A tool for gold standard data generation

Since we needed gold standard datasets to evaluate the accuracy of dose information extracted from RxStar and clinical notes, we developed a tool to assist manual annotation to generate gold standard datasets. We wrote a script using the R programming language, which generates a template pre-populated with subject ID, date, note ID, note type, note subtype, an excerpt containing dose information, and the whole note content and additional columns for dose extraction. The excerpt is a short text that includes 30 characters before each drug mention and up to 300 characters after that drug mention, where this window was empirically selected during development of the tool. The number of columns for dose extraction was adjusted depending on medications. For lamotrigine, we used 4 columns for dose amount at given intake time as lamotrigine is given up to 4 times a day, but for most medications, 3 columns would be sufficient. The lamotrigine template also included a column for "XR" indicating extended release or immediate release, and a column for "comment" that can be used for any comments by an annotator if needed (e.g., dose change). This tool helped us to avoid the major time consuming step of searching through all notes in the SD.