Abstract ID: 91679

Student: KUPPASSERY ABDULNAZAR AKHILA NAZ

Area of Research: Computational and structural science

PhD Programme: PhD Advanced Medical Biomarker Research (AMBRA)

Semester: 3

SAP-BERT based Token N-Gram Normalization using SNOMED-CT

AKHILA NAZ KUPPASSERY ABDULNAZAR

Normalization, i.e., context-aware mapping of medical terms to semantic identifiers of a terminology standard such as SNOMED CT, is of immense importance in making clinical data interoperable. Known obstacles are idiosyncratic, compact, ambiguous and often faulty language in clinical narratives. This work aims at improving the normalization of token n-grams (words and word sequences) for n = 1 (“cholecystectomy”) up to n = 5 (“removal of gallbladder by laparoscopy”) from a clinical document. These n-grams are mapped to their corresponding SNOMED CT codes using embeddings, i.e., representations in a low-dimensional vector space. Different normalization variants such as contextualized, non-contextualized and fine-tuned ones were compared, using the SAP-BERT model. Evaluation on N2C2 data revealed that non-contextualized normalization with SAP-BERT performed best with an accuracy of 0.85 on validation data and 0.819 on test data. N-gram based normalization yields a still acceptable accuracy of 0.787 accuracy, due to a large number of false positives. Future work will optimize filtering during normalization and fine-tuning of term recognition.