

# Statistical Machine Translation Gains Respect

David Geer

**T**he demand for machine-translation technology is rising as business, finance, education, and the Internet become increasingly international and multilingual.

Since the 1950s, universities, research institutions, and vendors have developed translation technologies, most using detailed rules based on a sophisticated knowledge of linguistics.

Relatively few researchers worked on approaches that compare and analyze documents and their already-available translations to determine statistically, without prior linguistic knowledge, the likely meanings of phrases. These statistical systems use this information to translate new documents. For years, because processors were not fast enough to handle the extensive computation these systems require, many experts considered statistical systems inferior to rules-based systems.

However, when the Speech Group of the US National Institute of Standards and Technology's Information Access Division recently tested 20 machine translation technologies, a statistical system developed by Google finished in first place.

The NIST test results' significance is that Google and other organizations will invest more time, money, and talent into researching this approach, said Dimitris Sabatakakis, CEO of translation vendor Systran, whose software



appears in many search engines, online translators, and other products. Meanwhile, faster processors and other advances are making statistical translation technology more accurate and thus more useful.

However, the approach must still clear several hurdles—such as still-inadequate accuracy and problems recognizing idioms—before it can be useful for mission-critical tasks.

## 2005 NIST TESTS

NIST has conducted machine-translation evaluations since 2002 to help researchers determine the effectiveness of their systems, said Mark Przybocki, a computer scientist with the agency and coordinator of the tests.

NIST tests noncommercial statistical and rules-based technologies. Companies with commercial products submit untrained, noncommercial research systems for evaluation.

This year, participating systems could conduct Chinese-to-English and/or Arabic-to-English translations of the same 100 newswire documents.

NIST analysts evaluated systems that learned how to translate by train-

ing only with material provided by the Linguistic Data Consortium (LDC)—a nonprofit group of universities, companies, and government research laboratories—and systems trained with any additional material the developers chose.

NIST used IBM's Bleu metric to measure translation quality. Bleu compares the number of *N*-grams that a translation has in common with one or more already-completed, high-quality translations of the document in question, according to Przybocki. An *N*-gram is a type of phrase within a document that has a set number of words (*N*) established by the developer in a certain order with a specific relationship to one another. *N*-grams are the basic linguistic unit with which a statistical translation system works.

The results of the NIST test, in order of accuracy, were as follows:

**Arabic-to-English.** Systems trained with LDC documents: Google, the University of Southern California's Information Sciences Institute (ISI), IBM, University of Maryland, Johns Hopkins University/University of Cambridge, University of Edinburgh, Systran, Mitre Corp., and Fitchburg State College. Systems trained with additional material: Google, Sakhr Software, and the US Army Research Laboratory.

**Chinese-to-English.** Systems trained with LDC documents: Google, ISI, University of Maryland, Rheinisch-Westfälische Technische Hochschule Aachen, Johns Hopkins University/University of Cambridge, IBM, University of Edinburgh, Istituto Trentino di Cultura-Il Centro per la Ricerca Scientifica e Tecnologica, National Research Council of Canada, NTT Communication Science Laboratories, the Advanced Telecommunications Research Institute International's Spoken Language Translation Research Laboratories, Systran, Saarland University, and Mitre. Systems trained with additional material: Google, the Chinese Academy of Sciences' Institute of Computing Technology, and the Harbin Institute of Technology's

Machine Intelligence and Translation Laboratory.

## STATISTICAL SYSTEMS

IBM pioneered statistical translation in the early 1990s. Widespread research began in 1999 after a workshop sponsored by the US National Science Foundation and Johns Hopkins University's Center for Language and Speech Processing.

### How they work

Statistical translation systems start by "learning" how various languages work. They begin with minimal dictionary and language resources. Users then train the systems before they can handle extensive translations, as Figure 1 shows.

