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) in one language (Spanish initially) 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.

In the initial phase, the CRL's objectives are 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.

With respect to developing tools for extracting information from on-line resources, during the first year we are focusing on providing general classificatory information for use in constructing the lexicons of the Spanish analysis and English generation components. Here we are building on work on automatically constructing interlingual word sense specifications from Longman's which we are extending and adapting to language particular lexical entries. With respect to the Spanish analysis component, the objective is to modify and extend the Spanish component of the CRL's multilingual machine translation system both in terms of coverage and robustness.

RECENT RESULTS

At this point, the three sites have established the central infrastructure for the Pangloss project. For the first year system, the formats of the interlingua, the ontology, and the knowledge base have been set and initial design of the interlingua for the second phase version is underway.

As a preliminary to the work on extracting information from on-line resources, we continue to gather resources in the form of monolingual and bilingual dictionaries and monolingual and bilingual corpora. We have obtained *Collins English Dictionary*, *Collins Bilingual Spanish-English Dictionary* and are looking into a Spanish monolingual dictionary and Japanese monolingual and bilingual dictionaries. We have identified a source for Spanish texts in the financial domain and are seeking further sources as well as sources of Japanese texts.

The central resource to date, however, is the *Longman Dictionary of Contemporary English* and the information being extracted relates to word formation, syntactic category, syntactic subcategorization, and semantic selection restrictions. Procedures are currently being developed to provide this information in a form that can be used directly by the generation component and indirectly by the analysis component.

The effort to develop the Spanish component has primarily focussed on the analysis of sample Spanish newspaper articles in the financial domain in order to identify, first, the lexical items and constructions that are currently beyond the scope of the system, and, secondly, the types of "ill-formedness" which we are likely to face. Work on modifying and extending the core system began in December along with the development of procedures for diagnosing "ill-formedness" in the input string prior to analysis and for relaxing constraints during processing.

PLANS FOR THE COMING YEAR

In the coming year, our intention is to produce a prototype workbench for testing and evaluation in September. That will be follow by the extension of the prototype to a second source language (Japanese). In addition, greater amounts of automation will be introduced into each of the components of the MAT system and into the process of constructing lexical items and ontological entries. Lastly, in phase 2 we will be introducing pragmatics into the analysis and generation components.