Ein Datensatz für Winograd Schema mit Polizeibezug

Johannes Fähndrich und Fabian Glock

Die maschinelle Verarbeitung von Text wird immer komplexer. So werden heute nicht mehr nur syntaktische Elemente analysiert, sondern semantische. Dies wird sich in der nächsten Zeit auf pragmatische Elemente fortsetzen [1]. Beim Übergang von syntaktischer zu semantischer Analyse, werden nicht mehr nur Äquivalente gesucht (was syntaktisch in den Texten gleich ist), sondern den Elementen auch Bedeutung zugeordnet. Semantik beschreibt dabei die kontextfreie Bedeutung. Diese Form der Bedeutung ist nicht genug, um den kommunikativen Sinn [7] maschinell erfassen zu können. Zusätzlich braucht es Informationen aus dem Kontext, in den der Text eingebettet, um dessen tiefere Bedeutung zu erschließen. Ein typisches Beispiel hierfür sind ambige Pronomen. Zum Beispiel kann der Satz: „Er hat sie geschlagen und dann an den Tisch gefesselt“, nur schwer auflösbar sein ohne weiteren Kontext. Wer *er* und *sie* ist, bleibt hier ungenannt. Eine Weiterentwicklung des Kontextbezogenen ist das Schlussfolgern aus dem Zusatzwissen. Das Schlussfolgern mit „normalen Menschenverstand“1 ist jedoch keine neue Idee [8]. Zu dieser Form des Schließens gehört die Verwendung von Hintergrundwissen, welche nicht im Kontext gegeben ist. Also beispielweise Wissen über die Welt, Muster und Regel, die meist kontextunabhängig verwendet werden können. Dazu gehören unter anderem zeitliche Zusammenhänge: Wenn eine Zugfahrt von Berlin nach Hamburg um 12:00 Uhr endet, war der Zug wahrscheinlich vor 12:00 in Berlin2.

Werden nun alle diese Arten von Textanalysen zusammengenommen, so bekommen wir eine bestimmte Art von Sprachstruktur. Diese Sprachstruktur wird Winograd Schema genannt [3]. Sollte ein System in der Lage sein, solche Winograd Schemas zu lösen, so wird dies als nächsten Schritt im Bereich Textverständnis einer Künstlichen Intelligenz verstanden.

Das die Menge des zu verarbeitenden Textes auch in der polizeilichen Arbeit zu nimmt [2] und von Menschen schon länger nicht mehr ohne maschinelle Unterstützung handhabbar ist, motiviert eine Anwendung von automatischer Auswertung. Die Herausforderung

[1] Im Englischen „reasoning with common sense”

[2] Hier ist der Kontext auch ein örtlicher: der Leser muss wissen wo Berlin und Hamburg liegen, das dies Städte sind, was eine Zugfahrt ist und wie Zeitgemessen wird. Aber auch logische zusammenhänge wie: Eine Zugfahrt endet erst frühestens nachdem sie angefangen hat.

ist dabei, dass die meisten Datensätze für maschinelle Lernverfahren und damit die meisten Ansätze auf und für englische Texte optimiert sind und somit bei anderen Sprachen keine ausreichende Qualität liefern.

Ziel dieser Veröffentlichung ist nun einen solchen Datensatz mit polizeilichem Kontext zu veröffentlichen und Wissenschaftler:Innen die Möglichkeit zu geben, maschinellen Ansätze zu schaffen, die ihr Sprachverständnis auch an diesem Datensatz messen können.

In der Belletristik Life 3.0 beschreibt Tegmark [4], dass „Moravec's landscape metaphor" welches die verschiedenen Herausforderungen der Künstlichen Intelligenz in ein Verhältnis setzt und deren Entwicklungen beschreibt.

Abb. 1 Moravec's landscape metaphor welches Winograd Schemas als noch nicht gelöst ansieht.

Abb. 1 zeigt mit dem Wasserstand, welche Herausforderungen KI schon mindestens so gut löst wie Menschen. Die Themen sind Thematisch geordnet und bauen meist aufeinander auf, so sind Theoremverifikationen einfacher und notwendig für die Beweise von Theoremen. Für Winograd Schemas als Textverständnis, bilden diese beispielweise die Grundlage für kognitive höhere Herausforderungen wie Programmieren oder das Schaffen von Wissen.

