IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

PRESS RELEASE

7 Nov 2023

Languages with more speakers tend to be harder for machines to learn

Just a few months ago, many people would have found it unimaginable how well artificial intelligence-based "language models" could imitate human speech. What ChatGPT writes is often indistinguishable from human-generated text. A research team at the Leibniz Institute for the German Language (IDS) in Mannheim, Germany have now used text material in 1,293 different languages to investigate how quickly different computer language models learn to "write". The surprising result: languages that are spoken by a large number of people tend to be more difficult for algorithms to learn than languages with a smaller linguistic community.

Language models are computer algorithms that can process and generate human language. A language model can recognize patterns and regularities in large amounts of textual data and thus gradually learns to predict future text. One particular language model is the so-called "Transformer" model, on which the well-known chatbot service, ChatGPT, has been built. As the algorithm is fed human-generated text, it develops an understanding of the probabilities with which word components, words and phrases occur in particular contexts. This learned knowledge is then used to make predictions, i.e. to generate new texts in new situations.

For example, when a model analyzes the sentence "In the dark night I heard a distant ...", it can predict that words like "howl" or "noise" would be appropriate continuations. This prediction is based on some "understanding" of the semantic relationships and probabilities of word combinations in the language.

In a new study, a team of linguists at the IDS has investigated how quickly computer language models learn to predict by training them on text material in 1,293 languages. The team used older and less complex language models as well as modern variants such as the Transformer model mentioned above. They looked at how long it took different algorithms to develop an understanding of patterns in the different languages. The study found that the amount of text an algorithm needs to process in order to learn a language – that is, to make predictions about what will follow – varies from language to language. It turns out that language algorithms tend to have a harder time learning languages with many native speakers than languages represented by a smaller number of speakers.

However, it is not as simple as it sounds. To validate the relationship between learning difficulty and speaker population size, it is essential to control for several factors. The challenge is that languages that are closely related (e.g., German and Swedish) are much more similar than languages that are distantly related (e.g., German and Thai). However, it is not only the degree of relatedness between

languages that needs to be controlled for, but also other effects such as the geographical proximity between two languages or the quality of the text material used for training. "In our study, we used a variety of methods from applied statistics and machine learning to control for potential confounding factors as tightly as possible," explains Sascha Wolfer, one of the two authors of the study.

However, regardless of the method and the type of input text used, a stable statistical correlation was found between machine learnability and the size of the speaker population. "The result really surprised us; based on the current state of research, we would have expected the opposite: that languages with a larger population of speakers tend to be easier for a machine to learn," says Alexander Koplenig, lead author of the study. The reasons for this relationship can only be speculated about thus far. For example, an earlier study led by the same research team demonstrated that larger languages tend to be more complex overall. So maybe the increased learning effort "pays off" for human language learners: because once you have learned a complex language, you have more varied linguistic options available to you, which may allow you to express the same content in a shorter form. But more research is needed to test these (or other explanations) out. "We're still relatively at the beginning here," Koplenig points out. "The next step is to find out whether and to what extent our machine learning results can be transferred to human language acquisition."

Original publication:

Koplenig, Alexander & Wolfer, Sascha. 2023. Languages with more speakers tend to be harder to (machine-)learn. Scientific Reports 13(1). 18521. DOI: https://doi.org/10.1038/s41598-023-45373-z

Contact for scientific information:

Dr. Sascha Wolfer
Leibniz Institute for the German Language
R 5, 6-13
D-68161 Mannheim
Tel.: +49 621 / 1581 – 439
Email: wolfer@ids-mannheim.de