Faulty machine translations litter the web

January 22 2024, by Peter Grad

Credit: CC0 Public Domain

Near the end of the last century, Bill Gates saw the prospect of unifying citizens of nearly 200 countries, speaking more than 7,000 languages, coming together in common dialogue through the suddenly burgeoning web community.

"The Internet is becoming the town square for the global village of tomorrow," he declared.

The Internet certainly has since drawn the world closer and has enriched global communications, commerce, research and entertainment immeasurably.

But a recent report reminds us—as if we really needed reminding—that along with progress sometimes come problems.

Researchers at Amazon Web Services Artificial Intelligence Lab and the University of California, Santa Barbara, say that after examining more than 6 billion sentences across the web, they have found that more than half had been translated into two or more different languages. The translations, they found, were often poor. And with each successive translation into other languages, some up to eight or nine, the results became worse.

The report, "A Shocking Amount of the Web is Machine Translated: Insights from Multi-Way Parallelism," was uploaded to the preprint server *arXiv* Jan. 11.

"The low quality of these … translations indicates they were likely created using machine translation," the authors report. "Our work raises serious concerns about training models such as multilingual large language models on both monolingual and bilingual data scraped from the web."

The researchers said texts are not only being translated by artificial intelligence but are being created by AI as well. They observed rates of AI-generated translations were highest among lower-resource languages, such as Wolof and Xhosa, African languages.

"We find that highly multi-way parallel translations are significantly lower quality than two-way parallel translations," the authors continue.

That means that as trillions of bits of data are ingested for AI training operations, regions under-represented on the web, such as African nations and other countries with more obscure languages, will face greater challenges in establishing reliable—and grammatical—large language models. With few native resources to draw upon, they must heavily rely on tainted translations flooding the market.

Mehak Dhaliwal, a former applied science intern at Amazon Web Services, told Motherboard in an interview, "We actually got interested in this topic because several colleagues who work in machine training and are native speakers of low resource languages noted that much of the internet in their native language appeared to be machine training generated… Everyone should be cognizant that content they view on the web may have been generated by a machine."

The Amazon researchers found bias in selection of content used for AI training.

They state, "Machine generated, multi-way parallel translations not only dominate the total amount of translated content on the web in lower resource languages, it also constitutes a large fraction of the total web content in those languages."

Such content, they suggested, tends to be simpler, lower-quality passages "likely produced to generate ad revenue." Since fluency and accuracy are lower for machine-trained material, numerous translations will lead to even less accurate content and increase the odds of AI hallucination.

Sometimes, computer-generated translations over the years have led to unintentionally humorous or embarrassing interpretations.

Google misinterpreted a phrase "Russia is a great country" and referred instead to Mordor, a fictional village in J.R.R. Tolkien's "The Lord of the Rings." Facebook's translation software in 2019 inadvertently referred to China's President Xi Jinping as "Mr. S***hole" several times in an English article translated from Burmese text. Facebook immediately apologized and blamed the mishap on a "technical error."

And a medical prescription translation tool for Armenian speakers provided some unfortunate advice for a patient with a headache.

English: "You can take over-the-counter ibuprofen as needed for pain."

Translation to Armenian: "You may take anti-tank missile as much as you need for pain."

More information: Brian Thompson et al, A Shocking Amount of the Web is Machine Translated: Insights from Multi-Way Parallelism, *arXiv* (2024). DOI: 10.48550/arxiv.2401.05749