The Economist

Machine translation

Tongues of the web

With its proliferating number of tongues, the Internet is giving MT—the use of computers to translate languages—a much needed shot in the arm

Mar 14th 2002 | from PRINT EDITION

SINCE its earliest days, machine translation—the use of computers to translate documents from one language to another automatically—has suffered from exaggerated claims and impossible expectations. One characteristic (but apocryphal) tale tells of an American military system designed to translate Russian into English, which is said to have rendered the famous Russian saying “The spirit is willing but the flesh is weak” into “The vodka is good but the meat is rotten.”

This sort of joke prompts a hollow laugh from those in the machine-translation (MT) business. It does so because it demonstrates both the difficulty of getting computers to understand human languages, and the high expectations that must be met if MT is to be taken seriously. Over the years, there have been a number of promising new approaches in the field, and ever-cheaper processing and storage technology have helped improve things. But progress has been painfully slow, and the decisive breakthrough that will transform the fortunes of MT has never appeared.

Now the Internet has given MT a much needed shot in the arm. This is odd because the ability to transmit information quickly and cheaply would not, on the face of it, appear to make the process of translation any easier. Yet, although the underlying technology of MT is still the same as it ever was, the rise of the Internet changes the way in which technology is perceived and the way it is used. And there are signs that, in future, it could improve the way it works as well.

The idea of automating the process of translation using computers goes back to the late 1940s. Warren Weaver of the Rockefeller Foundation in New York wrote a memorandum suggesting that the code-breaking successes of the second world war, combined with electronic computers and the new “information theory” laid out by Claude Shannon, might form the basis of an automatic translation system. This prompted research at several American universities, and the first public demonstration of MT—the result of a collaboration between IBM and Georgetown University—took place in 1954. This early system, based on a simple bilingual dictionary with a few rules to determine word order, caused a surge of enthusiasm and funding.

For the next decade, MT researchers tried to overcome the limitations of simple dictionary-based systems using more complex approaches which analysed the source text using grammatical rules. “Today, the computer, or electronic brain, is well along toward picking up the burden of machine translation,” declared the *Atlantic Monthly* in 1959. But despite such optimism, progress was slow, and in 1964 the American government established a committee to examine the prospects for MT. Its report, issued two years later, concluded that, compared with human translators, MT systems were slower, less accurate, and twice as expensive.

Instead, the committee recommended that research should concentrate on devising systems to assist human translators, rather than trying to replace them altogether. As a result, American funding for pure MT research dried up.

In some fields, however, it was recognised that even a rough-and-ready translation was better than none at all. Systran, a company established by Peter Toma, a researcher at the California Institute of Technology in Pasadena, sold a Russian-to-English translation system to the United States Air Force in 1970, and the same system was subsequently adopted by the European Commission. During the 1970s, demand for translation systems began to emerge in the business community.

During the 1980s, the combination of rapid falls in the cost of computing power and increasing demand from governments and multinational companies caused a revival of interest in MT, spurring renewed research. New systems were developed. Many of them worked by translating the source text into an intermediate language or symbolic representation, from which it could be translated into any of several other languages. As computers became more powerful and storage became cheaper, other new approaches emerged in the 1990s: analysis of parallel texts (the same text in two languages) led to new statistical-translation systems, which did not rely on any underlying grammatical rules, and to example-based systems which translated one sentence at a time by searching a database for examples of similar sentences whose translations were known.

Even so, the quality of MT has not really improved very much over the past three decades, says John Hutchins, an expert on the history of machine translation at the University of East Anglia, in Britain. “If you look at quality of output now, compared with 1970, in many cases you can't see much improvement,” he says. What has changed is that MT systems have now been plugged into the Internet. That changes the way they are used, and the expectations of them.

The network of Babel

The Internet has democratised MT and boosted demand dramatically, as users around the world struggle to understand pages in languages other than their own. And as companies set up increasingly elaborate websites, they have become aware of the need to maintain multiple sites in different countries and serve customers in different languages. Of America's 100 largest firms, 33 had multilingual websites at the end of 1999, and 57 did a year later. A study by Aberdeen Group, a management consultancy, found that, on average, users spend up to twice as long at a site, and are four times more likely to buy something from it, if it is presented to them in their own language. Another study by IDC, a technology consultancy, found that only 5% of the 50 top websites responded appropriately to e-mail queries in a foreign language; most simply asked for the message to be resent in English. All of which highlights the need for MT systems to provide on-the-fly translations, and for elaborate publishing systems that can manage multilingual websites.

