Grosse Sprachmodelle

Siegfried Handschuh

Der Artikel gibt einen umfassenden Überblick über den aktuellen Stand der Forschung zur generativen KI und insbesondere grossen Sprachmodellen (Large Language Models, LLMs). Es werden die Architektur, das Training und die emergenten Fähigkeiten von LLMs wie GPT-3 erläutert. Grosse Sprachmodelle basieren auf neuronalen Netzen und werden auf riesigen Textdatenmengen trainiert. Dabei lernen sie, basierend auf dem bisherigen Textverlauf das jeweils nächste Wort vorherzusagen. Obwohl dies eine einfache Aufgabe ist, ermöglicht dies komplexe sprachliche Fähigkeiten. Mit zunehmender Modellgrösse zeigen LLMs dabei unerwartete emergente Fähigkeiten wie Textzusammenfassung, mathematische Operationen oder räumliches Denken.
Allerdings haben LLMs auch Schwächen wie die Tendenz zum Fabulieren bei Wissenslücken und mangelnde Kohärenz. Aktuell gibt es rasante Fortschritte durch neue Modelle wie GPT-3 und ChatGPT. Zukünftige Entwicklungen müssen ethische Aspekte berücksichtigen. Insgesamt eröffnen grosse Sprachmodelle faszinierende Möglichkeiten, aber weitere Forschung ist nötig. Der Artikel liefert eine umfassende Übersicht zu Chancen und Herausforderungen dieses rasanten Technologiefeldes.

L'article donne un aperçu complet de l'état actuel de la recherche sur l'IA générative, en particulier sur les grands modèles de langage (Large Language Models, LLMs). Il explique l'architecture, l'apprentissage et les capacités émergentes des LLM comme GPT-3. Les grands modèles linguistiques sont basés sur des réseaux neuronaux et sont entraînés sur d'énormes quantités de données textuelles. Ils apprennent ainsi à prédire le mot suivant en se basant sur le déroulement du texte en amont. Il s'agit d'une tâche simple, mais elle permet d'acquérir des compétences linguistiques complexes. Avec l'augmentation de la taille du modèle, les LLM montrent des capacités émergentes inattendues telles que les résumés de textes, les opérations mathématiques ou le raisonnement spatial.
Toutefois, les LLM présentent aussi des faiblesses, comme la tendance à fabuler en cas de lacunes dans les connaissances et le manque de cohérence. Actuellement, les progrès sont rapides grâce à de nouveaux modèles comme GPT-3 et ChatGPT.

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 8 (2024) — 10.18755/iw.2024.3

2 Architektur grosser Sprachmodelle

In den vergangenen zwei Dekaden hat sich die linguistische Forschung zunehmend auf statistische Methoden der Semantikanalyse fokussiert[11]. Im Gegensatz zu früheren Ansätzen, welche Sprache als regelbasiertes System modellierten, hat sich gezeigt, dass sich linguistische Phänomene effektiver mit probabilistischen Modellen abbilden lassen, bei denen Wörter (streng genommen sind es Subwörter bzw. Subtokens, aber wir sprechen hier der Einfachheit halber von Wörtern) als Einheiten betrachtet werden, die sich mit einer bestimmten Wahrscheinlichkeit verbinden. Über die Analyse von Co-Occurrence-Mustern, also dem gemeinsamen Auftreten von Wörtern in sprachlichen Kontexten, können semantische Strukturen modelliert werden.

Diese Co-Occurrence-Modelle nennen wir Distributional-Semantics[12], und die wissenschaftliche Community hat begonnen, im grossen Stil mit solchen Vektorräumen zu arbeiten. Wenn ein Wortschatz beispielsweise 25.000 Wörter umfasst, hat dieser Vektorraum 25.000 Dimensionen. Wir Menschen sind auf drei Dimensionen beschränkt, weshalb wir uns 25.000 Dimensionen kaum vorstellen können. Oft werden auch komprimierte Vektorräume mit reduzierten Dimensionen erstellt; eine solche Form sind Word Embeddings[13], die von 300 (Word2Vec), über 1024 (BERT-Large) bis 12,288 (GPT-3) Dimensionen reichen können.

