To Appear in *WordNet: An Electronic Lexical Database and Some of its Applications*, Christiane Fellbaum (Ed.), MIT Press.

Automated Discovery of WordNet Relations

Marti A. Hearst

1 Introduction

The WordNet lexical database is now quite large and offers broad coverage of general lexical relations in English. As is evident in this volume, WordNet has been employed as a resource for many applications in natural language processing (NLP) and information retrieval (IR). However, many potentially useful lexical relations are currently missing from WordNet. Some of these relations, while useful for NLP and IR applications, are not necessarily appropriate for a general, domain-independent lexical database. For example, WordNet's coverage of proper nouns is rather sparse, but proper nouns are often very important in application tasks.

The standard way lexicographers find new relations is to look through huge lists of concordance lines. However, culling through long lists of concordance lines can be a rather daunting task (Church and Hanks, 1990), so a method that picks out those lines that are very likely to hold relations of interest should be an improvement over more traditional techniques.

This chapter describes a method for the automatic discovery of WordNet-style lexico-semantic relations by searching for corresponding lexico-syntactic patterns in large text collections. Large text corpora are now widely available, and can be viewed as vast resources from which to mine lexical, syntactic, and semantic information. This idea is reminiscent of what is known as "data mining" in the artificial intelligence literature (Fayyad and Uthurusamy, 1996), however, in this case the ore is raw text rather than tables of numerical data. The Lexico-Syntactic Pattern Extraction (LSPE) method is meant to be useful as an automated or semi-automated aid for lexicographers and builders of domain-dependent knowledge-bases.

The LSPE technique is light-weight; it does not require a knowledge base or complex interpretation modules in order to suggest new WordNet relations.

Instead, promising lexical relations are plucked out of large text collections in their original form. LSPE has the advantage of not requiring the use of detailed inference procedures. However, the results are not comprehensive; that is, not all missing relations will be found. Rather, suggestions can only be made based on what the text collection has to offer in the appropriate form.

Recent work in the detection of semantically related nouns via, for example, shared argument structures (Hindle, 1990), and shared dictionary definition context (Wilks et al., 1990) attempts to infer relationships among lexical items by determining which terms are related using statistical measures over large text collections. LSPE has a similar goal but uses a quite different approach, since only one instance of a relation need be encountered in a text in order to suggest its viability. It is hoped that this algorithm and its extensions, by supplying explicit semantic relation information, can be used in conjunction with algorithms that detect statistical regularities to add a useful technique to the lexicon development toolbox.

It should also be noted that LSPE is of interest not only for its potential as an aid for lexicographers and knowledge-base builders, but also for what it implies about the structure of English and related languages; namely, that certain lexico-syntactic patterns unambiguously indicate certain semantic relations.

The following section describes the lexico-syntactic patterns and the LSPE acquisition technique. This is followed by an investigation of the results obtained by applying this algorithm to newspaper text, and an analysis of how these results map on to the WordNet noun network. This is followed by a discussion of related work and a chapter summary.

2 The Acquisition Algorithm

Surprisingly useful information can be found by using only very simple analyses techniques on unrestricted text. Consider the following sentence, taken from *Grolier's American Academic Encyclopedia* (Grolier, 1990):

(S1) `Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use.`

Most readers who have never before encountered the term *"Gelidium"* will nevertheless from this sentence infer that *"Gelidium"* is a kind of *"red algae"*. This is true even if the reader has only a fuzzy conception of what red algae is. Note that the author of the sentence is not deliberately defining the term, as would a dictionary or a children's book containing a didactic sentence like *Gelidium is a kind of red algae.* However, the semantics of the lexico-syntactic construction indicated by the pattern:

(1a) NP_0 *such as* NP_1 $\{, NP_2$ *... , (and | or)* $NP_i\}$ $\quad i \geq 1$

are such that they imply

(1b) *for all* NP_i, $i \geq 1$, *hyponym*(NP_i, NP_0)

Thus from sentence (S1) we conclude

hyponym("Gelidium","red algae").

This chapter assumes the standard WordNet definition of *hyponymy*; that is: a concept represented by a lexical item L_0 is said to be a hyponym of the concept represented by a lexical item L_1 if native speakers of English accept sentences constructed from the frame *An* L_0 *is a (kind of)* L_1. Here L_1 is the *hypernym* of L_0 and the relationship is transitive.

Example (1a-b) illustrates a simple way to uncover a hyponymic lexical relationship between two or more noun phrases in a naturally-occurring text. This approach is similar in spirit to the pattern-based interpretation techniques used in the processing of machine readable dictionaries, e.g., (Alshawi, 1987; Markowitz, Ahlswede, and Evens, 1986; Nakamura and Nagao, 1988). Thus the interpretation of sentence (S1) according to (1a-b) is an application of pattern-based relation recognition to general text.

There are many ways that the structure of a language can indicate the meanings of lexical items, but the difficulty lies in finding constructions that frequently and reliably indicate the relation of interest. It might seem that because free text is so varied in form and content (as compared with the somewhat regular structure of the dictionary), it may not be possible to find such constructions. However, there is a set of lexico-syntactic patterns, including the one shown in (1a) above, that indicate the hyponym relation and that satisfy the following desiderata:

(i) They occur frequently and in many text genres.

(ii) They (almost) always indicate the relation of interest.

(iii) They can be recognized with little or no pre-encoded knowledge.

