[From: *Wired* 14(12) 2006; pp.210-213]

Me Translate Pretty One Day

Spanish to English? French to Russian? Computers haven't been up to the task. But a New York firm with an ingenious algorithm and a really big dictionary is finally cracking the code.
By Evan Ratliff

JAIME CARBONELL, CHIEF science officer of Meaningful Machines, hunches over his laptop in the company's midtown Manhattan offices, waiting for it to decode a message from the perpetrators of a grisly terrorist attack. Running software that took four years and millions of dollars to develop, Carbonell's machine – or rather, the server farm it's connected to a few miles away – is attempting a task that has bedeviled computer scientists for half a century. The message isn't encrypted or scrambled or hidden among thousands of documents. It's simply written in Spanish: *"Declaramos nuestra responsabilidad de lo que ha ocurrido en Madrid, justo dos años y medio después de los atentados de Nueva York y Washington."*

I brought along the text, taken from a Spanish newspaper transcript of a 2004 al Qaeda video claiming responsibility for the Madrid train bombings, to test Meaningful Machines' automated translation software. The brainchild of a quirky former used-car salesman named Eli Abir, the company has been designing the system in secret since just after 9/11. Now the application is ready for public scrutiny, on the heels of a research paper that Carbonell – who is also a professor of computer science at Carnegie Mellon University and head of the school's Language Technologies Institute – presented at a conference this summer. In it, he asserts that the company's software represents not only the most accurate Spanish-to-English translation system ever created but also a major advance in the field of machine translation.

My test alone won't necessarily prove or disprove those claims. Carbonell, a native Spanish speaker with a froggy voice, curly gray beard, and rumpled-professor chic style, could translate it easily. But throw the line into Babel Fish, a popular Web translation site that uses software from a company called Systran – the same engine behind Google's current Spanish translation tool – and it comes out typically garbled: "We declared our responsibility of which it has happened in Madrid, just two years and means after the attacks of New York and Washington."

Carbonell's laptop churns for a minute and spits out its own effort, which he reads aloud from the screen. "'We declare our responsibility for what happened in Madrid' – a somewhat better translation would be 'We acknowledge our responsibility'" he interjects – "'just two and a half years after the attacks on New York and Washington.' So, no interesting errors there," he concludes. "It got it right."

LANGUAGE TRANSLATION is a tricky problem, not only for a piece of software but also for the human mind. A single word in one language, for example, may map into three or more in another. Carbonell likes to cite bank, with its utterly divergent uses for the place you keep your money, the edge of a river, and what an airplane might do. Then there are the dramatic differences in grammar and structure across languages. Arabic, for example, uses very little punctuation compared with English; Chinese contains no conjugations or plurals. For human translators, these problems are most often resolved through context or personal experience. There's no rule that says "between a rock and a hard place" isn't literal. We just know.

Machine translation is even trickier, and Carbonell's "interesting errors" line is a good encapsulation of its history. Perhaps no technological endeavor has been more defined by its failures than the attempts over the last 60 years to use computers to convert one language into another. "It's one of the earliest computer science problems to be attacked, and it has proven to be the one that's most difficult," says Nizar Habash, a research scientist at the Center for Computational Learning Systems at Columbia University.

From its genesis at the post-World War II dawn of computing – when ambitious researchers believed it would take only a few years to crack the language problem – until the late 1980s, machine translation, or MT, consisted almost entirely of what are known as rule-based systems. As the name implies, such translation engines required human linguists to combine grammar and syntax rules with cross-language dictionaries. The simplest rules might state, for example, that in French, adjectives generally follow nouns, while in English, they typically precede them. But given the ambiguity of language and the vast number of exceptions and often contradictory rules, the resulting systems ranged from marginally useful to comically inept.

Over the past decade, however, machine translation has improved dramatically, propelled by the relentless march of Moore's law, a spike in federal funding in the wake of 9/11, and, most important, a new idea. The idea dates from the late 1980s and early 1990s, when researchers at IBM stopped relying on grammar rules and began experimenting with sets of already-translated work known as parallel text. In the most promising method to emerge from the work, called statistical-based MT, algorithms analyze large collections of previous translations, or what are technically called parallel corpora – sessions of the European Union, say, or newswire copy – to divine the statistical probabilities of words and phrases in one language ending up as particular words or phrases in another. A model is then built on those probabilities and used to evaluate new text. A slew of researchers took up IBM's insights, and by the turn of the 21st century the quality of statistical MT research systems had drawn even with five decades of rule-based work.

