The Elusive Goal of MACHINE TRANSLATION

Statistical methods hold the promise of moving computerized translation out of the doldrums

By Gary Stix

> Natrium Nepal Asia legend: The lion, the sorceress, the evil spirit wardrobe "already lack" the evil spirit abstains the trilogy "rich in poetic and artistic flavor, also has not let" the Harley baud "the series novel have the infinite pleasure the undercurrent to be turbulent.

The preceding gibberish was brought to you by a Chinese-to-English translation carried out by Altavista's Babelfish, the popular Internet-based translator. In coherent English, from a bilingual page on the Web site of Taiwan's *China Post*, it reads:

> "The Chronicles of Narnia" doesn't come near the poetic vision of "The Lord of the Rings" trilogy, and it doesn't have the dark undercurrents that makes the "Harry Potter" series endlessly fascinating.

This passage illustrates that machine translation, or MT, as it is known, remains one of the more challenged subdisciplines of the blighted field of artificial intelligence. A proper name or a few well-crafted phrases suffice to throw the software off track. In the past few years, though, a new research approach has fueled a revival for machine translation: brute-force computing methods—which gauge the probability that a word or phrase in one language matches that in another—are at last bringing MT closer to human performance, in the estimation of developers of this software.

Tougher Than Chess

THE EVER INCREASING POWER of hardware and software algorithms today has propelled the computer past the chess grandmaster. (Recall that IBM's Deep Blue supercomputer triumphed over Garry Kasparov in 1997.) But on the whole, machine translation has experienced only halting progress in achieving humanlike capabilities in its more than 50-year history—and some critics would classify even that characterization as overly generous.

In 1954 IBM and Georgetown University demonstrated the translation of more than 60 sentences from Russian into English. The IBM press release, dated January 8, 1954, glowed: "Russian was translated into English by an electronic 'brain' today for the first time." The military defense community and computer scientists expected routine machine translation within five years, but it never materialized.

In 1966 the U.S. government–sponsored Automatic Language Processing Advisory Committee reported that humans could perform faster, more accurate translation at half the cost. "There is no immediate or predictable prospect of useful machine translation," its study concluded.

SLIM FILMS

Funding dried up, and only modest advances came in subsequent decades. In the late 1960s the U.S. Air Force supplied support to a small company that created the machine translator called Systran—the Internet version of which provided the first paragraph of this article—to cope initially with voluminous demands to translate Russian documents into English.

Systran is based on rules about the source and target languages, as was IBM's original "brain" system, which relied on six rudimentary rules that govern syntax, semantics and the like. For example, the word "*o*" in Russian could be translated by an IBM 701 computer as either "about" or "of." If "*o*" followed the word "*nauka*" (science), it looked for the appropriate rule that told it to translate "*o*" as "of"—in other words, the "science of," not the "science about."

The Paris-based Systran company ranks as the biggest machine translation company in the world. Even with customers that include Google, Yahoo and Time Warner's AOL, its annual revenues were just $13 million for 2004—in an overall market for translations of all varieties that is estimated worldwide to total nearly $10 billion. "We're so small, and we're the largest," says Dimitris Sabatakakis, Systran's chairman and chief executive officer.

No More Rules

FOR RULE-BASED SYSTEMS, language experts and linguists in specific languages have to painstakingly craft large lexicons and rules related to grammar, syntax and semantics to generate text in a target language. Commercial systems contain tens of thousands of grammar rules for a corpus that is made up of hundreds of thousands of words.

Beginning in the late 1980s, IBM created a system for translating French into English called Candide that required knowledge of neither grammar nor syntax. It eschewed rules in favor of taking substantial bodies of already translated text, matching words between the two languages (more recent systems use whole phrases) and finally deriving probabilities—based on Bayes's theorem—to estimate whether an English word was a correct translation from the French.

Another analysis that relied solely on large English texts assessed whether the word translated into English fit in grammatically with surrounding words. The word or phrase in the target language accorded the highest probability could then be used to "decode" future texts—and multiple words could be linked to build entire documents. If the statistics showed that the word "*pouderie*" usually equated to "blowing snow," that, in principle, was all that was needed.

