Automatically deriving structured knowledge bases from on-line dictionaries[1]

William Dolan
billdol@microsoft.com

Lucy Vanderwende
lucyv@microsoft.com

Stephen D. Richardson
steveri@microsoft.com

Microsoft Corp.
One Microsoft Way
Redmond, WA 98052-6399
(206) 882-8080

Abstract

keywords: computational lexicography; lexical knowledge bases

We describe an automated strategy which exploits on-line dictionaries to construct a richly-structured lexical knowledge base. In particular, we show how the Longman Dictionary of Contemporary English (LDOCE) can be used to build a directed graph which captures semantic associations between words. The result is a huge and highly interconnected network of words linked by arcs labeled with semantic relations such as *Hypernym*, *Part_of*, *Location*, and *Purpose*. We argue that this knowledge base provides much more detailed information about word meanings than can be obtained using standard lexical lookup procedures or by relying on statistical measures of semantic associations among words.

1. Introduction

This paper describes an automated strategy which exploits on-line dictionaries to automatically construct a richly-structured lexical knowledge base. This effort is part of our long-term aim of producing a natural language understanding and generation system for English, mapping out a development methodology which will allow the extraction of information from dictionaries of other languages.

In order to achieve true broad-coverage, detailed syntactic and semantic information is required for tens (and eventually hundreds) of thousands of words, including those which are infrequent, technical, informal, slang, and so on. One current attempt to build a highly detailed semantic resource for natural language processing (NLP) is Dahlgren (1988). Similarly, the ongoing Cyc project (Lenat and Guha, 1989) has sometimes been positioned as a lexical resource for NLP systems. Wordnet (Miller et al., 1990), which has been receiving increasing attention within the NLP community, is a lexical database which attempts to capture a psychologically real model of the semantic interconnections within a significant portion of the vocabulary of English. What these resources have in common is that they were created by hand, a fact which we regard as a significant problem for our goal of achieving broad-coverage, multi-language NLP. Constructing the necessary lexical databases by hand represents a hugely expensive and time-consuming task. Even Wordnet, with its tens of thousands of lexical entries, is incapable of supporting true broad-coverage NLP.

Much more promising, in our view, are automated methods for building large computational lexicons. For instance, statistical methods can be used to acquire information about the syntactic and semantic properties of words from large corpora (e.g., Basili et al., 1992; Grefenstette, 1992). Other work attempts to extract semantic information from a (partial)

[1]We would like thank the other members of the Microsoft Natural Language group: Joseph Pentheroudakis, Karen Jensen, George Heidorn, and Diana Peterson.

analysis of free text (Grishman and Sterling, 1992; Hearst, 1992). Currently, however, none of these techniques appears capable of providing lexical information in sufficient detail.

Another area of current interest is the use of on-line versions of existing dictionaries for NLP tasks, exploiting definition text and example sentences. Some dictionaries are huge, describing hundreds of thousands of words, and new editions are continually being produced to keep up with new coinage, changes in usage, and so on.

The goal of this paper is to show that standard dictionaries implicitly define a highly connected network of words linked through basic semantic relations such as *Hypernym*, *Part_of*, and *Location*. We show that this network can be made explicit by analyzing the definitions to obtain the semantic relations and then displaying them as a directed graph. The resulting graph provides much more detail about word meanings than has been exploited by previous work involving the use of on-line dictionaries for NLP.

We begin by reviewing some of the work which has been done on extracting semantic information from on-line dictionaries, and by describing our own approach to this problem. We then respond to the possible objection that dictionaries are inadequate sources of lexical and world knowledge for a broad-coverage NLP system, and we show how representing the dictionary as a directed graph substantially reduces the scale of this problem. Finally, we present examples of how our graph-structured dictionary can be exploited by an understanding component to perform complex inferences which would not be possible using an ordinary computational lexicon or dictionary.

2. Semantic Information from Dictionaries

The earliest effort to exploit machine readable dictionaries for NLP applications was Amsler (1980), which explored the possibility of constructing taxonomies using computational methods. Although most of the ideas represented in this work were not actually implemented, this dissertation anticipated many of the issues which continue to confront researchers in computational lexicography. Chodorow et al. (1985) relied on string-matching to automatically extract genus terms for nouns and verbs from the on-line version of Webster's Seventh New Collegiate Dictionary (Webster 7). Markowitz et al. (1986) expanded on this general approach by attempting to discover "defining formulae" or "significant recurring patterns" in the text of definitions—that is, syntactic or lexical patterns which appear to have been used in a consistent way by lexicographers to express a specific semantic relationship. In addition, Calzolari (1984, 1988) used string matching procedures in order to extract both genus and differentiae information from the text of dictionary definitions.

