An Application of Lexical Semantics to Knowledge Acquisition from Corpora

Peter Anick
Computer Science Department
Brandeis University
Waltham, MA 02254

James Pustejovsky
Computer Science Department
Brandeis University
Waltham, MA 02254

Abstract

In this paper, we describe a program of research designed to explore how a lexical semantic theory may be exploited for extracting information from corpora suitable for use in Information Retrieval applications. Unlike with purely statistical collocational analyses, the framework of a semantic theory allows the automatic construction of predictions about semantic relationships among words appearing in *collocational systems*. We illustrate the approach for the acquisition of lexical information for several classes of nominals.

Keywords: Knowledge Acquisition, Information Retrieval, Lexical Semantics.

1 Introduction

The proliferation of on-line textual information has intensified the search for efficient automated indexing and retrieval techniques. Full-text indexing, in which all the content words in a document are used as keywords, is one of the most promising of recent automated approaches, yet its mediocre precision and recall characteristics indicate that there is much room for improvement [Croft, 1989]. The use of domain knowledge can enhance the effectiveness of a full-text system by providing related terms that can be used to broaden, narrow, or refocus a query at retrieval time ([Thompson and Croft 1989], [Anick et al, 1989] [Debili et al, 1988]. Likewise, domain knowledge may be applied at indexing time to do word sense disambiguation [Krovetz & Croft, 1989] or content analysis [Jacobs, 1989]. Unfortunately, for many domains, such knowledge, even in the form of a thesaurus, is either not available or is incomplete with respect to the vocabulary of the texts indexed.

The tradition in both AI and Library Science has been to hand-craft domain knowledge, but the current availability of machine-readable dictionaries and large text corpora presents the possibility of deriving at least some domain knowledge via automated procedures [Amsler, 1980] [Maarek and Smadja, 1989] [Wilks et al, 1988]. The approach described in this paper outlines one such experiment.

We start with: (1) a lexicon containing morpho-syntactic information for approximately 20,000 common English words; (2) encodings of English morphological paradigms and a morphological analyzer capable of producing potential citation forms from inflected forms; (3) a bottom-up parser for recognizing sub-sentential phrasal constructions; and (4) a theory of lexical semantics embodying a collection of powerful semantic principles and their syntactic realizations.

The aim of our research is to discover what kinds of knowledge can be reliably acquired through the use of these methods, exploiting, as they do, general linguistic knowledge rather than domain knowledge. In this respect, our program is similar to Zernik's (1989) work on extracting verb semantics from corpora using lexical categories. Our research, however, differs in two respects: first, we employ a more expressive lexical semantics for encoding lexical knowledge; and secondly, our focus is on nominals, for both pragmatic and theoretical reasons. For full-text information retrieval, information about nominals is paramount, as most queries tend to be expressed as conjunctions of nouns. From our theoretical perspective, we believe that the contribution of the lexical semantics of nominals to the overall structure of the lexicon has been somewhat neglected (relative to that of verbs) [Pustejovsky and Anick, 1988], [Pustejovsky 1989]. Indeed, whereas Zernik (1989) presents metonymy as a potential obstacle to effective corpus analysis, we believe that the existence of motivated metonymic structures provides valuable clues for semantic analysis of nouns in a corpus.

Our current work attempts to acquire the following kinds of lexical information without domain knowledge:

- Part of speech and morphological paradigms for new words and new uses of old words;
- Bracketing of noun compounds;
- Subclass relations between nouns;
- Lexical semantic categorization of nouns;
- Clustering of verbs into semantic classes based on the collections of nouns they predicate.

While such information is still inadequate for natural language "understanding" systems, it vastly simplifies the task of knowledge engineering, should one desire to hand-code lexical items. Furthermore, such information can be put to use directly in full-text

information retrieval systems, fulfilling some of the roles typically played by thesauri and faceted classifications [Vickery, 1975].