IBM eventually dropped its effort. At the end of the 1990s it could take an entire day for a machine translation of a single page. But then things began to stir. The Internet produced a rapid growth in the number of large, bilingual bodies of text. The Web also created demand for translation that could never be met by humans.

In 1999 the National Science Foundation held a workshop at Johns Hopkins University to construct a software tool kit that could be readily disseminated to the scientific community, an action that drew attention and spurred new activity. In 2002 one of the workshop organizers, Kevin Knight of the University of Southern California, and Daniel Marcu, also at U.S.C., founded Language Weaver, the only statistical machine-translation company. It now claims to be capable of translating at least 5,000 words a minute back and forth between English and Arabic, Farsi, French, Chinese and Spanish.

Google Is a Winner

ANOTHER ALUMNUS of both the workshop and U.S.C., Franz Och, was hired by Google. Last summer the still experimental Google system engineered by Och bested competitors such as IBM to win every category in a competition organized by the National Institute of Standards and Technology to translate 100 newswire documents from Arabic

STATISTICAL MACHINE TRANSLATION

INPUTTING ALREADY-TRANSLATED TEXTS

Existing translated texts from various sources form the foundation of the automated translations.

Translated archives — Dictionary and glossaries — Internet — Human translators

Statistical methods have proved to be more effective than other types of automated machine translations based on rules crafted by human translators.

The new methods take advantage of the brute-force calculating power of machines to crunch through existing translated texts to determine the probability that a word or a phrase in one language matches that in another.

PREPROCESSING

The texts are scanned, aligned and formatted.

`Que hambre tengo yo.`

Preprocessor

`que hambre tengo yo`

PHRASE MATCHING IN TRANSLATED TEXTS

A translation model picks out two- or three-word phrases from the source language (in this case, Spanish) that match the target language (English).

Source language: *Spanish*

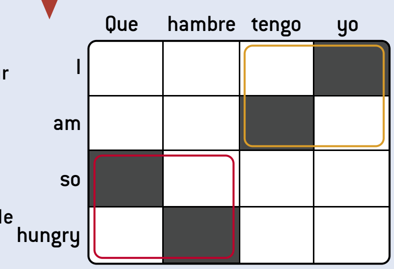
`Este guiso tradicional se ennoblece con el bogavante, la viera y el rodaballo. Que hambre tengo yo.`

Target language: *English*

`This traditional stew is refined with scallops, lobster and turbot. I am so hungry`

TRANSLATION MODEL

By using statistics to measure how often and where words occur in a given phrase in both languages, the model derives a template for word reordering. It also takes advantage of other techniques, such as reducing multiple Spanish words to a single translated word (*not shown*).



LANGUAGE MODEL

Working from its own statistical analyses of English-only texts, a language model attempts to predict the most likely word and phrase ordering for the already-translated text. Greater frequency of a phrase's occurrence increases the probability that it is correct.

`I am so baffled by Modern and Postmodern art.`

`The boy is so thirsty and the mother so sad.`

`I am so pleased to meet you.`

`"I am so hungry to see everything, and to know everything," she said to herself.`

`He cried, "I am so hungry! Will you give me food?"`

`I have so many people to thank.`

`What strength have I that I should endure?`

I am so > Have I that
I am so > I have so
So thirsty > Thirsty
Am so hungry > What hunger have

DECODER

When a new sentence gets inputted—one that can differ slightly or substantively from the text already processed (only *sed* substitutes for *hambre* here)—the decoder develops several hypothetical translations and picks the one with the highest probability.

INPUT TEXT

Que sed tengo yo.

DECODER

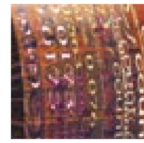
I am so thirsty	$P = 0.13$
What thirst have I	$P = 0.09$
Have I what thirst	$P = 0.07$
Thirsty I am so	$P = 0.00$

TRANSLATION

I am so thirsty.

