Natural language processing: Entity Linking auf ICD-10 Codes

NLP: Entity Linking auf ICD-10 Codes

Entity Linking (EL) ist eine klassische Aufgabe im Bereich des Natural Language Processing (NLP), in der zu einzelnen, identifizierten Textstellen ein korrespondierender Eintrag in einer Datenbank gefunden werden soll. Beispielsweise kann zu der Phrase “Störung der Nierenfunktion” der ICD-10-Code “C4376” zugeordnet werden.
Da manche Begriffe nicht exakt gematched werden können, muss statt einer klassischen Suche eine semantische Suche eingesetzt werden, die moderne NLP-Modelle einsetzen kann, um eine intelligente Suche zu ermöglichen.

Aufgabenstellung

Das Ziel der Arbeit ist es, ein NLP-Modell mit zugehörigem Pre- und Postprocessing-Code zu entwickeln, das für eine vorliegende kurze Textphrase den korrespondierenden ICD-10-Code identifizieren kann (bzw. bei bedeutungslosen Phrasen auch keinen Match liefert).
Das NLP-Modell kann von bestehenden Modellen aus der Huggingface-Plattform übernommen werden und ggf. optional nachtrainiert werden.
Der Fokus liegt auf deutschen Texten, aber dies ist auch anpassbar.

Konkret inkludiert die Arbeit folgende Schritte:
- Aufbau einer Liste an ICD-10-Codes und der zugehörigen Begriffe mittels Parsen der ICD-10-Dateien - Encodierung der Begriffe in eine Vektordatenbank (z.B. mittels BERT-Embeddings in pgvector)
- [Optional] Fine-tuning des BERT-Modells
- [Optional] Erweiterung der bestehenden ICD-10-Begriffe um weitere Synonyme zur Verbesserung der Trefferquote
- Auswertung der Qualität des EL & Ermitteln wichtiger Hyperparameter (z.B. Linking-Threshold)

Verfügbare Datenquellen:
- MIMIC-III/IV mit ICD-10 codes (Englisch)
- Unstrukturierte Wikipedia- & strukturierte WikiData-Daten (mehr Infos dazu im persönlichen Gespräch)
- ICD-10 (Deutsch)

Mögliche weitere Aspekte:
- Analyse der Qualität bei Cross-lingual-Modellen (Verbessert Englisch+Deutsch sogar das Ergebnis?)
- Wie sehr verbessert Fine-tuning die Qualität des EL?

Voraussetzungen:

- Sichere Programmierkenntnisse in Python
- Bereitschaft zur eigenständigen Arbeit
- Interesse an Deep Learning-Techniken wie Transformer-Netzwerken
- Interesse an Verarbeitung von natürlicher Sprache

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