NLP

Das Akronym *NLP* steht für *Natural Language Processing* und bezieht sich auf die Verarbeitung natürlicher Sprache durch Computersysteme. Dabei speist sich NLP aus dem Wissen und aus Verfahren verschiedener Disziplinen. Die zentrale Disziplin dürfte die *Computerlinguistik* als solche sein.[1] Aber auch Verfahren des maschinellen Lernens bzw. der künstlichen Intelligenz spielen eine Rolle, ebenso wie Aspekte der Datenverarbeitung, z.B. in Form der Datenbanktheorie oder durch die *Online*-Verarbeitung von *Streams* (Datenströmen), also die Verarbeitung von Daten ohne (dauerhafte) Speicherung.

Unterbereiche des NLP

Da NLP, wie zuvor erwähnt, jegliche computergestützten Verarbeitung (bzw. explizit zu nennen auch Generierung) menschlicher Sprache umfasst, sind die Teilbereiche der NLP vielfältig. Zu den wichtigsten gehören:[2]

- Konvertierungen *Text-to-Speech* und *Speech-to-Text*
- *Tokenisierung*, also die Zerlegung eines Textes in *Tokens*
- *Lemmatisierung*, also die Umwandlung von *Token* in entsprechende *Lemmata* (*Types*)
- *PoS-Tagging*, also die Bestimmung grammatischer Informationen für *Token*
- *Sentiment Analysis*, also die Ermittlung der *Polarität* (*positiv – negativ*) von textuellen Elementen
- *Machine Translation* (*MT*), also die automatisierte Übersetzung
- *Natural Language Generation* (*NLG*), also die Erzeugung menschlicher Sprache aus Daten
- v.m.

Technische Umsetzung: regelbasierter Ansatz vs. maschinelles Lernen

NLP kann technisch betrachtet v.a. auf zwei Wegen operieren. Der ‚ältere' Ansatz ist der *regelbasierte* (*rule based*), der versucht, sprachliche Prozesse bzw. Analyseverfahren für sprachliche Daten mithilfe von definierten Algorithmen zu beschreiben. Der zweite Ansatz beruht, grob gesagt, auf der Verarbeitung menschlicher Sprache ausgehend von der Berechnung von Wahrscheinlichkeiten, wofür heutzutage in der Regel *maschinelle Lernverfahren* eingesetzt werden, häufig sogar *neuronale Netze* bzw. *Deep Learning*. Während der regelbasierte Ansatz in der konkreten Anwendung (also sobald der zumeist äußert komplexe Algorithmus ermittelt wurde) meist recht einfach ist, benötigen maschinelle Lernverfahren mehr Vorbereitung.[3]

Zum einen müssen Textkorpora erstellt werden, die für maschinelles Lernen geeignet sind. Zum anderen müssen im Zuge des maschinellen Lernens sogenannte Modelle trainiert werden, die ausgehend von bekannten Referenzdaten neue Daten verarbeiten können. Die entsprechenden Modelle können äußert umfangreich werden: Modelle zur *NLG* (*Natural Language Generation*) können beispielsweise schnell so umfangreich werden, dass sie von handelsüblichen Heimcomputern kaum oder nur mit extrem langen Rechenzeiten generiert (oder genauer trainiert) werden können. Somit kann man nicht pauschal aussagen, welcher Ansatz der ‚bessere' ist. Sowohl regelbasierte Verfahren als auch maschinelle Lernverfahren sind auch heutzutage als gleichberechtigt zu sehen. Es muss immer im konkreten Anwendungsfall entschieden werden, welcher Ansatz der geeignetere ist.

Literaturnachweise

1. Vgl. zu den *Basics* der Computerlinguistik weiterhin CARSTENSEN, Kai-Uwe u.a. (Hg.) 2010: *Computerlinguistik und Sprachtechnologie*, Heidelberg.[↩]
2. Vgl. u.a. zu den einzelnen Bereichen CLARK, Alexander/FOX, Chris/LAPPIN, Shalom (Hg.) 2013: *The Handbook of Computational Linguistics and Natural Language Processing*, West Sussex. Dabei gehen Clark/Fox/Lappin davon aus, dass NLP die „engineering domain" der Compuerlinguistik darstellt (CLARK,

Alexander/FOX, Chris/LAPPIN, Shalom (Hg.) 2013: *The Handbook of Computational Linguistics and Natural Language Processing*, West Sussex, S. 1).[↩]

3. Vgl. hierzu weiterführend auch MMONTEJO-RÁEZ, Arturo/JIMÉNEZ-ZAFRA, Salud María 2022: „Current Approaches and Applications“, in: *Appl. Sci.* 2022, 12(10), 4859, 6 Seiten, online unter: https://www.mdpi.com/2076-3417/12/10/4859.[↩]