The Automatic Component of the LINGSTAT Machine-Aided Translation System*

Jonathan Yamron, James Cant, Anne Demedts, Taiko Dietzel, and Yoshiko Ito

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

PROJECT GOALS

LINGSTAT is an interactive machine-aided translation system designed to increase the productivity of a translator. It is aimed both at experienced users whose goal is high quality translation, and inexperienced users with little knowledge of the source whose goal is simply to extract information from foreign language text. The system makes use of statistical information gathered from parallel and single-language corpora, but also draws from linguistic sources of knowledge. The first problem to be studied is Japanese to English translation, and work is progressing on a Spanish to English system.

RECENT RESULTS

In the newest version of LINGSTAT, the user is provided with a draft translation of the source document, which may be used for reference or modified. The translation process in LINGSTAT consists of the following steps: 1) tokenization and morphological analysis; 2) parsing; 3) rearrangement of the source into English order; 4) annotation and selection of glosses.

After tokenization and de-inflection (described in an earlier report), a source document passes through a two-stage parsing process. The first stage implements a coarse probabilistic context-free grammar of a few hundred human-supplied rules acting on parts of speech. Because of this coarseness, some parsing ambiguities remain to be resolved by the second-stage parser, which implements a simple, lexicalized, probabilistic context-free grammar trained on word co-occurrences in unlabeled Japanese sentences without human input.

The next step in the translation process is a transfer of the parse of each Japanese sentence into a corresponding English parse, giving an English word ordering. This is accomplished through the use of English rewrite rules encoded in the Japanese grammar, which implies an English grammar. The rewrite process just consists of taking the Japanese parse and expanding in this English grammar.

The final step is to apply a trigram model to select the best gloss from among the many candidates for each word, by making the choices that maximize the average probability per word. The trigram model used was trained on Wall Street Journal text.

The January 1994 ARPA machine translation evaluation has recently been completed. In this test, Dragon used the same translators as in the May 1993 evaluation and provided them with essentially the same interface and online tools. The difference in this evaluation was that the translators were also provided an automatically generated English translation of the Japanese document as a first draft. Manual and machine-assisted translation times were measured, and the automatic output was also submitted for separate evaluation.

Preliminary timing results show a speedup by a factor of 2.4 in machine-assisted vs. manual translation, which is essentially unchanged from the May 1993 result. This suggests that the draft translation was of no significant help to the translators in this evaluation, probably because the quality of automatic output is not high enough to be relied upon. A quality measurement of the automatic output is not yet available.

PLANS FOR THE COMING YEAR

The two independent grammars (probabilistic context-free and lexicalized) that comprise the current syntactic analysis must be merged and trained together. Attempts to do this have so far resulted in an unacceptable increase in training and parsing time due to the complexity of the algorithm.

This newest version of the system must be ported to Spanish for the next evaluation, scheduled for June. This will require improvements to the Spanish dictionary and de-inflector, an update of the Spanish grammar from the older Spanish system, a lexicalized grammar trained on Spanish text, and Spanish rewrite rules. We intend to use the parallel Spanish-English component of the UN data to provide gloss information.