THE PENMAN PROJECT ON KNOWLEDGE-BASED MACHINE TRANSLATION

Eduard Hovy, Principal Investigator

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

PROJECT GOALS

The joint development, together with the ULTRA project at New Mexico State University and the Center for Machine Translation at Carnegie Mellon University, of an integrated knowledge-based machine-aided translation system called PANGLOSS. The ISI-specific work includes the development of English sentence generation and sentence planning capabilities and the construction of an Ontology of concepts to act as the semantic lexicon for all modules of the system as a whole. In addition, we continue to enhance Penman's existing generation technology, to collect and develop ancillary knowledge sources and software (such as grammars or bilingual dictionaries and lexicons for German, Japanese, Spanish, and Chinese), and to maintain and distribute Penman.

RECENT RESULTS

During the past year, the generation component of PANGLOSS was installed; PANGLOSS was tested during the first DARPA MT evaluation. This work necessitated the development of code to transfer the output of New Mexico's ULTRA parser to a form suitable for Penman.

More recently, Penman Project members have been working on the semi-automated construction and acquisition of an Ontology for PANGLOSS. A high-level taxonomy of the basic concepts required for the processing of ULTRA, the CMU software, and Penman was synthesized out of several sources; this 400-odd node taxonomy we call the Ontology Base (OB). Current work involves migrating wordsense names from LDOCE into WordNet using several automatic techniques and then taxonomizing fragments of WordNet under the OB; at the present time, approx. 11,000 concepts have been so taxonomized and another 10,000 are awaiting final placement. Our goal is an Ontology organized under the OB of approx. 50,000 items. Toward this goal we acquired WordNet from Princeton and an online copy of Roget's thesaurus.

Ramping up toward making the Ontology support processing of other languages, we have been collecting multilingual resources of various types. We have acquired an online Japanese-English dictionary (approx. 50,000 entries with phrases), several Chinese-English online dictionaries (approximately equal total size), and are in the process of acquiring the Collins bilingual Spanish-English dictionary. We have also established X-windows based display capabilities for Japanese and Chinese, including a Japanese emacs editor and dictionary access interface.

In other work, the core mapping engine of the Sentence Planning module of PANGLOSS has been constructed and is currently being debugged. The Sentence Planner converts representations of texts written in the Pangloss Interlingua into SPL expressions suitable for Penman.

PLANS FOR THE COMING YEAR

Three principal efforts are planned for the coming year: the construction of the 50,000-node Ontology, the development of English, Japanese, and Spanish lexicons associated with the Ontology, and the development and implementation of several microtheories for use in sentence planning.

The main problem in Ontology construction is the automated acquisition under Ontology nodes of semantic information, as used during semantic analysis and lexical selection. A number of methods of extracting such information from dictionaries, text corpora, and other resources are being developed, as well as a system to assist the acquisition of remaining information by humans.

A problem in automatically constructing lexicons of various languages is the association of a wordsense in a dictionary with its correct Ontology item (if such exists) or the creation of a new Ontology item and its correct placement in the Ontology. Variations of the algorithms used for associating LDOCE wordsenses with WordNet items will be used for this task, operating on the bilingual dictionaries we have collected.

The main problems facing the Sentence Planner are the development of microtheories for lexical selection, reference (including pronominalization), and theme development, to ensure high quality and coherent output.