2 A Framework for Lexical Semantics

The framework for lexical knowledge we will be assuming is that developed by Pustejovsky (1989), who proposes a theory of lexical semantics which explores the internal structure of lexical items from a computational perspective. In this theory, lexical and conceptual decomposition is performed *generatively*. That is, rather than assuming a fixed set of primitives, we assume a fixed set of rules of composition and generative devices. Thus, just as a formal language is described more in terms of the productions of the grammar rather than in terms of its accompanying vocabulary, a semantic language should be defined by the rules generating the structures for expressions, rather than the vocabulary of primitives itself. For this reason, a dictionary of lexical items and the concepts they derive can be viewed as a *generative lexicon*.[1]

Such a theory of lexical meaning specifies both a general methodology and a specific language for expressing the semantic content of lexical items in natural language. The aspect of this theory most relevant to our own concerns is a language for structuring the semantics of nominals. Pustejovsky (1989) calls this the *Qualia Structure* of a noun, which is essentially a structured representation similar to a verb's argument structure. This structure specifies four aspects of a noun's meaning: its constituent parts; its formal structure; its purpose and function (i.e. its Telic role); and how it comes about (i.e. its Agentive role). For example, *book* might be represented as containing the following information:

```
book(*x*,*y*)
    [Const:  information(*y*)]
    [Form: bound-pages(*x*) or disk(*x*)]
    [Telic: read(T,w,*y*)]
    [Agentive: artifact(*x*)
     & write(T,z,*y*)]
```

This permits us to use the same lexical representation in very different contexts, where the word seems to refer to different qualia of the noun's meaning. For example, the sentences in (2)-(3) refer to different aspects (or *qualia*) of the general meaning of *book*.

1: This book weighs four ounces.
2: John finished a book.
3: This is an interesting book.

Sentence (1) makes reference to the Formal role, while sentence (3) refers to the Constitutive role. Example (2), however, can refer to either the Telic or the Agentive aspects given above. The utility of such knowledge for information retrieval is readily apparent. This theory claims that noun meanings should make reference to related concepts and the relations into which they enter. The qualia structure, thus, can be viewed as a kind of generic template for structuring this knowledge.

To further illustrate how objects cluster according to these dimensions, we will briefly consider three object types: (1) containers (of information), e.g. *book*, *tape*, *record*; (2) instruments, e.g. *gun*, *hammer*, *paintbrush*; and (3) figure-ground objects, e.g. *door*, *room*, *fireplace*. Because of how their qualia structures differ, these classes appear in vastly different grammatical contexts.

As with containers in general, information containers permit metonymic extensions between the container and the material contained within it. Collocations such as those in (4) through (7) indicate that this metonymy is grammaticalized through specific and systematic head-PP constructions.

4: *read a book*
5: *read a story in a book*
6: *read a tape*
7: *read the information on the tape*

Instruments, on the other hand, display classic agent-instrument causative alternations, such as those in (8) through (11).

8: *... smash the vase with the hammer*
9: *The hammer smashed the vase.*
10: *... kill him with a gun*
11: *The gun killed him.*

Finally, figure-ground nominals permit perspective shifts such as those in (12) through (15). [2]

12: *John painted the door.*
13: *John walked through the door.*
14: *John is scrubbing the fireplace.*
15: *The smoke filled the fireplace.*

That is, *paint* and *scrub* are actions on physical objects while *walk through* and *fill* are processes in spaces. These collocational patterns, we argue, are systematically predictable from the lexical semantics of the noun, and we term such sets of collocated phrases *collocational systems*.[3]

To make this point clearer, let us consider a specific example of a collocational system. Because of the particular metonymy observed for a noun like *tape*, we will classify it as a 'container.' In terms of the semantic representation presented here, we can view it as a relational noun, with the following qualia structure:

```
tape(*x*,*y*)
  [Const:  information(*y*)]
  [Form:   phys-object(*x*)]
  [Telic:  hold(S,*x*,*y*)]
  [Agent:  artifact(*x*) & write(T,w,*y*)]
```

This simply states that any semantics for *tape* must logically make reference to the object itself (F),

[1] For elaboration on this idea and how it applies to various lexical classes, see Pustejovsky (forthcoming).