LUCY READING-IKKANDA; SOURCES: KEVIN KNIGHT *Language Weaver* AND PHILIPP KOEHN *University of Edinburgh*

Can a machine translator provide more than "gisting," a rough idea of the contents of a foreign-language text?



or Chinese into English. Och has mentioned that feeding the machine-translation software with text that equated to one million books was key to performance improvements. He contrasted Google's current Chinese-to-English MT system (Systran) with the experimental statistical one crafted by him and his co-workers:

> Google/Systran: "Doctor indicates, the bright kernel prearranges recuperates the about one month."
>
> Google Research: "Doctors said Akihito is scheduled to rest for about a month."

The buzz about statistical machine translation has put Systran on the defensive. "You need rules when learning a foreign language," Sabatakakis comments. "You don't learn a language with statistical methods." Systran uses statistical techniques when creating systems in very narrow domains, such as translating patent documents. But the current embrace of statistical methods is somewhat of a marketing technique, he says. The company still employs 50 people in research and development, among them linguists. "The major difference between Systran and Google is that Google claims that it doesn't need native Chinese people to develop Chinese [applications] because of the magic and beauty of this stuff," Sabatakakis says, adding, "If we don't have some Chinese guys, our system may contain enormous mistakes."

The distinction between the two camps has begun to blur a little as statistical MT researchers have started to incorporate techniques that account for the syntactical structure of a sentence. These methods forgo the intervention of a human linguist: a syntactic model might estimate the chance that an English adjective-noun phrase gets reordered after translation into French. Knight of Language Weaver says that relying on phrases instead of single words allows the statistics to deal with semantics as well, avoiding, for instance, having his surname translated as "*Caballero*."

Microsoft Research has a substantial natural-language group, which for the past six years has also worked on MT. The group first focused on rule-based systems. But it is increasingly incorporating statistical techniques. Recently Microsoft used primarily statistical approaches when translating its online customer-support Web sites into 12 new languages, including Russian, Arabic and Chinese. The text does not get edited afterward. "Some of it is admittedly pretty rough; other parts of it are quite good," notes Steve Richardson, a senior researcher in the natural-language processing unit. "The quality of the more statistical approaches is comparable to or beginning to exceed that of the rule-based systems that we used before."

Getting the Gist

ALL THESE TECHNIQUES, however, raise the question of whether the machine-translation equivalent of a Deep Blue, the IBM chess computer, will ever beat humans at their own game. Can a machine provide more than mere "gisting," a rough idea of the contents of a foreign-language text? Kevin Hendzel, a spokesman for the American Translators Association, says that the current optimism only promulgates decades' worth of overhyped claims—FAHQT, the idea of "fully automatic high-quality translation," for instance. Gisting can help sort through massive amounts of foreign-language texts as long as it is understood to be inherently unreliable, he notes. Even a rough translation has its perils. He cites one Arabic-to-English translation that mentioned two sides "going at" each other, a fragment that caught the attention of security officials. The reference turned out to be for a soccer game, not a terrorist attack or imminent battle.

Keith Devlin, executive director of Stanford University's Center for the Study of Language and Information, remarks that machine-based systems will never equal the human linguist. "The use of statistical techniques, coupled with fast processors and large, fast memory, will certainly mean we will see better and better translation systems that work tolerably well in many situations," Devlin says, "but fluent translation, as a human expert can do, is, in my view, not achievable."

Knight, the pioneer in statistical translation, disagrees and points to the progress achieved during this decade. He foresees no limit to the technology, which will ultimately achieve human-level translations for everything except possibly poetry. He has shown blind examples of human translations alongside those from a machine, and audiences have confused the two. "Let's not kid ourselves—there are lots of mistakes in human-level translations. The bar is not as high as you would imagine," he says. To prove that this round of translation tools is more than the perennial sales pitch, the statistics jocks who now lead the field must demonstrate that this time FAHQT is real. Only then will the technology go beyond, as Microsoft's Richardson puts it, mere "MT promises." SA

MORE TO EXPLORE

The History of Machine Translation in a Nutshell. Online at John Hutchins's Web site: **http://ourworld.compuserve.com/homepages/WJHutchins/nutshell.htm**

A Statistical MT Tutorial Workbook. Kevin Knight. Online at **www.isi.edu/natural-language/mt/wkbk.rtf**

The Candide System for Machine Translation. Adam L. Berger et al. Online at **http://acl.ldc.upenn.edu/H/H94/H94-1028.pdf**