Engineering a Machine Learning Pipeline for Automating Metadata Extraction from Longitudinal Survey Questionnaires

Suparna De*[1], Harry Moss[2], Jon Johnson[3], Jenny Li[3], Haeron Pereira[1], and Sanaz Jabbari[2]

[1]University of Surrey – United Kingdom
[2]Research IT Services, UCL – United Kingdom
[3]CLOSER, UCL Institute of Education – United Kingdom

Abstract

Data Documentation Initiative-Lifecycle (DDI-L) introduced a robust metadata model to support the capture of questionnaire content and flow, and encouraged through support for versioning and provenancing, objects such as BasedOn for the reuse of existing question items. However, the dearth of questionnaire banks including both question text and response domains has meant that an ecosystem to support the development of DDI ready CAI tools has been limited. Archives hold the information in PDFs associated with surveys, but extracting that in an efficient manner into DDI-Lifecycle is a significant challenge.
While CLOSER Discovery has been championing the provision of high-quality questionnaire metadata in DDI-Lifecycle, this has primarily been done manually. More automated methods need to be explored to ensure scalable metadata annotation and uplift.

This paper presents initial results in engineering a machine learning (ML) pipeline to automate the extraction of questions from survey questionnaires as PDFs. Using CLOSER Discovery as a 'training dataset', a number of machine learning approaches have been explored to classify parsed text from questionnaires to be output as valid DDI items for inclusion in a DDI-L compliant repository.
The developed ML pipeline adopts a continuous build and integrate approach, with processes in place to keep track of various combinations of the structured DDI-L input metadata, ML models and model parameters against the defined evaluation metrics, thus enabling reproducibility and comparative analysis of the experiments. Tangible outputs include a map of the various metadata and model parameters with the corresponding evaluation metrics' values, which enable model tuning as well as transparent management of data and experiments.

*Speaker