Item (i) indicates that the pattern will result in the discovery of many instances of the relation, item (ii) that the information extracted will not be erroneous, and item (iii) that making use of the pattern does not require the tools that it is intended to help build.

2.1 Lexico-Syntactic Patterns for Hyponymy

This section illustrates the LSPE technique by considering the lexico-syntactic patterns suitable for the discovery of hyponym relations. Since only a subset of the possible instances of the hyponymy relation will appear in a particular form, we need to make use of as many patterns as possible. Below is a list of lexico-syntactic patterns that indicate the hyponym relation, followed by illustrative sentence fragments and the predicates that can be derived from them (detail about the environment surrounding the patterns is omitted for simplicity):

(2) *such NP as* {*NP* ,}* {*(or* | *and)*} *NP*

`... works by such authors as Herrick, Goldsmith, and Shakespeare.`

⟹ *hyponym("author", "Herrick"),*
hyponym("author", "Goldsmith"),
hyponym("author", "Shakespeare")

(3) *NP* {, *NP*}* {,} *or other NP*

`Bruises, ..., broken bones or other injuries ...`

⟹ *hyponym("bruise", "injury"),*
hyponym("broken bone", "injury")

(4) *NP* {, *NP*} * {,} *and other NP*

`... temples, treasuries, and other important civic buildings.`

$\Longrightarrow$ *hyponym("temple", "civic building"),*
hyponym("treasury", "civic building")

(5) *NP* {,} *including* {*NP* ,} * {*or* | *and*} *NP*

`All common-law countries, including Canada and England ...`

$\Longrightarrow$ *hyponym("Canada", "common-law country"), hyponym("England", "common-law country")*

(6) *NP* {,} *especially* {*NP* ,} * {*or* | *and*} *NP*

`... most European countries, especially France, England, and Spain.`

$\Longrightarrow$ *hyponym("France", "European country"),*
hyponym("England", "European country")
hyponym("Spain", "European country")

When a relation *hyponym*(NP_0, NP_1) is discovered, aside from some lemmatization and removal of unwanted modifiers, the noun phrase is left as an atomic unit, not broken down and analyzed. If a more detailed interpretation is desired, the results can be passed on to a more intelligent or specialized language analysis component.

2.2 Discovery of New Patterns

Patterns (1) – (3) were discovered by hand, by looking through text and noticing the patterns and the relationships indicated. However, to make this approach more extensible, a pattern discovery procedure is needed. Such a procedure is sketched below:

1. Decide on a lexical relation, R, that is of interest, e.g. meronymy.

2. Derive a list of word pairs from WordNet in which this relation is known to hold, e.g., *"house-porch"*.

3. Extract sentences from a large text corpus in which these terms both occur, and record the lexical and syntactic context.

4. Find the commonalities among these contexts and hypothesize that common ones yield patterns that indicate the relation of interest.

This procedure was tried out by hand using just one pair of terms at a time. In the first case indicators of the hyponomy relation were found by looking up sentences containing *"England"* and *"country"*. With just this pair found new patterns (4) and (5), as well as patterns (1) – (3) which were already known. Next, trying *"tank-vehicle"* lead to the discovery of a very productive pattern, pattern (6). (Note that for this pattern, even though it has an emphatic element, this does not affect the fact that the relation indicated is hyponymic).

Initial attempts at applying Steps 1-3, by hand, to the meronymy relation did not identify unambiguous patterns. However, an automatic version of this algorithm has not yet been implemented and so the potential strength of Step 4 is still untested. There are several ways Step 4 might be implemented. One candidate is a filtering method like that of Manning (1993), to find those patterns that are most likely to unambiguously indicate the relation of interest, or the transformation-based approach of Brill (1995) may also be appropriate. Alternatively, a set of positive and negative training examples, derived from the collection resulting from Steps 1-3, could be fed to a machine-learning categorization algorithm, such as C4.5 (Quinlan, 1986),

On the other hand, it may be the case that in English only the hyponym relation is especially amenable to this kind of analysis, perhaps due to its "naming" nature, but this question remains open.

2.3 Parsing Issues

For detection of local lexico-syntactic patterns, only a partial parse is necessary. This work makes use of a regular-expression-based noun phrase recognizer (Kupiec, 1993), which builds on the output of a part-of-speech tagger (Cutting et al., 1991).[1] Initially, a concordance tool based on the Text DataBase (TDB) system (Cutting, Pedersen, and Halvorsen, 1991) is used to extract sentences that contain the lexical items in the pattern of interest, e.g., *"such as"* or *"or other"*. Next, the noun phrase recognizer is run over the resulting sentences, and their positions with respect to the lexical items in the pattern are noted. For example, for pattern (3), the noun phrases that directly precede *"or other"* are recorded as candidate hyponyms, and the noun phrase following these lexical items is the potential hypernym.[2] Thus, it is not necessary to parse the entire sentence; instead just enough local context is examined to ensure that the nouns in the pattern are isolated, although some parsing errors do occur.

