Natural language processing: Corpus-specific Entity Recognition Training

Eigennamenerkennung in biomedizinischen Publikationen

Eigennamenerkennung, auch als Named Entity Recognition (NER) bezeichnet, ist eine Technik im Bereich des Natural Language Processing (NLP), die darauf abzielt, Entitäten in einem Text zu identifizieren und zu klassifizieren. Entitäten können beispielsweise Medikamentennamen in einem Arztbrief sein. Die Eigennamenerkennung wird in der Regel mithilfe von maschinellen Lernalgorithmen und statistischen Modellen (häufig neuronale Netze) umgesetzt. Hierfür werden annotierte Trainingsdaten (sogenannte Corpora) benötigt. In verschiedenen Anwendungen existiert eine Vielzahl von annotierten Corpora, welche sich in Scope, Zielsprache, Format und Annotationsrichtlinien unterscheiden. In der Regel weiß man nicht auf welchem Corpus ein System trainiert werden soll und man möchte soviel wie möglich der verfügbaren Trainingsdaten verwenden. Aufgrund der Unterschiede der Corpora zueinander gestaltet sich dies leider als herausfordernd. Ziel dieser Arbeit ist es ein System zur Erkennung von Eigennamen zu entwickeln, welches für das Training Daten aus mehreren Systemen benutzt. Mögliche Anwendungsfälle sind multi-linguale Corpora zum Thema Medikamentenerkennung.

Umsetzung:

Für die Abschlussarbeit werden mehrere annotierte Corpora zur Verfügung gestellt. Die Arbeit umfasst die Umsetzung und Evaluation von transformerbasierten Verfahren zur Eigennamenerkennung, die Umsetzung von multi-corpus Training, sowie eine detaillierte Fehleranalyse. Für die Umsetzung der Maschinellen Lernverfahren sollen state-of-the-art Verfahren eingesetzt werden. Die Arbeit umfasst auch eine wissenschaftliche Exploration von relevanten Arbeiten und Technologien.

Quellen:

- https://aclanthology.org/2021.bsnlp-1.12.pdf
- https://aclanthology.org/P19-1014.pdf

Voraussetzungen:

- Gute Programmierkenntnisse in Python
- Erfahrung mit Deep Learning oder Transformer-Netzwerken
- Verarbeitung von natürlicher Sprache

Bei Interesse und/oder Fragen schicken Sie gerne eine Mail an: johann.frei@informatik.uni-augsburg.de