Arguably the best known online MT system is Babel Fish, which relies on Systran software to translate pages retrieved by the AltaVista search engine. Anyone who has used Babel Fish will be familiar with the unintentional hilarity of the results; one popular game involves scrambling the lyrics of pop songs by translating them from English into another language and then back again (a “round-trip” translation). Other MT systems are also in use online, providing rough-and-ready translations of chat-room conversations and e-mail messages. Demand for such services is likely to increase as the diversity of Internet users increases. At the end of 2000, 48% of Internet users were English speakers, but this figure is expected to fall

> The Internet changes the game for machine translation: users want speed, rather than quality, and are more likely to accept poor results

to 32% by the end of 2002.

Unfortunately, MT systems work best when they have been customised for a particular subject area, such as microbiology, aerospace or particle physics. This involves analysing typical documents and adding common words and technical terms to the system's dictionary. Using MT to translate Internet pages, which can be about anything at all, therefore produces terrible results, since no customisation is possible. To make matters worse, most MT systems were designed for use with high quality documents, whereas many web pages, chat-rooms and e-mails tend to involve slang, colloquial language and ungrammatical constructions.

Even so, Steve McClure, an analyst at IDC, notes that the Internet has “refocused”MT from being a tool that provides a first draft for translators to becoming a general tool “for gaining a quick, partial understanding of perishable texts in high-volume environments without human involvement in the translation process.” The Internet changes the game for machine translation: users want speed, rather than quality, and are more likely to accept poor results.

The fact that MT is now available to every Internet user is, however, a double-edged sword, says Laurie Gerber of the Association for Machine Translation in the Americas, an industry body that brings together MT researchers, vendors and users. People are now far more familiar with the technology, she says, and may have revised their attitudes towards it, as they become aware that rough-and-ready translations have their uses. But Ms Gerber warns that the Internet has also made enemies for MT. “They'll try to use it for something serious, or they'll do round-trip translation, and say ‘look how silly it is',” she says. “They get the impression that it's useless.”

Mr Hutchins makes a similar point. “At present, people are grateful for any translation,” he says. But he doubts that this will remain true for long. Soon, he suggests, Internet users will demand higher quality translations. Fortunately, there are several ways in which the Internet itself may be able to help improve the quality of machine translation.

Translate this

The biggest difference, says Dimitrios Sabatakakis, chief executive of Systran, is that the Internet makes it much easier to create a customised version of a machine-translation system. A typical system has around 300,000 custom entries, he says. What used to require three or four years of text-entry and analysis can now be done in three or four months—simply by sucking up documents from a company's intranet.

Another benefit is that large firms are increasingly imposing standardised style and terminology rules across all their internal and external documents, including reports and web pages. Once a machine translation system has been tuned appropriately, it can produce far more accurate results if the input text is more consistent. In some cases, says Mr Sabatakakis, firms are adopting standardised language with the specific intention of making documents easier to translate. A firm posting technical-support documents on its website, for example, might include a “translate this page” button on each page that feeds the page through a customised translation system. The resulting translation is then of far higher quality.

A pioneering example of this approach has been taken by Autodesk, a company based in San Rafael, California, that makes computer-aided design software. With over 60% of its revenues coming from outside the United States, the firm faces the daunting task of supporting 4m users of its software in many different parts of the world. It provides technical support in the form of an online database of some 10,000 documents, each about 1,000 words long, and each addressing a particular problem. A few hundred new documents are added each month, and around half a million documents are consulted by users each day.

Manually translating all of these documents was out of the question. But using off-the-shelf MT produced results of unacceptable quality. So, Autodesk opted for a customised Systran

system which translates search results into French, Spanish, German or Italian. (Support for Japanese is in the works.) By analysing the documents in Autodesk's database, and exploiting the fact that they are written in a consistent style, it was possible to tune the translation system to produce far more accurate results. The system was launched in July 2001.

One drawback, however, is that the initial search terms have to be in English. But translating search terms from other languages into English, performing the search, and then translating the results, is horribly inaccurate. Autodesk is investigating a number of possible solutions to this problem. Nonetheless, Mr McClure calls the Autodesk system a “milestone” in the commercial evolution of MT. Technical support represents an untapped market for MT, he says. Technical documents contain fewer ambiguities and translate well. “If Autodesk and Systran are successful, the floodgates may open for the use of MT for multilingual customer support,” he suggests. IBM has already launched a similar system.

