Applying Statistical Methods to Machine Translation

Peter F. Brown, Principal Investigator

IBM / T.J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

PROJECT GOALS

The goal of our project is to demonstrate the effectiveness of statistical techniques in machine translation by improving the state of the art in large-vocabulary French-to-English translation.

A common paradigm in machine translation is analysis, transfer, and synthesis. In French-to-English translation, for example, a French sentence is analyzed into an intermediate structure in which various ambiguities present in the surface form have been resolved. This structure is then transferred to a similar English structure. Finally, an English sentence is synthesized from the intermediate English structure. Analysis, transfer, and synthesis each require considerable linguistic insight for their successful dispatch.

The approach taken in this project is to incorporate models of statistical transfer into the analysis-transfer-synthesis paradigm. This will be done by constructing deterministic, invertible transformations of the surface forms in the bilingual data which improve the locality of the transfer process and which reduce its variety. An example of the former is part-of-speech labeling, which is best performed by an examination of the global properties of the sentence and simplifies transfer by distinguishing, for example, between 'le' the article, and 'le' the direct object. An example of the latter is morphological analysis which exposes the fraternity of different forms of the same root and obviates the discovery of separate statistics relating these different forms in the two languages.

RECENT RESULTS

Developed Translator's Workstation.
In the spring of 1992, we wrote TransMan, a post-editor's workstation, which permits a human translator to rapidly edit translations produced automatically by a machine. TransMan allows cutting and pasting of sections of machine translations, and rapid access to an on-line dictionary. In a July of 1992 DARPA evaluation, TransMan was used by human subjects to translate 35 percent more rapidly than they could translate without machine assistance.

Speech Recognition in Machine Translation.
It has been observed that humans can translate nearly four times as quickly with little loss in accuracy simply by dictating, as opposed to typing, their translations. We considered the integration of speech recognition into a translator's workstation. In particular, we showed how to combine statistical models of speech, language, and translation into a single system that decodes a sequence of words in a target language from a sequence of words in a source language together with an utterance of the target language sequence. We obtained results which demonstrate that the difficulty of the speech recognition task can be reduced by making use of information contained in the source text being translated.

PLANS FOR THE COMING YEAR

In the July evaluation of our system we found a number of sources of errors. Some of the major problems include limited vocabularies, translation of numbers, translation of names, morphological errors, errors in syntactic transformations, and limitations in our trigram language model.

We plan to construct name and number detectors and translators. These module will detect names and numbers in French text, and replace them by name and number markers. The markers will then be translated statistically into English name and number markers. The English markers will then be translated by rule into English names and numbers. Incorporating these modules into our system will permit our language model to 'see back through' multi-word names and numbers.

We plan to increase the number of surface forms in our French vocabulary from 60,000 to 280,000 and the number of surface forms in our English vocabulary from 40,000 to 70,000. We will also completely reconstruct our morphological tables for both French and English.

A number of new rule-based syntactic transformations will be added to account for problems encountered in the July evaluation.

Finally, we plan to incorporate a decision-tree language model into our system which can make predictions based on significantly more context than our existing trigram model.