LINGSTAT: AN INTERACTIVE, MACHINE-AIDED TRANSLATION SYSTEM*

Jonathan Yamron and James Baker

Dragon Systems, Inc., 320 Nevada Street, Newton, MA 02160

PROJECT GOALS

The goal of LINGSTAT is to produce an interactive machine translation system designed to increase the productivity of a user, with little knowledge of the source language, in translating or extracting information from foreign language documents. This system will make use of statistical information gathered from parallel and single-language corpora, and linguistic information at all levels (lexical, syntactic, and semantic). Initial efforts have been focused on the translation of Japanese to English, but work has also begun on a Spanish version of the system. As resources become available, particularly parallel corpora, the Spanish system will be further developed and work will be extended to include other European languages.

RECENT RESULTS

Productivity tests have been conducted on the rudimentary Spanish version of the workstation. This system incorporates a Spanish de-inflector, provides word for word translation to English, and has fast access to an online dictionary. On a scaled down version of the DARPA test of 7/92 (6 documents instead of 18, including 3 by hand and 3 with the aid of the system), a fluent speaker of Italian (a language very similar to Spanish) showed no productivity gain. At the other extreme, a user with no Spanish knowledge and no recent training in any European language was about 50% faster using the system's online tools than with a paper dictionary.

There are currently two programs underway to improve the translation system. The first is an effort to expand the Japanese and Spanish dictionaries, which requires not only adding words, but also glosses, pronunciations (for Japanese), and multi-word objects. Part of this task involves updating the Japanese and Spanish word frequency statistics, which will improve the performance of the tokenizer in Japanese and the de-inflector in both languages. Part of speech information is also being added, in anticipation of the use of grammatical tools.

*This work was sponsored by the Defense Advanced Research Projects Agency under contract number J-FBI-91-239

The second program is the development of a probabilistic grammar to parse the source and provide grammatical information to the user. This will supplement or replace the current rule-based finite-state parser currently implemented in the system. In the current phase, Dragon has chosen a lexicalized context-free grammar, which has the property that the probability of choosing a particular production rule in the grammar is dependent on headwords associated with each non-terminal symbol. Lexicalization is a useful tool for resolving attachment questions and in sense disambiguation. This grammar will be trained using the inside-outside algorithm on Japanese and Spanish newspaper articles.

PLANS FOR THE COMING YEAR

The grammar will be used to provide more accurate glossing of the source by making use of co-occurrence statistics among the phrase headwords. This requires developing an English word list with frequency and part of speech information, as well as constructing an English inflector-deinflector. These tools, along with an English grammar, will enable us to construct candidate translations of Japanese phrases and simple Spanish sentences.

For Japanese sentences and more sophisticated Spanish, Dragon plans to implement lexicalized tree-adjoining grammars in both source and target. Tree-adjoining grammars provide a rich framework for handling the difficult rearrangement of Japanese syntax into English. As in the case of context-free grammar, lexicalization helps keep the grammar small by resolving attachment and disambiguation questions. A translation can be constructed by transferring a parse in the source grammar into a parse in the (synchronized) target grammar.

Each of the analysis methods may produce several candidate translations of phrases or sentences in the target language. All of these candidates will then be rescored using a statistical language model in the target language, as well as translation and alignment probabilities with the source text. The maximum likelihood candidate will then be chosen for display to the user, who may then ask for more information or alternate translations.