[2] See Pustejovsky and Anick (1988) for details.

[3] This relates to Mel'čuk's lexical functions and the syntactic structures they associate with an element. See Mel'čuk (1988) and references therein. Cruse (1986) discusses the foregrounding and backgrounding of information with respect to similar examples.

what it can contain (C), what purpose it serves (T), and how it arises (A). This provides us with a semantic representation which can capture the multiple perspectives which a single lexical item may assume in different contexts. Yet, the qualia for a lexical item such as *tape* are not isolated values for that one word, but are integrated into a global knowledge base indicating how these senses relate to other lexical items and their senses. This is the contribution of inheritance and the hierarchical structuring of knowledge (e.g. [Brachman and Schmolze 1985] and [Bobrow and Winograd 1977]). In Pustejovsky (1989), it is suggested that there are two types of relational structures for lexical knowledge; a *fixed* inheritance similar to that of an ISA hierarchy (cf. Touretsky (1986))[4]; and a dymamic structure which operates *generatively* from the qualia structure of a lexical item to create a relational structure for *ad hoc* categories.

Let us suppose then, that in addition to the fixed relational structures, our semantics allows us to dynamically create arbitrary concepts through the application of certain transformations to lexical meanings. For example, for any predicate, Q — e.g. the value of a qualia role — we can generate its opposition, $\neg Q$. By relating these two predicates temporally we can generate the arbitrary transition events for this opposition. Similarly, by operating over other qualia role values we can generate semantically related concepts. The set of transformations includes: $\neg$, negation, $\leq$, temporal precedence, $\geq$, temporal succession, $=$, temporal equivalence, and *act*, an operator adding agency to an argument.

Intuitively, the space of concepts traversed by the application of such operators will be related expressions in the neighborhood of the original lexical item. We will call this the *Projective Conclusion Space* of a specific quale for a lexical item.[5] To return to the example of *tape* above, the predicates *read* and *copy* are related to the Telic value by just such an operation. Predicates such as *mount* and *dismount*, however, are related to the Formal role since they refer to the tape as a physical object alone.

It is our view that the approach outlined above for representing lexical knowledge can be put to use in the service of information retrieval tasks. On the one hand, the projective conclusion space, with its structured assembly of terms, clusterd about a nominal entity, can serve as a "virtual script", capable of homonym disambiguation ([Krovetz 1990], [Cullingford and Pazzani 1984]) and query reformulation. On the other hand, the qualia structure captures the inherent polysemy of many nouns. In the latter respect, our proposal can be compared to attempts at object classification in information science. One approach, known as "faceted classification" (Vickery (1975)) proceeds roughly as follows. Collect terms lying within a field. Then, group the terms into facets by assigning them to categories. Typical examples of this are state, property, reaction, device. However, each subject area is likely to have its own sets of categories, making it difficult to re-use a set of facet classifications in another domain. [6]

Even if the relational information provided by the qualia structure and inheritance would improve performance in information retrieval tasks, one problem still remains; namely that it would be very time-consuming to hand-code such structures for all nouns in a domain. Since it is our belief that such representations are generic structures across all domains, our long term goal is to develop methods for how these relations and values can be automatically extracted from on-line corpora. In the section that follows, we describe one such experiment which indicates that the qualia structures do, in fact, correlate with collocational systems, thereby allowing us to perform structure-matching operations over corpora to find these relations.

3 A Knowledge Acquisition Procedure

In this section, we outline our procedure for knowledge acquisition, implemented as part of the LINKS Lexicon/Corpus Management System.[7] Steps are illustrated with examples drawn from an analysis done on a Digital Equipment Corporation on-line corpus of 3000 articles containing VMS troubleshooting information. Briefly, the procedure consists of the following steps.

