Submitted for the ACL99 Demonstrations Program, University of Maryland, June 20-26, 1999.

Rapid Development of Translation Tools

Jan Amtrup, Karine Megerdoomian, Rémi Zajac
Computing Research Laboratory, New Mexico State University
Box 30001 Dept. 3CRL, Las Cruces NM 88003
{jamtrup,karine,zajac}@crl.nmsu.edu

See the CFP for demos at http://www.mri.mq.edu.au/conf/acl99/

1 Abstract

CRL is developing a technology that allows rapid deployment of translation capabilities. This technology is has been designed to handle low-density languages where resources (human or electronic) are scarce. The translation tools are built in an incremental fashion and some simple tools (bilingual dictionary, glosser) can be delivered early in the development to support analysis tasks. In particular machine readable dictionaries can be used for human translation and cross-language information retrieval. The original modular translation architecture also allows the delivery of more complex tools in several successive fully functional versions of increasing performance. We will demonstrate an application of this technology developed in the Shiraz Persian-English project as well as applications to Arabic, Japanese, Korean, Russian, Serbo-Croatian and Spanish.The set of translation tools that can be delivered in the framework are: A bilingual dictionary look-up tool (Dictionary Browser) for the human translator; A Glosser that look-up the dictionary for a whole text; A Spell-Checker that allow to correct a text before glossing or automatic translation; A Dictionary Editor that allow to enter new words in the system's dictionary; A Translator that produces an automatic translation of a source text at various levels of quality.

The Shiraz System (http://crl.nmsu.edu/shiraz/)

The aim of the project is to develop a machine translation system which will translate Persian text into English. The source material text is from on-line text or input from a Persian word processor and the output will be to a commercial English word processor. The system is running on Unix as well as Windows machines. The system uses an electronic bilingual Persian to English dictionary consisting of approximately 50,000 terms, plus a list of common proper names. This word and phrase lexicon is delivered as part of the machine translation system as well as a stand-alone tool running on PCs. The machine translation system is targeted at translating news material and miscellaneous types of documents. We collected a large Persian corpus of on-line material (approximately 10MB) and we have developed a bilingual tagged corpus to be used to test the MT system. This aligned test corpus includes morphological as well as syntactic tagging for both Persian and English. The dictionary has been built by a team of Persian lexicographers and includes single words, compounds and phrasal expressions. The dictionary contains information about the orthography, morphosyntactic category and syntactic properties of lexical items as well as the English word-sense equivalents.

English-Persian Dictionary

The system uses an electronic bilingual Persian to English dictionary consisting of approximately 50,000 terms that include single words, fixed idioms, compounds or phrases, some common proper names and functions words such as prepositions. The dictionary contains information about the orthography, morphosyntactic category and syntactic properties of lexical items as well as the English word-sense equivalents. The source dictionary is accessible through a Web-based dictionary browser. The compiled version of the dictionary is part of the machine translation system. The browser and the dictionary manager are written in Java and run on PC as well as Unix platforms.

The dictionary editor allow the user to enter new words in the dictionary. A file mode allows to edit a list of entries in a file as prepared by a linguist (or as produced by the spell-checker). After editing, the list entries can be revised or loaded in the dictionary proper. A database mode allows to edit entries directly in the dictionary. After edition, the dictionary can be compiled to produce the binary files used by the glosser, the spell-checker and the translator.

Glosser

The Glosser performs an automatic dictionary look-up for all words of a text. The Glosser uses a Persian tokenizer and a Morphological Analyzer to compute the citation form of inflected words. Input texts can be encoded in Unicode or Isiri. The results are presented using a Web-based interface. The source text and the glosses are presented in the left frame. The source text itself is presented vertically in the left column, and the English equivalents retrieved from the dictionary in the right column. The user can click on an English word to see the dictionary entry for that word (displayed in the right frame).

Spell-Checker

The spell-checker suggests alternatives for words which are not found in the dictionary. The user has the choice of correcting the word or marking it for future addition in the dictionary. The spell-checker uses a Web-based interface which currently uses a transcription scheme to display the text (a version supporting the Persian Script is under development).

Translator

The translator uses a robust and modular transfer-based architecture where a sentence is translated at the highest possible level of analysis, resorting to lower levels in case of failure at higher levels: the lowest possible level is a word-for-word translation. In comparison to the glosser, word-for-word translation maps morphological and lexical features of a source word to corresponding English morphological features. It then uses the English generator to produce the correct inflected form of words; finally, an English language model to select the most likely translation in a given context. The architecture allows incremental improvement of the system by adding or improving on various linguistic component without the need for modifying existing components. For example, the syntactic parser improves the translation by allowing a generic transfer component to produce the correct English word-order directly from the source parse tree.

Acknowledgments This research has been funded in part by DoD, Maryland Procurement Office, MDA904-96-C-1040.

References

Michelle Vanni and Rémi Zajac. 1997. “Glossary-Based MT Engines in a Multilingual Analyst’s Workstation Architecture”. Machine Translation 12, Special Issue on New Tools for Human Translators. pp131-157.

Rémi Zajac. 1998. “The Habanera Lexical Database Management System”. First International Conference Language Resources and Evaluation. Granada, Spain, 28-30 May 1998.

Rémi Zajac and Jan Amtrup. 1998. “Feature Structures, Unification and Finite-State Transducers”. Submitted for Computational Linguistics.

Rémi Zajac and Karine Megerdoomian. 1999. “Resolving Ambiguity Using Ontological Knowledge:the Case for Persian”. Submitted for TMI99.