
group	1163	institution, organization, society,
power	1075	authority, government, control
world	1027	european, american, country,
society	919	institution, education, culture, government
faith	835	religion, religious, tradition
member	810	group, year, student
life	804	history, development, education, society
right	788	freedom, power, authority
movement	744	group, education, development, society
court	730	authority, state, law, constitution
council	668	organization, association, movement, leader
service	611	program, education, system
people	609	society, student, child, education
practice	591	history, tradition, religion, study
tradition	587	architecture, art, history, culture
company	581	association, center, bank, government
president	517	member, year, legislature, director
control	503	power, authority, government
theory	489	idea, thought, view, history
museum	378	institute, art, theater
union	373	association, party, organization
level	316	education, teacher, student
professor	180	institute, department, faculty, director

Table 4: Terms Similar to Terms "once-removed" from 'institution'

WordNet as a Corpus Partitioner

Although neither of the language processing techniques described in this paper require knowledge bases for their operation, we have found that the knowledge implicit in WordNet plays an important role in the experiments described here.

We assumed that the WordNet synsets could be used for disambiguation since each sense of a polysemous word is associated with its own distinguishing hypernyms and hyponyms. This assumption allows us to extract a corpus that is coherent with respect to a semantic subdomain. For the 'Harvard' example discussed earlier, we extracted sentences that contained the terms found in the subtrees beneath 'institution'. We showed which terms from the corpus were found to be semantically closest to these institution terms. However, many words that are not found in the subtrees of 'institution' nevertheless occur frequently in these sentences. Furthermore, these terms tend to be highly ambiguous.

In Table 4, we show the results of performing our similarity analysis on these terms "once-removed." From this table, it appears that the institutional sense of these polysemous words has been identified, e.g., *member-group, right-freedom, service-program,* and *union-association*. In other words, it seems that by taking a WordNet synset as a filter through a large text, we have extracted a rather coherent corpus. In this case, the corpus is one with a bent towards the notion of institution. If we were to use a corpus associated with terms beneath the 'sports' synset, for example, we would expect different associations for 'service', 'movement', and 'court'.

This observation helps substantiate the claim that we can use existing knowledge structures to help coarse-level analysis techniques – in this case the thesaurus helps find a semantic partition on unrestricted textual data.

Conclusion

In this paper we've described an attempt to combine three different text analysis tools: a hand-built knowledge source, a statistical, syntactic similarity measurement, and a pattern-based relation extractor. The lexical relations are used to augment the lexical hierarchy, and the similarity measure is meant to determine which part of the hierarchy the relation belongs in, while simultaneously the lexical hierarchy selects which part of the corpus to feed to the similarity measure. Our results are preliminary; in order to improve them we need both to restructure the hierarchy and adjust the similarity measure to suit our needs more closely.

Acknowledgements. Hearst's research was sponsored in part by the University of California and Digital Equipment Corporation under Digital's flagship research project Sequoia 2000: Large Capacity Object Servers to Support Global Change Research, and in part by an internship at Xerox Palo Alto Research Center.

References

Brent, M. R. (1991). Automatic acquisition of subcategorization frames from untagged, free-text corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics.*

Calzolari, N. & R. Bindi (1990). Acquisition of lexical information from a large textual italian corpus. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, Helsinki.

Church, K. & P. Hanks (1990). Word association norms, mutual information, and lexicography. *American Journal of Computational Linguistics*, 16(1):22–29.

Evans, D. A., S. K. Henderson, R. G. Lefferts, & I. A. Monarch (1991). A summary of the CLARIT project. Technical Report CMU-LCL-91-2, Laboratory for Computational Linguistics, Carnegie-Mellon University.

Grefenstette, G. (1992a). SEXTANT: extracting semantics from raw text implementation details. Technical Report CS92-05, University of Pittsburgh, Computer Science Dept.

Grefenstette, G. (1992b). Use of syntactic context to produce term association lists for text retrieval. In *Proceedings of SIGIR '92*, Copenhagen, Denmark.

Grolier (1990). *Academic American Encyclopedia.* Grolier Electronic Publishing, Danbury, Connecticut.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, pages 539–545, Nantes, France.

Hindle, D. (1990). Noun classification from predicate-argument structures. *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 268–275.

Kelly, E. & P. Stone (1975). *Computer recognition of english word senses*, volume 13 of *North-Holland Linguistics Series.* North-Holland, Amsterdam.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, & K. J. Miller (1990). Introduction to WordNet: An on-line lexical database. *Journal of Lexicography*, 3(4):235–244.

Romesburg, H. C. (1984). *Cluster Analysis for Researchers.* Lifetime Learning Publications, Belmont, CA.

Smadja, F. A. & K. R. McKeown (1990). Automatically extracting and representing collocations for language generation. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 252–259.

Wilks, Y., L. Guthrie, J. Guthrie, & J. Cowie (1992). Combining weak methods in large-scale text processing. In P. S. Jacobs, editor, *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*, pages 35–58. Lawrence Erlbaum Associates.

Yarowsky, D. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, pages 454–460, Nantes, France.