IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

PRESSEMITTEILUNG

7.11.2023

Häufig gesprochene Sprachen - schwerer erlernbar für Maschinen?

Noch vor einigen Monaten war es für viele Menschen undenkbar, wie gut auf Künstlicher Intelligenz beruhende sogenannte „Sprachmodelle“ die menschliche Sprachfähigkeit imitieren können. Was zum Beispiel ChatGPT schreibt, ist oft nicht von menschen-generierten Texten zu unterscheiden. Anhand von Textmaterial in 1.293 verschiedenen Sprachen haben zwei Forscher des Leibniz-Instituts für Deutsche Sprache (IDS) in Mannheim nun untersucht, wie schnell verschiedene Computer-Sprachmodelle „schreiben“ lernen. Das überraschende Ergebnis der Studie: Sprachen, die von mehr Menschen gesprochen werden, sind für Algorithmen tendenziell schwieriger zu lernen als Sprachen, die eine kleinere Sprachgemeinschaft haben.

Sprachmodelle sind Computer-Algorithmen, die in der Lage sind, menschliche Sprache zu verarbeiten und zu generieren. Das Modell erkennt dabei Muster und Regelmäßigkeiten in einer großen Menge an Texten und lernt so nach und nach, zukünftige Textdaten vorherzusagen. Ein bekanntes Modell ist das sogenannte „Transformer“-Modell, das u.a. auch dem bekannten Chatbot ChatGPT zugrunde liegt. Während der Algorithmus mit menschlich generierten Texten gefüttert wird, entwickelt dieser ein gewisses Verständnis für die Wahrscheinlichkeiten, mit denen Wortbestandteile, Wörter und Phrasen in bestimmten Kontexten auftreten. Dieses erlernte Wissen wird dann zur Vorhersage, also zur Generierung von neuen Texten in neuen Situationen verwendet.

Wenn die Modelle zum Beispiel den Satz „In der dunklen Nacht hörte ich ein fernes …“ analysieren, können sie vorhersagen, dass Wörter wie „Heulen“ oder „Geräusch“ angemessene Fortsetzungen wären. Diese Vorhersage basiert auf einem gewissen „Verständnis“ für die semantischen Zusammenhänge und die Wahrscheinlichkeit von Wortkombinationen in der Sprache.

Sprachforscher des IDS haben nun in einer neuen Studie untersucht, wie schnell Computer-Sprachmodelle diese Art Vorhersage lernen, indem sie diese Textmaterial in 1.293 Sprachen lernen ließen. Dabei haben sie ältere und einfachere Modelle, aber auch moderne Varianten wie das oben angesprochene Transformer-Modell eingesetzt. Sie untersuchten, wie lange verschiedene Algorithmen brauchen, um ein Verständnis der Regelmäßigkeiten in den verschiedenen Sprachen zu entwickeln. Die Studie ergab, dass die Textmenge, die ein Algorithmus verarbeiten muss, um eine Sprache zu erlernen – also Vorhersagen darüber zu treffen, was als nächstes folgt – von Sprache zu Sprache unterschiedlich ist. Dabei zeigte sich, dass es Sprachalgorithmen tendenziell schwerer fällt, Sprachen mit vielen Muttersprachler/-innen zu erlernen, im Vergleich zu Sprachen, die von einer kleineren Sprecherzahl repräsentiert werden.

So einfach, wie diese Idee zu Beginn klingen mag, ist es aber doch nicht. Denn um wirklich sicherzugehen, dass Unterschiede in der Erlernbarkeit von Sprachen bestehen, müssen einige Dinge beachtet werden. Die Herausforderung liegt darin, dass Sprachen, die verwandt sind (z.B. Deutsch und Schwedisch) sich in vielen Aspekten viel ähnlicher sind als Sprachen, die sich verwandtschaftlich ferner sind (z.B. Deutsch und Thai). Doch nicht nur der Verwandtschaftsgrad von Sprachen muss kontrolliert werden, sondern auch andere Effekte wie die geographische Nähe zwischen zwei Sprachen oder die Qualität des Textmaterials, das zum Training verwendet wird. „In unserer Studie haben wir deshalb verschiedene Methoden aus der angewandten Statistik und dem maschinellen Lernen verwendet, um etwaige Einflussfaktoren möglichst streng kontrollieren zu können“, erklärt Sascha Wolfer, einer der beiden Autoren der Studie.

Unabhängig von der verwendeten Methode und über verschiedene Textsorten hinweg zeigte sich für die untersuchten Sprachmodelle eine stabile statistische Korrelation zwischen maschineller Erlernbarkeit und Größe der Sprecherpopulation. „Das Ergebnis hat uns wirklich überrascht, aufgrund des bisherigen Forschungsstands hätten wir eigentlich eher das Gegenteil erwartet, also dass Sprachen mit größerer Sprecherpopulation eher leichter maschinell erlernbar sind“, sagt Alexander Koplenig, Hauptautor der Studie. Über die Gründe für den Zusammenhang kann deshalb bisher nur spekuliert werden. In einer vorherigen Studie des Forscherteams zeigte sich etwa, dass größere Sprachen auch insgesamt komplexer sind. Es könnte also sein, dass sich der vermehrte Lernaufwand „lohnt“, wenn Menschen Sprache lernen: Denn hat man einmal eine komplexe Sprache erlernt, hat man vielfältigere sprachliche Ausdrucksmöglichkeiten zur Verfügung, die es erlauben könnten, den gleichen Inhalt in kürzerer Form auszudrücken. Um diese (oder andere) Erklärungen zu überprüfen, ist aber noch weitere Forschung nötig. „Wir stehen hier noch relativ am Anfang“, betont Koplenig. „In einem nächsten Schritt gilt es herauszufinden, ob und inwieweit sich unsere Ergebnisse für maschinelles Lernen überhaupt auf menschlichen Spracherwerb übertragen lassen.“

Originalarbeit:

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

Wissenschaftlicher Ansprechpartner:

Dr. Sascha Wolfer
Leibniz-Institut für Deutsche Sprache
R 5, 6-13
68161 Mannheim
Tel.: +49 621 / 1581 – 439
E-Mail: wolfer@ids-mannheim.de