

Language Processing and Commonsense Reasoning

Robert Wilensky, Principal Investigator
Computer Science Division
University of California at Berkeley
Berkeley, CA 94720 U.S.A.

PROJECT GOALS

The goal of this project is to develop the technology to construct intelligent, natural-language-capable agents. Such agents will have natural language and reasoning capabilities that facilitate interaction with untrained users who are seeking information pertinent to tasks in which they are engaged. Emphasis is given to techniques for creating such agents for new domains, focusing specifically on portability and knowledge acquisition. In particular, we have been developing DIRC (Domain Independent Retargetable Consultant), a kind of intelligent, natural-language-capable consultation shell, and a number of mechanisms for world and language knowledge acquisition.

RECENT RESULTS

Our implementation of DIRC is about 80% complete. We also continued our theoretical work on finding a probabilistic basis for text inference. Current probabilistic and statistical inductive models are difficult to bias to generalize from training data; specifically, the most common methods for inducing prior probability distributions — relative frequency priors and maximum entropy priors — are inadequate. To address this gap we propose a family of inductive methods, called the gamma-continuum. These are a parameterized set of methods for generating priors from a training set incorporating an a priori *abstractive bias* that causes the model to make generalizations. Our driving application is probabilistic pattern completion to support integrated natural language parsing and semantic interpretation, where the patterns combine lexical, syntactic, and semantic structures.

We have developed an accurate, relatively low-overhead method for the disambiguation of English noun homonyms using a large corpus of free text. The objective of the algorithm is the following: given an English sentence or sentence fragment containing the target noun, determine which of a set of predetermined senses should be assigned to the noun. This is accomplished by checking the context surrounding the target noun against that of previously recorded instances, and choosing the sense for which the most evidence is found. Initial results are promising.

We are continuing work on intelligent dictionary reading for natural language vocabulary acquisition. In particular, we have been studying noun/preposition patterns and methods for acquiring their meanings. In some cases, these patterns are defined directly in the dictionary. In other cases,

the semantics of the patterns can be derived by partially productive relations (subregularities) between the complement structures of the nouns and the verbs they are derived from. Thus we are exploiting our work on subregularities to aid in the interpretation of dictionary entries.

PLANS FOR THE COMING YEAR

Over the coming year we plan to finish DIRC, and continue to develop the various methods of lexical and world knowledge acquisition we have discussed previously. In addition, we are planning to use some of the technology we have developed to produce more intelligent text retrieval methods. Such methods are both useful additions to our intelligent agents approach to “help” systems, as well as ways of improving the performance of conventional information retrieval systems.