1. **Assign morphological paradigms to words in the corpus.**
2. **Generate of a set of bracketed noun compounds.** e.g. [TK50 [tape drive]], [[database management] system].
3. **Collect Noun Phrases related by prepositions** from the collocational systems for the desired lexical items. e.g. "file on tape", "format of tape".
4. **Hypothesize subclass relationships** on the basis of collocational information: e.g. If X and Y are nouns and the phrase XY appears in the corpus, and there is no phrase Y *Prep* X, then $ISA(X, Y)$. For example: From [TK50 [tape drive]] we can predict that $ISA(TK50, \textit{tape drive})$. However, the potential prediction from "tape drive" that $ISA(\textit{tape}, \textit{drive})$ is blocked by the existence of phrases like "tape in drive".
5. **Seek distributional verification of subclass relationships.** For each subclass so generated, seek distributional evidence to support the hypothesis. That is, is there a "substantial" inter-

[4] Thesaurus-like structures are similar within the IR community, cf. [National Library and Information Associations Council 1980].

[5] See Pustejovsky (1989) for details.

[6] This is reflected in the sublanguage work of Grishman et al (1986), whose automated discovery procedures are aimed at clustering nouns into domain-specific categories like "body-part," "symptom," etc.

[7] This is a system currently under development at Digital Equipment Corporation.

section between verbs collocated with the subclass and superclass terms?

6. **Attempt semantic classification into a known lexical category.** Try to match the set of syntactic constructions within which X appears with one of our diagnostic construction sets. This may involve searching for the set of constructions that contain nouns in other argument positions of the original set of constructions. For example, the set of expressions involving the word "tape" in the context of its use as a secondary storage device suggests that it fits the *container artifact* schema of the qualia structure, with "information" and "file" as its containees:
   (a) *read information from tape*
   (b) *write file to tape*
   (c) *read information on tape*
   (d) *read tape*
   (e) *write tape*
7. **Use heuristics to cluster predicates that relate to the *Telic* quale of the noun.** For example, the word "tape" is the object of 34 verbs in our corpus:

```
(require use unload replace mount
 restore time request control
 position dismount allocate off
 initialize satisfy contain create
 encounter get allow try leave be
 load read write have cause protect
  up perform enforce copy)
```

   Among these verbs are some that refer to the formal quale: mount, dismount and some which refer to tape in its function as an information container: *read*, *write*, and *copy*.

   One of the ways to tease these sets apart is to take advantage of the linguistic rule that allows a container to be referred to in place of the containee, i.e. the container can be used metonymically. The verbs which have "information" (previously identified as a likely "containee" for tape) as an object in the corpus are:

```
(check include display enter compare
list find get extract set be write fit
contain read recreate update return
provide specify see open publish give
insert have copy take relay lose gather)
```

   When we intersect the verb sets for "information" and "tape", we get a set that reflects the predicates appropriate to the telic role of tape, a container of information (plus several empty verbs):

```
(copy have read contain write be get)
```

   Thus, the metonymy between container and containee allows us to use set intersection to discriminate among predicates referring to the telic vs. formal roles of the container.

What results from this acquisition procedure is a kind of *minimal faceted analysis* for the noun *tape*, as illustrated below.

```
tape(*x*,*y*)
 [Const:  information(*y*), file(*y*)]
 [Form:   mount(w,*x*), dismount(w,*x*)]
 [Telic:  read(T,z,*y*), write(T,z,*y*),
  copy(T,z,*y*)]
 [Agent:  artifact(*x*)]
```

To illustrate this procedure on another semantic category, consider the term "mouse" in its computer artifact sense. In our corpus, it appears in the object position of the verb "use" in a "use-to" construction, as well as the object of the preposition "with" following a transitive verb and its object:

(a) use the mouse to set breakpoints
(b) use the mouse anywhere
(c) move a window with the mouse
(d) click on it with the mouse ...

These constructions are symptomatic of its role as an instrument; and the VP complement of "to" as well as the VP dominating the "with" PP's identify the telic predicates for the noun. Other verbs, for which "mouse" appears as a direct object are currently defaulted into the formal role, resulting in an entry for "mouse" as follows:

```
mouse(*x*)
  [Cont: button(*x*)]
  [Form: move(w,*x*), click(w,*x*),
   hold(w,*x*)]
  [Telic: set(*x*,breakpoint),
   move(*x*,window),
   click-on(*x*,window)]
  [Agent: inst(*x*)]
```

Thus, by bringing together the automatic construction of collocational systems with a notion of qualia structure for nouns, we have arrived at a fairly useful lexical representation for Information Retrieval tasks.

