

Erfahrungen meiner Datensammlung

Hallo und herzlich willkommen zu meiner ersten Folge von *BotTalks*, einer Podcast-Reihe von der Professur für Romanische Sprachwissenschaft (Französisch/Italienisch) der TU Dresden. Mein Name ist Claudia Rausch, ich arbeite am hiesigen Institut für Romanistik und in der heutigen Folge soll es um meine persönlichen Erfahrungen mit der Datensammlung zur sprachwissenschaftlichen Erforschung künstlich generierter Texte via ChatGPT gehen. Ist die KI wirklich immer zuverlässig? Warum brachte sie mich an den Rand der Verzweiflung? Und welche Auffälligkeiten können generierte Texte aufweisen, wenn wir die Resultate in verschiedenen Sprachen vergleichen? Dranbleiben lohnt sich, so viel sei gesagt...

Bevor wir jedoch richtig in das heutige Thema eintauchen, kurz vorab nochmal das Basiswissen: ChatGPT (*Chat Generative Pre-trained Transformer*) ist der Prototyp eines Chatbots, der auf künstlicher Intelligenz (kurz *KI*) beruht. ChatGPT wurde am 3. November 2022 veröffentlicht und kann Texte in unterschiedlichen Sprachen generieren, darunter Italienisch, Französisch, Spanisch, Portugiesisch, Englisch und Deutsch.

Trainiert wurde ChatGPT mit einer großen Menge an von Menschen verfassten Texten. Diese reichen jedoch nur bis in das Jahr 2021. Die Datengrundlage der KI umfasst Wikipedia Einträge, digitalisierte Bücher, Zeitungsartikel, Soziale Medien und Online-Foren. Durch das Training der KI mit diesen Daten lernt sie, wie menschliche Sprache funktioniert und wie es auf bestimmte Fragen oder Anfragen angemessen reagieren kann. So weit so gut – es kribbelte in den Fingern: Die Kompetenz des Tools musste ausgetestet werden...

Die grundlegende Idee war, mithilfe von ChatGPT ein Korpus zu erstellen, also eine wissenschaftlich präzise untersuchbare Sammlung sprachlicher Daten, die aus Texten über berühmte Frauen besteht, die hauptsächlich im 20. Jahrhundert gelebt haben, aus Italien, Frankreich oder Deutschland stammen und die einen Wikipedia-Eintrag besitzen. Ein Vorteil unserer Herangehensweise ist, dass durch eine so genaue Eingrenzung die Datengrundlage, aus der ChatGPT seine Informationen bezieht, bei allen Personen, egal in welcher Sprache der Text generiert wird, dieselbe ist, nämlich das Leben der jeweiligen Frauen. Ein derartiges Korpus kann somit als fundierte Grundlage für spätere Vergleiche dienen. Der Prompt, also die Anfrage oder die Aufgabe, die an ChatGPT gestellt wird, lautete in jeder der drei Sprachen gleichermaßen: „Schreibe einen Text von 1000 Wörtern, um zu erklären, wer (Name der berühmten Frau) ist“. Es sollte zudem zunächst zwei Versuche pro Person geben, jede in einem eigenen Chat. Das erreicht man, wenn man im Tool die Schaltfläche „New Chat“ benutzt, denn dadurch öffnet GPT jedes Mal eine neue Konversation und kann seine Antwort nicht an zuvor Geschriebenes anpassen – somit können auch Textdoppelungen vermieden werden.

Über dreieinhalb Wochen (07.01.-31.01.2023) entstanden somit in allen drei Sprachen jeweils zwei generierte Texte zu je 26 Frauen, sprich je 52 pro Sprache, 156 insgesamt. Wichtig war hierbei auch die Anzahl der tatsächlich generierten Wörter. Die Ergebnisse wurden systematisiert, nach den wichtigsten Informationen für einen Vergleich: Den Namen der Personen, dem Link zur Wikipedia-Seite über ebendiese, dem Datum der Textgenerierung sowie der Anzahl der Wörter eines jeden Versuchs.