More recently, semantic information has been extracted from on-line dictionaries in a two-step procedure, first parsing the dictionary text (the definition and/or example sentences), and then using this syntactic information to improve the accuracy of pattern identification. The first work of this kind was Jensen and Binot (1987), which involved parsing dictionary definitions using the PLNLP Grammar and then searching the resulting parse trees for combinations of syntactic and lexical features which could be reliably associated with semantic relationships like *Part_of* and *Instrument* as well as *genus* terms. Jensen and Binot show how the results of this extraction procedure can effectively help resolve the kinds of prepositional phrase attachment ambiguities encountered in free text. Related work includes Klavans et al. (1990), Ravin (1990), and Montemagni and Vanderwende (1992). Montemagni (1992), meanwhile, shows that this same general methodology can be used to acquire semantic information from on-line dictionaries of Italian.

An interesting aspect of the research program begun by Jensen and Binot is its claim that dictionary entries can be effectively analyzed by a parser designed for broad-coverage text analysis. In contrast, work such as Alshawi (1989) and Slator (in Wilks et al., 1989) has relied on specially-constructed parsers which exploit the idiosyncratic syntactic properties of LDOCE entries. The advantage of relying on a broad coverage parser is that the parser need not be modified or rewritten in the course of extending the approach to other dictionaries. We see this as an important consideration, given that the huge semantic resources needed for broad-coverage NLP can only be acquired through the merging of multiple on-line dictionaries, as well as the analysis of encyclopedia and other sources.

2.1. Identifying Semantic Relations

Our approach to semantic processing builds on the work of Jensen and Binot (1987) and Montemagni and Vanderwende (1992). The first step involves parsing the definitions of LDOCE entries using our broad-coverage grammar of English. The resulting parse structures are then subjected to a set of heuristic rules whose goal is to identify the occurrence of syntactic and lexical patterns which are consistently associated with some specific semantic relation, such as *Instrument* or *Location*. Consider, for example, the text of the following two LDOCE definitions:

1. **authority** (*n*,7) "a person, book, etc., mentioned as the place where one found certain information"

2. **storehouse** (*n*,1) "a place or person full of information"

In each of these definitions, a *Location* relation holds between the headword (i.e. the word being defined, shown in boldface) and the word "information," despite the fact that this relation is expressed quite differently in each case. The patterns which make it possible to identify the underlying semantic unity in these superficially different definitions can be roughly paraphrased as follows:

- if there is a relative clause and the relativizer is in the set {*where*, *in which*, *on which*}, then there is a *Location* relation between the headword and the verb of the relative clause (along with any of its arguments)

- if the genus term is in the set {*place*, *area*, *space*, …} and there is a PP containing the preposition *of*, then there is a *Location* relation between the headword and the noun of the PP (along with any of its modifiers).

Applying these patterns to the parsed definitions of "authority" and "storehouse" will yield, in part, the fact that each is the *Location* of "information." We currently recognize approximately 25 relation types for verbs and nouns, including *Location*, *Part_of*, *Purpose*, *Hypernym*, *Time*, (*Typical_*)*subject*, (*Typical_*)*object*, and *Instrument*.

2.2. Using Semantic Relations

Jensen and Binot (1987) show how information automatically extracted from Webster 7 entries can be used to determine the proper attachment of ambiguous prepositional phrases. Consider (3), in which "with bones" might plausibly be attached to either the verb "eat" or to "fish":

3. I ate a fish with bones.

The relevant semantic information from Webster 7 that allows a heuristic to decide that "with bones" should be attached to "fish" is:

4.	**bone** (*n*,1)	*Part-of*	skeleton, vertebrate
	fish (*n*,1b)	*Is-a*	vertebrate

While this approach is frequently successful in resolving ambiguities of this kind, it is also subject to failure when no match can be found between the sets of semantic relations extracted from the two lexical entries under consideration. For instance, if we replace the word "fish" in this sentence with any of its hyponyms, such as "salmon," our heuristic will now fail to find any semantic association between this prepositional phrase and the noun it modifies:

