DEEPSHORE

WHITE PAPER

WECKRUF CHATGPT: NATURAL LANGUAGE PROCESSING WIRD AUCH FÜR ENTERPRISE USE CASES DEUTLICH LEICHTER

WHITE PAPER

WECKRUF CHATGPT: NATURAL LANGUAGE PROCESSING WIRD AUCH FÜR ENTERPRISE USE CASES DEUTLICH LEICHTER

AUTOREN: FALK BORGMANN, NIKO KRASOWSKI

CHATGPT

Viele Menschen setzen bereits seit Jahren Assistenzsysteme im digitalen Raum ein. Die Nutzung einer Google-Suche ist beispielsweise völlig alltäglich geworden. In den vergangenen zwei Jahren haben darüber hinaus sprachgesteuerte Assistenten wie Alexa oder Siri einen festen Platz im Leben ihrer Anwender eingenommen. Aktuell überrascht das textbasierte Dialogsystem ChatGPT die Öffentlichkeit mit einer neuen inhaltlichen Gesprächsqualität und steht damit für eine neue Generation von anwendbaren KI-Sprachmodellen. Das gleichermaßen beeindruckende, beinahe schon beängstigende Leistungsvermögen von ChatGPT basiert dabei auf einer KI, die durch mehr als doppelt so viele Parameter spezifiziert ist als ein menschliches Gehirn Neuronen hat[1].

Das ist deshalb so interessant, weil einem echten produktiven Einsatz von KI bisher mehrere Monate, teilweise sogar Jahre der Datenkonsolidierung sowie deskriptiver und diagnostischer Analytik vorausgingen[2]. Nicht selten scheiterten KI-Projekte wegen mangelndem fachlichen Know-how und an hohen Entwicklungsaufwänden. Mit der nächsten

---

1 175 Milliarden vs. 86 Milliarden. Ein Vergleich dieser beiden Größen hat auf vielen Ebenen keinen Sinn, da es sich, trotz der Namensgebung „Künstliche Neuronale Netze“ um komplett unterschiedliche Systeme handelt. Eine Einordnung der Größenordnung der Modelle ist so aber durchaus möglich und legitim. Diese Veröffentlichung spezifiziert GPT-3: https://arxiv.org/abs/2005.14165 und dieser ist die Zahl der menschlichen Neuronen im Gehirn zu entnehmen: https://www.frontiersin.org/articles/10.3389/neuro.09.031.2009/full?source=post_page----------

2 Die Datentreiber haben diese Voraussetzung passend in dem Konzept des Analytischen Reifegrades zusammengefasst: https://www.datentreiber.de/en/method/analytics-maturity-canvas/

Stufe der technischen Evolution stellt sich die Frage, inwiefern bisherige Herangehensweisen im Unternehmensalltag überdacht und angepasst werden müssen.

Was ist zu beachten, um KI-Lösungen kosteneffizient einzusetzen? Welche neuen Use Cases werden durch die aktuelle Generation der Technologie erst möglich? Schauen wir uns jene technischen Aspekte etwas genauer an, die für die Beantwortung der aufgeworfenen Fragen relevant sind. Doch zunächst widmen wir uns dem Umfeld, in dem sich diese Entwicklungen abspielen.

WAS IST NATURAL LANGUAGE PROCESSING (NLP)?

Natural Language Processing (NLP) ist eines der zahlreichen Teilgebiete des maschinellen Lernens und von Künstlicher Intelligenz. NLP behandelt die Verarbeitung von gesprochener und geschriebener Sprache. Die meisten von uns nutzen NLP jeden Tag unbewusst, nämlich immer dann, wenn sie von ihren digitalen Assistenten Alexa oder Siri eine sinnvolle Antwort auf eine Frage erwarten. Nicht bei der Erkennung des gesprochenen Wortes, wohl aber im Bereich des semantischen Verständnisses hat ChatGPT (Generative Pre-trained Transformer) der breiten Öffentlichkeit die Leistungsfähigkeit eines modernen NLP-Systems vor Augen geführt.

WAS MACHT NLP MITTLERWEILE ANDERS?

Neben Neuerungen an der Architektur der KI-Sprachmodelle war auch das Paradigma des selbstüberwachten Lernens (Self-Supervised Learning) eine Voraussetzung der aktuellen Entwicklung. Self-Supervised Learning im Kontext von NLP beschreibt die Idee, ein KI-Sprachmodell zu trainieren, indem man es ein Wort vorhersagen lässt, das zuvor aus einem Text ausgeblendet wurde. Diese simple Methode macht prinzipiell den kompletten im Internet zugänglichen Text für Trainingszwecke nutzbar, da diese Trainingsdaten frei verfügbar sind.

