idw - Informationsdienst Wissenschaft
Nachrichten, Termine, Experten

Press release

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, DFKI
Andreas Schepers

04/11/2024
http://idw-online.de/en/news831784

Research projects, Transfer of Science or Research
Information technology
transregional, national

Occiglot – neue Open Source-Sprachmodelle für Europa veröffentlicht

Am Deutschen Forschungszentrum Künstliche Intelligenz (DFKI) sowie im Hessischen Zentrum für Künstliche Intelligenz (hessian.AI) wurde von Forschenden die Initiative Occiglot gegründet, die generative Open Source-Sprachmodelle für die europäischen Sprachen entwickelt. OcciGlot-LLM Release v.01 bereits verfügbar.

Seit dem ChatGPT-Moment werden generative Sprachmodelle u.a. eingesetzt, um das in ihnen enthaltene „Weltwissen“ in verständlicher Form zugänglich zu machen. Die Präzision der Sprachmodelle hängt vor allem von den Daten ab, auf die zugegriffen werden kann, sowie von der Rechenleistung, die investiert wird. Aufgrund der Dominanz der englischen Sprache im Internet funktionieren Sprachmodelle besser für Anfragen auf Englisch. Kapitalstarke Unternehmen sind zudem eher in der Lage, die notwendige Rechenleistung bereitzustellen.

Weniger verbreitete Sprachen und nichtkommerzielle Projekte benötigen stattdessen innovative Ansätze, um Nachteile auszugleichen und nicht profitabel erscheinende Gebiete zu erschließen. Hier setzt das Projekt Occiglot an, indem es eine Interessengemeinschaft von Forschenden, Sprachexperten, Software-Entwickelnden und Nutzenden bildet. Durch die Bündelung gemeinsamer Interessen sollen alle 24 Amtssprachen der Europäischen Union sowie weitere inoffizielle und regionale Sprachen im Sprachmodell berücksichtigt werden. Die erste Version von Occiglot wurde möglich durch die Nutzung von Rechnern des DFKI und des KI-Servicezentrums hessian.AISC, das vom Bundesministerium für Bildung und Forschung (BMBF) gefördert wird.

Der Parlamentarische Staatssekretär im BMBF, Mario Brandenburg, betont: „Exemplarisch zeigt sich hier der hohe Wert, den Wissenschaftsfreiheit in unserer Gesellschaft hat. Durch den freien Austausch unter Wissenschaftlerinnen und Wissenschaftlern aus den Disziplinen Künstliche Intelligenz (KI) und Sprachtechnologie ist eine Idee entstanden, die der europäischen Sprachsouveränität direkt dient. Ich wünsche dem Projekt Occiglot eine weite Verbreitung und die Mitwirkung von Engagierten mit vielfältigen Sprachhintergründen. Open Source ist der passende Rahmen für ein Projekt dieser Zielrichtung und dieser Entstehungsgeschichte.“

„Die Entwicklung europäischer Sprachmodelle ist der Schlüssel zum Erhalt der akademischen und wirtschaftlichen Wettbewerbsfähigkeit und der digitalen sowie KI-Souveränität Europas. Sie ist ebenso notwendig, um das angestrebte Ziel digitaler Sprachgerechtigkeit in Europa zu erreichen“ ergänzt Prof. Dr. Georg Rehm, Principal Researcher und Research Fellow am DFKI in Berlin.

Europäisches Forschungskollektiv und Aufruf zur Zusammenarbeit

Occiglot versteht sich als offenes europäisches Kollektiv Forschender von Organisationen und Initiativen wie dem Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI), Hessian.AI, der TU Darmstadt, der Katholischen Universität Leuven (Belgien), dem Barcelona Supercomputing Center BSC (Spanien) und einer Reihe weiterer Teams.

Die Occiglot-Initiative ist aktiv auf der Suche nach Kooperationen innerhalb der internationalen KI- und NLP-Community und nach Feedback von Anwendern und Anwenderinnen.

Unterstützt durch DFKI, Hessian.AI und BMBF

Ermöglicht wurde die Konzeption von Occiglot zu großen Teilen durch Forschende der DFKI-Labore in Darmstadt und Berlin. Das hessian.AI Innovation Lab (gefördert durch das Hessische Ministerium für Digitale Strategie und Innovation) und das hessian.AISC Service Center (gefördert durch das Bundesministerium für Bildung und Forschung, BMBF) unterstützen Occiglot durch die zur Verfügungstellung von Rechenzeit auf ihrem KI-Supercomputers fortytwo.

Kristian Kersting, Leiter des Forschungsbereichs Grundlagen der Systemischen KI am DFKI in Darmstadt und Co-Direktor von Hessian.AI betont den Erfolg des Zusammenwirkens im Netzwerk: “Zukünftige Sprachmodelle—ob größer als ChatGPT oder so klein, dass sie aufs Handy passen, ob offen oder proprietär—werden in ihrer Leistungsfähigkeit noch einige Überraschungen bereithalten. Damit wir das enorme Potential auch für Deutschland und Europa ausschöpfen können, brauchen wir mehr solcher Synergien. Es braucht ein starkes KI-Ökosystem mit entsprechender Rechen-Infrastruktur und Modellen, die auch für Unternehmen zugänglich und wirtschaftlich anwendbar sind.“

Die Kuration der Trainingsdaten wird zudem teilweise vom Bundesministerium für Wirtschaft und Klimaschutz (BMWK) über das Projekt OpenGPT-X (Projektnummer 68GX21007D) gefördert.

Occiglot-LLM Release v0.1

Zunächst wurden zehn Sprachmodelle jeweils in einer Größe von sieben Milliarden Parametern veröffentlicht. Diese Modelle sind die erste Version von einer Reihe von Sprachmodellen und konzentriert sich zunächst auf die fünf größten europäischen Sprachen: Englisch, Deutsch, Französisch, Spanisch und Italienisch.

Von Mistral-7B ausgehend – einem bereits für Englisch trainierten Open Source Modell – wurde zweisprachiges kontinuierliches Pretraining und anschließendes Instruction-Tuning für jede Sprache durchgeführt. Zusätzlich wurde ein mehrsprachiges Modell trainiert, das alle fünf Sprachen abdeckt.

Insgesamt wurden 700 Milliarden zusätzliche mehrsprachige Tokens während des kontinuierlichen Pretrainings und etwa eine Milliarden Tokens für das Instruction-Tuning verwendet.

Alle Sprachmodelle (mit und ohne Instruction-Tuning) sind unter Apache 2.0-Lizenz auf der Plattform Hugging Face verfügbar.

Roadmap

Das Hauptaugenmerk der Occiglot-Initative wird in den kommenden Monaten auf der Schaffung eines kohärenten Ansatzes zur Modellierung eines Sprachmodells liegen, das alle 24 offiziellen Sprachen innerhalb der Europäischen Union sowie mehrere inoffizielle und regionale Sprachen unterstützt.

Um dieses Ziel zu erreichen, wurden bereits ca. 1 Billion Token nicht-englischer Pre-Training-Daten gesammelt. Dieser Korpus wird kontinuierlich durch zusätzliche Daten, die von Occiglot-Community-Mitgliedern gesammelt werden, sowie durch weiteres Crawling im Internet erweitert.

Weitere Informationen

Weitere Informationen finden sich auf den Github-Seiten der Occiglot-Initiative unter
https://occiglot.github.io/occiglot/