5. I ate a salmon with bones.

6.	**bone** (*n*,1)	*part-of*	skeleton, vertebrate
	salmon (*n*,1a)	*is-a*	fish

Sometimes problems of this kind can be resolved by extending the search to include one or possibly two levels of hypernymy. For the sentence in (5) this will indeed solve the problem, since "salmon" has the *Hypernym* "fish," "fish" has the *Hypernym* "vertebrate," and, as we saw in (4), "bone" is known to be part of a "vertebrate." We can thus discover a complete path through the dictionary linking "bone" to "salmon." However, even if we allow such chains of *Hypernym* links to be exploited in processing, this approach sometimes fails to discover what might intuitively be an obvious semantic connection between two words. For example, in (7), no match can be found at any level for the headwords "research" and "chapter."

7. I researched the 19th century for this chapter.[2]

The prepositional phrase attachment in this sentence is unambiguous: any native speaker of English will infer first of all that writing/reading the chapter in question required researching the 19th century, and further, that the speaker is writing/reading a book. The apparent inability of a simple dictionary-based approach to provide evidence about whether the PP "for this chapter" should modify "researched" or "the 19th century" thus represents an area in which enhancements to the current techniques are necessary.

Some researchers have concluded that dictionaries are too impoverished a source of semantic information to ever serve as the lexical knowledge base for sophisticated semantic processing (e.g. Atkins, Kegl, and Levin, 1986). This pessimistic view seems to be supported by a casual examination of dictionary entries. Definitions frequently fail to express even basic facts about word meanings, facts which we would obviously want to include in a knowledge base which is to serve as the basis for understanding English. A typical case in LDOCE is the word "flower," whose primary sense is noteworthy more for the information it omits than for what it provides:

8. **flower** (*n*,1) "the part of a plant, often beautiful and coloured, that produces seeds or fruit"

Missing from this definition is any detailed description of the physical structure of flowers, information about what kinds of plants have flowers, and so on. Even the important fact that flowers prototypically have a pleasant scent goes unmentioned. We might, of course, try to increase our stock of information about this word's meaning by exploring the definitions of words used in *its* definition ("plant," "beautiful," etc.). In this case, however, such a strategy is not especially productive, yielding general information about plants but no specific details about flowers.

[2]A similar example is found in the American Heritage Dictionary, Third Edition, (AHD3) under the usage note for "research." "He spent a week at a funeral home researching mortuary procedures for his new novel" (p. 1534, AHD3).

3. The Dictionary as a Network

To a great extent, the apparent inadequacy of on-line dictionaries for semantic processing can be attributed to the way in which they have been used—what we might term the *forward-chaining* model of dictionary consultation. Given a dictionary in book form, the only way to find information about a given word involves looking it up, then exploring the semantic properties of any words mentioned in its definition, and so on. Once the data are available on-line, however, we exploit dictionary access strategies which involve not only *forward-chaining*, but also *backward-chaining*. That is, in looking up a word we might consult not just its own definition, but *also the definitions of any word which mentions it*. This possibility was noted, but not implemented, by Amsler (1980). This insight was also exploited by Chodorow et al. (1985) in developing a tool for helping human users disambiguate hyper/hyponym links among pairs of lexical items.

The important point is that much of the information about a given word's meaning is typically located not in the entry for that word itself, but rather in the entries for *other* words. For instance, it is relatively unusual to find the parts of some object described in the lexical entry for that object; instead, the relationship between the words for these parts and the larger object is defined only in the lexical entries describing the components themselves. Consider again the word "flower." Searching LDOCE for entries which mention "flower" in their definitions allows us to construct a highly detailed picture of this word's meaning. For instance, a number of entries (9a) describe components of flowers.

9a. **corolla** (*n*) "the part of a *flower* formed by the petals , usu. brightly coloured to attract insects"
petal (*n*) "any of the (usu. coloured) leaflike divisions of a *flower*"
style (*n*,7) "the rodlike part inside a *flower* which supports the stigma at the top"

We can also establish such facts as what time of year flowers bloom and are plentiful (9b), that they prototypically have a pleasant smell (9c), that bees collect nectar from them (9d), and so on. It is also possible to compile an exhaustive list of flowers and flowering plants, just two of which are given in (9e).

