

Wer dasselbe anders fragt, fragt etwas anderes

Das zumindest „denken“ Künstliche Intelligenzen wie ChatGPT & Co. und geben völlig unterschiedliche Antworten auf dieselbe Frage, je nachdem, in welchem Dialekt oder Stil sie gestellt wurde. Häufig ist das problematisch.

VON DR. ANNA WEGMANN

„Ich gehe nach Las Vegas!“ Wer das heute sagt, weckt sofort die Vermutung, sein Geld in Casinos verspielen zu wollen. Wenn vor 300 Jahren ein Mann in den Niederlanden sagte „Ich gehe nach Utrecht!“ hätte man ihm genauso unterstellt, dass er dort sexuellen Kontakt mit anderen Männern sucht. Weder Casinos noch Homosexualität werden in diesen Aussagen explizit erwähnt, oberflächlich geht es nur um den Besuch der beiden Städte. Aber geschichtliche Kenntnis um die Verhältnisse in Utrecht vor 300 Jahren und dass der Sprecher ein Mann ist, rückt die zweite Aussage in ein anderes Licht.

Kurz gesagt ist es komplexer Sprache zu verstehen, als es zunächst erscheinen mag. Trotz der beeindruckenden Leistungen von Systemen basierend auf Künstlicher Intelligenz (KI) sind ChatGPT & Co. oft noch nicht in der Lage die Vielfalt menschlicher Formulierungen, Kontexte und Nuancen zu verstehen und angemessen darauf einzugehen. Das ist problematisch, weil KI-Systeme in der Gesellschaft immer häufiger zum Einsatz kommen. Man kann ChatGPT jederzeit um Hilfe bitten. Bei gesundheitlichen Fragen, bei der Suche nach passenden Stellenanzeigen oder bei der Beurteilung von Bewerbern.

UTRECHTENAAR ODER UTRECHTER: KOMPLEXITÄT IN LOKALEN DIALEKTEN

Es ist verständlich, dass KI-Systeme Probleme damit haben, menschliche Formulierungen und Kontexte zu verstehen: In meiner Doktorarbeit habe ich herausgefunden, dass es auch uns Menschen schwerfällt, Sprache in

komplexen Kontexten zu verstehen und zu bestimmen, ob zwei unterschiedliche Aussagen das Gleiche bedeuten. Ein Grund, warum das so schwer ist, ist die Komplexität der Welt.

Ein Beispiel für diese Komplexität findet sich in Utrecht. Der Stadt in den Niederlanden, in der ich meine Doktorarbeit geschrieben habe und in der ich gerne Touristen herumführe. In Utrecht fegte 1674 ein verheerender Sturm durch die Stadt, der zum Einsturz des Domschiffs führte. Dessen Trümmer verblieben daraufhin über 150 Jahre auf dem Domplatz. Bald trafen sich dort Männer heimlich in den Schatten der Steine. In der Folge wurde der Domplatz und schließlich Utrecht selbst mit Männern assoziiert, die Sex mit anderen Männern haben. Dies führte sogar dazu, dass „Utrechtenaar“, der niederländische Begriff für Bewohnende von Utrecht, zu einer groben Beleidigung wurde. Im Laufe der Jahrhunderte setzte sich dann „Utrechter“ als die übliche Bezeichnung durch. „Utrechtenaar“ taucht daneben noch in lokalen Dialekten auf und wird neuerdings von der queeren Community in Utrecht wieder aufgegriffen.

Wenn wir vor diesem Hintergrund fragen, ob „Ich bin ein Utrechter“ und „Ich bin ein Utrechtenaar“ das Gleiche meinen, lässt sich das nicht so einfach mit Ja oder Nein beantworten. Beide Aussagen bezeichnen die sprechende Person vor allem als Bewohner*in von Utrecht. Sie sagen also das Gleiche, aber anders. Doch die Wahl zwischen „Utrechter“ und „Utrechtenaar“ ist nicht zufällig. Sie variiert je nach Kontext, also danach, wer die sprechende Person ist und an welchem Ort und in welcher Zeit wir uns befinden. Wenn jemand heute „Ich bin ein Utrechtenaar“ sagt, ist diese Person wahrscheinlicher queer, wurde in Utrecht geboren oder beides.

KI-Systeme kennen solche Nuancen der menschlichen Sprache kaum. Und selbst wenn sie diese erkennen, reagieren sie häufig problematisch.

DISKRIMINIERUNG AUFGRUND VON SPRACHE: EIN KLEINER UNTERSCHIED MIT GROSSEN FOLGEN

Nach meiner Promotion habe ich ChatGPT gefragt, welche Stellen es mir als „Utrechter“ vorschlägt. Es empfahl mir Forscherin an der Universität Utrecht zu werden. Die gleiche Frage stellte ich nochmal, diesmal bezeichnete ich mich als „Utrechtenaar“. Plötzlich landeten andere Berufe ganz oben: Kellnerin und Krankenpflegerin. Hier diskriminiert ChatGPT. Es

assoziiert meinen Bildungsstand mit dem dialektalen Begriff „Utrechtenaar“. Dieses Beispiel ist kein Einzelfall, sondern eine bekannte Assoziation von KI-Systeme mit Dialekten und abweichenden Schreibstilen. Studien zeigen zum Beispiel, dass KI-Systeme häufiger falsche Antworten auf Anfragen in einfacher Sprache geben und Sprechenden von afroamerikanischem Englisch weniger angesehene Jobs empfehlen.