11 Baroni/Lenci, 2010
12 Baroni/Lenci, 2010
13 Mikolov, Tomas; Sutskever, Ilya; Chen, Kai; Corrado, Greg S.; Dean, Jeffrey: Distributed Representations of Words and Phrases and Their Compositionality. In: Advances in Neural Information Processing Systems, Band 26, 2013, S. 3111-3119. und auch Pennington, Jeffrey; Socher, Richard; Manning, Christopher D.: GloVe: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, S. 1532–1543.

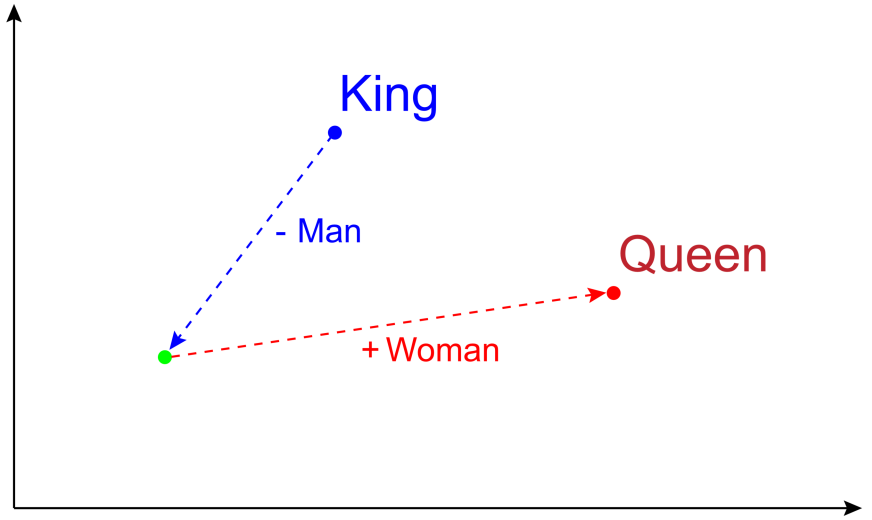

Abb. 1. Ähnlichkeiten zwischen Wörtern im Vektorraum

Es zeigt sich, dass dieser Vektorraum Ähnlichkeiten zwischen Wörtern gut darstellen kann, zum Beispiel dass "König" und "Königin" ähnliche Wörter (vgl. Abbildung 1) sind. Ein solcher Vektorraum kann auch Assoziationen gut abbilden, zum Beispiel dass "König" mit "Schloss" und "Krönung" verbunden ist und dass "Charles" ein König ist (siehe Abbildung 2). Mit diesem Modell können bestehende sprachliche Phänomene hervorragend dargestellt werden. Dies ermöglicht eine Verallgemeinerung unseres Wissens über Sprache. Dabei werden verschiedene Arten von sprachlichen Phänomenen generalisiert, einschliesslich des Weltwissens, des grammatikalischen Wissens (siehe Abbildung 3) und des Wissens über Metaphern. Es ist jedoch zu beachten, dass eine solche Generalisierung nur dann stattfindet, wenn diese Metaphern in den Daten in ausreichender Menge vorhanden sind.

