WHAT'S NEXT

From Uzbek to Klingon, the Machine Cracks the Code

By CHRISTOPHER JOHN FARAH

In the summer of 1999, at a workshop on statistical machine translation at Johns Hopkins University, Kevin Knight passed out a copy of an advertisement to each member of the research team he was leading. In the center of the ad was a picture of a yellowed, frayed parchment covered in Japanese characters. "To most people, this looks like a secret code," the ad announced. "Codes are meant to be broken."

The ad was for a product yet to be created called the Decoder. "Pour in a new bunch of text," said the ad's text, alongside a picture of a software box. "We think you'll be surprised."

The Decoder was meant to be a motivational tool. At the time, the field of statistical machine translation was all but dead. In the four years that have passed since that workshop, Dr. Knight, the head of machine translation research at the University of Southern California's Information Sciences Institute, is amazed by just how prophetic the ad has proved. "Here we are," he said. "It's no joke anymore."

Statistical machine translation - in which computers essentially learn new languages on their own instead of being "taught" the languages by bilingual human programmers - has taken off. The new technology allows scientists to develop machine translation systems for a wide number of obscure languages at a pace that experts once thought impossible.
In addition to the release of Egypt/Giza in 1999, the spread of the Internet has led to an explosion of translated texts in far-flung languages, greatly aiding the team’s research. Researchers have also benefited from a much faster means of evaluating the outcome of translation experiments: a computerized technique developed by I.B.M. enables researchers to test 10 to 100 new approaches for cracking languages each day.

The technique, known as the Bleu Metric, compares machine translations with a "gold standard" based on human translations. Instead of waiting for human beings to assign a score to the quality of a machine translation, the Bleu Metric does so almost instantly through a statistical comparison. This provides scientists with a fast, objective measurement that they can use to note improvement and saves them from having to review every unsuccessful experiment.

"Before Bleu, it was really a bad state of affairs," said Alex Fraser, a doctoral student at U.S.C. "You look at broken couplets of English for a long time, and eventually you start to accept it more and more."

Despite the progress being made in statistical machine translation, some researchers remain skeptical, preferring to focus their efforts on language-specific translation techniques. Ophir Frieder, a professor of computer science at the Illinois Institute of Technology, is working on a search system exclusive to Arabic text.
"Yes, N-grams work on any language, but as a search technique they work poorly on every language," he said. "It's a basic novice solution."

Dr. Knight acknowledges that statistical machine translation is far from perfect. In its latest efforts, his team has sought to combine the statistical and traditional approaches to achieve maximum accuracy and to produce translations that the average computer user can understand. The best machine translation systems today, while capable of yielding a passage's general meaning, are better known for their muddled syntax than their accuracy. By applying the principles of statistical translation to varying grammatical structures, Dr. Knight hopes to resolve some of these basic problems.

"N-grams are one of those things where you don't know how much you need it until you take it away," he said. "The way our imaginations work, we need help."
Dr. Knight and others said the progress and accuracy of statistical machine translation had recently surpassed that of the traditional machine translation programs used by Web sites like Yahoo and BabelFish. In the past, such programs were able to compile extensive databanks of foreign languages that allowed them to outperform statistics-based systems.

Traditional machine translation relies on painstaking efforts by bilingual programmers to enter the vast wealth of information on vocabulary and syntax that the computer needs to translate one language into another. But in the early 1990's, a team of researchers at I.B.M. devised another way to do things: feeding a computer an English text and its translation in a different language. The computer then uses statistical analysis to "learn" the second language.

Compare two simple phrases in Arabic: "rajl kabir" and "rajl tawil." If a computer knows that the first phrase means "big man," and the second means "tall man," the machine can compare the two and deduce that rajl means "man," while kabir and tawil mean "big" and "tall," respectively. Phrases like these, called "N-grams" (with N representing the number of terms in a given phrase) are the basic building blocks of statistical machine translation.

Although in one sense it was more economical, this kind of machine translation was also much more complex, requiring powerful computers and software that did not exist for most of the 90's. The Johns Hopkins workshop changed all that, yielding a software application package, Egypt/Giza, that made statistical translation accessible to researchers across the country.

"We wanted to jump-start a vibrant field," Dr. Knight said. "There was no software or data to play with."

Today researchers are racing to improve the quality and accuracy of the translations. The final translations generally give an average reader a solid understanding of the original meaning but are far from grammatically correct. While not perfect, statistics-based technology is also allowing scientists to crack scores of languages in a fraction of the time, and at a fraction of the cost, that traditional methods involved.

A team of computer scientists at Johns Hopkins led by David Yarowsky is developing machine translations of such languages as Uzbek, Bengali, Nepali - and one from "Star Trek."

"If we can learn how to translate even Klingon into English, then most human languages are easy by comparison," he said. "All our techniques require is having texts in two languages. For example, the Klingon Language Institute translated 'Hamlet' and the Bible into Klingon, and our programs can automatically learn a basic Klingon-English MT system from that."

Dr. Yarowsky said he hoped to have working translation systems for as many as 100 languages within five years. Although the grammatical structures of languages like Chinese and Arabic make them hard to analyze statistically, he said, it will only be a matter of time before such hurdles are overcome. "At some point, we start encountering the same problems over and over," he said.
Get The Times from $2.90 a week

RELATED ARTICLES

- Job Listings Decline 20% At Colleges, (December 14, 2002) $
- The Way We Live Now: 11-18-01; Other Woes (November 18, 2001) $
- About New York; At N.Y.U., Languages Si, Elitism Non. (December 10, 1997) $
- Language Study Shifts Again: Chinese Is Up, Russian Down. (October 9, 1996) $

Find more results for Language and Languages and Colleges and Universities.

TOP TECHNOLOGY ARTICLES

- To: Mom and Dad. Re: Homesickness
- State of the Art: This Palmtop Knows Its Place
- S.E.C. Investigates Bulk Sales of AOL Internet Subscriptions
- Online Shopper: Finding (or Honing) the Cutting Edge
- What’s Next: From Uzbek to Klingon, the Machine Cracks the Code

Copyright 2003 The New York Times Company | Home | Privacy Policy | Search | Corrections | Help | Back to Top