KI-Systeme können subtile Signale in Sprache erkennen, die ihnen scheinbar etwas über die sprechende Person verraten. Dieses Erkennen ist in der Mehrheit der Fälle aber nicht geplant, sondern ein Nebenprodukt der Trainingsdaten und kann zu falschen Rückschlüssen und unangemessenen Reaktionen führen. Die Beispiele, anhand derer KI-Systeme Sprache und Verhalten lernen, kommen häufig aus Foren aus dem Internet. Diese sind durchzogen von Vorurteilen gegenüber andersprechenden Menschengruppen.

WIE ENTSTEHT DAS PROBLEM IN DER KI?

Um Vorurteile und Ungleichbehandlungen durch KI-Systemen zu reduzieren, müssen wir ihnen grundsätzlich beibringen, sprachliche Vielfalt besser zu verstehen. Nur wenn Systeme verstehen, wie Sprache variiert, können sie verstehen, wann zwei Fragen prinzipiell die gleiche Antwort bekommen sollten. Allerdings gibt es mehrere Gründe, warum KI-Systeme noch nicht sonderlich gut darin sind.

Ein Grund liegt ganz am Anfang in der Berechnung der Antwort eines KI-Systems. Ein KI-System kann Wörter nicht wie wir lesen und mit Bedeutung verknüpfen. Es muss Wörter stattdessen erst in Zahlen umwandeln, die eine Bedeutung für das System haben. Allerdings werden verschiedene Weisen das Gleiche auszudrücken, wie „Utrechter“ und „Utrechtenaar“ von KI-Systemen komplett anders in Zahlen übersetzt. Ich fand in meiner Forschung heraus, dass es für KI-Systeme dadurch grundsätzlich schwierig ist zu lernen, beide Wörter gleich zu behandeln.

Ein weiterer Grund liegt darin, wie KI-Systeme verbessert werden: Systematische Tests zeigen Entwicklern, was ein System noch nicht gut genug kann. Und genau dort wird nachgebessert. Allerdings wird dieses Können aktuell vor allem auf standardisiertem amerikanischen Englisch getestet. Häufig wird getestet, ob ein KI-System die richtige Antwort auf

faktische Fragen gibt oder ein Matherätsel richtig löst. Diese Fähigkeiten sind wichtig, und in den letzten Jahren wurden hier beeindruckende Fortschritte erzielt. Aber KI-Systeme werden üblicherweise nicht darauf getestet, ob sie die gleiche Antwort geben, wenn beispielsweise im Utrechter Dialekt nach der Hauptstadt von Deutschland gefragt wird, oder ob sie erkennen, dass Wörter wie „Utrechenaar“ und „Ureg“ Teil des gleichen Dialekts sind. Und solange nicht getestet wird, ob KI-Systeme verstehen, wie vielfältig Sprache ist, wird es hier auch keinen systematischen Fortschritt geben.

EIN TEST AUF SPRACHLICHE ÄHNLICHKEIT: WIE KÖNNEN WIR KIS VERBESSERN

In meiner Doktorarbeit habe ich einen neuen Ansatz entwickelt, um zu bewerten, wie gut KI-Systeme sprachliche Nuancen verstehen. Die zugrunde liegenden Tests prüfen, ob ein KI-System erkennt, wenn zwei Aussagen sprachlich ähnlich sind. Basierend auf diesen Tests konnte ich selbst ein KI-System entwickeln, das den Schreibstil von Texten vergleichen kann. Mein System wurde mittlerweile fast zwei Millionen Mal heruntergeladen. Es wird heute unter anderem genutzt, um zu erkennen, ob ein Text von einer KI generiert wurde. Das kann zum Beispiel dabei helfen, „KI-Slop“ zu entfernen, also nichtssagende Texte, die KI-Systeme massenhaft produzieren. Mein System hilft außerdem dabei, den Schreibstil von Texten gezielt an Benutzer*innen anzupassen. Zum Beispiel damit ein KI-System automatisch in einfacher Sprache antwortet, wenn es erkennt, dass jemand selbst in einfacher Sprache schreibt.

Die Gleichbehandlung verschiedener Dialekte und Sprachstile konnte ich mit meiner Arbeit leider noch nicht erreichen. Aber mein System kann dabei helfen, Missstände zu erkennen. Zum Beispiel lässt sich untersuchen, ob ein KI-System schlechtere Entscheidungen trifft, wenn Benutzer*innen nicht im standardisierten amerikanischen Englisch schreiben. Heute arbeite ich an der Universität Utrecht weiter daran, KI-Systemen beizubringen, wie vielfältig sich Menschen ausdrücken.