Session 10: THE LEXICON

Ralph Grishman

Department of Computer Science
New York University
New York, NY 10003

Work in natural language processing has been moving rapidly towards the creation of large-scale systems addressed to real tasks. One aspect of this has been a rapid increase in the vocabulary size of these systems. "Toy" lexicons of 500 or 1000 words are no longer adequate; several tens of thousands lexical entries will be required, at a minimum. Developers of machine translation systems — who have confronted the problems of "real," largely unrestricted, text much longer than most other natural language researchers — have long recognized the central role of large, high quality lexicons.

Such broad-coverage lexical resources are of course costly and time-consuming to develop. Fortunately, however, there seems a reasonable prospect that they can be developed as shared resources. Current lexicons record for the most part relatively shallow (simply structured) information about the pronunciation, syntax, and semantics of words. There appears to be a general agreement between different system developers on at least some of the features to be captured in the lexicon, even though these features may be represented very differently in the various systems. The agreement seems to be clearest regarding syntactic information, but there is reason to believe that at least a partial consensus can be reached regarding pronunciation and possibly for semantic information as well.

All of the presentations in this session addressed the need for broad-coverage lexical resources. In addition to the papers included in this volume, there were presentations by Prof. Mark Liberman of the Univ. of Pennsylvania and Prof. Makoto Nagao of Kyoto Univ.

Prof. Liberman discussed some of the plans of the Linguistic Data Consortium. The Linguistic Data Consortium was created in 1992 with a combination of government and private funds in order to create a rich repository of resources for research and development of natural language systems. As part of its mandate, the Consortium intends to assemble a range of lexical resources including pronunciation, syntactic, and semantic information, under the general heading of COMLEX (a COMmon LEXicon). Among these efforts, the work on a syntactic lexicon — COMLEX Syntax — is furthest advanced; the paper by the group at New York University describes the status of this project.

These works are small in scale when compared to the dictionary efforts in Japan, which were summarized in Prof. Nagao's presentation. The largest of these efforts is the EDR Project of the Japan Electronic Dictionary Research Institute. This project is producing a collection of interrelated dictionaries, including a Japanese dictionary and an English dictionary (each of about 300,000 entries) whose entries are both linked to a "concept dictionary".

Prof. George Miller and his associates at Princeton University have for the past several years been constructing a lexical knowledge base called WordNet. In WordNet, English nouns, verbs, and adjectives are organized into synonym sets ("synsets"), each representing one underlying lexical concept; these synsets are connected by various semantic relations, such as antonymy, hyponymy, and meronymy. A word may have several meanings and so be assigned to several synsets; a word with its synset can thus be used to identify a particular sense of a word. The paper "A Semantic Concordance" describes an ongoing effort to "tag" a corpus by identifying, for each content word (noun, verb, adjective, and adverb), the synset to which it belongs in that context.

Corpus tagging can be even more valuable if the same corpus is tagged for several different lexical characteristics. For example, the COMLEX Syntax group is considering the possibility of tagging the verbs in a corpus according to the subcategorization frame used in each context. Although the COMLEX Syntax Lexicon will initially not be sense distinguished, correlating the subcategorization tags with WordNet sense tags would give some indication of the correspondence between subcategorizations and word senses.

Identifying the general vocabulary — nouns, verbs, adjectives, ... — is only part of the battle in lexical analysis. Many texts are replete with proper nouns (names). Although we can include the most frequent of these in our lexicon, the list can never be complete. A good lexicon must therefore be complemented by effective strategies for identifying and classifying proper nouns, which typically involve some combination of pattern matching with information from the lexicon. The final paper in this session, from Syracuse University, describes an approach to proper noun identification and and evaluation of this approach on a sample from the Tipster corpus.