How far can you get with machine translation?

Lost in translation? Online editor Brian Pellot looks at the joys and follies of machine translation and explains how Google Translate has expanded Free Speech Debate's multilingual reach.

??????????????????????

Can you understand the above comment by FSD user Tsukamoto Takanobu? If you don't speak Japanese, the answer is probably no. Tsukamoto also stumped Google Translate, which interprets the text as: "Anonymous the ? responsibility ? ? ? ? the ? test ? ?? on talk ? ? ki." Lucky for us, Tsukamoto often comments in both English and Japanese. The user added to the same post in English, "So I think that anonymous tends to become cynical and unresponsivel." A more accurate translation might be, "So I think that anonymity tends to be cynical and irresponsible."

Free Speech Debate's main editorial content does not rely on machine translation. Our international team of Oxford graduate students has carefully translated each principle and many of the relevant discussion pieces, interviews and case studies into 13 languages - languages of which they are native speakers. These languages have been chosen because between them they have the potential to reach more than 80 percent of the estimated two billion people online. That is the carefully considered core of our multilingual offering. In time, we hope to have all our editorial content translated into those languages, and then have the translations checked by experts.

That still leaves the rest of the world's languages. We would love to translate into each and every one of them carefully by hand, but that is obviously beyond our resources. On the other hand, we think it is essential that people feel free to comment in their own languages. For this, we use Google Translate.

Machine translation is far from perfect (as the above gibberish makes clear), but it's getting better all the time. As you can see from this comment by user Peregrino, Google Translate automatically detects the original language (in this case Spanish) and translate comments into whichever version of the site you're viewing. Although it usually works well, Translate failed to detect this comment's source language, leaving it in the original Croatian.

The Next Web recently evaluated some of the best online translation services, praising Linguee, Worldwide Lexicon, Babelverse and Google Translate for their various strengths. Scott A Hale mentioned in a recent piece for FSD that services like Doulingo and Monotrans2 enable monolingual users to translate content and learn a new language at the same time. Although these services are all quite similar, Google Translate is the clear market leader.

As the New York Times stated in 2010, "Google's quick rise to the top echelons of the translation business is a reminder of what can happen when Google unleashes its brute-force computing power on complex problems". So how does it work? Rather than teaching computers grammatical structures and vocabularies, a pioneering technique for many translation services, Google uploads millions of human-translated documents from international institutions like the United Nations and European Parliament along with reputable websites and scanned books. The documents are then analysed for statistically significant patterns to generate reliable translations.

Fewer translated source documents for a particular language means fewer patterns and usually poorer translations. So while Google Translate is great for most of our 13 languages, especially those among the six official UN languages and 23 official EU languages, it's far from perfect for the likes of Farsi, Hindi and, as Tsukamoto's comment demonstrates, Japanese. It's even worse for languages with only a few million native speakers like Afrikaans and Macedonian. The fact that Google Translate even supports these small languages reflects the company's mission to "organize the world's information and make it universally accessible and useful". The fact that it's 64th language was Esperanto, an international constructed language boasting less than 1,000 native speakers, is more likely a testament to Google's quirky corporate identity. Building upon

original source documents, Google Translate relies on its 200 million monthly active users to suggest better results as they translate the equivalent of one million books per day.

What Google Translate and other services still notably lack is the ability to translate ASCII Latin transliterations of non-Latin script languages. What does that mean? ASCII, the American Standard Code for Information Interchange, represents the standard Latin/Roman script keyboards found on laptops and mobile phones in the West. Because many non-Latin scripts like Arabic, Farsi and Urdu only recently became supported across technology platforms, users who wished to write in these languages devised workarounds to get their message across in Latin script. This resulted in the emergence of ASCII Arabic (and other language derivatives), which uses a somewhat fluid mixture of Latin letters and numbers to transliterate words. If an Arabic speaker wanted to send an SMS several years ago saying ??? ????? — “Congratulations” in Arabic — he or she would have written “2alf mabrouk” or “2alf mabroooooooook!!” for extra emphasis. ASCII conventions remain common across Facebook, blogs, SMS and anywhere informal computer-mediated communication takes place.

Until machine translation makes sense of highly flexible ASCII writing conventions, services like Google Translate should be wary of claiming they’ve created the omnilingual Babel fish. Human translators might be slow and expensive by comparison, ßu† | $tiII d0n+ †#!nk m@¢h!ne§ w!II ???® u^der$+ªnd †#!$ (Translation: But I still don’t think machines will ever understand this).

We’re committed to reaching as many of you online as possible. In addition to our 13 languages, Free Speech Debate followers have translated our 10 draft principles into Estonian and Polish. If you would like to translate them into your language, please let us know.

Published on:September 20, 2012