Delimiters such as commas are important for making boundary determinations. One boundary problem that arises involves determining the referent of a prepositional phrase. In the majority of cases the final noun in a prepositional phrase that precedes *"such as"* is the hypernym of the relation. However, there are a fair number of exceptions, as can be seen by comparing (S2a) and (S2b):

(S2a) `The component parts of flat-surfaced furniture, such as chests and tables, ...`

(S2b) `A bearing is a structure that supports a rotating part of a machine, such as a shaft, axle, spindle, or wheel.`

In (S2a) the nouns in the hyponym positions modify *"flat-surfaced furniture"*, the final noun of the prepositional phrase, while in (S2b) they modify *"a rotating part"*. So it is not always correct to assume that the noun directly preceding *"such as"* is the full hypernym if it is preceded by a preposition. It would be useful to perform analyses to determine modification tendencies in this situation, but a less analysis-intensive approach is to simply discard sentences in which an ambiguity is possible.

Pattern type (4) requires the full noun phrase corresponding to the hypernym *"other important civic buildings"*. This illustrates a difficulty that arises from using free text as the data source, as opposed to a dictionary –

[1] All code described in this chapter is written in Common Lisp and run on Sun SparcStations.

[2] In earlier work (Hearst, 1992) a more general constituent analyzer was used.

often the form that a noun phrase occurs in is not the form which should be recorded. For example, nouns frequently occur in their plural form but should be recorded in their singular form (although not always – for example, the algorithm finds that *"cards"* is a kind of *"game"* – a relation omitted from WordNet 1.5, although *"card game"* is present.). Adjectival quantifiers such as *"other"* and *"some"* are usually undesirable and can be eliminated in most cases without making the statement of the hyponym relation erroneous. Comparatives such as *"important"* and *"smaller"* are usually best removed, since their meaning is relative and dependent on the context in which they appear.

The amount of modification desired depends on the application for which the lexical relations will be used. For building up a basic, general-domain thesaurus, single-word nouns and very common compounds are most appropriate. For a more specialized domain, more modified terms have their place. For example, noun phrases in the medical domain often have several layers of modification which should be preserved in a taxonomy of medical terms.

3 Some Results

In an earlier discussion of this acquisition method (Hearst, 1992), hyponymy relations between simple noun phrases were extracted from *Grolier's Encyclopedia* (Grolier, 1990) and compared to the contents of an early version of WordNet (Version 1.1). The acquisition method found many useful relations that had not yet been entered into the network (they have all since been added). Relations derived from encyclopedia text tend to be somewhat prototypical in nature, and should in general correspond well to the kind of information that lexicographers would expect to enter into the network. To further explore the behavior of the acquisition method, this section examines results of applying the algorithm to six months worth of text from *The New York Times*.

When comparing a result *hyponym*(N_0,N_1) to the contents of WordNet's noun database, three kinds of outcomes are possible:

- Both N_0 and N_1 are in WordNet, and the relation *hyponym*(N_0,N_1) is already in the database (possibly through transitive closure).

- Both N_0 and N_1 are in WordNet, and the relation *hyponym*(N_0,N_1) is *not* (even through transitive closure); a new hyponym link is suggested.

- One or both of N_0 and N_1 are not present; these noun phrases and the corresponding hyponym relation are suggested.

As an example of the second outcome, consider the following sentence and derived relation, automatically extracted from *The New York Times*:

(S3) `Felonies such as stabbings and shootings, ...`
$\Longrightarrow$ *hyponym("shootings", "felonies")*,
hyponym("stabbings", "felonies")

The text indicates that a shooting is a kind of felony. Figure 1 shows the portion of the hyponymy relation in WordNet's noun hierarchy (Version 1.5) that includes the synsets *felony* and *shooting*. In the current version, despite the fact that there are links between *kidnapping, burglary*, etc., and *felony*, there is no link between *shooting, murder*, etc., and *felony* or any other part of the crime portion of the network. Thus the acquisition algorithm suggests the addition of a potentially useful link, as indicated by the dotted line. This suggestion may in turn suggest still more changes to the network, since it may be necessary to create a different sense of *shooting* to distinguish "shooting a can" or "hunting" (which are not necessarily crimes) from "shooting people".

Not surprisingly, the relations found in newspaper text tend to be less taxonomic or prototypical than those found in encyclopedia text. For example, the relation *hyponym("milo", "crop")* was extracted from the *New York Times*. WordNet classifies *"milo"* as a kind of *sorghum* or grain, but grain is not entered as a kind of crop. The only hyponyms of the appropriate sense of *crop* in WordNet 1.5 are *catch crop, cover crop,* and *root crop*. One could argue that hyponyms of *crop* should only be refinements on the notion of crop, rather than lists of types of crops. But in many somewhat similar cases, WordNet does indeed list specific instances of a concept as well as refinements on that concept. For example, hyponyms of *book* include *trade book* and *reference book*, which can be seen as refinements of *book*, as well as *Utopia*, which is a specific book (by Sir Thomas More).

By extracting information directly from text, relations like *hyponym("milo", "crop")* are at least brought up for scrutiny. It should be easier for a lexicographer to take note of such relations if they are represented explicitly rather than trying to spot them by sifting through huge lists of concordance lines.