4 Discussion

Previous investigators involved in corpus analysis using weak methods have documented limited successes and warned of many pitfalls (e.g. [Grishman *et al* 1986] and [Zernik 1989]). The approach described here differs from previous efforts in its combination of diagnostic collocational systems with a generic target representation for nouns. While our limited experiments with the acquisition algorithm show some promise, it is too early to tell how well this approach will do in a larger corpus containing a greater range of senses for terms. One danger is for the algorithm to be overly optimistic in matching a set of occurrences to a diagnosis. Given the rampant ambiguity of prepositions and the potential for verb object combinations that can spuriously suggest metonymic

relationships, we have found the algorithm as it stands to be too susceptible to jumping to false conclusions. We are looking to improve precision by increasing our repertoire of both positive and negative diagnostics, as well as by incorporating information theoretic statistics (as in Church and Hindle (1990)).

Likewise, we have been investigating ways to reduce misses - cases in which evidence of relationships between terms known to be related is not detected by our current set of heuristics. One case in point regards our analysis of "disk", which we initially expected to behave similar to "tape" in its telic quale. However, the intersection of predicate sets for "disk" and "information" yielded the terms

`(copy specify set be have)`

Missing are "read and "write", the telic predicates for *tape*. This example reveals the subtleties present in the container metonymy. Specifically, the container can stand in for its contents only in those situations where one refers to the contents as a whole. While one typically "reads" an entire tape, one usually reads only parts of a disk at a time. "Copying" a whole disk is more typical, however, and hence shows up in our corpus. Reading and writing still apply to disks; however, since they do not apply *holistically*, we find instead constructions with the prepositions *to* and *from*, e.g. *read/write from the disk*.

This example illustrates the pitfalls that are lurking if the linguistic rules are too coarsely defined, but it also shows that such rules are not domain specific, and thus, once properly formulated, could function in a general purpose diagnostic context. It remains an empirical question how well weak methods can be employed to discriminate among the quale of a noun. While this constitutes the primary focus of our current research, we also believe that the above methods complement well other ongoing research in the construction of word-disambiguated dictionaries (e.g. [Ravin 1990]).

5 Conclusion

We contend that using lexical semantic methods to guide lexical knowledge acquisition from corpora can yield structured thesaurus-like information in a form amenable for use within information retrieval applications. The work reported here, though preliminary, illustrates the applicability of this approach for several important classes of nominals. Future work includes refining the discovery procedures to reduce misses and false alarms and extending the coverage of the lexical semantics component to allow the testing of such techniques on a greater range of terms. Finally, we intend to apply the results of the analysis within an experimental information retrieval system to test their effectiveness as indexing and retrieval aids.

Acknowledgements

The authors wish to thank Bran Boguraev for useful discussion, and Jeff Brennan, Rex Flynn, and David Hanssen, members of Digital Equipment Corporation's AI-STARS Information Retrieval group, for their contributions to the development of the software used to conduct this research, as well as for many discussions around the applications of natural language processing to textual information retrieval.

6 Bibliography

Amsler, Robert (1980) *The Structure of the Miriam Webster Pocket Dictionary*, Ph.D. Dissertation, University of Texas at Austin, 1980.

Anick, Peter, Jeff Brennan, Rex Flynn, David Hanssen, Bryan Alvey, and Jeffrey Robbins (1990) A Direct Manipulation Interface for Boolean Information Retrieval via Natural Language Query, to appear in Proceedings of SIGIR '90.

Bobrow, D. G. and T. Winograd (1977) "An Overview of KRL, a Knowledge Representation Language," *Cognitive Science*, 1.1.