9b. **spring** (*n*, 2) "the season between winter and summer in which leaves and *flowers* appear"
summer (*n*, 1) "the season between spring and autumn when (…) there are many *flowers*"

9c. **fragrant** (*a*,1) "having a sweet or pleasant smell (esp. of *flowers*)"
sweet (*a*,5) "having a light pleasant smell, like many garden flowers"

9d. **nectar** (*n*,3) "the sweet liquid collected by bees from *flowers*"

9e. **aster** (*n*) "a garden *flower* with a bright yellow center"
alyssum (*n*) **"**a type of low-growing plant with yellow or white *flowers*"

On-line dictionaries thus represent formidable sources of "common sense" knowledge about the world. In order to exploit this information, however, we must tease out the network structure which is implicit in the text of dictionary definitions.

The idea that dictionaries define a huge interconnected network underlies work by Plate (in Wilks et al., 1989), who used co-occurrence patterns between words in LDOCE to produce a graphically-displayed statistical measure of semantic relatedness. Similarly, Veronis and Ide (1993) describe how statistical techniques can be used to transform a portion of the Collins English dictionary into a weighted neural network. A drawback to both of these approaches, however, is that the networks which they produce reveal only that two words are statistically correlated with one another in a dictionary; no information about the semantic *nature* of this relationship is available.

3.1. Implementing the Network

This section describes the steps involved in automatically constructing a highly-detailed semantic database from the on-line version of LDOCE. We then go on to illustrate how the resulting network can be exploited by inferencing strategies which will allow us to address certain NLP problems which appear to require "common-sense" world knowledge.

The first step in creating the lexical network involves parsing LDOCE definitions and then subjecting the resulting structure to the pattern-based search for semantic relations described above. The result for each entry is a set of one or more relations linking the headword and the words in its definition. For instance, processing the definition of "knowledge (*n*,3)" yields an attribute-value structure which (in part) associates this headword with the word "information" through the relation *Hypernym*. The definition of "inquire (*v*,2)" yields two relations, one a *Hypernym* link to "ask," and the other an *Is_for* link to "information."

10a. **knowledge** (*n*,3) "familiarity with; information about"
Hypernym information

11b. **inquire** (*v*,2) "to ask for information"
Hypernym ask
Purpose information

In many instances, complex syntactic relationships among words in a definition make it necessary to identify deeper levels of semantic relatedness between the headword and definition words. For example, recall the definition of "authority (*n*,7)" (1), repeated in (11), and compare it to the set of relations (12) identified for it by our system. In addition to recognizing that one *Hypernym* of "authority" is "person" (i.e., is [+human]) and that another is "book," we discover an attribute *Location* which has as its value the verb "find." This value is itself modified by a pair of relations: the *Typical_subject* (i.e., *agent*) of "finding" in (11) is "one" (i.e., [+human]), while the *Typical_object* (*patient*) of "finding" is "information."

12. **authority** (*n*,7) "a person, book, etc., mentioned as the place where one found certain information"

13. **authority** (*n*,7)
Hypernym person (= [+human])
Hypernym book
Location find
Typical_subject [+human]
Typical_object information

As each semantic relation is identified, it is added to the attribute-value structure of the sense entry for this definition in our on-line dictionary. In addition, each relation is added to the entry *for the word which is the value of the relation*. Thus, identifying the semantic relations in (11) does not simply result in new relations being added to the entry for "authority

(*n*, 7)," but also to the entries for the words "person," "book," "find," and "information." Once the entire dictionary has been processed in this way, each lexical entry contains a record of every word which mentions that word in its definition, along with a (potentially complex) description of the semantic relationship which holds between the two words.

Rounding numbers of entries to the nearest thousand, of the 75,000 definitions in LDOCE, we currently analyze the 33,000 single word noun definitions and the 12,000 single word verb definitions (45,000 definitions total) in a process that takes about 11 hours on our 486/66 PCs. (The exclusion of the 15,000 phrasal entries and other entries for adjectives and adverbs is temporary while we focus on refining the basic methods used by our system.) This procedure can be easily iterated each time significant change is made to the grammar rules or to the semantic feature extraction rules. The total number of top-level relations currently extracted is over 94,000. (In addition, several tens of thousands of secondary relations—like the two modifying "find" in (12)—are identified.).

