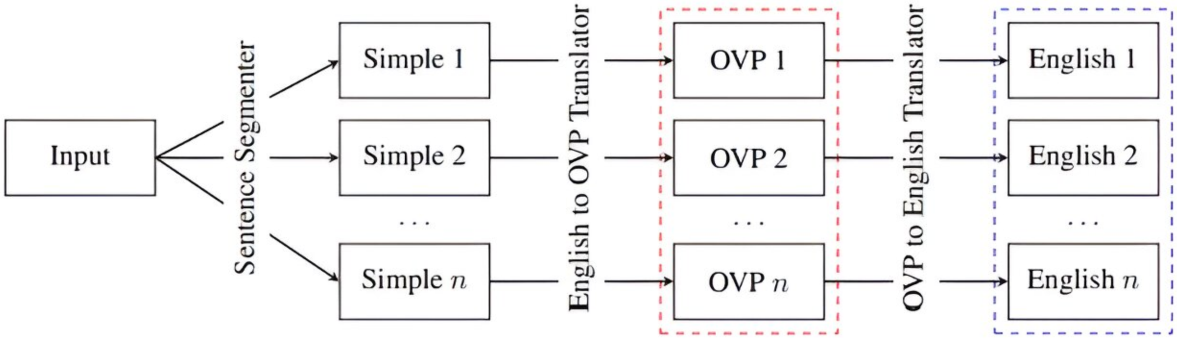
Student builds AI tool to revitalize endangered Indigenous language

June 21 2024, by Greg Hardesty



The entire English to OVP translation process. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2405.08997

Jared Coleman, who recently earned his Ph.D. in computer science, and his supervisor, Bhaskar Krishnamachari, are bound by a shared love of languages—both human and computer.

Krishnamachari grew up in India speaking Tamil, Hindi and English, and started learning French and Mandarin Chinese in college. Coleman, who was raised Anglophone, loved Spanish in high school and learned Portuguese from his now-wife and friends in college.

During the pandemic, Coleman started taking online classes in a lesser-known language: Owens Valley Paiute. Coleman is a member of the Big

Pine Paiute Tribe of Owens Valley—his father, David, grew up on the tribe's reservation in Big Pine, CA, and Paiute is his ancestral language.

ChatGPT and other large language models (LLMs) exhibit human-level performance on many natural-language tasks in English because one-fifth of the world speaks English. The same is true of other widely used tongues. But Paiute is deemed a "no-resource language," meaning there are no publicly available Paiute sentences translated into English on which to train a machine learning model.

In a new paper, "LLM-Assisted Rule-Based Machine Translation for Low/No-Resource Languages," appearing on the pre-print server *arXiv*, Coleman and Krishnamachari propose a machine translation approach called LLM-RBMT (Rule-Based Machine Translation) to help people learn no-resource languages. The paper's co-authors are Khalil Iskarous, a USC Dornsife associate professor of linguistics, and Ruben Rosales, an independent researcher.

Their approach consists of a more "old school" rule-based translator tools and a more advanced, natural language-based LLM. In the researchers' method, the LLM does not translate into or from Owens Valley Paiute. Instead, it helps to guide the rule-based translators, which rely on grammatical and vocabulary rules to translate between languages.

"Essentially, the LLM acts as a sophisticated intermediary, using its advanced understanding of language to make sure the rule-based system produces accurate translations," said Coleman.

The translation tool simplifies complex sentences and uses placeholders (in this case, English words) for unknown words. While this process loses some meaning, it still produces understandable and grammatically correct translations.

This method, said Coleman, mirrors how language learners naturally speak by mixing known and unknown words, making it a practical tool for real-world use.

"The tool is smart enough, given a few hints, to be able to do a lot of the translation on its own," adds Krishnamachari.

Personal satisfaction

Coleman also built and maintains a suite of digital tools related to language revitalization, named Kubishi or 'brain' in Paiute, including an online dictionary and a sentence-builder and translation system enabled by this research.

Overall, the paper, which will be presented at NAACL's AmericasNLP workshop, found that LLM's remarkable general-purpose language skills make them a promising tool in helping revitalize critically endangered languages.

For his part, Coleman credits his tribe's members, past and present, for paving the path. "A lot of people in my tribe have been working for a long time on different language revitalization efforts, including classes, dictionaries, recordings," said Coleman. "So as excited about this research as I am, I know it is one piece of a much larger puzzle."

Indeed, the paper points to many directions for future work, including adding more complex sentence structures to test the limits of the methodology outlined in his paper. Beyond that, it's both a personal and academic achievement for Coleman, who will join Loyola Marymount University as an assistant professor in computer science this fall.

"My dad did not grow up speaking the language—like many families, it was forced out of use by boarding schools where speaking the language was forbidden," said Coleman.

"I'm lucky my great-grandparents sat down with linguists to document the language and to create recordings so I can hear their voices and words. And now, to listen to my great-grandfather and know what he is saying, there's something very personally satisfying about that."

More information: Jared Coleman et al, LLM-Assisted Rule Based Machine Translation for Low/No-Resource Languages, *arXiv* (2024). DOI: 10.48550/arxiv.2405.08997

Provided by University of Southern California

Citation: Student builds AI tool to revitalize endangered Indigenous language (2024, June 21) retrieved 26 June 2024 from https://techxplore.com/news/2024-06-student-ai-tool-revitalize-endangered.html