EXTRACTING SEMANTIC HIERARCHIES FROM A LARGE ON-LINE DICTIONARY

Martin S. Chodorow

Department of Psychology, Hunter College of CUNY
and
I.B.M. Thomas J. Watson Research Center
Yorktown Heights, New York 10598

Roy J. Byrd
George E. Heidorn

I.B.M. Thomas J. Watson Research Center
Yorktown Heights, New York 10598

ABSTRACT

Dictionaries are rich sources of detailed semantic information, but in order to use the information for natural language processing, it must be organized systematically. This paper describes automatic and semi-automatic procedures for extracting and organizing semantic feature information implicit in dictionary definitions. Two head-finding heuristics are described for locating the genus terms in noun and verb definitions. The assumption is that the genus term represents inherent features of the word it defines. The two heuristics have been used to process definitions of 40,000 nouns and 8,000 verbs, producing indexes in which each genus term is associated with the words it defined. The Sprout program interactively grows a taxonomic "tree" from any specified root feature by consulting the genus index. Its output is a tree in which all of the nodes have the root feature for at least one of their senses. The Filter program uses an inverted form of the genus index. Filtering begins with an initial filter file consisting of words that have a given feature (e.g. [+human]) in all of their senses. The program then locates, in the index, words whose genus terms all appear in the filter file. The output is a list of new words that have the given feature in all of their senses.

1. Introduction.

The goal of this research is to extract semantic information from standard dictionary definitions, for use in constructing lexicons for natural language processing systems. Although dictionaries contain finely detailed semantic knowledge, the systematic organization of that knowledge has not heretofore been exploited in such a way as to make the information available for computer applications.

Amsler(1980) demonstrates that additional structure can be imposed upon a dictionary by making certain assumptions about the ways in which definitions are constructed. Foremost among these assumptions is that definitions consist of a "genus" term, which identifies the superordinate concept of the defined word, and "differentia" which distinguish this instance of the superordinate category from other instances. By manually extracting and disambiguating genus terms for a pocket dictionary, Amsler demonstrated the feasibility of generating semantic hierarchies.

It was our goal to automate the genus extraction and disambiguation processes so that semantic hierarchies could be generated from full-sized dictionaries. The fully automatic genus extraction process is described in Section 2. Sections 3 and 4 describe two different disambiguation and hierarchy-extraction techniques that rely on the genus information. Both of these techniques

are semi-automatic, since they crucially require decisions to be made by a human user during processing. Nevertheless, significant savings occur when the system organizes the presentation of material to the user. Further economy results from the automatic access to word definitions contained in the on-line dictionary from which the genus terms were extracted.

The information extracted using the techniques we have developed will initially be used to add semantic information to entries in the lexicons accessed by various natural language processing programs developed as part of the EPISTLE project at IBM. Descriptions of some of these programs may be found in Heidorn, et al. (1982), and Byrd and McCord(1985).

2. Head finding.

In the definition of *car* given in Figure 1, and repeated here:

car : a *vehicle* moving on wheels,

the word *vehicle* serves as the genus term, while *moving on wheels* differentiates cars from some other types of vehicles. Taken as an ensemble, all of the word/genus pairs contained in a normal dictionary for words of a given part-of-speech form what Amsler(1980) calls a "tangled hierarchy". In this hierarchy, each word would constitute a node whose subordinate nodes are words for which it serves as a genus term. The words at those subordinate nodes are called the word's "hyponyms". Similarly, the words at the superordinate nodes for a given word are the genus terms for the various sense definitions of that word. These are called the given word's "hypernyms". Because words are ambiguous (i.e., have multiple senses), any word may have multiple hypernyms; hence the hierarchy is "tangled".

Figure 1 shows selected definitions from *Webster's Seventh New Collegiate Dictionary* for *vehicle* and a few related words. In each definition, the genus term has been italicized. Figure 2 shows the small segment of the tangled hierarchy based on those definitions, with the hyponyms and hypernyms of *vehicle* labelled.

vehicle: (n) (often attrib) an inert *medium* in which a medicinally active agent is administered

vehicle: (n) any of various other *media* acting usu. as solvents, carriers, or binders for active ingredients or pigments

vehicle: (n) an *agent* of transmission : CARRIER

vehicle: (n) a *medium* through which something is expressed, achieved, or displayed

vehicle: (n) a *means* of carrying or transporting something : CONVEYANCE

vehicle: (n) a piece of mechanized *equipment*

ambulance: (n) a *vehicle* equipped for transporting wounded, injured, or sick persons or animals

bicycle: (n) a *vehicle* with two wheels tandem, a steering handle, a saddle seat, and pedals by which it is propelled

car: (n) a *vehicle* moving on wheels

tanker: (n) a cargo *boat* fitted with tanks for carrying liquid in bulk

tanker: (n) a *vehicle* on which a tank is mounted to carry liquids; also : a cargo *airplane* for transporting fuel

Figure 1. Selected dictionary definitions.

