The Penman Natural Language Project

Systemics-Based Machine Translation

Eduard Hovy
Information Sciences Institute of USC
4676 Admiralty Way
Marina del Rey, CA 90292-6695

OBJECTIVE

The development of an integrated knowledge-based machine-aided translation system based on Systemic Linguistics. Parts of the system are to function as modules to be incorporated in the MAT system being codeveloped with CMU and CRL. Our work involves the enhancement of Penman's existing parsing technology to match the level of the language generation system; the development of ancillary knowledge sources and software (such as bilingual lexicons and interlingua/transfer structures); the maintenance and continued distribution of the sentence generator; and the embedding of all these parts into the joint system.

PROGRESS AND RECENT ACCOMPLISHMENTS

The past few months saw the beginning of a new three-year collaboration in Machine-Aided Translation with CMU and CRL.

We have held the first meeting to discuss questions of evaluation, representational Interlingua, protocols for communication between modules developed at the three sites, etc. Work on the German grammar is continuing (at our German partner institution) and an early version will be incorporated into the system next month. Work is underway to implement the prototype parser to full-fledged form. The parser is being applied, debugged, and tested on increasingly larger portions of the grammar. Since it uses classification instead of unification as the primary inference mechanism, the parser required that the Loom knowledge representation language be given the ability to perform classificatory inference over disjunction. This has been completed. In addition, Loom's facility to manage alternative worlds has been completed and is being used for the treatment of ambiguity.

In addition, we are hosting Dr. Kenneth Church of AT&T Bell Laboratories this year. He has completed the first phase in a project to automatically construct large multilingual dictionaries from parallel multilingual texts, using trilingual banking texts from Switzerland and the bilingual Canadian Parliamentary Hansard. Aspects of this work is reported in this Proceedings.

Under separate funding, members of the project have also been involved in the development of a new text planner (based on Rhetorical Structure Theory and extensions) to plan out coherent multisentence texts. Several visitors from Europe are participating in this project.

The maintenance and distribution of the Penman sentence generator continues. The system incorporates one of the most extensive computational grammars of English for generation in the world, being the result of over 25 person-years' effort. Penman has been adapted to MacIntosh-II, Sun 3, TI Explorer and Symbolics computers, and has been distributed to over 50 research sites in the US, Europe, Canada, and Australia.