Since then, researchers have tweaked their algorithms and the Web has spawned an explosion of available parallel text, turning the competition into a rout. The lopsidedness is best seen in the results from the annual MT evaluation put on by the National Institute of Standards and Technology (NIST), which uses a measurement called the BiLingual Evaluation Understudy (BLEU) scale to assess a system's performance in Chinese and Arabic against human translation. A high-quality human translator will likely score between 0.7 and 0.85 out of a possible 1 on the BLEU scale. In 2005, Google's stat-based system topped the NIST evaluation in both Arabic (at 0.51) and Chinese (at 0.35). Systran, the most prominent rule-based system still in operation, languished at 0.11 for Arabic and 0.15 for Chinese.

The success of statistical systems, however, comes with a catch: Such algorithms do well only when applied to the same type of text on which they've been trained. Statistical MT software trained on English and Spanish translations of the BBC World Service, for example, excels with other news articles but flops with software manuals. As a result, such systems require large amounts of parallel text for not just every language pair they intend to translate – which may not be available for, say, Pashto – but different genres within those language pairs as well. "For a lot of practical reasons, we have to find ways around our need for parallel text," says Philip Resnik, a professor of linguistics and computer science at the University of Maryland. "That is what Meaningful Machines is doing."

WHEN MEANINGFUL MACHINES first tested its Spanish-English engine on the BLEU scale in spring 2004, "it came in at 0.37," recalls the company's CEO, Steve Klein. "I was pretty dejected. But Jaime said, 'No, that's pretty good for flipping the switch the first time.'" A few months later, the system had jumped above 0.60 in internal tests, and by the time of Carbonell's presentation in August, the score in blind tests was 0.65 and still climbing. Although the company didn't test the passage with any statistical-based systems, when it tested Systran and another publicly available rule-based system, SDL, on the same data, both scored around 0.56, according to Carbonell's paper. Meaningful Machines was in stealth mode at the time, protecting its ideas. But Carbonell was itching to talk about his results. He didn't just have an engine that he says earned the highest BLEU score ever recorded by a machine. He had an engine that had done it without relying on parallel text.

Instead, the Meaningful Machines system uses a large collection of text in the target language (in the initial case it's 150 Gbytes of English text derived from the Web), a small amount of text in the source language, and a massive bilingual dictionary. Given a passage to translate from Spanish, the system looks at each sentence in consecutive five- to eight-word chunks. The al Qaeda message analysis, for example, might start with *"Declaramos nuestra responsabilidad de lo que ha ocurrido."* Using the dictionary, the software employs a process called flooding to generate and store all possible English translations for the words in that chunk.

Making this work effectively requires a dictionary that includes all of the possible conjugations and variations for every word. *Declaramos*, for example, offers up "declare," "declared," "declaring," "stating," and "testifying," among others. Meaningful Machines' Spanish-to-English dictionary, a database with about 2 million entries (20 times more than a standard Merriam-Webster's), is a lexical feat in and of itself. The company outsourced the task to an institute run by Jack Halpern, a prominent lexicographer. The result is one of the largest bilingual dictionaries in the world.

The options spit out by the dictionary for each chunk of text can number in the thousands, many of which are gibberish. To determine the most coherent candidates, the system scans the 150 Gbytes of English text, ranking candidates by how many times they appear. The more often they've actually been used by an English speaker, the more likely they are to be a correct translation. "We declare our responsibility for what has occurred" is more likely to appear than, say, "responsibility of which it has happened."

Next, the software slides its window one word to the right, repeating the flooding process with another five- to eight-word chunk: *"nuestra responsabilidad de lo que ha ocurrido en."* Using what Meaningful Machines calls the decoder, it then rescores the candidate translations according to the amount of overlap between each chunk's translation options and the ones before and after it. If "We declare our responsibility for what has happened" overlaps with "declare our responsibility for what has happened in" which overlaps with "our responsibility for what has happened in Madrid," the translation is judged accurate.

So what happens if the dictionary is missing words or if the overlap technique can't find a match? A third process, called the synonym generator, is used to search for unknown terms in the smaller Spanish-only set. When it finds them, it drops the original term and searches for other sentences using the surrounding words. The process is easiest to understand with an example in English. When run through the synonym generator, the phrase "it is safe to say" might turn up results like "it is safe to say that within a week" or "it is safe to say that even a blind squirrel ..." By removing "it is safe to say" from each sentence and then searching for other terms that fit the surrounding words, the generator suggests results like "it is important to note" or "you will find" – instead of, for example, "it is unhurt to speak."

The system, Carbonell tells me, is "simple … anybody can understand it." It's so simple, in fact, that Carbonell is peeved that he didn't think of it first. **BORN IN URUGUAY,** Jaime Carbonell moved to Boston with his family when he was nine. He later enrolled at MIT, where he found part-time work translating Digital Equipment Corporation computer manuals into Spanish to help pay tuition. In an attempt to speed up the translation process, he built a small MT engine that ran the documents through a glossary of common DEC terms, substituting the translations automatically. The little system worked so well that Carbonell continued to dabble in it while earning his computer science doctorate at Yale University. After coauthoring a paper outlining a new type of rule-based MT, he was offered a professorship at Carnegie Mellon. There he helped develop a successful commercial rule-based translation system. Then he hopped on the wave of text-based MT in the '90s.