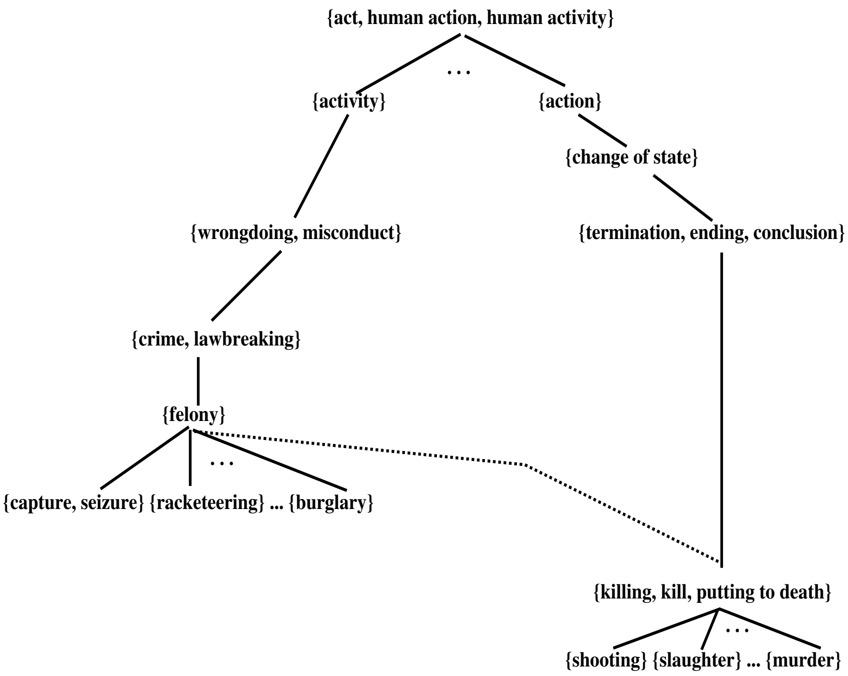

Figure 1: A portion of WordNet Version 1.5 with a new link (dotted line) suggested by an automatically acquired relation: *hyponym("shooting", "felony")*. Many hyponym links are omitted in order to simplify the diagram.

Often the import of relations found in newspaper text is more strongly influenced by the context in which they appear than those found in encyclopedia text. Furthermore, they tend more often to reflect subjective judgments, opinions, or metaphorical usages than the more established statements that appear in the encyclopedia. For example, the asserting that the movie *"Gaslight"* is a *"classic"* might be considered a value judgment (although encyclopedias state that certain actors are stars, so perhaps this isn't so different), and the statement that *"AIDS"* is a *"disaster"* might be considered more a metaphorical statement than a taxonomic one.

Tables 1 and 2 show the results of extracting 50 consecutive hyponym relations from six months worth of *The New York Times* using pattern (1a), the *"such as"* pattern. The first group in Table 1 are those for which the noun phrases and relation are already present in WordNet. The second group cor-

Description	**Hypernym**	**Hyponym(s)**
Relation and terms already appear in WordNet 1.5	fabric	silk
	grain	barley
	disorders	epilepsy
	businesses	nightclub
	crimes	kidnappings
	countries	Brazil India Israel
	vegetables	broccoli
	games	checkers
	regions	Texas
	assets	stocks
	jurisdictions	Illinois
Terms appear in Wordnet 1.5, relations do not	crops	milo
	wildlife	deer raccoons
	conditions	epilepsy
	conveniences	showers microwaves
	perishables	fruit
	agents	bacteria viruses
	felonies	shootings stabbings
	euphemisms	restrictees detainees
	goods	shoes
	officials	stewards
	geniuses	Einstein Newton
	gifts	liquor
	disasters	AIDS
	materials	glass ceramics
	partner	Nippon
Relation and term does not appear in WordNet 1.5 (proper noun)	companies	Volvo Saab
	institutions	Tufts
	airlines	Pan USAir
	agencies	Clic Zoli
	companies	Shell

Table 1: Examples of useful relations suggested by the automatic acquisition method, derived from *The New York Times*.

Description	**Hypernym**	**Hyponym(s)**
Does not appear in WordNet 1.5		
but is perhaps too general	things	exercise
	topics	nutrition
	things	conservation
	things	popcorn peanuts
	areas	Sacramento
Context-specific relations,		
so probably not of interest	Others	Meadowbrook
	facilities	Peachtree
	categories	drama miniseries comedy
	classics	Gaslight
	generics	Richland
Misleading relations resulting		
from parsing errors	tendencies	aspirin anticoagulants
(usually not detecting	competence	Nunn
the full NP)	organization	Bissinger
	children	Headstart
	titles	Batman
	companies	sports
	agencies	Vienna
	jobs	computer
	projects	universities

Table 2: Examples of less felicitous relations also derived from *The New York Times*.

responds to the situation in which the noun phrases are present in WordNet but the hyponymy relation between them is absent. The third group shows examples in which at least one noun phrase is absent, and so the corresponding relation is necessarily absent as well. In this example these relations all involve proper noun hyponyms.

Some interesting relations are suggested. For example, the only hyponyms of *euphemism* in Wordnet 1.5 are *blank, darn*, and *heck* (euphemisms for curse words). *Detainee* also exists in WordNet, as a hyponym of *prisoner, captive* which in turn is a hyponym of *unfortunate, unfortunate person*. If nothing else, this discovered relation points out that the coverage of euphemisms in WordNet 1.5 is still rather sparse, and also suggests another category of euphemism, namely, government designated terms that act as such. The final decision on whether or not to classify *"detainee"* in this way rests with the lexicographers.

Another interesting example is the suggestion of the link between *"Einstein"* and *"genius"*. Both terms exist in the network (see Figure 2), and *"Einstein"* is recorded as a hyponym of *physicist* which in turn is a hyponym of *scientist* and then *intellectual*. One of the hypernyms of *genius* is also *intellectual*. Thus the lexicographers have made *genius* and *scientist* siblings rather than specifying one to be a hyponym of the other. This makes sense, since not all scientists are geniuses (although whether all scientists are intellectuals is perhaps open to debate as well). The Figure also indicates that *philosopher* is also a child of *intellectual*, and individual philosophers appear as hyponyms of *philosopher*. Hence it does not seem unreasonable to propose a link between the *genius* synset and particular intellectuals so known.

