Developing a machine translation tool to help asylum seekers at the border

August 29 2022, by Caitlin Dawson

Credit: CC0 Public Domain

Imagine fleeing persecution at home, surviving a difficult journey, arriving in a new country to seek asylum, only to be turned away at the border because nobody speaks your language. This is the reality for hundreds of migrants coming into the United States from remote areas of Central America who do not speak common languages, such as Spanish or Portuguese.

A shortage of translators for Indigenous asylum seekers speaking

traditional languages means many must wait for months or even years in Mexico to apply for asylum, creating a long backlog in an already overwhelmed immigration system.

"The U.S. immigration system is set up to handle English and Spanish," said Katy Felkner, a Ph.D. student in computer science at the USC Viterbi School of Engineering, "but there are several hundreds of people a year who are minority language speakers, in particular, speaking Indigenous languages from Mexico and Central America, who are not able to access any of the resources and legal aid that exists for Spanish-speaking migrants."

In other cases, people are unable to explain the threats to their lives in their hometowns, which could be the basis for asylum. When migrants cannot understand or be understood, there is no way to establish the threat to their safety during a "credible fear interview" conducted by the U.S. Department of Homeland Security.

The statistics are staggering: asylum-seeking immigrants without a lawyer prevailed in only 13 percent of their cases, while those with a lawyer prevailed in 74 percent of their cases, according to a study in the Fordham Law Review.

Felkner, who conducts her research at the USC Information Sciences Institute (ISI) under Jonathan May, a research associate professor, is working on developing a solution: a machine translation system for Mexican and Central American Indigenous languages that can be used by organizations providing legal aid to refugees and asylum-seekers.

"People are being directly adversely impacted because there aren't interpreters available for their languages in legal aid organizations," said Felkner. "This is a concrete and immediate way that we can use natural language processing for social good."

"People are being directly adversely impacted because there aren't interpreters available for their languages in legal aid organizations." Katy Felkner.

Giving asylum seekers a fair chance

Felkner is currently working on a system for a Guatemalan language, which is one of the 25 most common languages spoken in immigration court in recent years, according to The New York Times.

"We're trying to provide a rough translation system to allow nonprofits and NGOs that don't have the resources to hire interpreters to provide some level of legal assistance and give asylum seekers a fair chance to get through that credible fear interview," said Felkner.

Felkner's interest in languages began during her undergraduate degree at the University of Oklahoma, where she earned a dual degree in computer science and letters, with a focus on Latin. During her first year of college, she worked on a project called the Digital Latin Library, writing Python code to create digital versions of ancient texts.

"That's what got me thinking about language technology," said Felkner. "I taught myself some basics of natural language processing and ended up focusing on machine translation because I think it's one of the areas with the most immediate human impact, and also one of the most difficult problems in this area."

While Felkner and May are currently focused on developing a text-to-text translator, the end goal, years from now, is a multilingual speech-to-speech translation system: the lawyer would speak English or Spanish, and the system would automatically translate into the asylum seeker's Indigenous language, and vice-versa.

Pushing the lower bound

Translation systems are trained using parallel data: in other words, they learn from seeing translation pairs, or the same text in both languages, at the sentence level. But there is very little parallel data in Indigenous languages, including K'iche', despite it being spoken by around one million people.

That's because parallel data only exists when there is a compelling reason to translate into or out of that language. Essentially, said Felkner, if it's commercially viable—Disney dubbing films from English to Spanish, for instance—or stemming from a religious motivation.

In many cases, due to the influence of missionaries throughout Latin America, the only parallel data source—the same text in both languages—is the Bible, which doesn't give researchers much to work with.

"Imagine you're an English speaker trying to learn Spanish, but the only Spanish you're ever allowed to see is the New Testament," said Felkner. "It would be quite difficult."

That's bad news for the data-hungry deep learning models used by language translation systems that take a quantity over quality approach.

"The models have to see a word, phrase, grammatical construction a bunch of times to see where it's likely to occur and what it corresponds to in the other language," said Felkner. "But we don't have this for K'iche' and other extremely low resource Indigenous languages."

The numbers speak for themselves. From English to K'iche', Felkner has roughly 15,000 sentences of parallel data, and 8,000 sentences for Spanish to K'iche'. By contrast, the Spanish to English model she trained

for some baseline work had 13 million sentences of training data.

"We're trying to work with essentially no data," said Felkner. "And this is the case for pretty much all low-resource languages, even more so in the Americas."

One tactic in existing low-resource work uses closely related, higher resource languages as a starting point: for instance, to translate from English into Romanian, you would start training the model in Spanish.

But since Indigenous languages of the Americas developed separately from Europe and Asia, the majority are low resource, and most of them are extremely low resource, a term Felkner coined to describe a language with less than around 30,000 sentences of parallel data.

"We're really trying to push the lower bound on how little data you can have to successfully train a machine translation system," said Felkner.

Creating something from nothing

But Felkner, with her background in linguistics, was undeterred. Over the past two years, she has worked on creating language data for the models using some tricks of the trade in natural language processing.

One tactic involves teaching the model to complete the abstract task of translation and then setting it to work on the specific language in question. "It's the same principle as learning to drive a bus by learning to drive a car first," said Felkner.

To do this, Felkner took an English to Spanish model, and then fine-tuned it for K'iche' to Spanish. It turned out, this approach, called transfer learning, showed promise even in an extremely low resource case. "That was very exciting," said Felkner. "The transfer learning

approach and pre-training from a not-closely-related language had never really been tested in this extremely low resource environment, and I found that it worked."

She also tapped into another resource: using grammar books published by field linguists in the mid-to-late 70s to generate plausible synthetic data that can be used to help the models learn. Felkner is using the grammar books to write rules that will help her construct syntactically correct sentences from the dictionaries. The technical term for this is bootstrapping or data augmentation—or colloquially, "fake it 'til you make it."

"We use this as pre-training data, to essentially teach the models the basics of grammar," said Felkner. "Then, we can save our real data, such as the Bible parallel data, for the fine-tuning period when it will learn what's semantically meaningful, or what actually makes sense."

Finally, she's testing a technique that involves parsing nouns in the English and K'iche' sides of the Bible, replacing them with other nouns, and then using a set of rules to correctly inflect the sentences for grammar.

For example, if the training data has the sentence: 'the boy kicked the ball,' the researchers could use this approach to generate sentences like 'the girl kicked the ball', 'the doctor kicked the ball', 'the teacher kicked the ball,' which can all become training data.

"The idea is to use these synthetically-generated examples to essentially build a rough version of the system, so that we can get a lot of use out of the small amount of real data that we do have, and finetune it to exactly where we want it to be," said Felkner.

Immediate humanitarian impact

Working in extremely low-resource language translation is not easy, and it can be frustrating at times, admits Felkner. But the challenge, and the potential to change lives, drive her to succeed.

Within the next year, she plans to undertake a field trip to observe how legal aid organizations are working at the border, and where her system could fit into their workflow. She is also working on a demo website for the system, which she hopes to unveil in 2023, and once developed, she hopes the system could one day be applied to other Indigenous languages.

"Hill climbing on high resource languages can make your Alexa, Google Home or Siri understand you better, but it's not transformative in the same way," said Felkner. "I'm doing this work because it has an immediate humanitarian impact. As JFK once said, we choose to go to the moon not because it is easy, but because it is hard. I often think the things that are worth doing are difficult."

Provided by University of Southern California