Unabhängig vom konkreten Aufbau der KI-Sprachmodelle benötigt ein solches Modell einen gewissen Spielraum, um Zusammenhänge zwischen Worten, Wortbedeutungen, Grammatik etc. zu lernen. Die Möglichkeit, auf große Textmengen zum Training zurückgreifen zu können, erzeugt also nur dann einen Mehrwert, wenn dieser Spielraum gegeben ist. Wie viel Spielraum Modelle haben, lässt sich an

deren Parameteranzahl bemessen[3] – gemeint sind unabhängige Stellschrauben, die das Verhalten des Modells bestimmen. Einfach gesagt, lernt die Software also auf Basis der im Internet verfügbaren Daten. Sie lernt, dass „Couch“ und „Sofa“ oft synonym verwendet werden und dass die Bedeutung einer „Bank“ stark vom Kontext abhängt. Dieses semantische Verständnis spiegelt sich technisch in einem sogenannten Encoding wider. Die inhaltliche Bedeutung wird hierbei in einer Zahlenfolge kodiert. Das Training der Sprachmodelle beinhaltet das Erlernen der Regeln, nach denen dieses Encoding funktioniert. Kurzum: Die Software lernt inhaltliche Bedeutung (Semantik).

Um zu verstehen, welchen Unterschied ein semantisches Verständnis in der Praxis machen kann, werfen wir einen Blick auf das Anwendungsbeispiel der automatischen Verschlagwortung von Dokumenten. Eventuell muss ein als Bild dargestellter Text erst durch OCR (Buchstabenerkennung aus Rastergrafiken) verarbeitbar gemacht werden. Klassischerweise hat man die Möglichkeit, mittels einer Klartextsuche bestimmte Schlagworte anzufügen. Im Dokument ist beispielsweise die Rede vom Ernten eines Apfelbaumes. Der Begriff „baum“ kann auch in klassischen Systemen bereits automatisch als relevantes Schlagwort für eine Klartextsuche identifiziert werden. Dass aber die oben genannte Passage mit dem Apfelbaum auch durch andere Suchbegriffe wie Obst, Frucht oder Streuobstwiese auffindbar wird, ist erst durch das semantische Verständnis eines modernen Sprachmodells möglich.

Das Potenzial der Abstraktion

Der strategisch relevante Aspekt von modernem NLP ist, dass die durch das Vorhersagen ausgeblendeter Worte trainierten Modelle die Fähigkeit zur Verallgemeinerung besitzen. Während man klassischerweise anwendungsspezifische Maschine-Learning-Lösungen betreibt, verallgemeinern die neuen Sprachmodelle sowohl über unterschiedliche Arten von Texten als auch über verschiedene fachliche Aufgabenstellungen[4]. Um die Ergänzung eines ausgeblendeten Wortes in einem Text zu bewerkstelligen, muss ein Verständnis von Wortbeziehungen,

---

3 Viele Parameter allein machen natürlich noch kein nützliches Modell. Erst eine sinnvolle Kombination dieser Parameter zu einer großen Funktion, also die Architektur der Modelle, ermöglicht sinnvolles Verhalten.

4 Hier https://medium.com/@shahrukhx01/multi-task-learning-with-transformers-part-1-multi-prediction-heads-b7001cf014bf wird beschrieben, wie der selbe Modellkern für unterschiedliche Aufgaben eingesetzt werden kann.

Grammatik und semantischer Bedeutung entwickelt werden. Das kann dann auch die Basis zur Lösung von Aufgaben (wie die Erstellung einer Zusammenfassung) sowie die Beantwortung von Fragen zum Text oder ähnlichem sein.

Zero-Effort-KI? – ChatGPT ist nicht allein

Die ausgeprägte Fähigkeit zur Verallgemeinerung ermöglicht Anwendungsfälle, die durch den nativen Einsatz eines vortrainierten Modells „von der Stange" abgebildet werden können. KI lässt sich mittlerweile wie ein gewöhnlicher Softwarebaustein einsetzen, was vor einigen Jahren in dieser Form noch nicht möglich war. Die entsprechenden Bibliotheken sind heute so ausgereift[5], dass initial[6] kein Machine-Learning-Expertenwissen vorausgesetzt werden muss, um die Modelle zu verwenden.

Sind damit die Arbeitsplätze von Machine Learning Engineers in Zukunft auf wenige große Tech-Giganten beschränkt? Bei weitem nicht. Die wenigsten tatsächlichen Use Cases werden eins zu eins von vortrainierten Modellen abgedeckt.