Table 2 illustrates some of the less useful discovered relations. The first group lists relations that are probably too general to be useful. For example, various senses of *"exercise"* are classified as *act* and *event* in Wordnet 1.5, but *"exercise"* is described as a *"thing"* in the newspaper text. Although an action can be considered to be a thing, the ontology of WordNet assumes they have separate originating synsets. A similar argument applies to the *"conservation"* example.

The next group in Table 2 refers to context-specific relations. For example, there most likely is a facility known as the *"Peachtree facility"*, but this is important only to a very particular domain. Similarly, the relationship between *"others"* and *"Meadowbrook"* most likely makes sense when the reference of "others" is resolved, but not out of context. On the other hand, the

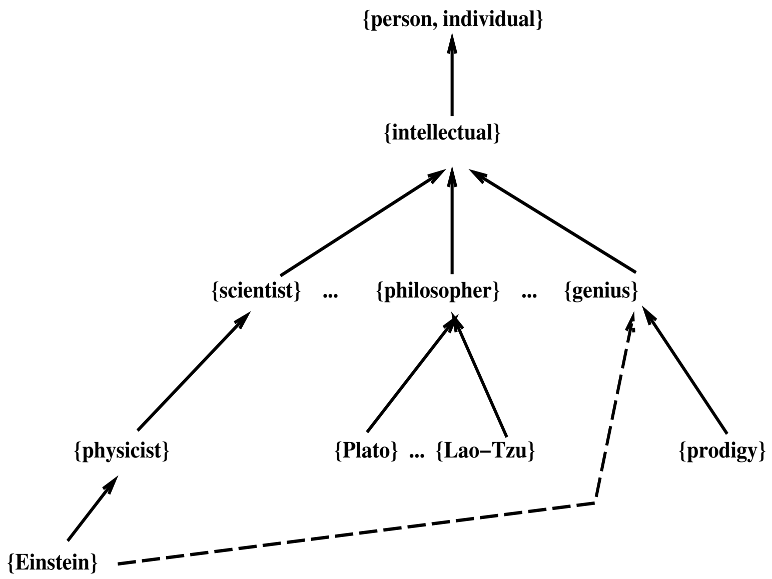

Figure 2: A portion of WordNet Version 1.5 with a new link (dotted line) suggested by an automatically acquired relation: *hyponym("Einstein", "genius")*. Many hyponym links are omitted in order to simplify the diagram.

hyponym("Gaslight", "classic") relation is inappropriate in this exact form, it may well be suggesting a useful omitted relation, *"classic films"*. Of course, the main problem with such a relation is the subjective and time-sensitive nature of its membership.

Most of the terms in WordNet's noun database are unmodified nouns or nouns with a single modifier. For this reason, the analysis presented here only extracts relations consisting of unmodified nouns in both the hypernym and hyponym roles (although determiners are allowed and a very small set of quantifier adjectives: *"some", "many", "certain", and "other"*). This restriction is also useful because, as touched on above, the procedure for determining which modifiers are important is not straightforward. Furthermore, for the purposes of evaluation, in most cases it is easier to judge the correctness of the classification of unmodified nouns than modified ones.

Although not illustrated in this example, the algorithm also produces many suggestions of multi-word term relations, e.g., *hyponym("data base search", "disk-intensive operation")*. To further elucidate the performance

Frequency	Explanation
38	Some version of the NPs and the corresponding relation were found in WordNet
31	The relation did not appear in WordNet and was judged to be a very good relation (in some cases both NPs were present, in some cases not)
35	The relation did not appear in WordNet and was judged to be at least a pretty good relation (in some cases both NPs were present, in some cases not)
19	The relation was too general
8	The relation was too subjective, or contained unresolved or inappropriate referents (e.g., "these")
34	The NPs involved were too long, too specific and/or too context-specific
12	The relations were repeats of cases counted above
22	The sentences did not contain the appropriate syntactic form (e.g., "all of the above, none of the above, or other")

Table 3: Results from 200 sentences containing the terms *"or other"*.

of the algorithm, 200 consecutive instances of pattern (3), the *"or other"* pattern, were extracted and evaluated by hand. Sentences that simply contained the two words *"or other"* (or *"or others"* were extracted initially, regardless of syntactic context. The most specific form of the noun phrase was looked up first; if it did not occur in WordNet, then the leftmost word in the phrase was removed and the phrase that remained was then looked up, and the process was repeated until only one word was left in the phrase. In each case, the words in the phrase was first looked up as is, and then reduced to their root form using the morphological analysis routines bundled with WordNet.

The results were evaluated in detail by hand, and placed into one of eight categories, as shown in Table 3. The judgments of whether the relations were "very good" or "pretty good" are meant to approximate the judgment that would be made by a WordNet lexicographer about whether or not to place the relation into WordNet. Of course this evaluation is very subjective,

Hypernym	**Hyponym**
fish	wildlife
birth_defect*	health_problem
money_laundering*	crime
strobe*	light
food	aid
grandchild	heir
deduction	write-off
name	ID
takeover	reorganization
shrimp	shellfish
canker_sore*	lesion

Table 4: Examples of good relations found using pattern (3) that do not appear in Wordnet 1.5. An asterisk indicates that the noun phrase does not appear in WordNet 1.5.

so an attempt was made to err on the side of conservativeness. Using this conservative evaluation, 104 out of the 166 elible sentences (those which had the correct syntax and were not repeats of already listed relations), or 63%, were either already present or strong candidates for inclusion in WordNet.

