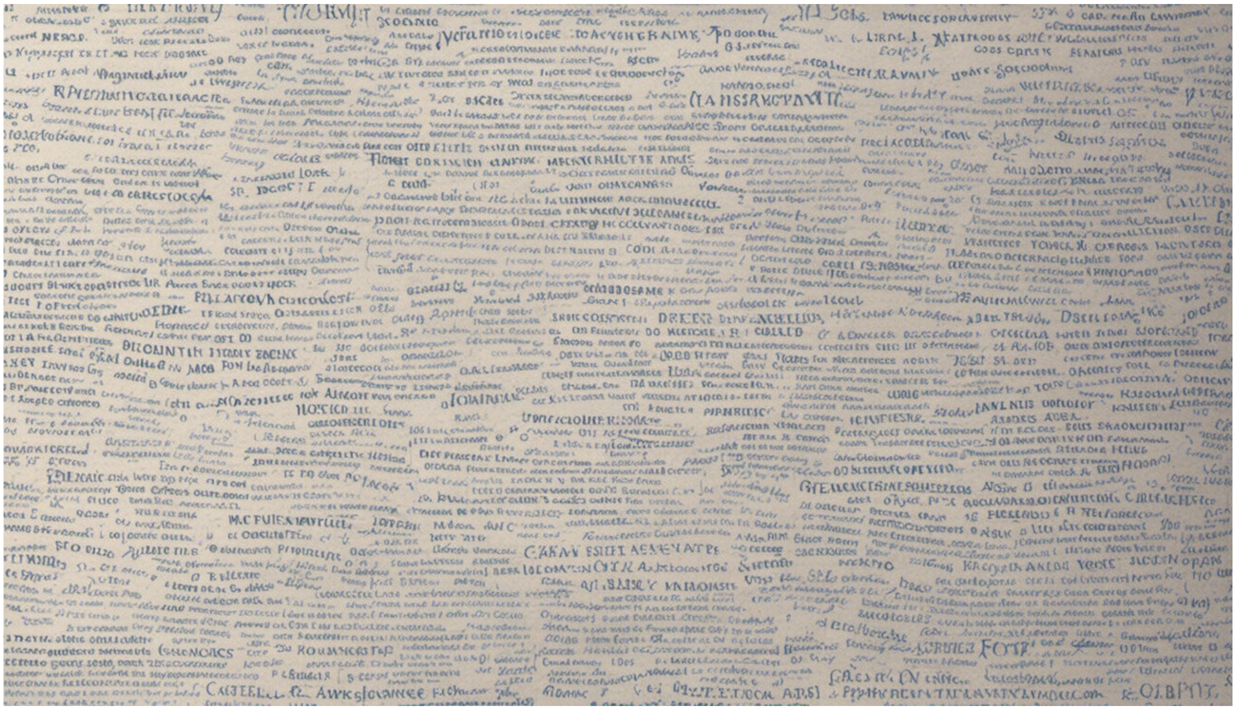
Sexist online translators get a little gender sensitivity training

March 31 2021, by Stefanie Ullmann and Danielle Saunders



Credit: AI-generated image (disclaimer)

Online translation tools have helped us learn new languages, communicate across linguistic borders, and view foreign websites in our native tongue. But the artificial intelligence (AI) behind them is far from perfect, often replicating rather than rejecting the biases that exist within a language or a society.

Such tools are especially vulnerable to gender stereotyping, because some languages (such as English) don't tend to gender nouns, while others (such as German) do. When translating from English to German, translation tools have to decide which gender to assign English words like "cleaner." Overwhelmingly, the tools conform to the stereotype, opting for the feminine word in German.

Biases are human: they're part of who we are. But when left unchallenged, biases can emerge in the form of concrete negative attitudes towards others. Now, our team has found a way to retrain the AI behind translation tools, using targeted training to help it to avoid gender stereotyping. Our method could be used in other fields of AI to help the technology reject, rather than replicate, biases within society.

