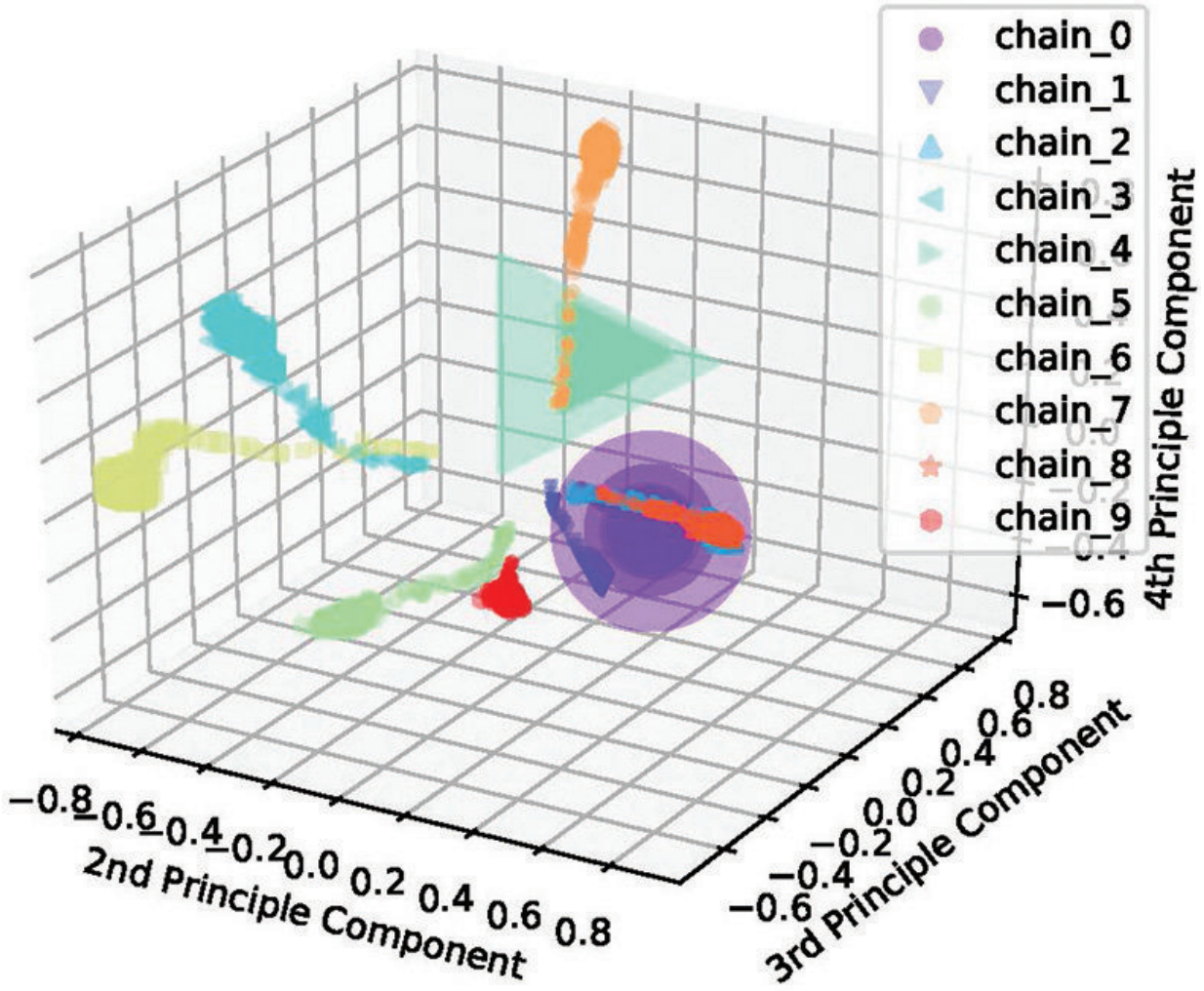
Linguistics team using Ohio Supercomputer Center to translate lesser-known languages

November 21 2017, by Ross Bishoff



This graph displays an algorithm that explores the space of possible probabilistic grammars and maps out the regions of this space that have the highest probability of generating understandable sentences. Credit: Ohio Supercomputer Center

Off the top of your head, how many languages can you name? Ten? Twenty? More?

It is estimated there are more than 7,000 languages worldwide. For those involved in disaster relief efforts, the breadth and variety of that number can be overwhelming, especially when addressing areas with low resources.

William Schuler, Ph.D., a linguistics professor at The Ohio State University, is part of a project called Low Resource Languages for Emergent Incidents (LORELEI), an initiative through the Defense Advanced Research Projects Agency (DARPA). The LORELEI program's goal is to develop technology for languages about which translators and linguists know nothing.

As part of LORELEI, Schuler and his team are using the Ohio Supercomputer Center's Owens Cluster to develop a grammar acquisition algorithm to discover the rules of lesser-known languages, learning the grammars without supervision so disaster relief teams can react quickly. "We need to get resources to direct disaster relief and part of that is translating news text, knowing names of cities, what's happening in those areas," Schuler said. "It's figuring out what has happened rapidly, and that can involve automatically processing incident language."Schuler's team is working to build a Bayseian sequence model based on statistical analysis to discover a given language's grammar. It is hypothesized this parsing model can be trained to learn a language and make it syntactically useful.

"The computational requirements for learning grammar from statistics are tremendous, which is why we need a supercomputer," Schuler said. "And it seems to be yielding positive results, which is exciting."

On a powerful single server, Schuler's team can analyze 10to15 categories of grammar, according to Lifeng Jin, a Ph.D. student who oversees the computational aspects of the project. But using the GPUs on OSC's Owens System allows Jin to increase the number of categories greatly.

GPUs - graphics processing units - are more powerful and cost-efficient than CPUs - central processing units. CPUs are the brains of a computer and are composed of just a few cores with plenty of cache memory. GPUs are a complementary processing unit to CPUs composed of hundreds of cores that can handle thousands of threads simultaneously. GPUs have the capability to quickly execute computations important in engineering analysis and simulation.

"We can increase the complexity of the model exponentially, so we can use 45 to50 categories and get results in an even shorter amount of time," Jin said. "It's a more realistic scenario of imitating what humans are doing. The models are really big, so memory is crucial.

"The statistical model is also very complicated. In order to train it, we have to do a lot of computation. Say we have 20,000 sentences from a given language, we use that to train the grammar. That's where OSC comes in. In the first stage, we tried to train the grammar using CPUs, but they're too slow. So we refactored our code to use GPUs for sampling, and it's sped up our process greatly."

Speed is critical in the project because the LORELEI goal is quick response to disaster relief, meaning high performance computing is critical. In August, DARPA organized a trial run to simulate two real disasters in Africa. Schuler's group used 60 GPUs on the Owens Cluster for seven days for four grammars of two languages, illustrating the importance of OSC's resources to the project.

Jin said that as they begin using more realistic configurations for grammars, the size of the grammars and the computation required to explore them will be even greater, giving OSC an even greater future role as the research evolves.

"For rapid grammar acquisition, when minutes count you need lots of power in a hurry," Schuler said.

"We're answering these fundamental questions about what it means to be human and have language and be the animal that talks to each other.

The ability to ask these kinds of questions and get answers is a relatively recent innovation that requires the high performance computing infrastructure OSC gives us. It's really a game-changer."

Provided by Ohio Supercomputer Center

Citation: Linguistics team using Ohio Supercomputer Center to translate lesser-known languages (2017, November 21) retrieved 30 April 2024 from https://phys.org/news/2017-11-linguistics-team-ohio-supercomputer-center.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.