We have hand-checked a random sample of 250 semantic relations across our dictionary and judged their overall accuracy to be 78%. Using common statistical techniques we estimate that this rate is representative of the entire dictionary (all 94,000 relations) with a margin of error of +/- 5%. About half of the relations in the sample were of the type *Hypernym*, and these relations were judged accurate 87% of the time. While this may be seen as inflating the overall accuracy rate, it is counteracted by the currently dismal accuracy of the *Part_of* relation (only 15%). Removing the *Part_of* numbers from the tallies raised the accuracy of relations other than *Hypernym* from 68% to 78%. Our immediate plan for improving the overall accuracy of the relations identified by our system involves computing *Hypernym* relations in a first pass through the dictionary, and then using that information to aid in the identification of other relations (such as *Part_of*) during a second pass. We also have specific plans to improve our parser and other aspects of our structural pattern matching.

The result of this processing is a significantly enlarged on-line version of LDOCE. In one sense, all that we have done is to copy and redistribute the original information throughout the set of entries. This seemingly simple change in format, however, provides us with important new ways to explore the information which is available. In particular, we can view the dictionary as a huge, directed graph whose nodes correspond to headwords and senses. These nodes are interconnected by arcs labeled with semantic relations. A small fragment of the graph surrounding the word "book" is shown in (13). (For the sake of legibility, we have omitted most of the 173 top-level relations linking LDOCE headwords to "book.")

(14) A small fragment of the LDOCE graph surrounding "book." Headwords point to one or more individual senses, which are, in turn, connected to other words, including "book," via arcs labeled with semantic relations. All arcs are directed from left to right.

4. Discussion

Hand-constructing lexicons and knowledge bases for NLP can be extraordinarily hard. While it may be relatively simple to make decisions about how to capture words representing concrete concepts—it is intuitively obvious, for instance, that "petal" is a part of a "flower," which in turn is a part of (some types of) "plant"—adequately capturing the meaning of more abstract words can be much more problematic, involving difficult and sometimes arbitrary decisions about what semantic properties of a concept might be relevant for NLP tasks. Frequently, representing some problematic concept or word can force wholesale changes in the ontology or in the set of semantic features which are assumed. All of this work is time-consuming and costly, demanding the skills of linguists or knowledge engineers.

A great strength of our approach, then, is that it requires no hand-coding. Each dictionary entry can be thought of as implicitly defining its own semantic frame of the kind familiar from traditional AI approaches (Minsky, 1975). Furthermore, the set of semantic relation-types which can associated with a given lexical item is extremely rich, rivaled only by hand-coded efforts like Wordnet. Our graph-structured lexicon is thus far richer in semantic information than any produced by previous efforts to derive taxonomies by parsing on-line dictionaries, which have focused almost exclusively on identifying hypo/hypernymy relationships between words.

Most importantly, we believe that our network captures exactly the semantic partitioning of the English word space which is relevant to NLP tasks. The structure of this network reflects lexicographers' intuitions about the relationship between the English lexicon and a model of the conceptual structure of the world. Each definition reflects the influence of a long

tradition of lexicographic work aimed at discovering how prose descriptions of English can be formulated to most economically distinguish English words from one another.

In effect, we view the process of identifying semantic relationships between a headword and the *genus* and *differentiae* in these prose descriptions a form of data compression—abstracting away from syntactic structure to arrive at a more concise description formulated in purely semantic terms. The resulting model of the lexicon is a Hjelmslev-style view in which a word's meaning is defined entirely by the set of structural oppositions it enters into with other words.

5. Inferencing over the Network

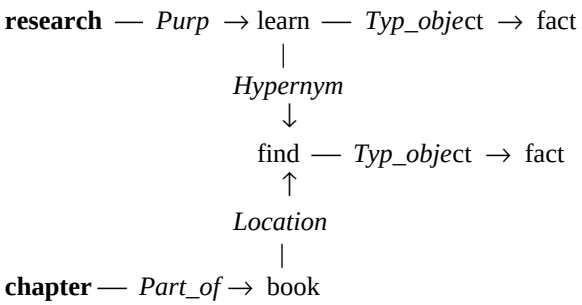
The real importance of our LDOCE-derived network stems from its use in resolving semantic ambiguities in text. In this section we describe a few of the problems which benefit from the inferencing capabilities which become possible once we have detailed information about *how* words are semantically related.