Mr Sabatakakis says that his firm has noticed a recent surge in interest from multinational companies. Just as the construction of America's highways meant that Goodyear sold many more tyres, he hopes the Internet will have a similar effect on the machine-translation business. Rather than seeing MT as a product they can simply buy off the shelf, large firms are now realising that MT systems must be customised and integrated into their document-management processes. Mr Sabatakakis draws an analogy with databases. When you buy software from Oracle, he says, you expect to have to spend some time setting it up and customising it before you can start using it.

Thanks for the memory

Another way in which the Internet may be able to improve the effectiveness of MT is through the use of shared “translation memories”. A translation memory is a parallel database of previously translated content. Human translators use such databases to speed up their work. Given a sentence that needs to be translated, it may be that the same sentence, or a very similar sentence, is in the translation memory already. (A translator who translates technical manuals, for example, may find that many sentences are common to manuals for a range of similar products, such as cameras or printers.) Some translation systems combine translation memories with MT systems to provide first drafts of sentences that do not appear in the memory.

Over the past few months, however, a number of firms producing translation-memory software have suggested that translators might wish to pool their work. By establishing a vast, online translation memory, translators would be saved the hassle of translating sentences that had already been translated by others. In theory, MT systems could tap into these vast memories, too.

Rather than seeing MT as a product they can simply buy off the shelf, large firms are now realising that MT systems must be customised

It sounds like a great idea. But there are a number of practical problems with shared translation memories. One is quality. How can you be sure that a sentence plucked from a translation memory is an accurate translation? A bigger problem is ownership. Mr Hutchins points out that translators may be unwilling to give away their work. And Ms Gerber notes that if you translate a document for a large company, that company will probably be reluctant to make that document, and its translated version, available to all-comers. If one computer firm has gone to the trouble of translating its manuals into Chinese, why should a competing firm reap the benefit? Another problem when translating technical materials is that many firms have their own proprietary terminology, which they use to distinguish themselves from their competitors.

These are all valid objections, but there may be ways around them. Prolyphic, a firm based in San Diego, has set up a royalty system whereby translators are rewarded each time someone

else uses some of their work. Prolyphic also operates a vetting system to guarantee the quality of translated content. And fussiness about special terminology, or proprietary information, is something that only bothers really large firms, says Ms Gerber. She suggests that using shared translation memories might appeal to smaller firms that have not invested so heavily in their corporate images, and simply want to sell a few widgets.

By boosting usage, changing user perceptions, and encouraging the standardisation and pooling of resources, the Internet has improved the prospects for MT enormously, even though the technology itself is unchanged. What will happen next? Mr Hutchins predicts that specialist translation systems will proliferate, as users in particular fields demand higher quality free translations. Trade publications or scientific journals, for example, might set up customised translation systems to make their content more easily accessible. But the pitfall of having to search in the original document's language remains. Mr McClure says he would like to see more research into “cross-language retrieval” to address this problem, by allowing a database in one language to be searched and accessed by someone who does not speak that language. His firm has found that even those who speak a second language very well prefer to access documents in their own language when possible.

Spectrum of tools

But perhaps the greatest impact of the Internet is that there is now a broader perception that MT technology provides a spectrum of tools for use in different circumstances. Free, fast translations are only good enough to get the gist of a document. Better translations require fancier systems, and more effort on the part of users to accommodate their vagaries.

It seems unlikely, however, that MT will ever shrug off its reputation for making silly mistakes. Ms Gerber notes that, even with the best human translators, some people are unhappy with the results, because of inappropriate use of style or terminology. The quality of translation is, in other words, highly subjective. “MT gets a bad rap for not being ‘human quality',” she says. But this means it is being held to unreasonably high standards. Even so, the Internet has made MT technology more useful; and MT technology has done the same for the Internet. The two are interdependent, says Mr Sabatakakis. Perhaps symbiotic is more like it. Both technologies need each other—and now stand to prosper from each other's progress.

Readers wishing to discuss issues raised in this article are encouraged to post their comments on the online forum at www.economist.com/forums/tq (http://www.economist.com/forums/tq)