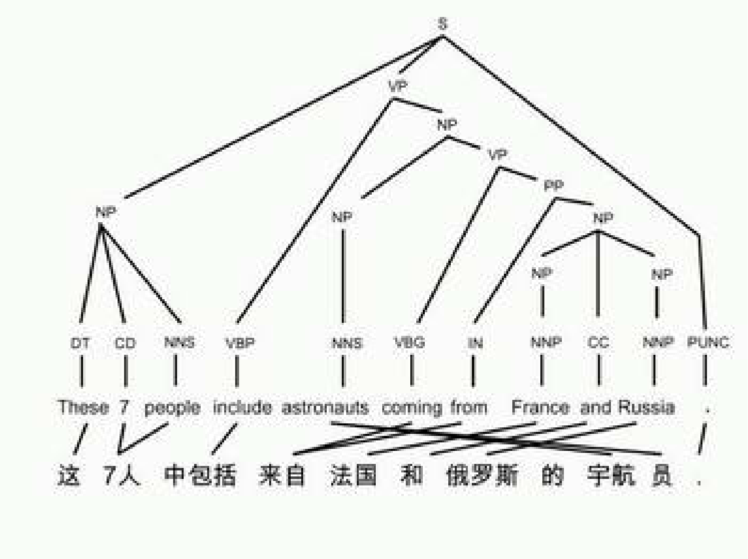
Grammar Lost Translation Machine In Researchers Fix Will

September 9 2005



The makers of a University of Southern California computer translation system consistently rated among the world's best are teaching their software something new: English grammar.

Image: A Tree Grows in Translation Grammatical structure, long in second place, is emerging as a key to better English in the finished product.

Most modern "machine translation" systems, including the highly rated one created by USC's Information Sciences Institute, rely on brute force correlation of vast bodies of pre-translated text from such sources as newspapers that publish in multiple languages.

Software matches up phrases that consistently show up in parallel fashion — the English "my brother's pants" and Spanish "los pantalones de mi hermano," — and then use these matches to piece together translations of new material.

It works — but only to a point. ISI machine translation expert Daniel Marcu (left) says that when such a system is "trained on enough relevant bilingual text ... it can break a foreign language up into phrasal units, translate each of them fairly well into English, and do some re-ordering. However, even in this good scenario, the output is still clearly not English. It takes too long to read, and it is unsatisfactory for commercial use."

So Marcu and colleague Kevin Knight (right), both ISI project leaders who also hold appointments in the USC Viterbi School of Engineering department of computer science, have begun an intensive $285,000 effort, called the Advanced Language Modeling for Machine Translation project, to improve the system they created at ISI by subjecting the texts that come out of their translation engine to a follow-on step: grammatical processing.

The step seems simple, but is actually imposingly difficult. "For example, there is no robust algorithm that returns 'grammatical' or 'ungrammatical' or 'sensible' or 'nonsense' in response to a user-typed sequence of words," Marcu notes.

The problem grows out of a natural language feature noted by M.I.T. language theorist Noam Chomsky decades ago. Language users have literally a limitless ability to nest and cross-nest phrases and ideas into intricate referential structures — "I was looking for the stirrups from the saddle that my ex-wife's oldest daughter took with her when she went to Jack's new place in Colorado three years ago, but all she had were Louise's second-hand saddle shoes, the ones Ethel's dog chewed during the fire."

Unraveling these verbal cobwebs (or, in the more common description, tracing branching "trees" of connections) is such a daunting task that programmers long ago went in the brute force direction of matching phrases and hoping that the relation of the phrases would become clear to readers.

With the limits of this approach becoming clear, researchers have now begun applying computing power to trying to assemble grammatical rules. According to Knight, one crucial step has been the creation of a large database of English text whose syntax has been hand-decoded by humans, the "Penn Treebank."

Using this and other sources, computer scientists have begun developing ways to model the observed rules. A preliminary study by Knight and two colleagues in 2003 showed that this approach might be able to improve translations.

Accordingly, for their study, "We propose to implement a trainable tree-based language model and parser, and to carry out empirical machine-translation experiments with them. USC/ISI's state-of-the-art machine translation system already has the ability to produce, for any input sentence, a list of 25,000 candidate English outputs. This list can be manipulated in a post-processing step. We will re-rank these lists of candidate string translations with our tree- based language model, and we plan for better translations to rise to the top of the list."

One crucial trick that the system must be able to do is to pick out separate trees from the endless strings of words. But this is doable, Knight believes -- and in the short, not the long term.

Referring to the annual review of translation systems by the National Institute of Science and Technology, in which ISI consistently gains top scores, "we want to have the grammar module installed and working by the next evaluation, in August 2006," he said.

Knight and Marcu are cofounders and, respectively, chief scientist and chief technology and operating officer of a spinoff company, Language Weaver.

Source: University of Southern California

Citation: Grammar Lost Translation Machine In Researchers Fix Will (2005, September 9) retrieved 15 August 2024 from https://phys.org/news/2005-09-grammar-lost-machine.html