Brachman, R. J. and J. Schmolze (1985) "An Overview of the KL-ONE Knowledge Representation System," *Cognitive Science* 9.2.

Church, Kenneth and Donald Hindle (1990) Collocational Constraints and Corpus-Based Linguistics. In Working Notes of the AAAI Symposium: Text-Based Intelligent Systems.

Croft, W. B. (1989) Automatic Indexing. in INDEXING: The State of Our Knowledge and the State of Our Ignorance, edited by Bella Hass Weinberg, Learned Information, Inc., Medford, N. J., pp. 87-100.

Croft, W. B. and R. H. Thompson (1987) I3R: A New Approach to the Design of Document Retrieval Systems. JASIS, 38(6):389-404.

Cullingford R. and Pazzani M. (1984) "Word-Meaning Selction in Multiprocess Language Understanding Programs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 6.4.

Debili, Fathi, Christian Fluhr, Pierre Radasoa (1988) About Reformulation in Full-Text IRS. RIAO 88 proceedings, pp. 343-357.

Grishman, Ralph, Lynette Hirschman, Ngo Thanh Nhan (1986) Discovery Procedures for Sublanguage Selectional Patterns: Initial Experiments. Computational Linguistics, Vol. 12, Number 3, pp. 205-215.

Jackendoff, Ray (1983) *Semantics and Cognition*, MIT Press, Cambridge, MA.

Jacobs, Paul (1989) Paper presented at the *First International Workshop on Lexical Acquisition*, IJCAI-1989, Detroit.

Krovetz, Robert and W. B. Croft (1989) Word Sense Disambiguation Using Machine-Readable Dictionaries. Proceedings of SIGIR '89, pp. 127-136.

Krovetz, Robert (1990) "Information Retrieval and Lexical Ambiguity" In Working Notes of the AAAI Symposium: Text-Based Intelligent Systems.

Maarek, Y.S. and F. Z. Smadja (1989) Full Text Indexing Based on Lexical Relations, An Application: Software Libraries. Proceedings of SIGIR '89, pp. 127-136.

Mel'čuk, I. (1988) *Dependency Syntax*, SUNY Press. Albany, New York.

Moravcsik, J. M. (1975) Aita as Generative Factor in Aristotle's Philosophy, *Dialogue*.

National Library and Information Associations Council (1980) *Guidelines for Thesaurus Structure, Construction, and Use*, New York: American National Standards Institute.

Pustejovsky, James (1989a) Type Coercion and Selection, *Proceedings of West Coast Conference on Formal Linguistics*, Vancouver, 1989.

Pustejovsky, James (1989b) The Generative Lexicon. ms. Brandeis University.

Pustejovsky, James (forthcoming) *The Generative Lexicon: A Theory of Computational Lexical Semantics*, MIT Press, Cambridge, MA.

Pustejovsky, James and Peter Anick (1988) The Semantic Interpretation of Nominals, COLING '88.

Ravin, Yael (1990) "Heuristics for Disambiguating and Interpreting Verb Definitions," *Proceedings of 1990 ACL*, Pittsburgh, PA.

Thompson, R.H. and W.B. Croft (1989) "Support for Browsing in an Intelligent Text Retrieval System," *Internation Journal of Man-Machine Studies*, 30:639-668.

Touretzky, David S. (1986) *The Mathematics of Inheritance Systems*, Morgan Kaufmann, Los Altos, CA.

Vickery, B. C. (1975) *Classification and Indexing in Science*. Butterworth and Co., Ltd. London, England.

Wilks, Yorick (1975) An Intelligent Analyzer and Understander for English. Comm ACM, 18, 264-274.

Wilks, Yorick A, Dan C. Fass, Cheng-Ming Guo, James E. McDonald, Tony Plate, and Brian M. Slator (1988) Machine Tractable Dictionaries as Tools and Resources for Natural Language Processing. Proceeding of COLING-88, Budapest, Hungary.

Zernik, Uri (1989) Lexicon Acquisition: Learning from Corpus by Exploiting Lexical Categories. Proceedings of IJCAI 89.