WORDNET: A LEXICAL DATABASE FOR ENGLISH

George A. Miller, Principal Investigator

Cognitive Science Laboratory
Princeton University
Princeton, NJ 08542

PROJECT GOALS

Work under this grant is intended to provide lexical resources for research on natural languages. The principal product is WordNet, a lexical database for English whose organization is inspired by current psycholinguistic theories of human lexical knowledge. Lexicalized concepts are organized by semantic relations for nouns, verbs, adjectives, and adverbs.

The principal goal of the project is to upgrade WordNet and make it available to interested users. A secondary goal is to explore practical applications of WordNet; its possible use in the resolution of word senses in context (semantic disambiguation) is viewed as a necessary precursor for many other applications.

RECENT RESULTS

WordNet Upgrade. The basic elements of WordNet are sets of synonyms (synsets), which are taken to represent lexicalized concepts. The database is organized by bidirectional pointers between synsets which correpond to familiar semantic relations: (synonymy, antonymy, hyponymy, meronymy, troponymy, entailment, etc.). During the past year the number of entries (words and collocations) in WordNet has grown from 62,700 to 82,600; the number of lexicalized concepts (synsets) from 50,300 to 62,700; the number of unique word-sense combinations from 98,300 to 117,300; and the number of synsets that include definitional glosses from 20,700 to 39,400. The syntactic category of adverb has been added; there are now 1,200 adverb synsets.

Software Development and Distribution. WordNet is available to the research community via anonymous ftp or on diskettes for PC users. A fourth version of the database, WordNet 1.3, was released 9 December 1992; notices were sent to 175 individuals and laboratories who had previously expressed interest. An X-windows interface that includes a program to recognize inflectional morphology is available for Sun SPARCstations; interfaces are also available for DECstations, NeXT, Microsoft Windows, and MacIntosh. Address inquiries to wordnet@princeton.edu.

Sense Resolution. A corpus of sentences using the noun "line" in different senses has been compiled and used to compare different methods of automatically determining from context which sense was intended. Preliminary efforts have been made to explore the use of WordNet to generalize the contexts that emerge from such studies.

Semantic Tagging. In support of the studies of sense resolution, a portion of the Brown Corpus has been tagged with pointers to WordNet. The result is a semantic concordance: a textual corpus and a lexicon so combined that every substantive word in the text is linked to its appropriate sense in the lexicon. To facilitate the task, an X-windows interface, ConText, has been developed that displays the text along with the WordNet entry for each content word; the user selects the appropriate sense, or inserts a comment that is used by the lexicographers to upgrade WordNet.

PLANS FOR THE COMING YEAR

WordNet will continue to be expanded and improved, and made available to interested users. There is interest in including more syntactic information, although that step would probably require formulating WordNet as a relational database. We have been considering that move for some time, but have lacked the necessary personnel. We are also considering how to incorporate more syntactic information in the output of ConText, and have secured permission from IBM to use McCord's English Slot Grammar on an experimental basis.

We plan to collect more systematic data on the adequacy of WordNet. The use of WordNet for semantic tagging provides a valuable test of its coverage and precision; the percentage of words in a text that are either missing or are used to express senses not included in WordNet fluctuates radically depending on the topic of a passage, but the average should decline steadily as the work proceeds. By summer 1993 we expect to be able to release an initial corpus of semantically tagged passages to authorized users of the Brown Corpus.

We are also developing a method to analyze the cooccurrences of concepts in the same sentences, in the expectation that such information will facilitate sense resolution. The value of such analyses will guide us in deciding how much farther to proceed with the construction of a semantic concordance based on the Brown Corpus.