CASE STUDY: Responsible AI

A Dataset for Evaluating Gender Bias in ML Translation Models

The Research

Researchers on the Translate team recently developed a new dataset for studying and preventing gender bias in machine learning in alignment with AI Principle #2, “avoid unfair bias.” This research explored gender translation between English and Spanish and English and German.

The research leverages the ways different languages employ gender markers to investigate potential gender bias in translation models. Spanish is a “pro drop” language, which means subject pronouns are optional. Both Spanish and German have grammatical gender, so they mark gender on adjectives that modify people and objects. Spanish has a single possessive

pronoun for his, her, and their, but English and German have separate pronouns for each. These grammatical gender differences across languages can pose a challenge for machine translation systems. This challenge is especially difficult when translating from a language without subject pronouns (such as Spanish) to one with required gendered subject pronouns (such as English).

Traditional neural machine translation (NMT) methods translate sentences one by one, but gender information often is not explicitly stated in every sentence. Seeking a novel way to address this challenge, the researchers built a new "context-aware" model that incorporated context from surrounding sentences or passages to be translated to improve gender accuracy when personal pronouns are translated.

When translating between languages with and without grammatical gender, the responsible AI challenge lies in training machine learning (ML) systems to choose the appropriate pronoun or maintain gender agreement between references throughout the content. Gender translation mistakes can be especially harmful errors, given that gender markers often convey a person's gender identity. The Translate team's new dataset was built to test the performance of this context-aware model, using gender differences across English, Spanish, and German to "challenge" the model to correctly translate people's genders across multiple sentences.

The Approach

The researchers applied for an AI Principles review of their dataset and proactively requested fairness testing. Reviewers and testers assessed the team's rationale for using Wikipedia biographies as a source for data. The researchers chose Wikipedia biographies because the entries are well-written, geographically diverse, contain multiple sentences, and refer to subjects in the third person, using many pronouns. The reviewers and testers also looked at the researchers' strategy to prioritize equal representation of feminine and masculine identities within the dataset, while acknowledging that there were not as many biographies for non-binary people available on Wikipedia. The researchers used articles about groups (which are referred to using gender-neutral "it" or "they" in English) to train ML models not to incorrectly generate gendered pronouns. In addition, the reviewers and testers looked at the researchers' decision to investigate gender translation accuracy for non-western names by sourcing Wikipedia biographies about people from 90 different nations spread across the world.

The Outcome

The result is the Translated Wikipedia Biographies dataset, which can be used to evaluate gender bias in translation models. This dataset enables a novel method of evaluation to help reduce gender bias in machine translations. Because each instance refers to a person with a known gender, the researchers could use the dataset to compute the model accuracy of the

gender-specific translations that refer to that person. This dataset provided useful performance measurements for the new context-aware models; using the dataset, researchers determined that context-aware models made 67% fewer gender translation errors than previous models that translated sentence by sentence. You can find examples of the kinds of improvements the context-aware model showed in the blog post about this research.

In alignment with AI Principle #4, “Be accountable to people,” the AI Principles reviewers recommended that the researchers publish a data card, which is a structured document offering details about how the dataset was created and tested. With respect to AI Principle #6, “Uphold high standards of scientific excellence,” the researchers decided to share the dataset publicly in order to support long-term improvements on ML systems focused on pronouns and gender in translation. The researchers make it clear that the dataset focuses on a *specific* problem related to gender bias and doesn't aim to cover all challenges of NMT, nor to be prescriptive in determining the optimal approach to address gender bias. This dataset and the research behind it aim to foster progress on this challenge across the global research community.