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

Y. Wilks, 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 a user in the translation of newspaper articles in the area of finance (mergers and acquisitions, followed by joint ventures) in one language (Spanish initially, followed by Japanese) into a second language (English). At its core is a multilingual, knowledge-based, interlingual, interactive, machine-assisted translation system consisting of a source language analysis component, an interactive augmentor, and a target language generation component.

During the initial phase, the CRL's objectives were to develop tools for constructing lexical items and ontological entries automatically from on-line resources, to develop the initial Spanish analysis component, and, jointly with CMT and ISI, to establish the infrastructure for the three site project, develop the formats and initial content of the interlingua, the ontology, and the knowledge base, and to prepare design documents for the second phase versions of the analysis and generation components, the augmentor, and the translator's workstation.

The second phase is a two-year program, and we are currently in the first six months of this phase. Building on the results of the initial phase, the second phase calls for the construction of a new analysis component, with a considerably broader base than the first year analysis sytem, and incorporating a wider variety of fail-soft techniques; development and incorporation of the ontology into the year-two system; use of the jointly-developed interlingua; continued emphasis on automatic acquisition of lexical and semantic information.

RECENT RESULTS

At this point, a first version of the year-two analysis system has been completed and is undergoing expansion and refinement. The initial ontology has been constructed, and a large number of sense-tokens have been incorporated. A phrasal lexicon has been extracted from on-line dictionary resources and is being prepared for incorporation into the Translator's Workstation.

The second year analysis sytem begins with a dictionary-based part-of-speech tagger, followed by a component which chunks the tagged text into small syntactic sections. These are analyzed and semantic/lexical information accessed and incorporated into the representation by a constituent parser. These smaller constituents will be grouped into predicates and arguments, which are then further grouped into clausal structures. At each level, possible readings are rated for syntactic and semantic likelihood.

With respect to automatic acquisition, several advances have occurred in the past year. We have provided the ontology with a sense-disambiguated hierarchy of nominal word senses drawn from *Longman's Dictionary of Contemporary English*. The hierarchy is derived from disambiguated genus terms and rooted in the semantic categories provided by Longman's. We have gathered from our bilingual dictionaries a large number of phrases correlated with translations for that phrase. Verb classes for Spanish verbs, used as the basis for the morphological analysis program that feeds the part-of-speech tagger, were provided by *Collin's Spanish-English/English-Spanish Dictionary*. In a final note, the CRL has contracted with EFE (Spanish newswire service) for a continuous line feed.

PLANS FOR THE COMING YEAR

By next fall we hope to extend the project to a second source language (Japanese). During the remainder of phase 2, we will deepen and broaden the coverage of the analysis system. Broadening will include further addition of domain-specific lexical items (and a new domain) and the inclusion of proper name recognizers of various types (gazetteers, company names, personal names and titles). Deepening will involve working with the ontology to sharpen the semantic coherence judgments–providing fewer but more likely analyses for each input sentence.