Herausforderungen KI auf deutsch

Diese in Abb. 1 dargestellten Herausforderungen werden international bearbeitet und sowohl an Universitäten wie auch in Firmen beforscht. Die Ergebnisse werden in internationalen Evaluationen verglichen. Beispielweise bei der Sprachverarbeitung in [8]. Der Stand der Technik wird dort meist auf Englisch getestet. Eine Anwendung auf andere Sprachen zieht weitere Aufgaben mit sich. Eine Pipeline zu Bearbeitung von Sprachverarbeitungsaufgaben (Englisch: Natural Language Processing (NLP)) besteht aus mehreren Elementen. So gibt es beispielsweise mehrere grundlegende sprachspezifische Aufgaben wie die Erkennung einer Wortart [9]. All diese Komponenten einer NLP-Pipeline müssen in diesem Fall auf andere Sprachen angepasst werden, damit eine NLP-Aufgabe wie Fragen-Beantwortung oder Textzusammenfassung bearbeitet werden können.
Dazu kommt die rechtliche Lage, die bestimmt, welche Daten verwendet werden können. Damit Datensätze mit möglichst wenig Bias erstellt werden können, muss hier noch einiges an manueller Arbeit geleistet werden, bevor größere Sprachmodelle trainiert werden können. Die akademische Evaluation an Güte ist dabei jedoch nicht ausreichend, die Anwendung in echten Verfahren hat den Zweck zu zeigen, wo noch Verbesserungen vorgenommen werden sollten.

Was sind Winograd Schemas?

Auf der Suche nach einer Alternative zum Turing-Test entwickelten Levesque et al. 2012 ein Testverfahren, mit dem sich die Sprachverständnisfähigkeiten von Maschinen messen lassen sollten. Das Ergebnis war die nach dem KI-Forscher Terry Winograd benannte Winograd Schema Challenge, zu deren Bewältigung mehrdeutige Pronomen zum richtigen Bezugswort zugeordnet werden müssen.
Ein solches Winograd Schema besteht aus einer Phrase mit zwei Parteien und einem Pronomen, welches grammatikalisch zu beiden Parteien passt. Eine inhaltliche Zuordenbarkeit des Pronomens entsteht aber erst bei der Einsetzung eines „special word" bzw. dessen Gegenstücks, des „alternate word".

> Ich gieße Milch aus der Flasche in die Tasse, bis sie [voll/leer] ist. Was ist [voll/leer]?
>
> 0: Flasche; 1: Tasse

Für den menschlichen Betrachter fällt eine Zuordnung des Pronomens „sie", wie in diesem Fall, leicht. Mühelos lässt sich eine inhaltliche Verbindung zu „Flasche" oder „Tasse" herstellen, je nachdem, ob „voll" oder „leer" eingesetzt wird. Doch genau in dieser vermeintlich einfachen Zuordnung liegt die Herausforderung für einen Computer. Während Menschen auf das ein Leben lang erlernte Allgemeinwissen über die Welt und, wie in diesem Fall, über das Verhalten von Flüssigkeiten zurückgreifen können, fehlt dem maschinellen Kandidaten das Weltwissen. Die notwendige Information ist nicht in der Äußerung enthalten. Die Austauschbarkeit des „special word" verhindert dabei, dass statistische Zusammenhänge zur Lösung herangezogen werden können, der Test wird quasi „google-proof". Mit der Winograd Schema Challenge lässt sich also die Fähigkeit zur Zuordnung von Pronomen messen, und damit ein nicht unerheblicher Teil des Verständnisses von Sprache. Ein gutes Abschneiden bei diesem Test sagt auch etwas über die Qualität sprachverarbeitender Modelle aus.

Nutzen für die Polizei

Jedes Winograd Schema erfordert Sprachverständnisfähigkeiten in speziellen Bereichen. Um zu beantworten, ob die Tasse beim Ausschütten voll oder leer wird, ist beispielsweise Wissen über das Verhalten von Flüssigkeiten und die Bedeutung von voll und leer in diesem Kontext gefragt. Dank der freien Gestaltbarkeit lassen sich folglich auch Winograd Schemas im polizeilichen Kontext erstellen, die Sprachverständnis in diesem Bereich prüfen. Im Land mit der bundesweit geringsten Polizeidichte[3] läuft die Auswertung strafprozessual beschlagnahmter Geräte schleppend. Smartphones, Computer, etc. können teils monatelang nicht bearbeitet werden.

Gleichzeitig ist der Trend einer Zunahme solcher digitalen Spuren zu beobachten, wie das Landeskriminalamt Baden-Württemberg bereits 2016 festhielt[4]. Eine schnellere und zielgerichteter Auswertung durch sprachverstehende Maschinen zur Bewältigung der Flut digitaler Spuren, könnte Ressourcen sparen. Darüber hinaus könnte diese Methode auch für die Auswertung von Ermittlungsberichten und Vernehmungsdokumenten verwendet werden. Hier verlockt die Anwendung von ermittlungsunterstützenden Assistenten, die sich

[3] https://www.gdp.de/gdp/gdpbw.nsf/id/BD8723445AF3C4FCC12584C5005D03F5/$file/2019-11-27-Polizeidichte-auf-dem-Stand-von-1981.pdf?open
[4] https://lka.polizei-bw.de/wp-content/uploads/sites/14/2017/06/Cybercrime_Digitale_Spuren.pdf

in Akten einlesen und beispielweise durch das beantworten von Fragen bei der Arbeit helfen können.

