PANGLOSS

Jaime Carbonell[1], David Farwell[2], Robert Frederking[1], Steven Helmreich[2],
Eduard Hovy[3], Kevin Knight[3], Lori Levin[1], Sergei Nirenburg[2]

1) Center for Machine Translation
Carnegie Mellon University
Schenley Park
Pittsburgh, PA 15213-3890

2) Computing Research Laboratory
New Mexico State University
Box 30001 / 3CRL
Las Cruces, NM 88003

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

The System

The PANGLOSS project is a three-way equal Machine Translation partnership funded since 1991 by the US Advanced Research Projects Agency (ARPA). The three participating partners are the Center for Machine Translation (CMT) at Carnegie Mellon University in Pittsburgh, the Computing Research Laboratory (CRL) at New Mexico State University in Las Cruces, and the Information Sciences Institute (ISI) of the University of Southern California in Marina del Rey.

Some central aspects of PANGLOSS are:

- The system contains three parallel translation engines for Spanish to English translation and two for Japanese to English translation. The three Spanish to English engines are:

  —lexical transfer (word-by-word and phrase-by-phrase substitution using a large bilingual glossary);

  —example-based MT (identification of phrases or even passages from the source text that appear in the large pre-constructed bilingual corpus of examples, and subsequent production of the translated phrases or passages as the target text);

  —knowledge-based MT (the more traditional route involving source text parsing, analysis, and generation, using an internal text representation that is gradually being upgraded to an Interlingua).

  Each source text sentence is translated by all three engines, who produce one or more candidate translations plus rating for their goodness. A final translation fragment selection and text smoothing module integrates the best results from the various engines into a resulting target text.

- Initially, the system was conceived as a human assistant, with gradual improvements and extensions until it would become the primary translator requiring only occasional human

assistance and moderate postediting. A human interface called the Augmentor was built for use by translation assistants with a passing but not deep familiarity with the source languages. Due to contingencies of the ARPA evaluation paradigm, however, research effort has focused on the development of fully automated translation at significantly reduced quality.

- The system involves a mixture of knowledge-based and statistical modules, as well as knowledge resources built up in various ways, by hand, from online dictionaries and knowledge bases, and through statistical extraction of knowledge from text.
- The knowledge-based engine involves an Interlingua to represent the contents of the text being translated. Interlingua terms are defined in a taxonomy of approximately 80,000 concepts called the Ontology.
- For parsing, CRL's parser PANGLYZER and Spanish grammar are used. The parser's output contains a mixture of syntactic and semantic information, following the theory of preference semantics. CRL is also responsible for the creation of the Spanish lexicon and the collection of other useful textual resources. The PANGLYZER is written in Quintus Prolog.
- For the remaining semantic analysis and construction of the Interlingua statements corresponding to the input, software from ISI is used. The Ontology is being constructed at ISI with assistance from CRL and CMT.
- For generation, ISI's PENMAN system is used. The software is written in Common Lisp.
- CMT is responsible for the system architecture, the operator interface (including the Augmentor, the CMAT human-assistance editor, the inclusion of WordPerfect and Emacs text editing tools, etc.) and the successful integration of all the modules. CMT software is written in Common Lisp.
- The Example-Based translation engine and resources are developed and built at CMT and CRL.
- The lexical transfer engine and its bilingual glossary was developed at CMT.

Despite the differences in theoretical approach and implementation, our work on PANGLOSS makes clear that systems as different as the PANGLYZER, PENMAN, the various knowledge-based MT components at CMT can share knowledge resources in a meaningful way.

The Presentation

This presentation will demonstrate the operation of the various engines of PANGLOSS on Spanish to English translation of newspaper texts. Since the complete system requires three Sun SPARC-Stations, only a scaled-down version of the system will be displayed. This version will however contain representative examples of the bilingual resources required by the EBMT and lexical transfer engines as well as all the intermediate steps produced in the chain of KBMT engine modules. The Ontology and various statistical routines will also be demonstrated. Finally, the operation of the CMAT editor and its function within the PANGLOSS Translator's Workstation will be shown.