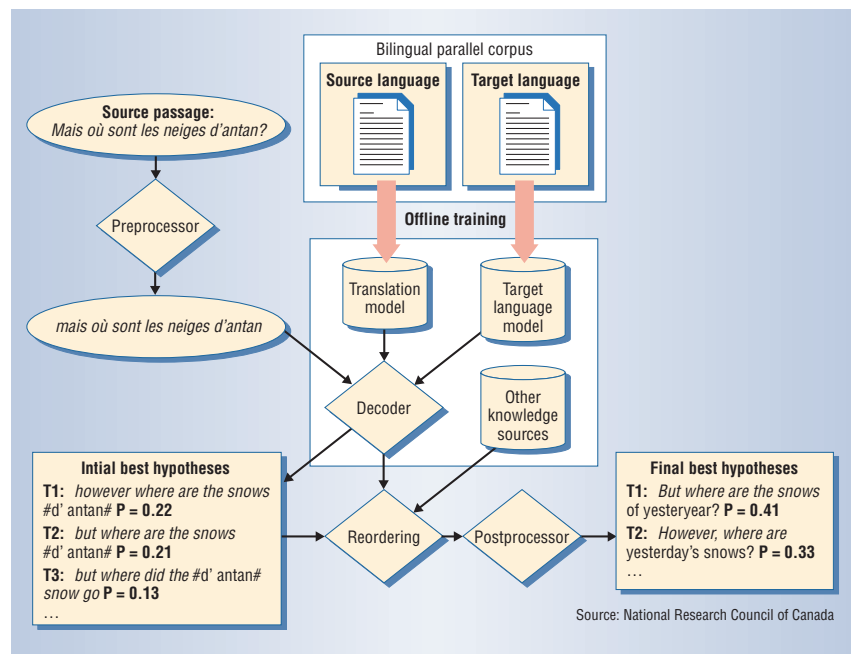
During the training, researchers feed the system documents in a source language and their high-quality human translations in a target language, explained Philip Resnik, associate professor in the University of Maryland's Department of Linguistics and its Institute for Advanced Computer Studies.

The system uses its existing resources to guess at documents' meanings. An application compares the guesses to the human translations and returns the results to improve the system's performance.

Statistical systems generally work by dividing sentences into *N*-grams. Analyzing *N*-grams improves accuracy and performance because, while a word may have many definitions, it has fewer potential meanings when part of a phrase.

Most systems translate based on trigrams, explained Franz Och, head of Google's Machine-Translation Group. Three-word groups are big enough to make the process efficient and are likely to be repeated often enough in documents to be useful for analysis, he said. Larger sets, on the other hand, are less likely to be repeated and require more computing power to analyze and translate.

While training, statistical systems track common *N*-grams, learn which



**Figure 1. A statistical-translation system starts with minimal dictionary and language resources. Developers then train it by feeding documents and their high-quality human translations from a bilingual parallel corpus. The system uses the training to develop translation models. When handling a source text, in this case a French passage, a preprocessor converts the material into a form suitable for statistical analysis. The system then processes the text, producing sets of best hypothetical translations in order of their probable accuracy.**

translations are most frequently used, and apply those meanings when finding the phrases in the future. They also statistically analyze the position of *N*-grams in relation to one another within sentences, as well as words' grammatical forms, to determine correct syntax.

After their training, the systems process phrases and string them together to translate entire documents, Resnik noted.

Even after training, they continue to update their mathematical analysis models based on how accurate they turn out to be.

Faster processors have enabled the level of computation necessary to make statistical translation effective, noted Robert Frederking, a senior systems scientist and chair of graduate programs with Carnegie Mellon University's (CMU's) Language Technologies Institute.

And developers can use the increasing number of available online documents in various languages to train the applications more thoroughly. In addition, better and higher-capacity storage technology helps systems store a larger volume of important data more effectively, according to Resnik. He said the algorithms the systems use are also better, largely because of improved and more efficient training and translation models.

Statistical translation toolkits are generally written in C++, Java, and Perl, which are cross-platform languages. They thus enable translation systems to run on most OSs, Frederking noted.

Because these applications work statistically, they can learn multiple language domains and are thus good for general-purpose translation. Rules-based systems operate with detailed rules for a particular language domain. They are thus best for specialized

domains, such as medical terminology, which have many words with meanings peculiar to the domain.

Vendors can sell translation systems as packaged software for use on a customer's PC or as an online service.

### Strengths and weaknesses

With repeated, intensive training, developers can make statistical systems increasingly accurate. Compared to rules-based systems, said Stephan Vogel, research associate with CMU's Language Technologies Institute, "when putting the same amount of money into generating translations, the statistical system will give better translations."

While complex rules-based systems for language pairs can require years and considerable expense to create, researchers can train statistical systems to produce translations in weeks or days. This makes the statistical approach less labor intensive and more useful for time-critical corporate and government applications.

However, the approach also requires powerful computers to handle the training and subsequent translations.

Because they don't require input from linguistic experts, statistical techniques can be effective for translating to and from obscure languages, as long as many documents and translations are available for training.

The approach has trouble with some languages, such as Chinese, that require systems to have high levels of linguistic knowledge, explained Miles Osborne, professor in the University of Edinburgh's Statistical Machine Translation Group.

### Projects tested by NIST

The projects that NIST tested are examples of several statistical translation approaches. For example, IBM's technology can perform multiple analyses of a sentence and pick the parsing with the highest probability of being correct, explained Brian Garr, program director and segment manager for contact center solutions in the company's Software Group.

