KI übertrifft Mensch bei der Spracherkennung

Computersystem macht weniger Fehler beim Transkribieren gesprochener Dialoge

Bei direkter Ansprache funktionieren Sprachassistenten schon gut, aber Dialoge und Alltagsgespräche überfordern die meisten Spracherkennungsprogramme noch. © metamorworks/ iStock

Besser als der Mensch: Forscher haben ein Computersystem entwickelt, das spontan gesprochene Sprache besser erkennen kann als ein Mensch. Bei einem aus tausenden Telefongesprächen mitgeschnittenen Test erreicht die künstliche Intelligenz eine Fehlerrate von fünf Prozent, bei menschlichen Probanden sind es 5,5 Prozent. Die Latenzzeit der KI lag nur bei gut einer Sekunde – sie reagierte also nahezu in Echtzeit.

Ob Siri, Alexa oder Cortana: Spracherkennungssysteme ermöglichen es heute, akustisch mit Computersystemen zu kommunizieren. Auch Übersetzungen oder die Transkription gesprochener Texte sind möglich. Dahinter stehen künstliche neuronale Netzwerke – lernfähige Systeme, die darauf trainiert werden, die akustischen Sprachlaute einer Bibliothek von Silben und Wörtern zuordnen. Bei gelesenem Text oder direkter Ansprache erreichen diese Spracherkennungssysteme schon verblüffend gute Leistungen.

Stottern, Pausen und Genuschel

Doch bei alltäglichen Gesprächen oder Telefongesprächen stoßen sie an ihre Grenzen. „Wenn Menschen miteinander sprechen, gibt es Abbrüche, Stotterer, Fülllaute wie ‚äh' oder ‚hm' und auch Lacher oder Huster", erklärt Alex Waibel vom Karlsruher Institut für Technologie (KIT). „Oft werden Worte zudem noch undeutlich ausgesprochen." Schon für Menschen ist es manchmal schwer, von einem solchen informellen Dialog eine akkurate Transkription anzufertigen.

„Einer KI fiel dies bislang noch schwerer", sagt Waibel. Ein alltägliches Gespräch zu verfolgen und genau wiederzugeben, gilt daher als eine der größten Herausforderungen für die künstliche Intelligenz. Einem Forscherteam um Waibel ist es nun erstmals gelungen, ein Computersystem zu entwickeln, das diese Aufgabe besser erledigt als Menschen und schneller als andere Systeme.

Genau und trotzdem schnell

Das neue System baut auf einem automatischen Live-Übersetzer auf, der Universitätsvorlesungen aus dem Deutschen oder Englischen überträgt. Die Spracherkennung beruht auf sogenannten Encoder-Decoder-Netzwerken, die die akustischen Laute verarbeiten und zuordnen. „Die Erkennung spontaner Sprache ist die wichtigste Komponente in diesem System", erläutert Waibel. „Denn Fehler und Verzögerungen bei der Erkennung machen die Übersetzung schnell unverständlich."

Dieses Programm haben die Forscher nun weiterentwickelt und dabei auch die Latenzzeit des Systems verringert. Denn gerade bei Echtzeit-Übersetzungen ist es wichtig, den Nachlauf des Programms so klein wie möglich zu halten, ohne dabei die Präzision der Erkennung zu opfern. Um das zu erreichen, kombinierten Waibel und seine Kollegen einen auf der Wahrscheinlichkeit bestimmter Wortkombinationen basieren Ansatz mit zwei weiteren Erkennungsmodulen.

Weniger Fehler als der Mensch

Um die Leistung des Systems zu ermitteln, unterzogen die Forscher es einem standardisierten Benchmark-Test. Bei diesem hört die Spracherkennung Gesprächsausschnitte, die aus einem Pool von rund 2.000 Stunden an Mitschnitten von Telefongesprächen stammen. Aufgabe war es, diese Dialoge zu transkribieren. „Die menschliche Fehlerrate liegt hier bei um die 5,5 Prozent", berichtet Waibel. „Unser System erreicht nun

5,0 Prozent."

Damit ist dies das erste Computersystem, das den Menschen beim Erkennen solcher spontan gesprochenen Sprache übertrifft – und dies mit nur minimaler Verzögerung zum Sprechen. Denn die Latenzzeit des Spracherkennung lag im Schnitt bei 1,63 Sekunden. Menschen benötigen rund eine Sekunde für die Aufgabe – also kaum weniger.

Aufbauend auf der neuen Technologie könnten Dialog-, Übersetzungs- und weitere KI-Module künftig schneller und mit größerer Genauigkeit sprachliche Interaktion ermöglichen. (Preprint, arXiv:2010.03449)

Quelle: Karlsruher Institut für Technologie

23. Oktober 2020
- Nadja Podbregar