Microsoft Research India is creating tools to help preserve fast disappearing languages

Synopsis

Every two weeks, a language is lost somewhere in the world, as the number of people who speak that dwindles.

In 2010, Bo, a **language** of the Andaman Islands that is at least 65,000 years old became extinct when the only person who spoke this pre-Neolithic tongue died.

This isn’t an isolated case.

Every two weeks, a language is lost somewhere in the world, as the number of people who speak that dwindles.

A lack of digital representation of these languages means non-written languages such as Idu Mishmi, spoken by a community of 11,000 in Arunachal Pradesh, may not be accessible for future generations.

In 2015, the **Microsoft Research** lab in India (MSR) started a project to build a digital presence for these low-resource languages.

Project Ellora — Enabling Low Resource Languages – aims to preserve a language for posterity and ensure that communities that speak these languages can participate in the digital world.

“Ellora started when Microsoft Research India was looking at **Indian languages** and how it could get more data and techniques to build tech and applications for Indian languages,” said Kalika Bali, principal researcher at Microsoft Research India, who leads this project.

So far, MSR has worked on three Indian languages - Gondi, Mundari and Idu Mishmi.

Mainstream Indian languages do have a digital presence, but MSR is focusing on these low-resource languages.

The internet is dominated by English, and seven other widely spoken languages including Chinese and Spanish.

This means 88% of the world’s languages do not have enough of a presence on the internet. About 1.2 billion people, or 20% of the world’s population, can’t use their own language to navigate the digital world.

The lack of data, said Bali, was the biggest hurdle in creating tools and techniques to preserve these languages.

Microsoft Research works with partner organisations to identify communities and create the database with native speakers that will then be used to build AI technologies for them.

For instance, for Mundari, a language spoken by the Mundas, a one million strong community spread across parts of West Bengal, Odisha and Jharkhand, MSR collaborated with IIT-Kharagpur and undertook a study on what the community needs to keep the language alive.

This has now evolved from a simple vocabulary game for school children to a Hindi-to-Mundari text translation tool and a speech recognition model that will provide the community access to more content in the language.

The team at MSR has also collaborated with Karya, a digital work platform, for capturing and annotating data to build ML and AI models that create the text-to-speech dataset.

Creating the framework is a challenging and time-consuming process, with the IIT professors first working with members of the community to manually translate sentences from Hindi to Mundari. To speed up the translation, MSR researchers developed a new technology called Interneural Machine Translation (INMT), which helps predict the next word when someone is translating between languages.

“It (INMT) allows for humans to translate from one language to another more effectively. If I’m translating from Hindi to Mundari, when I start typing in Mundari, it gives me predictive suggestions in Mundari itself. It’s like the predictive text you get in smartphone keyboards, except that it does it across two languages,” Bali said.

The tool has been designed to work across languages and has been used by Translators without Borders, a non-profit organisation, to translate from French to Congolese Swahili.

The tool is also part of the Karya app.

MSR India is also working closely with Microsoft Africa Research Institute to see how this tool can be used by them in Africa.

The next step, said Bali, would be to make these tools available to other communities.

“One of the things people keep talking about is that it is very resource intensive to build systems for all the languages, so we are trying to answer that question and see how much data, and money, you really need to build an efficient system,” she said.