The Economist

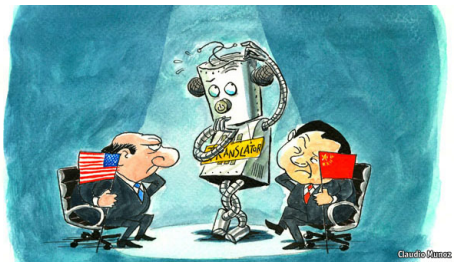
Machine translation

Conquering Babel

Simultaneous translation by computer is getting closer

Jan 5th 2013 | Seattle |From the print edition

IN "STAR TREK", a television series of the 1960s, no matter how far across the universe the Starship *Enterprise* travelled, any aliens it encountered would converse in fluent Californian English. It was explained that Captain Kirk and his crew wore tiny, computerised Universal Translators that could scan alien brainwaves and simultaneously convert their concepts into appropriate English words.



Science fiction, of course. But the best sci-fi has a habit of presaging fact. Many believe the flip-open communicators also seen in that first "Star Trek" series inspired the design of clamshell mobile phones. And, on a more sinister note, several armies and military-equipment firms are working on high-energy laser weapons that bear a striking resemblance to phasers. How long, then, before automatic simultaneous translation becomes the norm, and all those tedious language lessons at school are declared redundant?

Not, perhaps, as long as language teachers, interpreters and others who make their living from mutual incomprehension might like. A series of announcements over the past few months from sources as varied as mighty Microsoft and string-and-sealing-wax private inventors suggest that workable, if not yet perfect, simultaneous-translation devices are now close at hand.

Over the summer, Will Powell, an inventor in London, demonstrated a system that translates both sides of a conversation between English and Spanish speakers—if they are patient, and speak slowly. Each interlocutor wears a hands-free headset linked to a mobile phone, and sports special goggles that display the translated text like subtitles in a foreign film.

In November, NTT DoCoMo, the largest mobile-phone operator in Japan, introduced a service that translates phone calls between Japanese and English, Chinese or Korean. Each party speaks consecutively, with the firm's computers eavesdropping and translating his words in a matter of seconds. The result is then spoken in a man's or woman's voice, as appropriate.

Microsoft's contribution is perhaps the most beguiling. When Rick Rashid, the firm's chief research officer, spoke in English at a conference in Tianjin in October, his peroration was translated live into Mandarin, appearing first as subtitles on overhead video screens, and then as a computer-generated voice. Remarkably, the Chinese version of Mr Rashid's speech shared the characteristic tones and inflections of his own voice.

Que?

Though the three systems are quite different, each faces the same problems. The first challenge is to recognise and digitise speech. In the past, speech-recognition software has parsed what is being said into its constituent sounds, known as phonemes. There are around 25 of these in Mandarin, 40 in English and over 100 in some African languages. Statistical speech models and a probabilistic technique called Gaussian mixture modelling are then used to identify each phoneme, before reconstructing the original word. This is the technology most commonly found in the irritating voice-mail jails of companies' telephone-answering systems. It works acceptably with a restricted vocabulary, but try anything more free-range and it mistakes at least one word in four.

The translator Mr Rashid demonstrated employs several improvements. For a start, it aims to identify not single phonemes but sequential triplets of them, known as senones. English has more than 9,000 of these. If they can be recognised, though, working out which words they are part of is far easier than would be the case starting with phonemes alone.

Microsoft's senone identifier relies on deep neural networks, a mathematical technique inspired by the human brain. Such artificial networks are pieces of software composed of virtual neurons. Each neuron weighs the strengths of incoming signals from its neighbours and send outputs based on those to other neighbours, which then do the same thing. Such a network can be trained to match an input to an output by varying the strengths of the links between its component neurons.

One thing known for sure about real brains is that their neurons are arranged in layers. A deep neural network copies this arrangement. Microsoft's has nine layers. The bottom one learns features of the processed sound waves of speech. The next layer learns combinations of those features, and so on up the stack, with more sophisticated correlations gradually emerging. The top layer makes a guess about which senone it thinks the system has heard. By using recorded libraries of speech with each senone tagged, the correct result can be fed back into the network, in order to improve its performance.

Microsoft's researchers claim that their deep-neural-network translator makes at least a third fewer errors than traditional systems and in some cases mistakes as few as one word in eight. Google has also started using deep neural networks for speech recognition (although not yet translation) on its Android smartphones, and claims they have reduced errors by over 20%. Nuance, another provider of speech-recognition services, reports similar improvements. Deep neural networks can be computationally demanding, so most speech-recognition and translation software (including that from Microsoft, Google and Nuance) runs in the cloud, on powerful online servers accessible in turn by smartphones or home computers.

Quoi?

Recognising speech is, however, only the first part of translation. Just as important is converting what has been learned not only into foreign words (hard enough, given the ambiguities of meaning which all languages display, and the fact that some concepts are simply untranslatable), but into foreign sentences. These often have different grammatical rules, and thus different conventional word orders. So even when the English words in a sentence are known for certain, computerised language services may produce stilted or humorously inaccurate translations.

Google's solution for its Translate smartphone app and web service is crowd-sourcing. It compares the text to be translated with millions of sentences that have passed through its software, and selects the most appropriate. Jibbigo, whose translator app for travellers was spun out from research at Carnegie Mellon University, works in a similar way but also pays users in developing countries to correct their mother-tongue translations. Even so, the ultimate elusiveness of language can cause machine-translation specialists to feel a touch of *Weltschmerz*.

For example, although the NTT DoCoMo phone-call translator is fast and easy to use, it struggles—even though it, too, uses a neural network—with anything more demanding than pleasantries. Sentences must be kept short to maintain accuracy, and even so words often get jumbled.

Microsoft is betting that listeners will be more forgiving of such errors when dialogue is delivered in the speaker's own voice. Its new system can encode the distinctive timbre of this by analysing about an hour's worth of recordings. It then generates synthesised speech with a similar spread of frequencies. The system worked well

in China, where Mr Rashid's computerised (and occasionally erroneous) Mandarin was met with enthusiastic applause.

A universal translator that works only in conference halls, however, would be of limited use to travellers, whether intergalactic or merely intercontinental. Mr Powell's conversation translator will work anywhere that there is a mobile-phone signal. Speech picked up by the headsets is fed into speech-recognition software on a nearby laptop, and the resulting text is sent over the mobile-phone network to Microsoft's translation engine online.

One big difficulty when translating conversations is determining who is speaking at any moment. Mr Powell's system does this not by attempting to recognise voices directly, but rather by running all the speech it hears through two translation engines simultaneously: English to Spanish, and Spanish to English. Since only one of the outputs is likely to make any sense, the system can thus decide who is speaking. That done, it displays the translation in the other person's goggles.

At the moment, the need for the headsets, cloud services and intervening laptop means Mr Powell's simultaneous system is still very much a prototype. Consecutive, single-speaker translation is more advanced. The most sophisticated technology currently belongs to Jibbigo, which has managed to squeeze speech recognition and a 40,000-word vocabulary for ten languages into an app that runs on today's smartphones without needing an internet connection at all.

Nani?

Some problems remain. In the real world, people talk over one another, use slang or chat on noisy streets, all of which can foil even the best translation system. But though it may be a few more years before "Star Trek" style conversations become commonplace, universal translators still look set to beat phasers, transporter beams and warp drives in moving from science fiction into reality.