The many voices of the web

The internet: New combinations of human and computer translation are making web pages available in foreign languages.

THE web connects over a billion people, but it is fragmented by language. Anglophone web-users have as many pages to choose from as Chinese speakers, and there are roughly as many blogs in Japanese as there are in English. And although the Arabic blogosphere got off to a late start, it is now booming. But each of these groups of users is walled off from the others by language.

What might the web look like without such linguistic barriers? Imagine if internet users everywhere could have content automatically, smoothly and accurately translated into their own languages. A Chinese web-surfer could then visit an English newspaper website and read all the content in excellent Mandarin, before moving on to read blog entries written in Malagasy or Twitter posts in Galician.

This fantasy is still just that, but bits of it are starting to look plausible. Start with the translation part. Thanks to the internet, this is now a relatively flexible and cheap process. At the base of the translation hierarchy are free services offered by Google and others. Such services "learn" by analysing collections of documents that have been translated by humans, such as the records of the European Parliament, which are translated into 11 different languages. These collections are so big, and the machines that analyse them so powerful, that automatic translation (known in the jargon as "machine translation") can usually convey the gist of a text, albeit it in a slightly garbled manner. Google and its rivals focus on widely spoken tongues, but academics are working on machine-translation services for more obscure languages.

An army of volunteer translators occupies the next level up in the hierarchy. Several prominent English-language publications, including this newspaper, are regularly translated into Mandarin by groups of unpaid volunteers for the benefit of other readers (see ecocn.org/bbs). More formal projects also exist. At Global Voices, a kind of polyglot bloggers' collective, around 200 volunteers select and translate their colleagues' posts. Items on Meedan, a social network dedicated to the discussion of Middle East news, are translated into English or Arabic by machine and can then be tidied up by readers.

Paid human translators, unsurprisingly, still produce the best results. But even here costs are coming down, as the translation industry is shifting from project-based to piecemeal working. The methods are inspired by Mechanical Turk, an online service operated by Amazon that companies use to farm out mundane tasks to a pool of online workers. SpeakLike, which launched in late 2009, has a pool of 3,000 translators and can supply a translation of a given text within hours for $0.05-0.15 a word, depending on turnaround time. SpeakLike will even translate Twitter posts and send them to a parallel account within minutes for $0.25 a pop.

All this activity can, at least in theory, take place out of sight of the reader. One way to make this happen is to use the Worldwide Lexicon (WWL), a series of interlocking pieces of free software created by Brian McConnell, a software developer based in San Francisco. WWL gives bloggers and media companies fine control over how their content is translated. A blogger can, for example, provide a machine-translated version of a post whenever the speaker of a different language visits his site. (Web browsers like Internet Explorer and Firefox specify the user's language when requesting pages.) WWL also provides a neat interface that, if enabled, allows readers to improve the translation of blog postings, for the benefit of subsequent visitors.

Commercial producers of content can use the software to create an initial machine translation and then send it to SpeakLike for further work. The WWL software can also wait until the hit count on an item exceeds a certain value, indicating that it is popular, before sending the machine-translated version out to a human. This combination of human and computer work—cyborg translation, as it were—takes place entirely behind the scenes; visitors are simply presented with a more or less readable article. Mr McConnell is working to integrate his system with WordPress, one of the most widely used blogging platforms. He says WWL is being used by several publishers, including the owners of a well-known technology magazine.

So how much closer is the dream of a unified web? Volunteer translators only cluster around popular sites, so the vast majority of blogs will remain untranslated, or only machine-translated. Most content producers are unable to pay for human translation, even at today's prices. That leaves them reliant on machine translation, too. It is getting better, but it still struggles with colloquialisms and idioms. As Ethan Zuckerman, co-founder of Global Voices and a researcher at Harvard University, puts it: "If you sound like an EU parliamentarian, we can translate you quite well." Until computers learn how to cope just as proficiently with the outbursts of self-absorbed teenage bloggers or snarky gossip columnists, machine-translated articles will struggle to attract readers. Clever technology can help lower the web's linguistic barriers, but cannot yet eliminate them.

Fonte: The Economist. Disponível em: <www.economist.com>. Acesso em: 5 mar. 2010.