TACITUS: RESEARCH IN TEXT UNDERSTANDING

Jerry R. Hobbs, Principal Investigator
Artificial Intelligence Center
SRI International
Menlo Park, California 94025

1 PROJECT GOALS

The aim of the TACITUS project is to develop a general, domain-independent capability for text understanding that allows for variable levels of analysis, depending on the requirements of the task. Four stages of processing are being developed: preprocessing, syntactic analysis, inferential pragmatics processing, and template generation. Template generation is a straightforward programming task, but the other three stages of processing consititute research issues.

Our specific goals in these three areas are as follows:

Preprocessing: To build as much of the processing as possible into linear-time, string-based, finite-state methods, in order to achieve fast, highly accurate processing.

Syntactic Analysis: To develop the capability of parsing full, syntactically complex sentences in real-world texts, choosing among alternative parses with a high degree of precision, and recovering most of the information in sentences in cases of parser failure.

Inferential Pragmatics Processing: To implement methods for dealing with various pragmatics problems, such as schema recognition, reference resolution, metonymy resolution, the interpretation of vague predicates, and the resolution of various ambiguities, using a scheme of weighted abductive inference.

2 RECENT RESULTS

Our recent successes in each of these areas are as follows:

Preprocessing: We implemented a keyword-based statistical filter to select the relevant sentences in the corpus. This eliminates 75% of the sentences from deeper analysis, but only eliminates about 10% of the relevant sentences. In addition, we developed highly successful methods for dealing with unknown words.

Syntactic Analysis: We developed a bottom-up, agenda-based parser which produces the best n parses, together with their logical forms. We were able to produce a correct or nearly correct parse for 75% of the sentences under 30 words in the MUC-3 corpus. We also developed a new method, called "terminal substring parsing" for parsing sentences of over 60 words, and heuristics for extracting whatever structure was found in cases of failed parses, allowing us to recover nearly 90% of the propositional content.

Inferential Pragmatics Processing: An inference component, based on the Prolog Technology Theorem Prover (PTTP) has been developed and refined. It uses a knowledge base encoding commonsense and domain-specific knowledge in the form of over 1000 axioms. This component allows us to extract information from some quite complex texts with a high degree of accuracy. In the MUC-3 evaluation, it resulted in our precision being higher than any of the other sites tested.

3 PLANS FOR THE COMING YEAR

Our plans for the coming year are as follows:

Preprocessing: We are beginning to implement a considerably more powerful preprocessing component, in an effort to build as much capability as possible into linear-time, string-based, finite-state methods. This component alone is intended to produce passable results in the template-generation task.

Syntactic Analysis: We will work primarily on incremental refinements.

Inferential Pragmatics Processing: We are improving the theorem prover in a number of ways and increasing the size of our knowledge base, as will be required in order to raise our recall. We are investigating how best to use the results of our enhanced preprocessing in the inferential processing—whether to use them as defaults in case inferential processing fails, to use them for helping guide the search for relevant information, or to have inferential processing attempt to refute them. The deeper processing provided by the syntactic and inference components is more reliable than preprocessing alone, but requires

substantially more time. We need to discover the best way to blend information from both sources.