Biased algorithms

To the dismay of their creators, AI algorithms often develop racist or sexist traits. Google Translate has been accused of stereotyping based on gender, such as its translations presupposing that all doctors are male and all nurses are female. Meanwhile, the AI language generator GPT-3—which wrote an entire article for the Guardian in 2020—recently showed that it was also shockingly good at producing harmful content and misinformation.

> Hungarian is a gender neutral language, it has no gendered pronouns, so Google Translate automatically chooses the gender for you. Here is how everyday sexism is consistently encoded in 2021. Fuck you, Google. pic.twitter.com/EPqkEw5yEQ
>
> — Dora Vargha (@DoraVargha) March 20, 2021

These AI failures aren't necessarily the fault of their creators. Academics and activists recently drew attention to gender bias in the Oxford English

Dictionary, where sexist synonyms of "woman"—such as "bitch" or "maid"—show how even a constantly revised, academically edited catalogue of words can contain biases that reinforce stereotypes and perpetuate everyday sexism.

AI learns bias because it isn't built in a vacuum: it learns how to think and act by reading, analysing and categorising existing data—like that contained in the Oxford English Dictionary. In the case of translation AI, we expose its algorithm to billions of words of textual data and ask it to recognise and learn from the patterns it detects. We call this process machine learning, and along the way patterns of bias are learned as well as those of grammar and syntax.

Ideally, the textual data we show AI won't contain bias. But there's an ongoing trend in the field towards building bigger systems trained on ever-growing data sets. We're talking hundreds of billions of words. These are obtained from the internet by using undiscriminating text-scraping tools like Common Crawl and WebText2, which maraud across the web, gobbling up every word they come across.

The sheer size of the resultant data makes it impossible for any human to actually know what's in it. But we do know that some of it comes from platforms like Reddit, which has made headlines for featuring offensive, false or conspiratorial information in users' posts.

New translations

In our research, we wanted to search for a way to counter the bias within textual data-sets scraped from the internet. Our experiments used a randomly selected part of an existing English-German corpus (a selection of text) that originally contained 17.2 million pairs of sentences—half in English, half in German.

As we've highlighted, German has gendered forms for nouns (doctor can be "der Arzt" for male, "die Ärztin" for female) where in English we don't gender these noun forms (with some exceptions, themselves contentious, like "actor" and "actress").

Our analysis of this data revealed clear gender-specific imbalances. For instance, we found that the masculine form of engineer in German (der Ingenieur) was 75 times more common than its feminine counterpart (die Ingenieurin). A translation tool trained on this data will inevitably replicate this bias, translating "engineer" to the male "der Ingenieur." So what can be done to avoid or mitigate this?

Overcoming bias

A seemingly straightforward answer is to "balance" the corpus before asking computers to learn from it. Perhaps, for instance, adding more female engineers to the corpus would prevent a translation system from assuming all engineers are men.

Unfortunately, there are difficulties with this approach. Translation tools are trained for days on billions of words. Retraining them by altering the gender of words is possible, but it's inefficient, expensive and complicated. Adjusting the gender in languages like German is especially challenging because, in order to make grammatical sense, several words in a sentence may need to be changed to reflect the gender swap.

Instead of this laborious gender rebalancing, we decided to retrain existing translation systems with targeted lessons. When we spotted a bias in existing tools, we decided to retrain them on new, smaller data-sets—a bit like an afternoon of gender-sensitivity training at work.

This approach takes a fraction of the time and resources needed to train models from scratch. We were able to use just a few hundred selected translation examples—instead of millions—to adjust the behaviour of translation AI in targeted ways. When testing gendered professions in translation—as we had done with "engineers"—the accuracy improvements after adapting were about nine times higher than the "balanced" retraining approach.

In our research, we wanted to show that tackling hidden biases in huge data-sets doesn't have to mean laboriously adjusting millions of training examples, a task which risks being dismissed as impossible. Instead, bias from data can be targeted and unlearned—a lesson that other AI researchers can apply to their own work.

This article is republished from The Conversation under a Creative Commons license. Read the original article.

Provided by The Conversation

Citation: Sexist online translators get a little gender sensitivity training (2021, March 31) retrieved 6 May 2024 from https://phys.org/news/2021-03-sexist-online-gender-sensitivity.html