Our automated mechanism for finding the genus terms is based on the observation that the genus term for verb and noun definitions is typically the head of the defining phrase. This reduces the task to that of finding the heads of verb phrases and noun phrases.

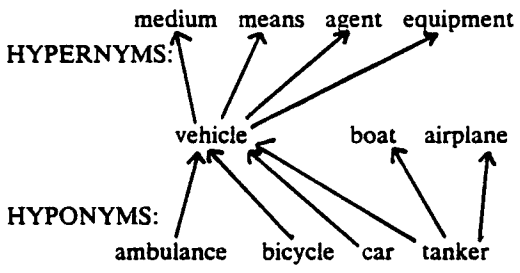

Figure 2. The tangled hierarchy around "vehicle".

The syntax of the verb phrase used in verb definitions makes it possible to locate its head with a simple heuristic: the head is the single verb following the word *to*. If there is a conjunction of verbs following *to*, then they are all heads. Thus, given the following two definitions for *winter*,

winter: (v) to pass the winter
winter: (v) to keep, feed, or manage during the winter

the heuristic would find four heads: *pass*, *keep*, *feed*, and *manage*.

Applying this heuristic to the definitions for the 8,000 verbs that have definitions in Webster's Seventh showed that 2225 distinct verbs were used as heads of definitions and that they were used 24,000 times. In other words, each genus term served as the hypernym for ten other verbs, on average. The accuracy of head finding for verbs was virtually 100 percent.

Head finding is much more complex for noun definitions because of their greater variety. At the same time, the magnitude of the task (over 80,000 defining noun phrases) demanded that we use a heuristic procedure, rather than a full parser, which would have been prohibitively expensive. We were able to take advantage of the fact that dictionary definitions are written in a special and predictable style, and that their analysis does not require the full power of an analyzer for general English.

The procedure used may be briefly described as follows. First the substring of the definition which *must* contain the head is found. This substring is bounded on the left by a word which obligatorily appears in prenominal position: a, an, the, its, two, three, ..., twelve, first, second, ... It is bounded on the right by a word or sequence that can only appear in postnominal position:

- a relative pronoun (introducing a relative clause)
- a preposition not followed by a conjunction (thus, introducing a complement to the head noun)
- a preposition-conjunction-preposition configuration (also introducing a complement)
- a present participle following a noun (thus, introducing a reduced relative clause)

The heuristic for finding the boundary on the right works because of certain restrictions on constituents appearing within a noun phrase. Emonds (1976, pp. 167-172) notes that an adjective phrase or a verb phrase must end with its phrasal head if it appears to the left of the head noun in a noun phrase. For example, in *the very old man*, the adjective phrase *very old* has its head adjective in final position; in *the quietly sleeping children*, the verb phrase *quietly sleeping* ends in its head verb. Another constraint, the Surface Recursion Restriction (Emonds, 1976, p. 19), prohibits free recursion of a node appearing within a phrase, to the left of the phrase head. This prevents prenominal modifying phrases from containing S and PP nodes. Taken together, the two restrictions specify that S, PP, and any other constituent which does not end in its head-of-phrase element cannot appear as a prenominal modifier and must, therefore, be postnominal. Lexical items or sequences that mark the beginnings of these constituents are used by the heuristic to establish the right boundary of the substring which must contain the head of the noun definition.

Once the substring is isolated, the search for the head begins. Typically, but not always, it is the rightmost noun in the substring. If however, the substring contains a conjunction, each conjunct is processed separately, and multiple heads may result. If the word found belongs to a small class of "empty heads" (words like *one*, *any*, *kind*, *class*, *manner*, *family*, *race*, *group*, *complex*, etc.) and is followed by *of*, then the string following *of* is reprocessed in an effort to locate additional heads.

Applying this procedure to the definitions for the 40,000 defined nouns in Webster's Seventh showed that 10,000 distinct nouns were used as heads of definitions and that they were used 85,000 times. In other words, each genus term served as the hypernym for 8.5 other verbs, on average. The accuracy of head-finding for nouns was approximately 98 percent, based on a random sample of the output.

3. Sprouting

Sprouting, which derives its name from the action of growing a semantic tree from a specified root, uses the results of head-finding as its raw material. This information is organized into a "hyponym index", in which each word that was used as a genus term is associated with all of its hyponyms. Thus, "vehicle" would have an entry which reads (in part):

vehicle: ambulance ... bicycle ... car ... tanker ...

For a given part-of-speech, the hyponym index needs to be built only once.

When invoking the sprouting process, the user selects a root from which a semantic tree is to be grown. The system then computes the transitive closure over the hyponym index, beginning at the chosen root. In effect, for each new word (including the root), all of its hyponyms are added to the tree. This operation is applied recursively, until no further new words are found.

The interactiveness of the sprouting process results from the fact that the user is consulted for each new word. If he decides that that word does not belong to the tree being grown, he may prune it (and the branches that would emerge from it). These pruning decisions result in the disambiguation of the tree. The user is assisted in making such decisions by having available an on-line version of Webster's Seventh, in which he may review the definitions, usage notes, etc. for any words of which he is unsure.

