idw - Informationsdienst Wissenschaft
Nachrichten, Termine, Experten

Press release

Universität Zürich
Melanie Nyfeler

01/15/2025
http://idw-online.de/en/news845747

Research results, Transfer of Science or Research
Information technology, Language / literature, Media and communication sciences, Social studies, Teaching / education
transregional, national

Im Kneipenlärm: Automatische Spracherkennung auf menschlichem Niveau

Wer kann Sprache besser erkennen: Mensch oder Maschine? In geräuschvoller Umgebung erreichen moderne Spracherkennungssysteme eine beeindruckende Präzision und übertreffen teilweise sogar Menschen. Doch während Maschinen riesige Mengen an Sprachdaten benötigen, eignen sich Menschen vergleichbare Fähigkeiten in kürzerer Zeit an.

Die automatische Spracherkennung (ASR) hat in den letzten Jahren bemerkenswerte Fortschritte gemacht, insbesondere bei global häufig verwendeten Sprachen wie Englisch. Während man noch vor 2020 davon ausging, dass menschliche Spracherkennungsfähigkeiten denen von automatischen Systemen weit überlegen sind, zeigen einige aktuelle Systeme eine vergleichbare Leistung. Ziel der Weiterentwicklung von ASR war, die Fehlerrate zu minimieren, unabhängig davon, wie gut Menschen in der gleichen Geräuschumgebung abschneiden. Denn auch Menschen erreichen in geräuschvollen Umgebungen keine perfekte Genauigkeit.

In ihrer aktuellen Studie verglich die UZH-Computerlinguistin Eleanor Chodroff gemeinsam mit ihrerz Kollegin Chloe Patman von der Cambridge University die Spracherkennungsleistung zweier beliebter Systeme – «wav2vec 2.0» von Meta und «Whisper» von OpenAI – direkt mit britischen Muttersprachler:innen. Sie testeten die ASR-Systeme unter Bedingungen wie sprachähnlichem Rauschen oder Kneipenlärm, jeweils mit und ohne Gesichtsmaske aus Baumwolle.

Neuestes OpenAI-System besser – mit einer Ausnahme

Die Ergebnisse zeigten, dass Menschen beiden ASR-Systemen überlegen waren. Allerdings übertraf das neueste OpenAI-System «Whisper large-v3» die menschliche Leistung in allen getesteten Bedingungen deutlich, ausser bei realistischem Kneipenlärm, wo es mit der menschlichen Leistung mithalten konnte.

«Whisper large-v3» bewies damit seine Fähigkeit, die akustischen Eigenschaften von Sprache zu verarbeiten und sie erfolgreich einem Satz zuzuordnen. «Dies war beeindruckend, als die getesteten Sätze aus dem Zusammenhang gerissen wurden und es auch schwierig war, ein Wort aus den vorhergehenden Wörtern vorherzusagen», erklärt UZH-Expertin Eleanor Chodroff.

Riesige Mengen an Trainingsdaten

Ein genauerer Blick auf die ASR-Systeme und ihre Trainingsmethoden zeigt, wie bemerkenswert die menschliche Leistung nach wie vor ist. Beide getesteten Systeme basieren auf Deep Learning, aber das leistungsstärkste System «Whisper» benötigt immense Mengen an Trainingsdaten. Während «wav2vec 2.0» von Meta mit 960 Stunden englischer Sprachdaten trainiert wurde, griff das Standardsystem von «Whisper» auf mehr als 75 Jahre Sprachdaten zurück. Das System, das die menschlichen Fähigkeiten tatsächlich übertraf, nutzte sogar mehr als 500 Jahre Sprachdaten. «Menschen erreichen diese Leistung in nur wenigen Jahren», betont Chodroff. «Ausserdem bleibt die automatische Spracherkennung in fast allen anderen Sprachen weiterhin eine grosse Herausforderung.»

Unterschiedliche Fehler

Die Studie zeigte auch, dass Menschen und ASR-Systeme unterschiedliche Fehler machen. Englische Hörer:innen bildeten fast immer grammatikalisch korrekte Sätze, schrieben aber häufig Satzfragmente, anstatt zu versuchen, für jeden Teil des gesprochenen Satzes ein geschriebenes Wort zu liefern. «wav2vec 2.0» hingegen produzierte unter schwierigsten Bedingungen häufig Kauderwelsch. «Whisper» lieferte zwar grammatikalisch korrekte Sätze, neigte aber dazu, Lücken mit völlig falschen Informationen zu füllen.

contact for scientific information:

Kontakt
Prof. Dr. Eleanor Chodroff
Institut für Computerlinguistik
Universität Zürich
+41 76 426 27 07
eleanor.chodroff@uzh.ch

Original publication:

Literatur
Chloe Patman, Eleanor Chodroff. Speech recognition in adverse conditions by humans and machines. JASA Express Lett. 4, 115204 (2024). DOI: https://doi.org/10.1121/10.0032473

URL for press release: https://www.news.uzh.ch/de/articles/media/2025/Spracherkennung.html