Das erste Ergebnis ließ nicht lange auf sich warten: Jeder von ChatGPT generierte Text umfasste niemals die über den Prompt gewünschten 1000 Wörter. Die Texte bestanden zu Beginn der Korpuserstellung – wir sprechen hier von Anfang Januar 2023 – im Durchschnitt aus nur ca. 300 Wörtern – etliche Texte lagen sogar weit darunter. Ein Teil der Aufforderung wurde also nicht befolgt. Die Länge änderte sich jedoch im Laufe der Zeit. Wir beschlossen in der ersten Februarwoche, das von uns *BioFem* getaufte Korpus um eines mit Texten über berühmte Männer zu erweitern – praktischerweise *BioMasc* genannt –. Auch in diesem Korpus galten erneut dieselben Grundkriterien: berühmt (egal, in welcher Branche), aus jeweils Italien, Frankreich oder Deutschland stammend und geboren überwiegend im 20. Jahrhundert. Erneut 52 Texte pro Sprache. In der ersten Februarwoche erstellte ChatGPT zunächst Texte, die in etwa dieselbe Dimension hatten, wie die zuvor generierten weiblichen Biografien. Zum 13. Februar 2023 gab es ein Update für die KI, so hieß es auf der Seite von Open.AI: „*We’ve updated performance of the ChatGPT model on our free plan in order to serve more users.*“ - also "Wir haben die Leistung des ChatGPT-Modells auf unserem kostenlosen Programm aktualisiert, um mehr Nutzer:innen bedienen zu können."

Schnell wurde klar, dass plötzlich andere Dimensionen im Sinne der Informationsanforderung erreicht wurden: Die ab dem 15. Februar 2023 zu Männern generierten Texte erreichten auf einmal (fast ausschließlich) Längen von weit über 400 Wörtern pro Text. Dieser Unterschied ist natürlich nicht enorm, doch er bestand.

Dies nahmen wir zum Anlass, Ende Februar/Anfang März 2023 alle Prompts aus dem weiblichen Korpus ein erneutes Mal durch ChatGPT laufen zu lassen. Würde sich auch da etwas ändern? Ja!: Es wurden plötzlich auch im Korpus zu Frauen Texte mit über 400 Wörtern generiert. Die verschiedenen GPT-Versionen gingen also mit für unsere Untersuchungen relevanten Veränderungen der Texte einher. Und so fiel die Entscheidung, beide Korpora, männlich und weiblich, stets mit den neuesten Versionen von ChatGPT zu erweitern, sodass beinahe Woche für Woche neue Biografien generiert wurden. Bis zum 14. Juni 2023 wurden somit für beide Geschlechter und in jeder Sprache jeweils Texte zu 35 Biografien generiert - insgesamt entstanden 900 Texte. Eine gute Grundlage für linguistische Studien.

Ein Unterschied fiel im Vergleich der drei Sprachen sowohl im Korpus BioFem als auch im Korpus BioMasc auf: Im Deutschen entstanden deutlich mehr Texte mit unter 400 Wörtern als im Italienischen oder Französischen. Ich fragte also bei der KI nach, woran dies liege. Mein Prompt lautete: „Warum schreibt ChatGPT längere Texte auf Französisch und kürzere auf Deutsch?“ Die Antwort war:

„Als künstliche Intelligenz ist ChatGPT darauf trainiert, auf Anfragen in verschiedenen Sprachen zu antworten, einschließlich Deutsch und Französisch. Die Länge einer Antwort hängt nicht von der Sprache ab, sondern von der Komplexität der Frage und der erforderlichen Tiefe der Antwort. Es kann sein, dass bestimmte Fragen, die auf Französisch gestellt werden, komplexer oder nuancierter sind und daher längere Antworten erfordern. In anderen Fällen kann es sein, dass die Fragen auf Deutsch einfacher oder direkter sind und daher kürzere Antworten erfordern.“

Zweifelhaft. Man muss nämlich bedenken, dass die Prompts identisch, lediglich übersetzt, sind, von unterschiedlicher Nuancierung – wie die KI es nannte - und Komplexität kann also nicht die Rede sein. Die generierte Antwort liefert demnach keine ausreichende Begründung für diese deutlichen Unterschiede in der Textlänge zwischen den kürzeren deutschen und den ausführlicheren italienischen und französischen Texten. Damit war ich nicht zufrieden.

Aber es gab noch mehr zu klären und so fragte ich erneut direkt bei der KI nach. Sie sollte mir nämlich eine klare Antwort bezüglich einer weiteren Auffälligkeit liefern: Im Zuge der Textgenerierung zur französischen Politikerin Christine Lagarde beispielsweise wurde nämlich einmal ein 605-Wörter Text generiert – aber nicht, so wie der Prompt es voraussetzte, auf Französisch, sondern auf Englisch. (Dies passierte in beiden Korpora, jedoch wirklich selten). Weshalb war dieser englische Text deutlich länger, als die Texte in anderen Sprachen? Meine Frage lautete: „Warum sind Antworten, die – wenn auch fälschlicherweise – auf Englisch generiert werden, länger, als Antworten in anderen Sprachen, wie Deutsch, Italienisch oder Französisch?“ Und die KI antwortete einleuchtend:

