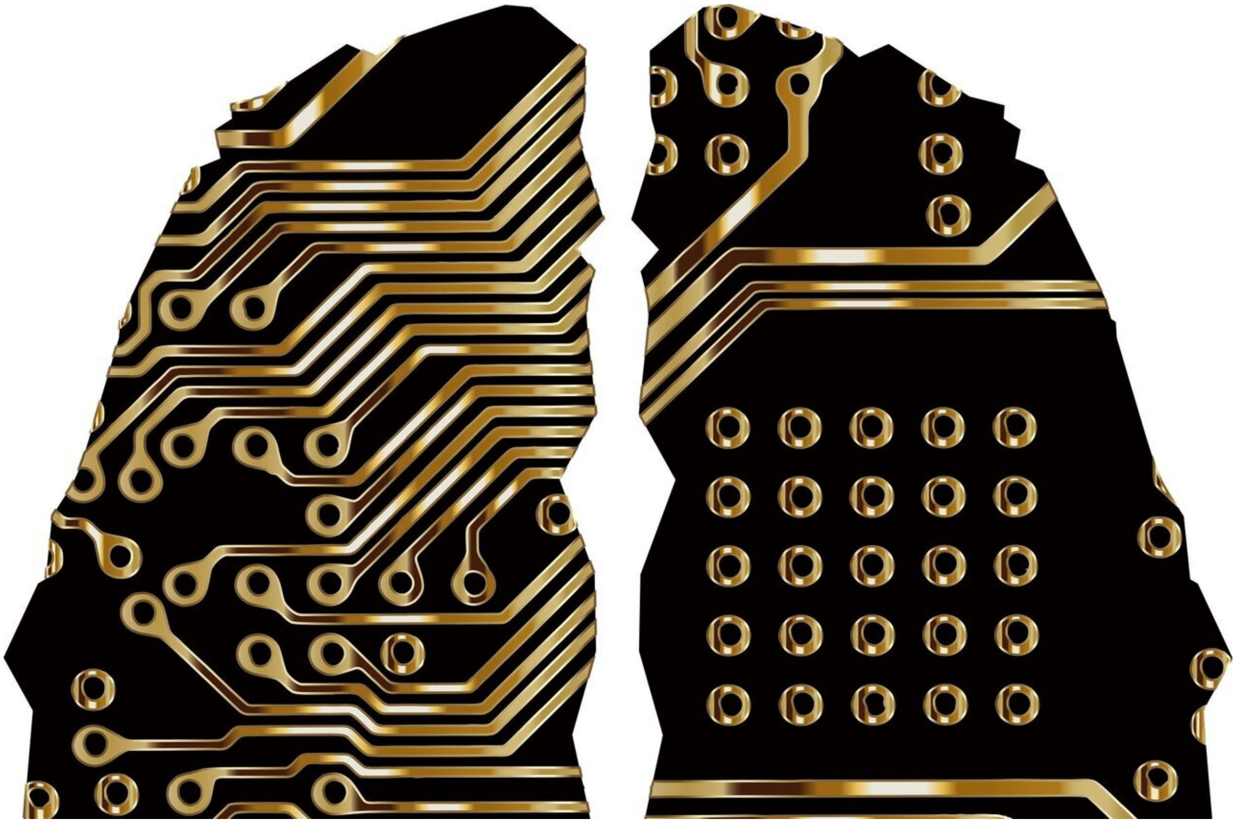
New AI brings the power of natural language processing to African languages

November 9 2021



Credit: Pixabay/CC0 Public Domain

Researchers have developed an AI model to help computers work more efficiently with a wider variety of languages.

African languages have received little attention from computer scientists, so few natural language processing capabilities have been available to large swaths of the continent. The new language model, developed by researchers at the University of Waterloo's David R. Cheriton School of Computer Science, begins to fill that gap by enabling computers to analyze text in African languages for many useful tasks.

The new neural network model, which the researchers have dubbed AfriBERTa, uses deep-learning techniques to achieve state-of-the-art results for low-resource languages.

The neural language model works specifically with 11 African languages, such as Amharic, Hausa, and Swahili, spoken collectively by more than 400 million people. It achieves output quality comparable to the best existing models despite learning from just one gigabyte of text, while other models require thousands of times more data.

"Pretrained language models have transformed the way computers process and analyze textual data for tasks ranging from machine translation to question answering," said Kelechi Ogueji, a master's student in computer science at Waterloo. "Sadly, African languages have received little attention from the research community."

"One of the challenges is that neural networks are bewilderingly text- and computer-intensive to build. And unlike English, which has enormous quantities of available text, most of the 7,000 or so languages spoken worldwide can be characterized as low-resource, in that there is a lack of data available to feed data-hungry neural networks."

Most of these models work using a technique known as pretraining. To accomplish this, the researcher presented the model with text where some of the words had been covered up or masked. The model then had to guess the masked words. By repeating this process, many billions of times, the model learns the statistical associations between words, which mimics human knowledge of language.

"Being able to pretrain models that are just as accurate for certain downstream tasks, but using vastly smaller amounts of data has many advantages," said Jimmy Lin, the Cheriton Chair in Computer Science and Ogueji's advisor. "Needing less data to train the language model means that less computation is required and consequently lower carbon emissions associated with operating massive data centres. Smaller datasets also make data curation more practical, which is one approach to reduce the biases present in the models."

"This work takes a small but important step to bringing natural language processing capabilities to more than 1.3 billion people on the African continent."

Assisting Ogueji and Lin in this research is Yuxin Zhu, who recently completed an undergraduate degree in computer science at Waterloo. Together, they present their research paper, Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resource languages, at the Multilingual Representation Learning Workshop at the 2021 Conference on Empirical Methods in Natural Language Processing.

Provided by University of Waterloo

Citation: New AI brings the power of natural language processing to African languages (2021, November 9) retrieved 20 March 2024 from https://techxplore.com/news/2021-11-ai-power-natural-language-african.html