As seen in the examples from *New York Times* text, many of the suggested relations are more encyclopedic, and less obviously valid as lexico-semantic relations. Yet as WordNet continues to grow, the lexicographers may choose to include such items.

In summary, these results suggest that automatically extracted relations can be of use in augmenting WordNet. As mentioned above, the algorithm has the drawback of not guaranteeing complete coverage of the parts of the lexicon that require repair, but it can be argued that some repair is better than none. When using newspaper text, which tends to be more unruly than well-groomed encyclopedia text, a fair number of uninteresting relations are suggested, but if a lexicographer or knowledge engineer makes the final decision about inclusion, the results should be quite helpful for winnowing out missing relations.

4 Related Work

There has been extensive work on the use of partial parsing for various tasks in language analysis. For example, Kupiec (1993) extracts noun phrases from encyclopedia texts in order to answer closed-class questions, and Jacobs and Rau (1990) use partial parsing to extract domain-depending knowledge from newswire text. In this section, the discussion of related work will focus on efforts to automatically extract lexical relation information, rather than general knowledge.

4.1 Hand-coded and Knowledge-Intensive Approaches

There have been several knowledge-intensive approaches to automated lexical acquisition. Hobbs (1984), when describing the procedure his group followed in order to build a lexicon/knowledge base for an NLP analyzer of a medical text, noted that much of the work was done by looking at the relationships implied by the linguistic presuppositions in the target texts. One of his examples,

> "[the phrase] 'renal dialysis units and other high-risk institutional settings' tells us that a renal dialysis unit is a high-risk setting" (Hobbs, 1984)

is a version of pattern (4) described above. This analysis was done by hand; it might have been aided or accelerated by an automatic analysis like that described here.

Coates-Stephens (1991; 1992) describes FUNES, a system that acquires semantic descriptions of proper nouns using detailed frame roles, a sophisticated parser, and a domain-dependent knowlege base/lexicon. FUNES attempts to fill in frame roles, (e.g., name, age, origin, position, works-for) by processing newswire text. This system is similar to the work described here in that it recognizes some features of the context in which the proper noun occurs in order to identify some relevant semantic attributes. For instance, Coates-Stephens mentions that *"known as"* can explicitly introduce meanings for terms, as can appositives. However, this is one small component in a complex, knowledge-intensive system.

Two more examples of acquisition techniques that make use of extensive domain knowledge are those of Velardi and Pazienza (1989), who use hand-coded selection restriction and conceptual relation rules in order to assign case roles to lexical items, and Jacobs and Zernik (1988), who use extensive domain knowledge to fill in missing category information for unknown words.

4.2 Automatic Acquisition from Machine-Readable Dictionaries

Researchers have attempted several approaches to acquisition of lexical and syntactic information from machine readable dictionaries. As mentioned above, dictionaries are extremely rich resources for lexico-semantic relations, but are inherently limited in their scope.

Much of the dictionary extraction work, e.g., (Boguraev et al., 1987), focuses on acquisition of part of speech and syntactic information such as subcategorization frames for verbs. Several of these projects also involve extracting lexical relations, the type of which differs with each project. The two main approaches to extraction are (i) using patterns tailored to dictionary definitions and (ii) syntactically parsing the definitions.

Several research groups use patterns to acquire lexical relation information from machine readable dictionaries. Alshawi (1987), in interpreting LDOCE definitions, uses a hierarchy of patterns which consist mainly of part-of-speech indicators and wildcard characters. Markowitz, Ahlswede, and Evens (1986), and Chodorow, Byrd, and Heidorn (1985) created a taxonomic hierarchy based on information extracted from patterns, as do Wilks et al. (1990) and Guthrie et al. (1990). Nakamura and Nagao (1988) also use patterns on LDOCE to extract relations such as taxonomy, meronymy, action, state, degree, and form.

Ahlswede and Evens (1988) compared an approach based on parsing with one based on pattern recognition for interpreting definitions from Webster's 7th. The pattern matcher was more accurate and much faster than the parser, although the authors speculated that if they had been extracting more complex relations the parser would probably have produced the better results. Montemagni and Vanderwende (1992), however, demonstrate why structural information is crucial to successful extraction of semantic relations such as location, color, and purpose. Jensen and Binot (1987) and Ravin (1990) explored the extraction of detailed semantic information using careful analysis of the possible semantic and syntactic configurations that appear in dictionary definitions. These analyses are somewhat brittle because they require many of the words in the definitions to have already been unambiguously identified. Vanderwende (1995) improves this procedure by bootstrapping with unambiguous words, and iterating through the contents of the dictionary, each time disambiguating new words based on those identified in the previous iteration and some straightforward similarity information.

