PANGLOSS: 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 goals of the PANGLOSS project are to investigate and develop a new-generation knowledge-based interlingual machine translation system, combining symbolic and statistical techniques. The system is to translate newspaper texts in arbitrary domains (though a specific financial domain is given preference) to as high quality as possible using as little human intervention as possible.

The project involves three sites (USC/ISI, New Mexico State University, and Carnegie Mellon University). NMSU is responsible for Spanish parsing and lexicon acquisition; CMU for glossary and example-based MT translation, interlingua specification, workstation development, and system integration, and ISI for Japanese parsing and analysis, Spanish analysis, English generation, Japanese and English lexicon acquisition, and semantic term lexicon (Ontology) acquisition.

Within PANGLOSS, it is the particular focus of ISI to strive toward large-scale system coverage by investigating the feasibility and utility of combined statistical and human acquisition techniques of grammars, lexicons, and semantic knowledge. To this end, we have acquired several large resources, especially of Japanese lexical information, and are developing methods to integrate this knowledge with the ongoing development of Japanese parsing and semantic analysis and Ontology term acquisition and taxonomization.

RECENT RESULTS

The most recent ARPA evaluations of several MT systems, including PANGLOSS, are not yet available. However, preliminary measurements indicate that translators performed around 40% more quickly using the system than translating manually (for Spanish to English; the Japanese effort is only 6 months old at this time).

In recent work, we have:

- continued the construction of the PANGLOSS Ontology, the taxonomy of terms used in the semantic interlingua representation (the Ontology now contains approx. 50,000 items);
- acquired and deployed the lexical analyzer JUMAN and the parser SAX, with their accompanying 130,000-item wordlist;
- acquired a bilingual Japanese-English dictionary of approx. 70,000 entries and fully decoded its contents;
- acquired several other Japanese lexicons of various sizes and amounts of information;
- developed algorithms for linking Japanese lexical items to the Ontology;
- developed an English lexicon for our Penman sentence generator that contains approx. 70,000 items;
- developed several mappers that convert the output of one module of PANGLOSS into the input of another (all these mappers employ the same bottom-up unification-based chart parser);
- developed a collection of 200,000 statistically-based rules that govern the inclusion of the articles "the" and "a" into English text without articles (which is how it would come from Japanese).

PLANS FOR THE COMING YEAR

Our major efforts for the next year fall in four areas:

1. Japanese parsing, analysis, and lexis: the continued extension and testing of the current systems and lexicons;
2. Spanish semantic analysis: the development of the current mapper from the NMSU parser output to interlingua form into a more powerful and robust semantic mapper;
3. Ontology enrichment: the extraction of concept features and interrelationships from online resources and text, and their inclusion into the Ontology;
4. Sentence planning and English generation: the enhancement of the current interlingua-to-Penman mapper into a true Sentence Planner and the continued extension of the Penman generator.