Da viele Sprachmodelle frei verfügbar sind[7], ist die Aufgabe einer Implementierung sinnhafterweise zweigeteilt:

1. Recherche des Modells, das durch Architektur und verwendeten Trainingsdatensatz am ehesten für die Lösung des entsprechenden Use Cases in Frage kommt.

2. Eine Nachjustierung (Finetuning) der Modell-Parameter. Hier kommt das klassische Handwerkszeug der Machine Learning Engineers und Data Scientists zum Einsatz.

Im Gegensatz zu vor drei Jahren muss man aber nicht mehr bei Null starten. Beherrscht ein Modell beispielsweise schon deutsche Vokabeln

---

5 Eine eindrucksvolle Demonstration, wie leicht zugänglich Sprachmodelle sind, findet sich in der Dokumentation der transformers library von huggingface: https://huggingface.co/docs/transformers/v4.26.0/en/main_classes/pipelines#_blank

6 Selbst in Fällen, in denen die Qualität der Ergebnisse out of the box den Anforderungen entspricht, hat es Sinn, die Produktivnahme nach MLOps-Prinzipien zu gestalten. So ist beispielsweise ein kontinuierliches Monitoring der Modell-Qualität und -Stabilität wichtig, um initiale Modellqualität auch zukünftig garantieren zu können. Einen ersten Eindruck vom Thema gibt https://towardsdatascience.com/a-gentle-introduction-to-mlops-7d64a3e890ff

7 Huggingface hat sich als entscheidender Hub zum Austausch von vortrainierten Modellen entwickelt https://huggingface.co/models

und Grammatik, dann muss es „nur“ noch in die Feinheiten juristischer Fachsprache eingeführt werden. Eine weitaus kleinere und damit auch kostengünstigere Aufgabe.

WIE VIEL IST GENUG? – LLMS UND COMPUTING POWER

Weitere Aspekte, die durch die neuen Entwicklungen vermehrt auch strategische Relevanz besitzen, sind Rechenpower und damit verbundene IT-Kosten oder schlussendlich auch die Auswirkungen auf die Flexibilität der Unternehmens-IT.

Rechenleistung und Datenspeicher werden zwar immer preiswerter, jedoch benötigen neuartige Modellarchitekturen überproportional viele Ressourcen, um von Grund auf trainiert zu werden. Die Annahme, dass man KI-Modellen durch eine größere Datenbasis immer menschenähnlicheres Verhalten entlocken kann[8], ist hier die treibende Kraft. In den letzten Jahren hat sich dafür eine eigene Bezeichnung Large Language Models (große Sprachmodelle, kurz LLMs) herausgebildet[9]. Der Trend zu größeren Modellen scheint auch nicht abzureißen und wird mit der öffentlichen Aufmerksamkeit für ChatGPT sogar eher befeuert. Rein empirisch wachsen die größten Modelle jährlich um den Faktor zehn[10].

Dieser LLMs-Trend fügt dem klassischen, oft selbstverständlichen Ziel einer Verbesserung der Modellqualität zwei weitere Dimensionen hinzu: Ressourcensparsamkeit und Flexibilität. Beispielsweise benötigt GPT-3.5, das Sprachmodell hinter ChatGPT, allein 800 GB Speicherplatz, um seine bis zu 175 Milliarden Parameter zu speichern[11]. Würde

---

8 Interessanterweise kann man zeigen, dass gewisse Fähigkeiten der Sprachmodelle (wie Sprachverständnis, die Fähigkeit zum logischen Schlussfolgern, die Fähigkeit zur Arithmetik, …) erst bei einer gewissen Parametergröße auftreten. Diese Veröffentlichung zeigt die Emergenz verschiedener Fähigkeiten: https://arxiv.org/pdf/2206.07682.pdf. Interessant ist, dass sich die unterschiedlichen Emergenzphänomene modellübergreifend ab gewissen Größen einstellen. Dass ein sinnvoller Gradientenabstieg hier überhaupt denkbar ist, ist nicht selbstverständlich. Jahre des Try and Errors und viel Ingenieurarbeit haben hierfür den Weg geebnet. Hier findet sich eine ausführliche Liste der Fähigkeiten, auf die Modelle getestet werden: https://github.com/google/BIG-bench/blob/main/bigbench/benchmark_tasks/keywords_to_tasks.md#summary-table

9 Google Trends erlaubt es, die Explosion der Modellgröße durch die Verwendung des Suchbegriffes „Large Language Model“ nachzuvollziehen: https://trends.google.com/trends/explore?date=2017-02-02%202023-02-02&q=large%20language%20model