5.1. PP Attachment Ambiguities

A classic problem in NLP is the attachment of ambiguous prepositional phrases, which was already illustrated by (7), repeated as (14):

15. I researched the 19th century for this chapter.

As we noted earlier, it is impossible to discover any straightforward association between "research" and "chapter" if we pursue a standard dictionary exploration strategy. However, once we have a network linking the set of words in LDOCE, "research" and "chapter" *can* be linked, through "find" and (indirectly) "authority" (see (12)):



Furthermore, we can use the labels on these links to begin to draw inferences about ideas which are not overtly mentioned in the input sentence. For instance, given this chain of associations, we can infer that the purpose of this research was *to learn facts* for a *book*.

During processing, the discovery of a sequence of links connecting two words in the network is crucially driven by evidence from the syntax of the input sentence. In (14), for instance, the preposition "for" in the PP "for this chapter" will bias the system to prefer paths which link the verb "research" to "chapter" through a *Purpose* relation.

5.2. Anaphora

Sentences in a discourse are sometimes linked by anaphora, with pronouns and definite NP's referring back to some entity introduced earlier in the discourse. An example of this kind is (15), where the word "chapter" in the first sentence implicitly introduces into the discourse both the larger concept "book" and, by extension, the related entities "bibliography" and "proofreading."

16. "I researched the 19th century for this chapter. *The book* is coming along really well, although it needs to be *proofread* and *the bibliography* is too long."

Our lexical network makes it possible to discover the links between these sentences, providing an explanation for the otherwise mysterious use of definite articles preceding "book" and "bibliography" in the second. The reason is that the word "chapter" in the first sentence invokes an entire schema of lexical items associated with "books":

chapter — *Part_of* → book

bibliography — *Part_of* → book

proofread — *Typ_object* → book

We have had only enough space here to begin to suggest how our lexical network can be used to make sophisticated inferences about how to interpret sentences and discourse.

6. Remaining Challenges

A number of significant challenges remain before we can fully exploit our lexical network. One of the

most significant of these stems from the fact that LDOCE is a learner's dictionary. One aspect of the resulting emphasis on simplicity is its use of a highly restricted set of vocabulary items in its definitions. This has dramatic effects on the structure of the semantic network which emerges from the dictionary: only about 2500 words of the approximately 40,000 headwords in LDOCE appear as the value of any relation, and those which do are often connected to a huge number of other words. Once we interpret it as a network, then, LDOCE provides a tremendous level of detail on the semantic properties of this small set of words, but relatively little (direct) information about the other words in the dictionary.

This fact is peculiar to LDOCE and similar learner's dictionaries, however, and we expect the pattern of connections among lexical items in our network to become more evenly distributed as we merge other on-line dictionaries into the existing knowledge base.

A second practical problem with our network involves sense ambiguity. Although the structure of the network accurately reflects *which sense* of a word points to some other word in the network, the word pointed to may itself have multiple senses. For instance, the LDOCE definition of "flower, *n*1," which includes the noun "plant," does not explicitly state which of the 4 different senses of this noun is intended. In some cases this information *is* included, as in (16), where the first sense of *hipster* points to the fourth, "fashionable," entry for *hip* and the second sense points to the "anatomical" entry. Unfortunately, however, this level of explicitness is usually reserved for cases of homophony rather than simple sense ambiguity, and even here it is by no means marked consistently in the dictionary. (The superscripts in the following definitions are used in LDOCE to indicate the sense.)

17. **hipster** (*n)* 1. "a person who is hip[4]"
    2. "[usu. pl.] an article of clothing which fits up to the hips[2]"

The result of all this ambiguity is that the arcs in our network link a specific definition to an ambiguous lexical entry, rather than to some specific sense of this entry. This is often problematic—obviously, we do not want the word "flower" to be linked to the sense of "plant" referring to "a thing, esp. stolen goods…."

This problem is not quite as serious as it appears, however, since various heuristics can exploited to help disambiguate these links. For instance, in the bulk of cases involving a definition which contains a polysemous word, it is safe to assume that the relevant sense is the first one in the LDOCE entry (considered to be the most frequent). Bruce and Guthrie (1992) found that in 76% of such cases in a sample of 520 noun word senses could be correctly regarded as pointing to the first sense in a polysemous word's entry. Nevertheless, this heuristic will sometimes go awry: while all 15 of the words in our reconstructed dictionary which point to the word "cell" refer to its biological sense (e.g., "mitosis," "germ," "leucocyte") this sense is the fifth listed in the LDOCE entry for this word. The development of other heuristics to help automate the process of sense disambiguation is a current focus of attention.

