IN-DEPTH KNOWLEDGE-BASED MACHINE TRANSLATION

Eduard Hovy, Principal Investigator

Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292-6695

PROJECT GOALS

The development of an integrated knowledge-based machine-aided translation system called PANGLOSS in collaboration with the Center for Machine Translation (CMT) at CMU and the Computing Research Laboratory (CRL) at New Mexico State University. The ISI part of the collaboration is focused initially on providing the system's output capabilities, primarily in English and then in other languages, including (some of) German, Chinese, and Japanese. Additional tasks are the maintenance and continued distribution of the Penman sentence generator and text planner and the development of ancillary knowledge sources and software.

RECENT RESULTS

Members of the project have participated in several aspects of the design and setting up of PANGLOSS and in the overall MT effort. Three major efforts are:

1. **Incorporation of language generation:** In the first-year version of PANGLOSS, the ULTRA analyzer of CRL is linked to the Penman generator, both being embedded in the Translator's Workstation (TWS) that includes several browsing, editing, and other user facilities. A process for converting ULTRA output to Penman input has been developed and is being debugged. Approximately 80 ULTRA output sentences (each with approximately 13 variant parses) have been used as test suite; at present the conversion+Penman system produces roughly 25% correct throughput, 35% identifiable errors (which will be trapped and sent to the user for correction), 15% Penman grammar shortcomings, and 25% miscellaneous problems, mostly involving representational inconsistencies. Current work is focusing on extending the grammar, developing ways of interacting with the user, and ironing out the inconsistencies. Also, work on acquiring the system substrate to support PANGLOSS at ISI has been performed; including software acquisition and various licensing requirements.

2. **Interlingua construction:** The PANGLOSS Interlingua Committee recently began constructing an Interlingua, using as starting point the terminologies developed by the three partners, namely ONTOS, the ontology developed at the CMT, IR, the Intermediate Representation terminology used at the CRL, and the Penman Upper Model. Both ONTOS and IR have already been used to support Interlingual machine translation, while variants of the Upper Model suited for German, Japanese, and Chinese are under construction at GMD/IPSI (Germany) and the University of Sydney (Australia). An initial specification of the Interlingua has been developed. A set of issues to be addressed next, including the notation and the substrate for the Interlingua, has been drawn up.

3. **Committee work:** An overall MT Coordinating Committee (MTCC) has been formed and a set of specialized committees with specific tasks have been created under its supervision. The first MTCC meeting will be held at IBM Yorktown Heights on February 26-27. The machine translation effort's Evaluation Committee has produced three documents, one outlining the general methodology of evaluation, one describing the particulars of the upcoming MT system evaluation, and one describing the particulars of the Dry Run evaluation held of the IBM system CANDIDE in February.

PLANS FOR THE COMING YEAR

Version Alpha of PANGLOSS will be completed, tested, and evaluated. An early test with real-world users will take place in March.

The aspects in which Penman needs strengthening to handle the needs of the domain will be addressed, including portions of the grammar and the lexicon. A lexicon of at least 5,000 words will be in place in Penman by June. The kinds of human assistance possible during input preparation and generation will be worked out in detail and incorporated in the TWS. Penman and its ancillary resources will be embedded into the TWS.

The first version PANGLOSS interlingua, including an initial domain model, will be put in place.

The MT system evaluation will be organized and take place around end-April or mid-May.