Das Fraunhofer-Institut für Naturwissenschaftlich-Technische Trendanalysen berichtet über neue Technologien

Natural Language Processing

Natural Language Processing (NLP) beschreibt computergestützte Techniken zur maschinellen Erkennung und Verarbeitung von natürlicher Sprache. Das Ziel ist dabei, die direkte Kommunikation zwischen Mensch und Computer auf der Basis natürlicher Sprache zu ermöglichen, die zwischenmenschliche Kommunikation durch maschinelle Übersetzung zu erleichtern und auch die enorm zunehmenden Mengen an Texten in natürlicher Sprache maschinell analysieren zu können. NLP verwendet hierfür Ergebnisse der Sprachwissenschaften sowie Methoden und Techniken der Künstlichen Intelligenz, speziell aus dem Bereich maschinelles Lernen und dem Unterbereich Deep Learning. In Kombination mit den stark gestiegenen Computerleistungen lassen besonders die aktuellen Fortschritte in den zuletzt genannten Technologiegebieten erwarten, dass im bereits seit Längerem etablierten Forschungsgebiet NLP in naher Zukunft mit entscheidenden Fortschritten zu rechnen ist.

Für die Verarbeitung von natürlichen Texten müssen moderne NLP-Algorithmen eine Reihe von Aufgaben erfüllen. In einem ersten Schritt müssen sie den Text in einzelne Sätze und anschließend in einzelne Wörter unterteilen. Falls wie im Deutschen oder Englischen Satz- und Leerzeichen zwischen einzelnen Wörtern verwendet werden, ist das relativ einfach. Etwas komplexer ist der nächste Schritt, in dem man den einzelnen Wörtern Wortarten (z.B. Nomen) zuordnet (part-of-speech tagging). Bei den heute verwendeten Verfahren unterscheidet man hierbei zwischen überwachtem und unüberwachtem maschinellen Lernen. Bei dem überwachten maschinellen Lernen „lernt" der Computer mithilfe von zahlreichen verfügbaren annotierten Texten. Gerade für Sprachen, für die wenig annotiertes Textmaterial zur Verfügung steht, bietet sich die Methode des unüberwachten maschinellen Lernens an. Hierbei erkennt der Algorithmus die Muster im Sprachgebrauch selber und definiert eigene Wortarten-Kategorien. Man konnte zeigen, dass diese Wortarten den manuell erstellten Wortarten erstaunlich ähnlich sind. Weitere wichtige Analyseschritte in dem Zusammenhang sind die Erkennung der Grundform des Wortes (z.B. sieht – sehen) sowie die Erkennung von Eigennamen.

Eine besonders komplexe Aufgabe für NLP-Systeme ist die Syntaxanalyse (Parsing), um die grammatische Struktur der Sätze zu verstehen. Wichtige Analysemodelle sind hierbei das Dependency Parsing, bei dem man die Beziehung der Wörter untereinander beschreibt (z.B. das Subjekt des zugehörigen Verbs) und das Constituency Parsing, bei dem man den Satz als Baumdiagramm aus verschiedenen Satzteilen darstellt (z.B. die Nominalphrase „der weiße Hase"). Dieser Bereich wird immer noch aktiv erforscht, so wurden in den letzten Jahren auch Deep-Learning-Ansätze auf der Basis von künstlichen neuronalen Netzen entwickelt.

Ein Spezialbereich des NLP ist die Sentiment-Analyse, in der Texte automatisiert bezüglich der positiven oder negativen Einstellung und emotionalen Haltung des Autors analysiert werden. Die Analyse erfolgt entweder statistisch durch Definition von positiven oder negativen Signalwörtern oder mithilfe des maschinellen Lernens durch Trainieren der Software anhand von Beispieltexten mit bekannter emotionaler Haltung. Verwendet wird die Sentiment-Analyse z.B. zur Analyse von Produktbewertungen oder auch zur Analyse der Stimmung an der Börse.

Auch gesprochene Sprache kann von den NLP-Systemen analysiert werden. In dem Fall wird der gesprochene Text in Clips im Millisekunden-Bereich aufgeteilt und auf vorhandene Phoneme hin untersucht. Die Erkennung der gesprochenen Sprache ist eine wichtige Basis für die chatbots, die heute bereits im Kundenservice oder den Smart-Home-Geräten eingesetzt werden. Besonders in den asiatischen Ländern haben sich auch Social Chatbots schnell verbreitet, die in erster Linie ein virtueller Freund sein sollen, die zuhören, Emotionen erkennen und darauf eingehen können. Herausragende Bedeutung erlangt die Erkennung gesprochener Sprache zunehmend im Zusammenhang mit der Sprachsteuerung von technischen Systemen durch den Menschen.

Eine weitere Anwendungsmöglichkeit für NLP bietet sich für den Wissenschaftsbereich an. Hier wächst die Anzahl der Publikationen in vielen Bereichen exponentiell, sodass man hier auf NLP-Systeme hofft, die diese Informationen zusammenfassen, wichtige Fakten herausstellen und erste Hypothesen aufstellen können. Ein anderer Bereich, in dem die Anzahl der vorhandenen Texte rasant wächst, sind die sozialen Medien. Hier nutzen NLP-Systeme die vorhandenen Daten, um Stimmungen in der Bevölkerung zu erkennen oder in Notfallsituationen den Einsatzkräften Informationen über die aktuelle Lage zur Verfügung zu stellen. Allerdings gibt es gerade im Bereich Social Media Mining auch wachsende Bedenken z.B. bezüglich des Datenschutzes. Eine wichtige Anwendung von NLP-Systemen sind auch die Übersetzungsprogramme, die in den letzten Jahren deutlich besser geworden sind.

NLP-Systeme gelten als einer der kompliziertesten Bereiche der Informatik, was daran liegt, dass die natürliche Sprache komplex und mehrdeutig ist und häufig Kontextinformationen notwendig sind, um einen Satz verstehen zu können. Eine besondere Herausforderung stellt dabei z.B. Ironie dar. Gleichzeitig ist die Nutzbarmachung auf natürlicher Sprache basierender Mensch-Maschine-Interaktion eine immer entscheidendere Voraussetzung für die weitere Verbreitung computerbasierter Unterstützungssysteme in den verschiedensten Bereichen unseres täglichen Lebens, weil alle anderen Steuerungsmechanismen als zu kompliziert erscheinen bzw. zu große Aufmerksamkeit verlangen würden. Auch das erklärt die enormen Forschungsanstrengungen, die auf diesem Gebiet inzwischen unternommen werden.

Weitere Anwendungsbereiche, denen für die nächsten Jahre beschleunigte Zuwachsraten für NLP-Systeme vorausgesagt werden, sind z.B. maschinelle Übersetzung, Frage-Antwort-Systeme (wie Chatbots), Sentiment-Analyse oder automatisierte Zusammenfassungen. Vielversprechend ist hierbei insbesondere die Kombination von NLP mit Technologien aus den Bereichen Big Data und Machine Learning.

Dr. Sonja Grigoleit