Pangloss: A Knowledge-based Machine Assisted Translation Research Project – Site 2

D. Farwell, Principal Investigator

Computing Research Laboratory
New Mexico State University, Las Cruces, New Mexico 88003

PROJECT GOALS

The Computing Research Laboratory (CRL) at New Mexico State University, jointly with the Center for Machine Translation (CMT) at Carnegie Mellon University and the Information Sciences Institute (ISI) at the University of Southern California, are developing a Translator's Workstation to assist users in the translation of newspaper articles in the area of finance from Spanish or Japanese into English. At its core is a multi-engine MT system, consisting of a knowledge-based, interlingual system, an example-based system, and an extensive glossary and bilingual dictionary system. Results from all systems are combined in a chart structure which selects the most reliable and complete translation. The KBMT system consists of a source language analysis component, a mapper, and a target language generation component.

During the first two years of the project, the CRL's objectives were to develop tools for constructing lexical items and ontological entries automatically from on-line resources, to develop the Spanish analysis component, and, jointly with CMT and ISI, to establish the infrastructure for the three site project and develop the formats and initial content of the interlingua, the ontology, and the knowledge base.

The second phase of the project is also for two years, and we are currently in the first six months of this phase. CRL's responsibilities continue as before, with primary responsibility for Spanish analysis and joint planning of inter-site cooperation, but with the addition of the task of fulfilling knowledge-acquisition needs for all three sites, both automatic and manual.

RECENT RESULTS

In analysis, three new modules (a proper name recognizer, a clause boundary identifier and a syntactic dependency analysis module) have been added during the past year. As a result, full sentence throughput is now possible within the KBMT system. Acquisition work is continuing both with respect to integrating sense tokens into the growing ontology and in reviewing and increasing the Spanish vocabulary in the system.

The analysis system begins with a dictionary-based part-of-speech tagger, followed by a component which groups the tagged text into small syntactic chunks. Chunks which are tagged as proper nouns are sent to the proper name recognizer for categorization. All chunks are then analyzed: semantic/lexical information is accessed and incorporated into the representation. These smaller constituents are then grouped into clause-level groups, which are then further analyzed to produce ranked possible syntactic dependency structures.

With respect to knowledge acquisition, CRL is currently integrating sense tokens (concepts) into the Ontology Base. These sense tokens are drawn both from *Longman's Dictionary of Contemporary English* (LDOCE) and from Collins *Spanish-English/English-Spanish Dictionary*. Work is underway to provide a large set of tagged Spanish texts both for work within the project and ultimately for the NLP community at large. The portion of the Ontology used in Spanish analysis has been incorporated into a data-base and merged with information in an LDOCE-based database. This allows for experimentation with lexicon content and format and rapid increase in vocabulary coverage.

PLANS FOR THE COMING YEAR

Major work is underway and expected to be completed this year for a final module in the analysis system which will rank semantic readings of an input text. This module will use semantic consistency (preference) information, as well as other information (possibly statistically-based), to help disambiguate word senses in the input text by ranking the likelihood of readings. In acquisition work, large-scale semi-automatic acquisition of Spanish verbs is expected, and the ability to use the information about object entries in the Ontology will be developed.