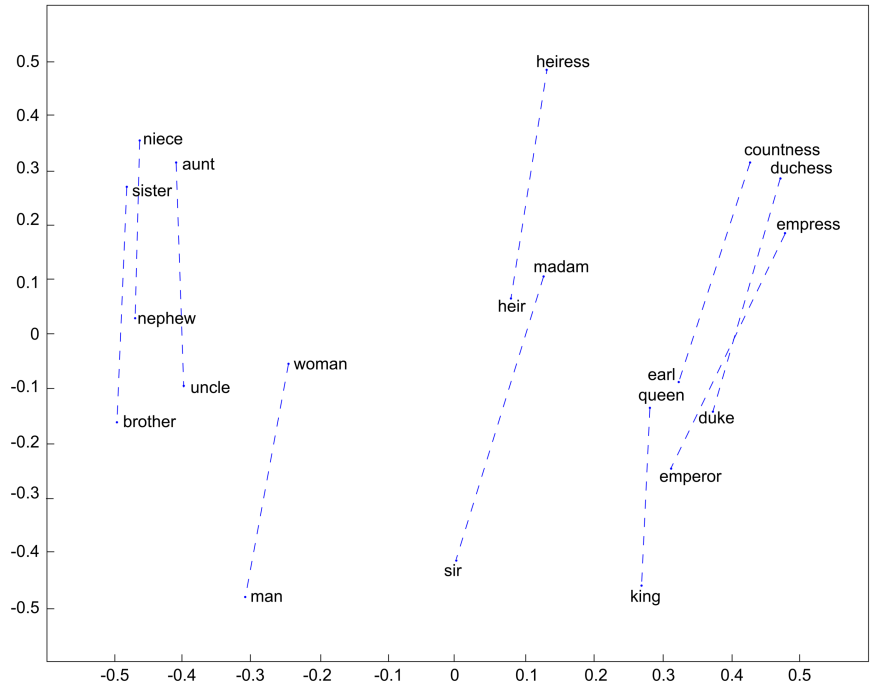

Abb. 2. Wortassoziationen und Wortähnlichkeiten im Vektorraummodell

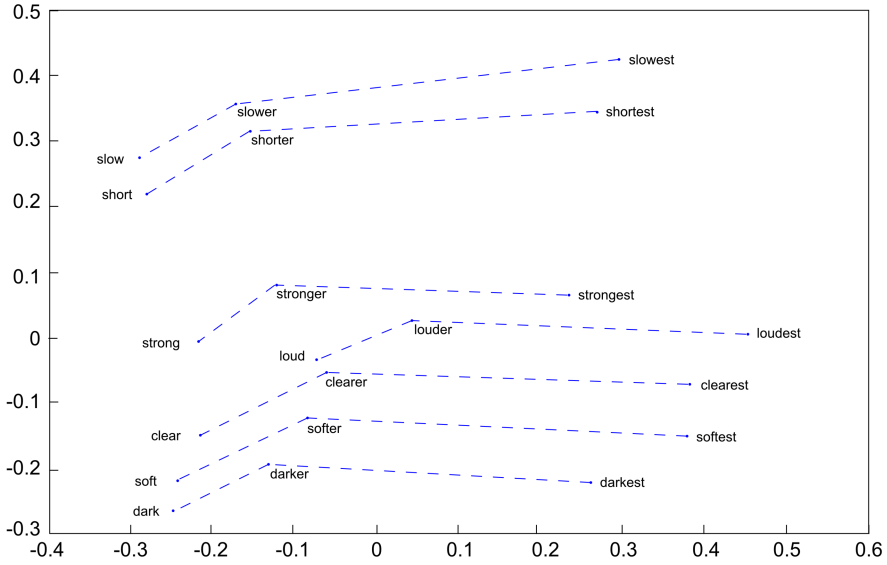

Abb. 3. Abbildung grammatikalischer Strukturen, hier Superlative, im Vektorraummodell

Aktuelle Forschung im Bereich der Computerlinguistik setzt vermehrt vektorbasierte Repräsentationen von Subwörtern ein, bei denen jedes Wort durch einen hochdimensionalen Vektor repräsentiert wird, der seine Bedeutung codiert[14]. Die Dimensionalität dieser Vektorräume ist dabei sehr hoch, da sie der Anzahl der verschiedenen Wörter entspricht. Jedoch hat sich gezeigt, dass sich die Komplexität durch Reduktion auf einige Hundert Dimensionen reduzieren lässt, ohne signifikanten Informationsverlust bezüglich linguistischer Phänomene. Neuronale Sprachmodelle wie ELMo[15], BERT[16]oder GPT-3[17] nutzen derartige vektorbasierte Wortrepräsentationen als Grundlage.

14 Mikolov et al., 2013.

15 Peters, Matthew E.; Neumann, Mark; Iyyer, Mohit; Gardner, Matt; Clark, Christopher; Lee, Kenton; Zettlemoyer, Luke: Deep Contextualized Word Representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, S. 2227-2237.

16 Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019