**Google.** According to Google's Och, there are many potential applications for which the company can use its translation technology. For example, he noted, the company already provides Web-page-translation services. Google says it has several other applications in mind but won't comment further.

Google has several advantages in developing translation technology, said Och. First, Google's large search database contains documents, many in languages other than English, on which to train its translation system. In addition to its own resources, Google has used United Nations documents, which are, of course, in many languages.

### Researchers can train statistical systems to produce translations in weeks or days.

Also, Google's server farm, generally acknowledged as one of the world's largest, provides the computing power necessary to effectively train and use its statistical system.

Because of its resources, Google says it could develop a more sophisticated analysis approach. For example, Och explained, while most systems translate based on trigrams, Google can also work with larger word groupings. The ability to collect statistics for larger groupings lets the company's application recognize and predict word patterns more accurately than other systems, according to Google.

**National Research Council of Canada.** The NRC's Portage system develops statistically likely translations by working with matching sentences in various language pairs, said Roland Kuhn, research officer in the Interactive Language Technologies Group of the council's Institute for Information Technology.

Statistical translation systems typically work from sentence to sentence, handling the *N*-grams within each sen-

tence. Portage uses an algorithm that looks at material outside a particular sentence being processed to clear up problematic translations within the sentence, particularly for words that can have multiple meanings, Kuhn explained.

**University of Maryland.** Unlike other statistical translation applications, associate professor Resnik said, the university's Hiero system models hierarchical relationships in a language—such as the parsing of sentences into phrases, parts of speech, and even words—rather than simply stringing together phrases sequentially. This lets Hiero capture linguistically rich aspects of syntactic behavior, he explained.

### ONGOING CHALLENGES

Most services translate individual words accurately but give only the gist, at best, of documents and Web sites. At worst, their results can be almost incomprehensible.

Research efforts to improve statistical translation's accuracy face several key challenges. For example, translation technology has trouble with constantly changing and growing vocabularies, which reflect new idiomatic uses and technical terminology.

In addition, translation technology can have problems recognizing proper nouns. "The translation of proper nouns is not about meaning; names are arbitrary," said Resnik. "Also, new proper nouns show up all the time, and systems aren't familiar with them."

Another challenge is that a statistical system's effectiveness is largely dependent on the type and quality of its training data. For example, Resnik explained, accuracy can suffer if a system is trained on one type of data, such as news articles, but is asked to translate another, such as literature or weblogs.

Eventually, rules-based and statistical methods will merge, predicted David Nahamoo, group manager for IBM's Human Language Tech-

nologies Group. He said this would let vendors add linguistic knowledge to statistical analysis, giving translation programs the benefits of both approaches. IBM's WebSphere Translation Server represents the beginnings of a hybrid system, according to the company's Garr.

Meanwhile, said CMU's Frederking, researchers might soon start offering translation technology for many of the world's languages they are not addressing now.

Despite all the improvements, noted the University of Edinburgh's Osborne, no one is ready yet to use only machine

translation for mission-critical tasks. Organizations might use the technology to determine whether a document is important enough to spend the time and money needed to have it manually translated, he explained.

However, added Ronald Rogowski, senior analyst with Forrester Research, a market analysis firm, "As the tools improve, they will be used increasingly for first-draft translations that can be fed directly to the editing process."

According to the NRC's Kuhn, machine translation won't compete with human translation for handling sophisticated texts in the near future,

but it may be widely used on e-mail notes and other texts for which style isn't important. ■

*David Geer is a freelance technology writer based in Ashtabula, Ohio. Contact him at david@geercom.com.*

Editor: Lee Garber, *Computer*,  
l.garber@computer.org

## MICROSOFT RESEARCH REQUESTING PROPOSALS—FALL 2005

for grants in the following areas:

**Digital Inclusion** will fund academic research focused on advancing computing technologies that address social and economic challenges in under-served communities. Being digitally connected has become ever more critical to economic, educational, and social advancement. This grant will tackle the tough research problems that must be solved to realize that vision—such as work in networking infrastructures, user interfaces, computing devices, and relevant applications.

**Compilation and Managed Execution** will fund cross-cutting research that examines and reconsiders the relationships between development tools, compilers, managed runtime environments, runtime code generation, and underlying operating systems. Work will be based on Phoenix, Microsoft's next-generation code generation and optimization framework, and SSCLI—the Shared Source Common Language Infrastructure.

Complete details of these funding opportunities available at <http://research.microsoft.com/ur/us/fundingopps/default.aspx>

*Microsoft Research and the Academic Research community—advancing the frontiers of computing.*