10 Diese Entwicklung und eine kritische Perspektive auf das zu erwartende Wachstum der Modellgrößen finden sich hier: https://huggingface.co/blog/large-language-models

11 „Bis zu“, da es verschiedene Varianten gibt: https://iq.opengenus.org/gpt-3-5-model/

man dieses Modell auf der AWS-Cloud hosten, müsste man mit Kosten von etwa 90.000 USD pro Jahr rechnen[12]. Und bei dieser Kalkulation sind noch nicht einmal Abfragen an das System eingerechnet. Große Modelle mit vielen Daten zu trainieren oder zu betreiben, kostet also auch viel Geld. Das klingt zwar logisch – und selbst diese Zahlen mögen im Kontext der Budgets großer Unternehmen lächerlich klingen –, jedoch wird die Infrastruktur solcher Modelle nicht nur zu einem echten finanziellen Faktor, sie wird aufgrund der reinen Datenmenge auch entsprechend träge[13]. Für eine Vielzahl von Anwendungsbeispielen aus der täglichen Unternehmenswelt ist es aber völlig unnötig, derartig große Modelle einzusetzen.

Noch bevor man sich ein vortrainiertes Modell aussucht oder gar selbst eines erstellen möchte, sollte man sich deshalb Gedanken über die Metriken und die Modellqualität im Kontext des anvisierten Use Case machen. Ziel ist es, Klarheit darüber zu gewinnen, welche Modellqualität für eine bestimmte Anwendung ausreichend ist. Es sollte der Grundsatz gelten, nur so komplex, groß oder kostenintensiv wie nötig zu werden, also dem ökonomischen Minimalprinzip zu folgen. Dies gilt sowohl für die initiale Architektur als auch für das Ausmaß des Finetunings. Ein 800-GB-Modell einzusetzen, um eine Klassifikation von Dokumenten durchzuführen, mag inhaltlich funktionieren, ist aber aus unternehmerischer Sicht nicht sonderlich zielführend.

FAZIT UND AUSBLICK

Zusammenfassend sind KI-Use-Cases, besonders im Umgang mit Texten im NLP-Umfeld, in ihrer Umsetzung deutlich einfacher geworden. Was früher eine langjährige Datenstrategie voraussetzte, ist heute in relativ kurzer Zeit implementierbar.

Das bedeutet aber nicht, dass langjährige Datenstrategien an Relevanz verlieren. Im Gegenteil. Die Menge an gespeicherten Daten steigt weltweit[14]. Dieser Trend ist auch in Unternehmen zu beobachten. Eine Datenstrategie erlaubt es, die Use Cases zu identifizieren, die einen

---

12 Neben der Berechnung der minimalen jährlichen Betriebskosten finden sich hier https://bdtechtalks.com/2020/09/21/gpt-3-economy-business-model/ auch Betrachtungen zu den Kosten des initialen Modell-Trainings.

13 Diese Trägheit und Inflexibilität entsteht schon allein durch den Effekt der Datengravitation https://deepshore.de/knowledge/2020-01-06

14 https://www.statista.com/statistics/871513/worldwide-data-created/

Mehrwert aus den entsprechenden Daten generieren. Die Möglichkeit, vortrainierte Sprachmodelle ohne Weiteres einzusetzen, bietet hierzu eine Ergänzung und kann neue, unstrukturierte Datenquellen erschließen.

Das Tech-Rad dreht sich im Bereich der KI und vor allem im Bereich des NLP derzeit rasant. Im Vergleich zum Jahr 2020 können Projekt- und Implementierungszeiten durch bausteinartige und frei verfügbare Modelle drastisch reduziert werden – und das bei qualitativ hochwertigen Ergebnissen. Selbst für die an Innovationen gewöhnte IT-Branche ist das Tempo derzeit atemberaubend. Dennoch tun sich Unternehmen schwer damit, die Potenziale zu erkennen, geschweige denn zu nutzen. Dies kann man auf den Mangel an Fachpersonal in diesem Bereich zurückführen. Gleichwohl sollte die Veröffentlichung von ChatGPT den größten Zweiflern vor Augen führen, dass der KI-Zug schon lange und mit hoher Geschwindigkeit rollt. Das Risiko, dass Unternehmen mit einer abwartenden Haltung in mindestens ebenso hoher Geschwindigkeit den Anschluss verlieren werden, ist – vorsichtig formuliert – mindestens proportional. Anders gesagt: Es besteht dringender Handlungsbedarf.

AUTOREN: FALK BORGMANN, NIKO KRASOWSKI

KONTAKT

Deepshore GmbH · Van-der-Smissen-Straße 9, 22767 Hamburg
+49 40 46664-296 · info@deepshore.de · www.deepshore.de