WORDNET: A LEXICAL DATABASE FOR ENGLISH

George A. Miller, Principal Investigator

Cognitive Science Laboratory
Princeton University
Princeton, New Jersey 08544

PROJECT GOALS

WordNet is a lexical database for English organized in accordance with current psycholinguistic theories. Lexicalized concepts are organized by semantic relations (synonymy, antonymy, hyponymy, meronymy, etc.) for nouns, verbs, and adjectives.

Work under this grant is intended to extend and upgrade WordNet, to make it generally available, and to develop it as a tool for use in practical applications. In order to make it available for information retrieval and machine translation, a system is being developed that accepts English text as input and automatically gives as output the same text augmented by syntactic and semantic notations that disambiguate all of the substantive words. Initially, the semantic tagging is being done manually so that we can (1) obtain extensive experience with the tagging process and (2) create a database of correctly tagged text for use in testing proposals for automatic sense disambiguation.

RECENT RESULTS

The work falls into four categories: (1) Preprocessing of textual corpora; (2) Development of ConText, an interface for semantically tagging text; (3) WordNet upgrade and; (4) Software development and distribution of WordNet.

Textual Corpora. Acquisition of several large text corpora and development of programs to preprocess them. Currently the preprocessor formats the text one sentence per line, searches the text for all WordNet collocations, tokenizes the text accordingly, and subsequently runs the text through a part-of-speech tagger. The search for WordNet collocations handles inflectional morphology.

ConText Interface. ConText is an X-windows interface to WordNet which takes the preprocessed text (described above) as input and accesses the WordNet entry for each content word (in the appropriate part of speech). The user selects the appropriate sense or, if that is not possible, indicates the nature of the problem encountered. ConText outputs the text with pointers to the selected WordNet senses (or, if no sense is chosen, the reason why). Current and previous sentences are displayed to give the user adequate context for disambiguation. The inflectional morphology component of WordNet is used by ConText so that the WordNet uninflected forms are found even though inflected forms are both input to and output from ConText.

WordNet Upgrade. As of January, 1992 the number of different character strings in WordNet is 62,726; the number of lexicalized concepts (synsets) is 50,318; the number of unique string-sense combinations is 98,300; the total number of semantic pointers is 71,126; and the number of synsets containing definitional explanations is 20,718. In addition, the grammatical category of relational adjectives has been added to WordNet. These adjectives are defined by unidirectional pointers to the related noun, encoding the distinction between the adjective "nervous" in "nervous student" and the relational adjective in "nervous condition."

Software Development. WordNet has been developed on Sun SPARCstations in the C language with an X-windows interface. MS-DOS and NeXT interfaces are also available. A MacIntosh interface is currently being developed.

PLANS FOR THE COMING YEAR

The organization of WordNet provides an estimation of semantic distances between lexicalized concepts, and research will focus on how this feature can be exploited to facilitate lexical disambiguation. Once a large enough body of semantically tagged text has been created using ConText, hypotheses can be tested. To facilitate retrieval of semantically related words from WordNet, it is likely that WordNet will be put into a relational database.

WordNet will continue to be extended and upgraded. The direction of this work will be determined, in part, by the results of the semantic tagging.