4.3 Automatic Acquisition from Corpora

There is a growing body of work on acquisition of semantic information from unrestricted text. In early work Church and Hanks (1990) used frequency of co-occurrences of content words to create clusters of semantically similar words and Hindle (1990) used both simple syntactic frames and frequency of occurrence of content words to determine similarity among nouns. For example, the nouns most similar to *"legislator"* in a 6 million word sample of AP newswire text were found to be *"Senate"*, *"committee"*, *"organization"*, *"commission"*, *"legislature"*, *"delegate"* and *"lawmaker"*. As can be seen from this example, these nouns represent a wide variety of relations to the target word *"legislator"*, including meronymy, synonymy, and general relatedness. Grefenstette (1994) takes this approach a bit further by using shallow syntactic analysis on local text contexts to determine semantic relatedness information.

More recent examples of algorithms that derive lexical co-occurrence information from large text collections include the work of Schütze (1993) and Resnik (1993). In Word Space (Schütze, 1993), statistics are collected about the contexts in which words co-occur and the results are placed in a term-by-term co-occurrence matrix which is reduced using a variant of multidimensional scaling. The resulting matrix can be used to make inferences about the closeness of words in a multidimensional semantic space. Hearst and Schütze (1996) show how the distributional association information of Word Space can be combined with word similarity information from WordNet to classify frequently-occurring proper nouns. A different approach is taken by Resnik (1993; 1995), who develops an information-theoretic model of word similarity based on frequency of occurrence in a corpus combined with the structural information available in WordNet.

Although the focus in this chapter has been on automatic acquisition of lexico-semantic information, it is appropriate to mention as well some recent work on automatically deriving syntactic association information. For example, the acquisition algorithm of Smadja and McKeown (1990; 1993) uses statistical techniques to derive well-formed collocation information from large text collections. Calzolari and Bindi (1990) use corpus-based statistical association ratios to determine lexical relations such as prepositional complementation and modification. More recent work by Basili, Pazienza, and Velardi (1992) uses a shallow syntactic analyzer to find binary and ternary syntactically-defined collocations (e.g., subject-verb, noun-preposition-noun, verb-adverb).

Brent (1990; 1993) describes methods for finding verb subcategorization frames by searching for simple syntactic patterns across large collections. The patterns all reflect well-known linguistic phenomena, e.g., in English there is a class of verbs that can take an infinitive argument, so try to find instances the verb of interest followed by the pattern *to INF-VERB*. Brent employs statistical filtering techniques in order to reduce the already small error rates. The fact that only unambiguous distinguishers are allowed to supply positive information ensures that this method is very accurage; however, its conservativeness inherently limits its scope and coverage. For example, it cannot discover all kinds of verb frames.

Manning (1993) describes an algorithm that is able to find a much larger dictionary of subcategorization frames than Brent's algorithms by filtering the results statistically, rather than requiring that every relation detected be unambiguously correct. Manning's algorithm makes use of a finite state parser run on the output of a stochastic part-of-speech tagger.

Like Brent's approach, LSPE is able to distinguish clear pieces of evidence from ambiguous ones. Unlike Brent's approach, however, it is at least potentially extensible, using the procedure for discovery of new patterns described above, and perhaps culling out ambiguous results using a statistical filtering pattern like that suggested by Manning.

5 Summary

This chapter has described LSPE, a low-cost approach for augmenting the structure and contents of WordNet. LSPE uses lexico-syntactic patterns to automatically extract lexico-semantic relations from unrestricted text collections. Since LSPE requires only a single specially expressed instance of a relation, it is complementary to those methods that infer semantic relations based on word co-occurrence or other statistical measures. However, there is a trade-off between the simplicity of the knowledge-free text interpretation and the sparseness of coverage that it offers.

The LSPE approach also suggests a linguistic insight: some semantic relations in English are indicated unambiguously by simple lexico-syntactic patterns. This idea merits further exploration, in at least two directions. First, are other relations besides hyponymy unambiguously indicated by lexico-syntactic patterns in English, or is the IS-A relation a special case? And second, do other languages exhibit similar behavior for hyponymy or other lexical relations? The answers to these questions may lead to further advances in automated lexical acquisition.

5.0.1 Acknowledgments

I would like to thank Geoff Nunberg for very helpful comments on the analysis of the results of the algorithm, Julian Kupiec for getting the NP recognition software into a form which I could use, Jan Pedersen for helpful comments on the paper, Robert Wilensky for earlier support of this work, and Christiane Fellbaum for unflagging enthusiasm, very helpful suggestions for the paper, and very gentle deadline reminders.

References

Ahlswede, Thomas and Martha Evens. 1988. Parsing vs. text processing in the analysis of dictionary definitions. *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 217–224.

Alshawi, Hiyan. 1987. Processing dictionary definitions with phrasal pattern hierarchies. *American Journal of Computational Linguistics*, 13(3):195–202.

Basili, Roberto, Maria Pazienza, and Paola Velardi. 1992. A shallow syntactic analyser to extract word associations from corpora. *Literary and Linguistic Computing*, 7(2):113–123.

Boguraev, Bran, Ted Briscoe, John Carroll, David Carter, and Claire Grover. 1987. The derivation of a grammatically indexed lexicon from ldoce. *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pages 193–200.

Brent, Michael. 1990. Semantic classification of verbs from their syntactic contexts: Automated lexicography with implications for child language acquisition. In *Proceedings of The 12th Annual Conference of the Cognitive Science Society*, pages 428–437.

Brent, Michael R. 1993. From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Lingustics*, 19(2):243–262.

Brill, Eric. 1995. Unsupervised learning of disambiguation rules for part of speech tagging. In David Yarowsky and Kenneth Church, editors, *Proceedings of the Third Workshop on Very Large Corpora*, Cambridge,MA.