7. Conclusions

This paper has described a strategy for constructing a rich source of lexical and common-sense information from on-line dictionaries. We feel that this knowledge base represents a potentially unique and substantial contribution to NLP. It contains richly structured semantic information for tens of thousands of words, representing one of the largest and most deeply processed lexical databases ever produced by automatic means.

As the coverage of both parser and semantic heuristics improve, so will the extent and quality of the semantic relations linking words in our network. Furthermore, given that our method exploits the power of a broad-coverage parser, we expect it to be readily extensible to other dictionaries of English, sources of information which can be merged into an existing graph. In the longer term, we see this methodology as pointing the way to the acquisition of information from free text.

8. References

Alshawi, H. 1989. Analysing the dictionary definitions. In *Computational lexicography for natural language processing*, ed. B.K. Boguraev and E.J. Briscoe, 153-170. London: Longman Group.

Amsler, R.A. 1980. The structure of the Merriam Webster Pocket Dictionary. Ph.D.diss., University of Texas, Austin.

Atkins, B. T., J. Kegl, and B. Levin. 1986. Explicit and Implicit Information in Dictionaries. *Proceedings of the Second Annual Conference of the University of Waterloo Centre for the Oxford English Dictionary: Advances in Lexicology*.

Basili, R., M.T. Pazienza, and P. Velardi. 1992. Combining NLP and statistical techniques for lexical acquisition. In *Proceedings of AAAI Fall Symposium Series: Probabilistic Approaches to Natural Language*, October 23-25, 1992, 1-9.

Bruce, R., and L. Guthrie. 1992. Genus disambiguation: a study in weighted preference. In *Proceedings of COLING92*, 1187-1191.

Calzolari, N. 1984. Detecting patterns in a lexical data base. In *Proceedings of COLING84*, 170-173.

-----. 1988. Acquisition of semantic information from an on-line dictionary. In *Proceedings of COLING88*, 87-92.

Chodorow, M.S., R.J. Byrd, G.E. Heidorn. 1985. Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd Annual Meeting of the ACL*, 299-304.

Dahlgren, K. 1988. *Naive Semantics for Natural Language Understanding*. Norwell, MA: Kluwer Academic Publishers.

Grefenstette, G. 1992. Finding Semantic Similarity in Raw Text: the Deese Antonyms. In *Proceedings of AAAI Fall Symposium Series: Probabilistic Approaches to Natural Language*, October 23-25, 1992, 61-65.

Grishman, R., and J. Sterling. 1992. Acquisition of selectional patterns. In *Proceedings of COLING92*, 658-664.

Hearst, M.A. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of COLING92*, 539-545.

Jensen, K., and J.-L. Binot. 1987. Disambiguating prepositional phrase attachments by using on-line dictionary definitions. *Computational Linguistics* 13, no.3-4, 251-260.

Klavans, J.L., M.S. Chodorow, and N. Wacholder. 1990. From dictionary to knowledge base via taxonomy. In *Proceedings of the 4th Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary: Electronic Text Research*, 110-132.

Lenat, D.B., and R.V. Guha. 1989. *Building large knowledge-based systems: representation and inference in the Cyc project*. Addison-Wesley Publishing Company, Inc.

Markowitz, J., T. Ahlswede, and M. Evans. 1986. Semantically significant patterns in dictionary definitions. In *Proceedings of the 24th Annual Meeting of the ACL*, 112-119.

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

Montemagni, S. 1992. Tailoring a broad-coverage grammar for the analysis of dictionary definitions. In *EURALEX '92: Papers submitted to the 5th International Euralex Congress on Lexicography*.

Montemagni, S., and L. Vanderwende. 1992. Structural patterns vs. string patterns for extracting semantic information from dictionaries. In *Proceedings of COLING92*, 546-552.

Ravin, Y. 1990. Disambiguating and interpreting verb definitions. In *Proceedings of the 28rd Annual Meeting of the ACL*, 260-267.

Veronis, J., and N. Ide. 1993. Large neural networks for the resolution of lexical ambiguity. Groupe Representation et Traitement des Connaissances/Centre National de lat Recherche Scientifique *Tech. Report* no. 593.

Wilks, Y., D. Fass, C. Guo, J. McDonald, T. Plate, and B. Slator. 1989. A tractable machine dictionary as a resource for computational semantics. In *Computational lexicography for natural language processing*, ed. B.K. Boguraev and E.J. Briscoe, 193-228. London: Longman Group.