Part-of-Speech-Tagging

Resch, Claudia; claudia.resch@oeaw.ac.at

Eine der häufigsten fachspezifischen Annotationen von Texten besteht in der Zuweisung linguistischer Information nach morphosyntaktischen Merkmalen. Das sogenannte *Part-of-Speech-Tagging* (PoS-Tagging) ist das Klassifizieren eines Textes nach Wortarten und stellt neben der Tokenisierung und Lemmatisierung einen wesentlichen Bestandteil der linguistischen Basisannotation dar: Dabei werden die in einem Text vorkommenden Wörter und Satzzeichen mit einem vordefinierten Inventar von verfügbaren Wortarten (TagSet) einer grammatikalischen Klasse zugewiesen, wodurch eine Suche nach abstrakten sprachlichen Phänomenen möglich wird. Auf diese Weise kann die Abfrage generalisiert werden, etwa indem nach bestimmten Wortarten oder Sequenzen von Wortarten gesucht wird. Die Abfrage kann damit jedoch auch weiter spezifiziert werden, zum Beispiel, wenn gezielt nach allen Belegen der Wortform 'sein' in der Funktion des Possessivpronomens gesucht wird, hingegen die Vorkommen von 'sein' als Auxiliarverb ausgeschlossen werden sollen.

Die Zuordnung der Wortformen zu einer Wortart kann manuell, halb-automatisch oder – bei sehr großen Textsammlungen – automatisch durch sogenannte *Part-of-Speech*-Tagger (kurz: PoS-Tagger) erfolgen. Deren Zuweisungen und Disambiguierungen basieren entweder auf Regeln (symbolische Tagger) oder auf statischen Verfahren bzw. maschinellen Lernverfahren (stochastische Tagger). Sogenannte hybride oder transformationsbasierte Tagger kombinieren beide Verfahren, indem sie bei der Disambiguierung mehrdeutiger Einheiten zunächst von der wahrscheinlichsten Wortart ausgehen, um diese dann durch kontextspezifische Regeln zu korrigieren (Perkuhn/Keibel/Kupietz 2012, S. 59). Automatische Tagger können an verschiedene Sprachen und Sprachstufen angepasst werden – eine Liste von frei verfügbaren PoS-Taggern bietet die Universität Stanford an.

Als Referenzsysteme und Grundlage für das Trainieren des PoS-Taggings werden Texte von bester Qualität (Goldstandard) herangezogen, deren automatische Annotation manuell überprüft und nachkorrigiert wurde. Generell ist die Automatisierung des PoS-Taggings für große, moderne Standardsprachen bereits sehr weit fortgeschritten – allerdings „bleibt hier für historische oder variante Spracherzeugnisse oder bestimmte literarische Genres noch viel zu tun" (Rapp 2017, S. 259).

Die Qualität von manuell durchgeführten Annotationen beruht letztlich auf interpretativen Entscheidungen, deren Zuverlässigkeit durch die Anwendung des *Inter-Annotator-Agreements* oder des *Intra-Annotatator-Agreements* gesichert werden soll. Im Sinne der Nachnutzbarkeit von Annotationen ist es wichtig, deren Qualität einzuschätzen, mögliche Fehlerquellen zu diskutieren und die Verwendung der Labels sowie getroffene Entscheidungen in den Tagging-Guidelines nachvollziehbar zu dokumentieren.

Literatur:

- Bański, Piotr; Haaf, Susanne; Mueller, Martin: Lightweight Grammatical Annotation in the TEI: New Perspectives. In: LREC 2018 – 11th International Conference on Language Resources and Evaluation. Japan, S. 1795–1802.
- Lemnitzer, Lothar; Zinsmeister, Heike: Korpuslinguistik. Eine Einführung. Tübingen: 2010.
- Perkuhn, Rainer; Keibel, Holger; Kupietz, Marc: Korpuslinguistik. Paderborn: 2012.
- Rapp, Andrea: Manuelle und automatische Annotation. In: Digital Humanities. Eine Einführung. Stuttgart: 2017, S. 253–267.
- Schiller, Anne; Teufel, Simone; Stöckert, Christine; Thielen, Christine: Guidelines für das Tagging deutscher Textcorpora mit STTS: 1999. URL: `http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf`.

Software:

CLARIN-mediated NLP-services, weblicht, Apache OPENNLP, CLAWS POS-Tagger for English, TreeTagger, RNNTagger, SoMeWeTa, spacy , Stanford Log-linear Part-Of-Speech-Tagger, Universal Dependencies

Verweise:

Lemmatisierung, NLP, SpaCy, Tagger, Tagsets, Textannotation, Weblicht, Elemente digitaler Editionen

Projekte:

Deutsches Textarchiv, Austrian Baroque Corpus (ABaC:us), travel!digital, Liste von Part-of-Speech-Taggern

Themen:

Einführung, Natural Language Processing

Zitiervorschlag:

Resch, Claudia. 2021. Part-of-Speech-Tagging. In: KONDE Weißbuch. Hrsg. v. Helmut W. Klug unter Mitarbeit von Selina Galka und Elisabeth Steiner im HRSM Projekt "Kompetenznetzwerk Digitale Edition". URL: https://gams.uni-graz.at/o:konde.156