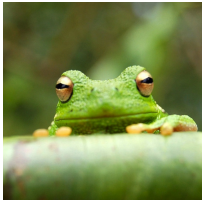


A history of NLP



Abstract

Das Natural-Language-Processing (NLP) bezeichnet ein Teilfeld der Forschung im Bereich der künstlichen Intelligenz an dem bereits seit den 1950er Jahren geforscht wird. Auch wenn mittlerweile Deep Learning mit LSTMs und Transformer-Modellen seine Anwendung hier findet, soll in diesem Artikel auf die statistischen

Herangehensweisen, die über die Jahre entwickelt wurden.

The start of NLP

Natural Language Processing (NLP) bezeichnet eine Schnittmenge der Computer Science und der Linguistik. Durch seine Vielfältigkeit gestaltet es sich eher schwierig sich auf eine einheitliche Definition des Feldes festzulegen, aber ich denke, dass E. D. Liddy hier eine recht umfassende und zugleich kurze Definition hat: "Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications."¹

Kurz und sinngemäß zusammengefasst soll das heißen, dass NLP die Anstrengung bezeichnet Maschinen natürliche Sprache auf die ein oder andere Art und Weise verarbeiten können soll. Hierzu wurden vor dem großen Siegeszug von Deep Learning vor allem statistische

Methoden hergenommen.

Markov Modelle

Ein Markov Modell ist ein Verfahren basierend auf dem momentanen Zustand einer Sache den nächsten Zustand vorhersagen zu können, unter der Annahme, das nur der aktuelle Zustand für den nächsten von Belang ist. Jedem möglichen Zustandswechsel wird eine Wahrscheinlichkeit, die bspw. empirisch erhoben wurde, eine sogenannte Markov-Kette, zusehen in Abb. 1. Diese Wahrscheinlichkeiten für die Transitionen können dann in einer Matrix gesammelt werden. Ausgehend vom momentanen Zustand (dargestellt als ein Vektor aller möglicher Zustände) kann dann per Matrixmultiplikation eine Vorhersage über einen beliebigen Zeitpunkt in der Zukunft getroffen werden.

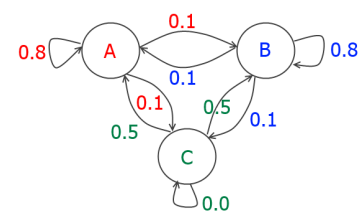


Abb. 1: Markov-Kette

Ein Hidden-Markov-Modell geht dabei noch einen Schritt weiter. Der Beobachter sieht nicht direkt die Transition zwischen den Zustände sondern eine Inzidenz von der wir wiederum mit einer gewissen Wahrscheinlichkeit auf den aktuellen Stand schließen können. Beispiel: Wenn wir jemand mit einem Regenschirm sehen können wir davon ausgehen, dass es regnet. Angewendet auf natürliche Sprache ist es somit möglich angefangen bei einem Wort "eine Näherung höherer Ordnung" nächster Wörter zu erzeugen, bzw. eine Wahrscheinlichkeit für unterschiedliche Sequenzen auszugeben. Dieses Verfahren war nicht unumstritten, weil man der Meinung war,

¹Liddy, E. D. (2001) SURFACE SURFACE Center for Natural Language Processing School of Information Studies (iSchool) 2001 Natural Language Processing Natural Language Processing Natural Language Processing 1. Available at: <https://surface.syr.edu/cnlp> (Accessed: 20 June 2021)

dass die (englische) Sprache nicht angemessen über dieses statistische Verfahren dargestellt werden kann. Trotzdem findet dieses immer noch in einigen Fällen Anwendung beim Part-of-Speech-Tagging. Hierbei stellen jetzt die Wortarten (Subjekt, Verb, usw.) unsere Zustände dar, die aufeinander folgen können, und die Wörter des Satzes sind unsere Inzidenz. Beispiel: Ein Satz beginnt mit einer Wahrscheinlichkeit von 70% mit einem Nomen oder mit 30% mit einem Verb und das erste Wort des Satzes ist "rennen", was sowohl ein Verb (90%) als auch ein Nomen sein kann (10%). Unter der Inzidenz "rennen" und dem aktuellen Stand "Satzbeginn" ist das erste Wort also der Wahrscheinlichkeit ein Verb.

Ein solches Vorgehen ist in seiner Performance logischerweise stark abhängig von der Art und Menge an Texten auf denen man trainieren kann und auch wenn dieses Verfahren bereits aus den 50er Jahren stammt erfuhr es immer wieder Schübe durch die Verfügbarkeit großer maschinenlesbarer Textkorpi, wie dem "Brown Corpus" und mittlerweile durch das Internet auch

zu den meisten noch so spezifischen Themen. Jedoch wird wohl immer das Problem bestehen bleiben, dass eigentlich nie genügend Daten vorhanden sind, egal wie groß der Textkorpus (Zipf's Law).

TF-IDF

Ein weiteres Verfahren des NLPs, beziehungsweise des Information Retrievals, dass ganz ohne ML-Modelle auskommt ist die Term Frequency – Inverse Document Frequency (TF-IDF). Term Frequenz ist schlicht wie oft ein Term in einem Dokument vorkommt. Oft wird hierbei noch gesehen auf die Länge des Dokuments normalisiert. Füllwörter haben dabei natürlicherweise höhere Werte als bedeutungsvolle. Um den entgegen zu wirken wird der TF mit dem IDF verrechnet: Dieser ist der Logarithmus aller Dokumente in denen das Wort vorkommt, durch die Anzahl aller Dokumente. Somit wird der Wert für Terme die in (fast) jedem Dokument vorkommen, wie "und" und "der" drastisch verringert. Dies ist beispielsweise hilfreich für ein so-

genanntes Stop-Word-Removal, um jene Füllwörter aus seinem Textkorpus zu filtern. Ein anderer Anwendungszweck ist bspw. das Erzeugen von Dokumentenvektoren basierend auf dem ermittelten TF-IDF, um über die Kosinusdistanz Ähnlichkeit zwischen unterschiedlichen Dokumenten oder Dokumenten und einer Suchanfrage zu ermitteln.

Modernes NLP

Auch wenn beide Techniken relativ "simpel" sind im Vergleich zu den heutigen Techniken im Bereich des NLP, mit Transformern, LSTMs und anderen Deep Learning Verfahren, so finden sie auch heute noch zahlreiche Anwendung, weil sie entweder die geeignetste Lösung sind und in diesem Anwendungsbereich nur schwer zu überbieten sind, oder weil sie einfach umzusetzen sind und dabei trotzdem eine hinreichend gute Lösung bieten. Ebenso bauen neuere Techniken wie "Bag of Words" auf diesen älteren Vorgehen auf und bilden somit eine wichtige Grundlage, die man kennen sollte wenn man sich mit diesem Feld der Forschung auseinander setzen möchte.