Ein neuer Datensatz

Zur Überprüfung der Sprachverständnisfähigkeiten, aber auch zum Trainieren neuer Sprachmodelle im Zuge des Maschinellen Lernens, wurde ein neuer Datensatz mit Winograd Schemas erstellt. Als Grundlage wurden kriminalpolizeiliche Vernehmungsdokumente aus Ermittlungsakten des Polizeipräsidiums Mannheim herangezogen, an deren Sprachgebrauch sich der Autor orientieren konnte. Auf natürliche Personen rückführbare Informationen sind selbstverständlich nicht in den Datensatz eingeflossen, alle Winograd Schemas wurden manuell anonymisiert bzw. pseudonymisiert. Das Ergebnis sind 110 Winograd Schemas wie bspw.:

> Die Nachbarn haben die Einbrecher bei der Tat gehört, weil sie [aufmerksam/unvorsichtig] waren. Wer war [aufmerksam/unvorsichtig]?
>
> 0: Nachbarn; 1: Einbrecher

Eine Suche nach bestehenden Datensätze mit solchen Winograd Schemas in einschlägigen Foren verlief ergebnislos. Ohnehin gibt es in der englischsprachig geprägten Domäne kaum deutsche Beiträge. Diese Lücke soll nun gefüllt werden, indem auch KI-Forscher weltweit zur Erstellung und Benutzung weiterer solcher Schemas motiviert und an das Thema „Law Enforcement" herangeführt werden sollen. Im Rahmen der Erstellung des Datensatzes stand der hier auf dem Vorgebirge angesiedelte Winograd Test im Vordergrund. Herausforderungen dabei sind die Ausgewogenheit der Daten, keinen Gender, rassistischen oder andere kulturelle Bias mit einzubringen.
Eine aktuelle Sammlung von Sprachmodelle und Datensätze ist die Plattform Huggingface (https://huggingface.co/). Auch hier zeigt sich, dass die meisten Datensätze z. B. zum Training von Sprachmodellen in der wissenschaftlichen Gemeinschaft der Künstlichen Intelligenz auf Englisch sind. Auf der Seite wird zudem der aktuelle Stand der Technik ersichtlich: Mehrsprachigkeit ist noch eine Herausforderung für NLP-Methoden [6][8]. Der hier veröffentlichte Datensatz wurde auch bei Huggingface eingereicht und wird vermutlich bald als Standarddatensatz aufgenommen.

Zusätzlich ist der Datensatz hier erreichbar:

http://www.kimanufaktur.de/kimanufaktur/DataSets/Winograd/winograd.txt[5]

Referenzen

[1] Cambria E. and B. White, “Jumping NLP Curves: A Review of Natural Language Processing Research,” IEEE Computational Intelligence Magazine, vol. 9, no. 2, pp. 48–57, 2014.

[2] Chen H. , Chung, W., Xu, J. J., Wang, G., Qin, Y., & Chau, M. (2004). Crime data mining: a general framework and some examples. computer, 37(4), 50-56.

[3] Levesque H., “The Winograd Schema Challenge.,” AAAI Spring Symposium - Logical Formalizations of Commonsense Reasoning, 2011.

[4] Löbner, S., Semantik: Eine Einführung. Walter de Gruyter GmbH & Co KG, 2013.

[5] Martelli, F., Kalach, N., Tola, G., & Navigli, R. (2021). SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC). In Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval-2021).

[6] Qi, F., Chang, L., Sun, M., Ouyang, S., & Liu, Z. (2020, April). Towards building a multilingual sememe knowledge base: Predicting sememes for BabelNet synsets. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 05, pp. 8624-8631).

[7] Sakaguchi, K., Le Bras, R., Bhagavatula, C., & Choi, Y. (2020, April). Winogrande: An adversarial winograd schema challenge at scale. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 05, pp. 8732-8740).

[8] Sun, R., “A connectionist model for commonsense reasoning incorporating rules and similarities,” Knowledge Acquisition, vol. 4, no. 3, pp. 293–321, Sep. 1992.

[9] Tegmark T. (2017). Life 3.0: Being human in the age of artificial intelligence. Knopf.

[10] Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., ... & Rush, A. M. (2020, October). Transformers: State-of-the-art natural

[5] Für Aktualisierungen oder Erweiterungen können sie sich gerne unter faehndrich@gmail.com melden.

language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 38-45).