Candide: A Statistical Machine Translation System

Stephen DellaPietra and Vincent DellaPietra, Principal Investigators

sdella@watson.ibm.com, vdella@watson.ibm.com
IBM T.J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

PROJECT GOALS

The Candide project has two objectives. First, we want to develop a fully-automatic, large vocabulary, French-to-English translation system. Second, we want to develop an interactive translator's workstation that will increase the speed and productivity of a human translator. The philosophy of the project is to combine, within a probabilistic framework, both statistical information acquired automatically from bilingual corpora and linguistic knowledge provided by human experts.

RECENT RESULTS

System Overview

The automatic system is organized around an analysis-transfer-synthesis architecture. A French sentence is first analyzed into an intermediate structure in which various linguistic components are identified. This structure is then transfered to a corresponding structure in English. Finally, an English sentence is synthesized from the intermediate English.

The heart of the system is the transfer component. We view transfer from an information-theoretic perspective and attack it with statistical modeling techniques. Candide's transfer component incorporates a stochastic language model that estimates the probability of an intermediate English structure; a stochastic translation model that estimates the conditional probability of a French structure given an English one; and a decoder that, given a French structure, searches for that English structure which maximizes the product of the language model and translation model probabilities.

Recent Improvements

Over the past year we improved Candide on several fronts. We increased the sizes of our English and French vocabularies to 70,000 and 280,000 respectively. We improved the analysis phase by overhauling the morphological tables, refining the treatment of numerical expressions and proper names, improving various syntactic transformations, and improving the statistical bilingual sense disambiguation module.

We enhanced the translation model by using maximum entropy techniques to make it more sensitive to context. We developed a new language model based upon the recently formulated notion of link grammar. The link grammar model takes into account long range correlations between words in a sentence, thus attacking a well-known limitation of our previous trigram model.

Finally, we re-trained all our stochastic models using more text, including 100 million words from the Wall Street Journal and UPI.

Current Performance

The table summarizes the gain in performance of our system between the July 1992 and July 1993 ARPA sponsored machine translation evaluations.

	Fluency		Adequacy		Time Ratio	
	1992	1993	1992	1993	1992	1993
Fully Automatic	.511	.580	.575	.670		
Machine Aided	.819	.838	.837	.850	.688	.625
Manual		.833		.840		

Fluency and adequacy are scored on scales from 0 to 1, with 1 being the highest score. The time ratio is the factor by which the use of the machine-aided system decreases the time spent by a human translator.

We also evaluate our automatic translation system on sentences of 15 words or less randomly chosen from a test corpus of Canadian parliamentary proceedings. The 1992 version of the system translates 45% of the sentences accurately, while the 1993 system translates 62% of the sentences accurately.

PLANS FOR THE COMING YEAR

Our strategy is two-pronged. First, we want to bring the current incarnation of Candide to its full potential, thus establishing a benchmark for our future research. Second, we want to begin developing a new version of Candide that uses more sophisticated intermediate structures.

For the current version, we will enlarge the vocabularies and make incremental corrections to several analysis and synthesis components. We will also speed-up the decoder so that we can perform tests more quickly. For the next version, we will continue work on the link grammar language model. We will also begin developing a new translation model that takes advantage of the enhanced intermediate structures.