SCIENCE & TECHNOLOGY

Facebook Develops Machine Translation System for 100 Languages

October 21, 2020

Facebook has developed the first machine learning model that can translate between any two of 100 languages without going into English first.

Facebook says the new multilingual machine translation model was created to help its more than two billion users worldwide. The company is still testing the translation system – which it calls M2M-100 - and hopes to add it to different products in the future.

The social media service says it has made the system open source -- meaning its computer code will be freely available for others to copy or change.

Angela Fan, a research assistant at Facebook, explained the new machine translation model this week on one of the company's websites. She said its development represented a "milestone" in progress after years of "foundational work in machine translation."

Fan said the model produces better results than other machine learning systems that depend on English to help in the translation process. The other systems use it as an intermediate step -- like a bridge -- to translate between two non-English languages.

One example would be a translation from Chinese to French. Fan noted that many machine translation models begin by translating from Chinese to English first, and then from English to French. This is done "because English training data is the most widely available," she said. But such a method can lead to mistakes in translation.

"Our model directly trains on Chinese to French data to better **preserve** meaning," Fan said. Facebook said the system outperformed English-centered systems in a widely used system that uses data to measure the quality of machine translations.

Facebook says about two-thirds of its users communicate in a language other than English. The company already carries out an average of 20 billion translations every day on Facebook's News Feed. But it faces a huge test with many users publishing massive amounts of **content** in more than 160 languages.

The development team trained, or directed, the new model on a data set of 7.5 billion sentence **pairs** for 100 languages. In addition, the system was trained on a total of 2,200 language directions. Facebook said this is 10 times the number on the best machine translation models in the past.

One difficulty the team faced was trying to develop an effective machine translation system for language combinations that are not widely used. Facebook calls these "low-resource languages." The data used to create the new model was collected from content available on the internet. But there is limited internet data on low-resource languages.

To deal with this problem, Facebook said it used a method called back-translation. This method can create "**synthetic** translations" to increase the amount of data used to train on low-resource languages.

For now, the company says, it plans to continue exploring new language research methods while working to improve the new model. No date has been set for launching the translation system on Facebook.

But Angela Fan said the new system marks an important step for Facebook, especially for the times we live in. "Breaking language barriers through machine language translation is one of the most important ways to bring people together, provide **authoritative** information on COVID-19, and keep them safe from harmful content," she said.

I'm Bryan Lynn.

Bryan Lynn wrote this story for VOA Learning English, based on reports from Facebook and Agence France-Presse. George Grow was the editor.

We want to hear from you. Write to us in the Comments section, and visit our Facebook page.

_____

Words in This Story

translate – *v.* change written or spoken words from one language to another

code – *n.* a set of rules used to instruct computers how to behave or do things

milestone - *n.* an event that reaches never before seen levels

intermediate – *adj.* between two different stages in a process

preserve – *v.* keep something the same or prevent it from being damaged of destroyed

pair – *n.* two things that look the same and are used together

content – *n.* information contained in a piece of writing, a speech, a movie or on the internet

synthetic – *adj.* not made from substances or in the usual way

authoritative – *adj.* respected and considered to be accurate