„Antworten von ChatGPT auf Englisch können länger sein, da Englisch eine der am häufigsten verwendeten Sprachen im Internet und in der Wissenschaft ist. Es gibt daher eine Fülle von Informationen und Daten in englischer Sprache, auf die ChatGPT zugreifen und aus denen es Informationen extrahieren und verwenden kann, um seine Antworten zu generieren. Darüber hinaus wurde der ChatGPT-Algorithmus ursprünglich in englischer Sprache trainiert, wodurch er eine tiefere und umfassendere Kenntnis der englischen Sprache hat als andere Sprachen.“

Abseits der Textlänge und des Generierens in der falschen Sprache, gab es allerdings auch andere ernüchternde Auffälligkeiten. Fehlende Satzzeichen und Rechtschreibfehler kommen zwar in beiden Korpora, ob männlich oder weiblich, vor, jedoch selten. Aber Margot Honecker war z.B. „eine prominenten politischen Figur“ und Uschi Glas wiederum „ein sehr bekanntes und beliebtes Medienstar“ – und nicht, wie laut Duden korrekt wäre, „ein sehr bekannter und beliebter Medienstar“.

Solche Fehler können ja jedem, besonders Nichtmuttersprachler: innen, mal passieren. Doch dabei blieb es nicht. Die KI erzeugte auch eine Reihe von Halluzinationen, d.h. Inhaltsfehler, insbesondere biographischer Natur. Manchmal wurde das falsche Geburtsjahr oder ein falscher Familienname angegeben, oder aber auch verschiedene Todeszeitpunkte und –umstände etc.

Andrea Berg, die deutsche Schlagersängerin, wurde einmal als Andrea Zellen (was korrekt ist), einmal aber auch als Andrea Ferber geboren.

Paul Klee wurde entweder in Münchenburgsee in der Nähe von Bern oder in München – ohne Burgsee – geboren.

Spannend bezüglich der Textstrukturierung wurde es ab dem 3. Mai 2023, ab Version 9 des Tools, denn da begann ChatGPT, für bestimmte Biografien, unabhängig vom Geschlecht, Aufmerksamkeit erweckende Überschriften zu den Texten zu generieren. So ergaben sich unter anderem folgende Beispiele:

V9, 16.05.2023:

„Reinhold Messner: L'Avventuriero Estremo che ha Dominato le Vette ('Der Extremabenteurer, der die Gipfel eroberte')“

V10, 24.05.2023:

„Hilde Weström: Die Architektin, die die moderne deutsche Architektur geprägt hat“

Aber auch während der Textproduktion von ChatGPT gab es Überraschungen. Es gab z.B. häufig Fehlermeldungen woraufhin man als User die Seite neu laden und den Prompt erneut eingeben musste, oftmals wiederholt. Außerdem stockte die KI immer häufiger mitten im Vorgang der

Textproduktion und kam gar nicht mehr voran, der Cursor blinkte, sodass man als User den Eindruck bekam, die KI suche lediglich nach Informationen und würde so schnell wie möglich weiter generieren, doch dies geschah auch nach Minuten nicht mehr, folglich musste man wieder den Vorgang abbrechen und den Prompt neu stellen. Frustrierend. So etwas konsumiert natürlich unnütz Zeit – genauso wie der Hinweis „Too many requests in 1 hour. Try again later“ – dieser blitzte immer dann auf, wenn die Beantwortung meines Prompts wegen vermeintlich zu vieler Anfragen meinerseits innerhalb einer Stunde abgebrochen wurde. Auf die Frage, wie viele Anfragen denn innerhalb einer Stunde zulässig seien, antwortete die KI ganz selbstbewusst „Es gibt keine Begrenzung für die Anzahl der Anfragen, die Sie an ChatGPT stellen können. Sie können so viele Fragen stellen, wie Sie möchten.“ Davon wurde ich allerdings ab Version 7 erlöst. Zum 14. März 2023 ging das GPT-4 Modell für alle Plus-Nutzer online. Seitdem lief die Anwendung auch für mich als Normal-Nutzer wie geschmiert: Keine Unterbrechungen, kein Warten wegen zu vieler Anfragen und meist sogar ein erstaunlich schnelles Erstellen der Texte – unabhängig von der Sprache, in der generiert wurde. Aber wer weiß, was noch kommt...

Zum Abschluss der Folge lässt sich also sagen: Es scheint im Umgang mit ChatGPT ähnlich wie im Umgang mit echten Individuen und Texten – eine gewisse Portion Skepsis ist immer anzuraten.

Vielen Dank fürs Zuhören sowie das Interesse an unserer Podcastreihe *BotTalks* – einer Podcast-Reihe der Professur für Romanische Sprachwissenschaft (Französisch/Italienisch) der TU Dresden.