Cracking the code of machine translation

20 June 2011 by **Jacob Aron**
Magazine issue 2817. **Subscribe and save**

AUTOMATIC translation services seem like magic. Input some foreign text and you instantly get a decent English version in return - unless your text happens to be in Farsi, Pashto or any number of other widely used languages that computers can't currently translate.

What if there's no Rosetta stone? (Image: John Brecher/Corbis)

That's because machine translation techniques rely on analysing the statistical properties of the same text written in two different languages - a Spanish-English dictionary, for example. "You have parallel data for common language pairs like French-English, but for rare or uncommon language pairs it's very difficult to find bilingual sources," explains Sujith Ravi, a computer scientist at the University of Southern California in Marina Del Rey, who is trying a new approach to the problem.

Ravi and his colleague Kevin Knight treat translation as a cryptographic problem, as if the foreign text were simply English written in an advanced cipher. Their software cracks the code by estimating the probability that a foreign word matches an English word based on the number of times it appears in the text - a frequently occurring word is more likely to mean "the" or "a" than "antidisestablishmentarianism".

To ensure the translation makes sense, the pair use another piece of software to evaluate the quality of English that comes out. This in turn tweaks the probabilities used in the translation software. They tested the system on a collection of short phrases such as "last year" and "the fourth quarter", attempting to translate the Spanish equivalents back into English, along with a number of movie subtitles that existed in both languages.

The resulting translations - known, confusingly, as "monolingual" translations - rated highly compared with standard computer translation techniques. But it remains to be seen whether the models can be scaled up from such short phrases to deal with longer, more complex texts.

Chris Callison-Burch of Johns Hopkins University in Baltimore, Maryland, says Ravi and Knight's method is "extremely promising" but adds that it hasn't proved itself yet. His team is also working on translation software that eschews parallel data. Their version crawls online texts and compares disparate texts from different languages - say, a collection of Spanish blog posts and news stories in English. For example, the word "tsunami" spiked in 2004 and 2011 following the Indian Ocean and Japanese events, as did the equivalent word in Spanish, *maremoto*, suggesting that they mean the same thing.

Ravi and Knight are also exploring how monolingual methods could help us crack long-lost languages or unknown ciphers (see "Machine versus the Zodiac killer"). But what about the ultimate unknown tongue - could their methods translate an alien language? "Totally," says Ravi. "You could also think of deciphering dolphin-speak."

Monolingual translation might also help soldiers or aid workers react quickly in countries with unfamiliar languages; responding to a bombing in Indonesia or an earthquake in Haiti, for instance.

Don't expect a Google Translate upgrade just yet, though. "They're trying to do something very ambitious," says Phil Blunsom, a machine translation researcher at the University of Oxford. "It's not something you're going to see popping up in commercial systems any time soon."

Machine versus the Zodiac killer

Coded messages apparently sent by a San Francisco serial killer in the late 1960s have baffled cryptanalysts ever since, but Ravi and Knight's translation model could help crack the cipher. The Zodiac killer's code replaced letters with strange symbols and sometimes used multiple symbols for the same letter, making it very hard to decipher.

The first three messages were decoded by hand, revealing them to be parts of a single message, but the fourth remains unsolved to this day. Ravi and Knight's model has successfully cracked the first messages - the first time this has been achieved without human intervention - and they now hope to decode the fourth.