Computer scientist helps preserve endangered language for future generations

January 12 2023

Gyalrong textbook. Credit: University of Sheffield

A Chinese language at risk of extinction is being kept alive for future generations with the help of Department of Computer Science research.

Using natural language processing (NLP)—computational processes designed to understand speech and text as humans can—the Gyalrong language and the rich cultural history it carries are being preserved.

Gyalrong, which is spoken by a very limited population in China's Sichuan Province, is estimated to date back over 1,000 years but is now thought to have fewer than 33,000 speakers.

Most native speakers are elderly and with many young people leaving the villages in which it is spoken to seek work in urban areas, fewer and fewer people have the opportunity to learn the language from elders.

It is estimated that the decline of the language—which has little in the way of written records and is considered very difficult to learn—will become irreversible over the next few decades.

Xutan Peng, a Ph.D. student at the University's Department of Computer Science, is using his research to speed up the production of a textbook to teach the endangered language to local schoolchildren.

"Many people say language is the DNA of a culture," said Xutan.

"If the language dies the memory of this rich culture is in danger of being lost forever. Things such as old stories passed to their children and grandchildren by elders will be no more, and it will be impossible for future generations to learn the culture and traditions."

His technique takes Gyalrong texts and summarizes them into Mandarin using an automated process. As such, language documentation work that could take a linguist months or years by immersing themselves in the culture can be done far more rapidly.

"One way to imagine it is that there are two libraries, side by side, with the same architecture and layout but with one exclusively supplying Mandarin texts, and the other Gyalrong," said Xutan.

"If two similar books, covering similar subject matter, are in the

corresponding location in both libraries and you move both buildings into one location, you can align the two to identify patterns.

"So, as long as we're able to master certain frequently used words, we can use this technique to make educated guesses to piece the jigsaw together."

You can read more about the process, known as cross-lingual word embedding (CLWE), in the papers "Cross-Lingual Word Embedding Refinement by $\ell 1$ Norm Optimisation" and "Understanding Linearity of Cross-Lingual Word Embedding Mappings." The technique used on documenting Gyalrong also draws on research from Xutan's earlier paper, "Summarising Historical Text in Modern Languages."

The results of Xutan's work are already bearing fruit, with a small group of Chinese schoolchildren, whose families can speak at least some Gyalrong, learning from and providing feedback on a textbook. It is hoped this first version will be followed by further volumes as more data is collected.

Its success has even caught the attention of documentary makers, who've featured the story on China Central Television.

"It's a unique and very satisfying project to work on," Xutan added.

"And although it may be limited in scope, we're making a real impact on society. It also suggests a very bright future for this type of technique in helping to preserve endangered languages."

Xutan plans to explore how the technique could be adapted to help document other endangered languages.

Dr. Mark Stevenson, a senior lecturer in the natural language processing

research group, said, "Endangered languages, like Gyalrong, face a real risk of extinction. This project shows how NLP, including work carried out within Sheffield's NLP research group, can help preserve them for future generations."

Provided by University of Sheffield

Citation: Computer scientist helps preserve endangered language for future generations (2023, January 12) retrieved 8 June 2023 from https://techxplore.com/news/2023-01-scientist-endangered-language-future-generations.html