Natural language processing to extract information from pathology reports for targeted cancer treatment.

Daniel Nyakas

MSc in Big Data

UNIVERSITY of STIRLING

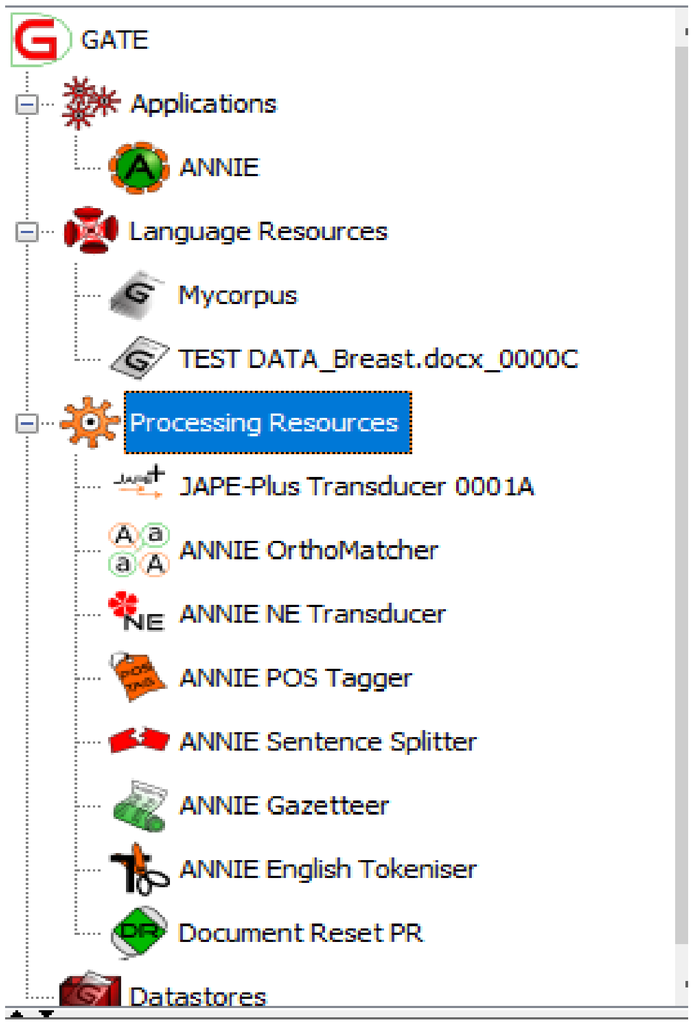
Background

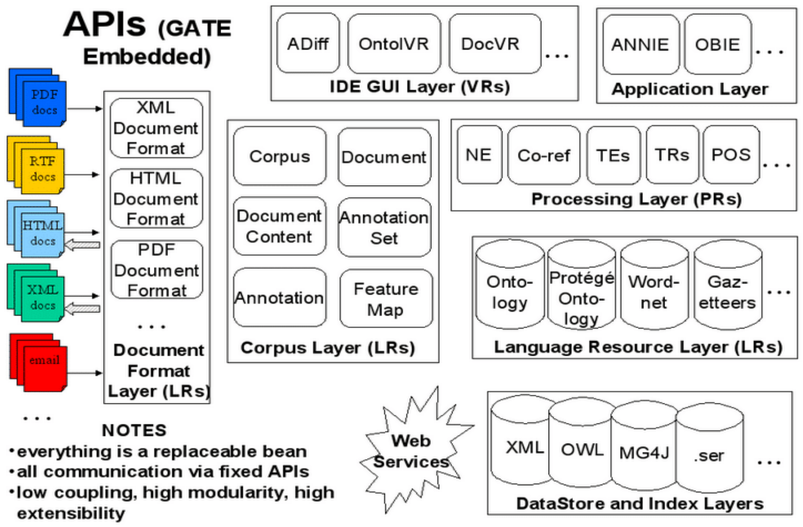
Within the NHS, patient's electronic clinical records contain a lot of knowledge that exists within free text reports. One of the largest areas of free text is the pathology reporting system which contains many 100,000's of records. The main objective of the project is to test the ability of the NLP using GATE to code pathology reports for breast cancer.

Methods

I obtained anonymised breast cancer pathology reports from the Western General Hospital and NHS Lothian. Iteratively, I developed a rule-based natural language processing (NLP) system to automate codification (SNOMED/UMLS) of clinical concepts within free-text narratives. I used GATE Embedded framework for implementation in Java.

Processing resources in GATE





Conclusion and future directions

It is possible to automate the identification of important information from pathology reports with NLP, although several challenges will need to be overcome. These include meeting requirement for satisfactory information governance and data security, implementing an efficient processing strategy for large volumes of data and the reporting of coded data in a manner meaningful for clinical interpretation and analysis with integration alongside other clinical datasets.