One afternoon in 2001, Carbonell got a cold call from Steve Klein, a lawyer, hotel investor, and occasional film writer and director. Klein said that he'd formed a partnership with an Israeli inventor named Eli Abir – a man with little school or technical training who previously ran a restaurant. Abir, according to Klein, had a new machine-translation idea they wanted Carbonell to evaluate. Klein had been one of the first people to take the garrulous Abir seriously when he began hitting up investors for a previous invention in 2000, often in jeans and a T-shirt, claiming credentials as "the worst student in the history of the Israeli school system." Abir, who is bilingual in Hebrew and English, also said he could solve several of the world's thorniest computer science problems, based in part on knowledge gained from three days of playing *SimCity*.

Suspicious but curious, Carbonell agreed to meet the pair. When they arrived in his office and Abir explained the concept for what is now called the decoder, Carbonell was floored by its elegance. "In the few weeks that followed, I kept wondering, 'Why didn't I think of that? Why didn't the rest of the field think of that?' Finally I said, Enough of this envy. If I can't beat them, join them."

With Carbonell on board, the new company set about building its Spanish system. Soon, however, Abir's peripatetic invention habits created conflicts. Klein, Carbonell, and the developers feared the company was losing focus. "Eli is a mad genius," Carbonell says. "Both of those words apply. Some of his ideas are totally bogus. And some of his ideas are brilliant. Eli himself can't always tell the two apart." Abir, determined to build a larger AI "brain" that would tackle not just MT but other problems, took little interest in the day-to-day engineering. Eventually he left the company and returned to Israel to be closer to his son and to work on a new venture, a data compression system that he says "violates the rules of math as we know them." Of Meaningful Machines, he says, "They all are my friends. I think they are very talented people. They will bring it home."

ON MY MORNING in Meaningful Machines' offices, Carbonell does eventually encounter his "interesting errors" in the Spanish terrorism translation: dropped subjects, misplaced modifiers, garbled phrases that reveal gaps in the dictionary and shortcomings in the software. A larger concern for Carbonell than perfect accuracy, however, is time: The software takes 10 seconds to translate each word, a number the company wants to shrink to one second in the next year. "That's the biggest single impediment to commercializing this technology," he says.

Speed, in fact, may determine whether the system ends up being truly useful. Meaningful Machines recently hired a translation company to compare its system's first translations of Spanish news articles against those of human professionals. The results – according to the company, which hasn't released the data publicly – sounded at first like a typical MT failure: The output from the automated system required twice as many human hours to clean up. But the experiment also showed that cleaning up errors takes only a small fraction of the time required for the initial human translation. Thus, even with slightly sloppier first drafts, replacing the initial translator with a machine cuts the total human-hours of paid work in half. With that data in hand, Meaningful Machines recently entered discussions with a global translation conglomerate to field a commercial version of its Spanish engine.

When they do get the system out, Carbonell and company will have to play catch-up. Language Weaver – a four-year-old firm based in Southern California that has successfully commercialized its statistical system – already offers its software in 32 language pairs. That's a significant lead. But Meaningful Machines has a different algorithm, its impressive BLEU score, and the ability to translate without parallel text. There's also room for more than one player. The commercial translation market is now roughly $10 billion annually, and the government market is getting a boost from global terrorism. Language Weaver, which got an investment from the CIA's venture firm In-Q-Tel in 2003, now has customers in intelligence agencies here and abroad. The software, CEO Bryce Benjamin says, "is being used day in and day out to catch bad guys."

Meaningful Machines has military connections, too. Right now, the Global Autonomous Language Exploitation program run by Darpa is aiming to complete an automated speech and text translation system in the next five years. Meaningful Machines is part of a team participating in that challenge, including the "surprise language" segment (in which teams are given a more obscure language and asked to build a translation system). The challenge sounds a lot like another attempt to create the sort of universal translator that has eluded MT for 60 years. But success seems much more plausible now than ever before.

Nothing works perfectly, of course. In Meaningful Machine's translation of my Spanish al Qaeda sentences, the speaker warns, "If you do not save your injustices, there will be more and more blood and these attacks are very little with what will be able to happen with what you call terrorism." For a second, I pause, thinking the software must not be that good after all. But then Carbonell translates it himself and shows that some of the fault lies in the original Spanish, which was itself probably translated by a human from formalized Arabic. "We do not improve upon the original," he tells me as he looks over the results. "Yet."

*Contributing editor Evan Ratliff (*eratliff@atavistic.org*) interviewed Larry Brilliant in issue 14.07.*