The output of a sprouting session, then, is a disambiguated tree extracted from the tangled hierarchy represented by the hyponym index. Actually, the output more nearly resembles a bush since it is usually shallow (typically only 3 to 4 levels deep) and very wide. For example, a tree grown from *vehicle* had 75 direct descendants from the root, and contained over 250 nodes in its first two levels, alone. The important aspect of the output, therefore, is not its structure, but rather the fact that the words it contains all have at least one sense which bears the property for which the root was originally selected. It is important to note that any serious use of sprouting to locate all words bearing a particular semantic feature must involve the careful selection and use of several roots, because of the variety of genus terms employed by the Webster's lexicographers. For example, if it were desired to find all nouns which bear the [+female] inherent feature, sprouts should at least be begun from *female*, *woman*, *girl*, and even *wife*.

4. Filtering

Filtering, like sprouting, results in lists of words bearing a certain property (e.g., [+human]). Unlike sprouting, however, filtering only picks up words all of whose senses have the property.
It is based on a "hypernym index" (the inversion of the hyponym index), in which each word is listed with its hypernyms, as in the example given here:

vehicle: agent equipment means medium

The filtering process begins with a "seed filter" consisting of an initial set of words all of whose senses bear some required property. The seed filter may be obtained in any manner that is convenient. In our work, this may be either from the semantic codes assigned to words by the Longman Dictionary of Contemporary English, or from morphological processing of word lists, as described in Byrd and McCord (1985). For example, morphological analysis of words ending in *-man*, *-sman*, *-ee*, *-er*, and *-ist* constitute a rich source of [+human] nouns. Given the filter, the system uses it to evaluate all of the words in the hypernym index. Any words, all of whose hypernyms are already in the filter, become candidates for inclusion in the filter during the next pass. The user is consulted for each candidate, and may accept

Pass#	Filter Size	New Words
1	2539*	1091
2	4113**	234
3	4347	43
4	4390	0
...		
5	4661***	49
Total	4710	

* Obtained from Longman Dictionary of Contempory English
** Includes 483 new words from morphological analysis
*** Includes 271 new words from morphological analysis

Figure 3. A Filtering of [+human] nouns.

or reject it. Finally, all accepted words are added to the filter, and the process is repeated until it converges.

An example of the filtering procedure applied to nouns bearing the [+human] inherent feature is given in Figure 3. It can be seen that the process converges fairly quickly, and that it is fairly productive, yielding, in this case, an almost two-for-one return on the size of the initial filter. For nouns with the [-human] inherent feature, an initial filter of 22,000 words yielded 11,600 new words on the first pass, with a false alarm rate of less than 1% based on a random sample of the output. From an initial filter of 15 [+time] nouns, 300 additional ones were obtained after three passes through the filter. These examples demonstrate another important fact about filtering: that it can be used to project the semantic information available from a smaller, more manageable source such as a learner's dictionary onto the larger set of words obtained from a collegiate sized dictionary.

As does sprouting, filtering also produces a list of words having some desired property. In this case, however, the resulting words have the property in all of their senses. This type of result is useful in a parsing system, such as the one described in Heidorn, et al. (1982), in which it may be necessary to know whether a noun *must* refer to a human being, not merely that it *may* refer to one.

5. Conclusion

This work receives its primary motivation from the desire to build natural language processing systems capable of processing unrestricted English input. As we emerge from the days when hand-built lexicons of several hundred to a few thousand entries were sufficient, we need to explore ways of easing the lexicon construction task. Fortunately, the tools required to do the job are becoming available at the same time. Primary among them are machine readable dictionaries, such as Webster's and Longman, which contain the raw material for analysis. Software tools for dictionary analysis, such as those described here and in Calzolari (1982), are also gradually emerging. With experience, and enhanced understanding of the information structure in published diction-

aries, we expect to achieve some success in the automated construction of lexicons for natural language processing systems.

References

Amsler, R. A. (1980), *The Structure of the Merriam-Webster Pocket Dictionary*, Doctoral Dissertation, TR-164, University of Texas, Austin.

Byrd, R. J. and M. C. McCord (1985), "The lexical base for semantic interpretation in a Prolog parser" presented at the CUNY Workshop on the Lexicon, Parsing, and Semantic Interpretation, 18 January 1985.

Calzolari, N. (1984), "Detecting Patterns in a Lexical Data Base," Proceedings of COLING/ACL-1984

Emonds, J.E. (1976), *A Transformational Approach to English Syntax*. New York: Academic Press.

Heidorn, G. E., K. Jensen, L. A. Miller, R. J. Byrd, and M. S. Chodorow (1982), "The EPISTLE Text-Critiquing System," *IBM Systems Journal*, 21, 305-326.

Longman Dictionary of Contemporary English (1978), Longman Group Limited, London.

Webster's Seventh New Collegiate Dictionary (1963), G. & C. Merriam, Springfield, Massachusetts.