Calzolari, Nicoletta and Remo Bindi. 1990. Acquisition of lexical information from a large textual italian corpus. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, Helsinki.

Chodorow, Martin S., Roy Byrd, and George Heidorn. 1985. Extracting semantic hierarchies from a large on-line dictionary. *Proceedings of the 23th Annual Meeting of the Association for Computational Linguistics*, pages 299–304.

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

Coates-Stephens, Sam. 1991. Coping with lexical inadequacy – the automatic acquisition of proper nouns from news text. In *The Proceedings of the 7th Annual Conference of the UW Centre for the New OED and Text Research: Using Corpora*, pages 154–169, Oxford.

Coates-Stephens, Sam. 1992. The analysis and acquisition proper names for the understanding of free text. *Computers and the Humanities*, 5-6:441–456, December.

Cutting, Douglass R., Julian Kupiec, Jan O. Pedersen, and Penelope Sibun. 1991. A practical part-of-speech tagger. In *The 3rd Conference on Applied Natural Language Processing*, Trento, Italy.

Cutting, Douglass R., Jan O. Pedersen, and Per-Kristian Halvorsen. 1991. An object-oriented architecture for text retrieval. In *Conference Proceedings of RIAO'91, Intelligent Text and Image Handling, Barcelona, Spain*, pages 285–298, April. Also available as Xerox PARC technical report SSL-90-83.

Fayyad, Usama M. and Ramasamy Uthurusamy, editors. 1996. *The First International Conference on Knowledge Discover and Data Mining*. AAAI Press, August.

Grefenstette, Gregory. 1994. *Explorations in automatic thesaurus discovery*. Kluwer international series in engineering and computer science. Kluwer Academic Publishers.

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

Guthrie, Louise, Brian M. Slator, Yorick Wilks, and Rebecca Bruce. 1990. Is there content in empty heads? In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, Helsinki.

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

Hearst, Marti A. and Hinrich Schütze. 1996. Customizing a lexicon to better suit a computational task. In Branimir Boguraev and James Pustejovsky, editors, *Corpus Processing for Lexical Acquisition*. MIT Press, Cambridge, MA.

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

Hobbs, Jerry R. 1984. Sublanguage and knowledge. In Ralph Grishman and Richard Kittredge, editors, *Proceedings of the Workshop on Sublanguage Description and Processing*, New York University, January.

Jacobs, Paul and Lisa Rau. 1990. SCISOR: Extracting information from On-Line News. *Communications of the ACM*, 33(11):88–97.

Jacobs, Paul and Uri Zernik. 1988. Acquiring lexical knowledge from text: A case study. In *Proceedings of AAAI88*, pages 739–744.

Jensen, Karen and Jean-Louis Binot. 1987. Disambiguating prepositional phrase attachments by using on-line dictionary definitions. *American Journal of Computational Linguistics*, 13(3):251–260.

Kupiec, Julian. 1993. MURAX: A robust linguistic approach for question answering using an on-line encyclopedia. In *Proceedings of the 16th Annual International ACM/SIGIR Conference*, pages 181–190, Pittsburgh, PA.

Manning, Christopher D. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual*

Meeting of the Association for Computational Lingusitics, pages 235–242, Columbus, OH.

Markowitz, Judith, Thomas Ahlswede, and Martha Evens. 1986. Semantically significant patterns in dictionary definitions. *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, pages 112–119.

Montemagni, Simonetta and Lucy Vanderwende. 1992. Sturctural patterns vs. string patterns for extracting semantic information from dictionaries. In *Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING)*, pages 546–552, Nantes, France, July.

Nakamura, Junichi and Makoto Nagao. 1988. Extraction of semantic information from an ordinary english dictionary and its evaluation. In *Proceedings of the Twelfth International Conference on Computational Linguistics*, pages 459–464, Budapest.

Quinlan, J. R. 1986. Induction of decision trees. *Machine Learning*, 1:81–106. Reprinted in Shavlik and Dietterich (eds.) *Readings in Machine Learning*.

Ravin, Yael. 1990. Disambiguating and interpreting verb definitions. *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 260–267.

Resnik, Philip. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania, December. (Institute for Research in Cognitive Science report IRCS-93-42).

Resnik, Philip. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-95)*, volume 1, pages 448–453, Montreal, Canada.

Schütze, Hinrich. 1993. Word space. In Stephen J. Hanson, Jack D. Cowan, and C. Lee Giles, editors, *Advances in Neural Information Processing Systems 5*. Morgan Kaufmann, San Mateo CA.

Smadja, Frank A. 1993. Retrieving collocations from text: Xtract˙ *Computational Linguistics*, 19(1):143–177.

Smadja, Frank A. and Kathleen 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.

Vanderwende, Lucy. 1995. Ambiguity in the acquisition of lexical information. In *Working Notes of the 1995 AAAI Spring Symposium on Representation and Acquisition of Lexical Knowledge*, pages 174–179.

Velardi, Paola and Maria Teresa Pazienza. 1989. Computer aided interpretation of lexical cooccurrences. *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 185–192.

Wilks, Yorick A., Dan C. Fass, Cheng ming Guo, James E. McDonald, Tony Plate, and Brian M. Slator. 1990. Providing machine tractable dictionary tools. *Journal of Computers and Translation*, 2.