Can the CCR Leverage NLP for Quality Control Activities?

S. Wood, B.A
California Cancer Reporting and Epidemiologic Surveillance (CalCARES) Program*
Institute for Population Health Improvement, UC Davis Health System

Background
As the California Cancer Registry (CCR) moves more and more towards automatic processes to link and consolidate data we need to identify new methods for quality control activities. Previously all incoming hospital abstracts uploaded into Eureka (CCR’s Database Management System) were “visually edited” to ensure that all coded values were supported by the incoming text related to the diagnosis. The sheer volume of data makes it impossible to manually perform this method of quality control on all cases. In automating tumor linkage, we have found that while all cases are linked properly based on the coded (edit free) data, further analysis suggests that for a small portion of the data a different outcome is required. These cases can be edit free but incorrectly coded based on the code to text comparison or linkage may be incorrect when viewed in a larger context of existing disease.

A small portion of the cases that automatically create new tumors are later found to be representative of metastatic disease and require that the newly created tumor be deleted and the case linked to a tumor that already existed. The data that informs this decision is found in the free form text fields of the new case record it is not feasible to introduce additional logic into the automated tumor linkage rules. Currently we are developing database queries to see how accurately we can identify whether the text definitively identifies whether a case is positive or negative for metastatic disease. At this point the queries are very basic string comparisons that look for the mention of metastatic disease without terms that negate metastatic disease.

Objective
Currently the CCR is looking to improve our method for quality control. We have already adjusted our process flow to allow for cases to complete (go through Linkage and Consolidation) before quality control (Visual Editing), rather than letting the case be held up for quality control. Our next step is to implement changes to the quality control process as well as the selection criteria for that process. Our quality control checks are primarily concerned with ensuring that the code is supported by the text. In order to help this process I want to see if there are any NLP tools that could help us identify cases that may be problematic, and target those. Will we be able to define what qualifies as text that supports the code and identify where cases fall out? Without major software development efforts can the CCR begin to implement a tool that identifies cases that may represent metastatic disease for target quality control activities.

Are there tools that exist that allow for a business analyst or quality control specialist to identify positive and negative and apply machine learning logic to flag such cases?

Methodology
I performed research about available Open Source NLP tools and utilities to perform a proof of concept on some example data, particularly final diagnosis text. The research pointed me into the direction of a few different projects:

- StanfordCoreNLP: provides a set of human language technology tools. It can give the base forms of words, their parts of speech, whether they are names of companies, people, etc., normalize dates, times, and numeric quantities, mark up the structure of sentences in terms of phrases and syntactic dependencies, indicate which noun phrases refer to the same entities, indicate sentiment, extract particular or open class relations between entity mentions, get the quotes people said, etc.
- Canary: an open-source Natural Language Processing (NLP) platform for information extraction. It is geared towards researchers and clinicians without an extraneous technical background with the aim of empowering them to conduct their own research independently.
- Apache OpenNLP: OpenNLP supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, language detection and coreference resolution

The graphical interface and apparent ease of use of Canary suggested that it would be the best candidate to try to configure and run small sets of sample data through.

Overall Results
Installing the tool and setting up a project in Canary was simple and straightforward. Although the interface is uncomplicated, it quickly became clear that the work involved in configuring a new project was a great deal more complex and time consuming than I had anticipated, especially considering the diversity and lack of consistency in the text fields involved, and required a significant commitment to perform even the most basic proof of concept.

Conclusion
The answer as to whether the CCR Leverage NLP for Quality Control Activities remains to be seen. Certainly NLP can be leveraged to help identify portions of the data that may require the attention of a quality control specialist in order to help ensure the highest data quality possible. Whether Canary is the best tool for this process also remains to be seen. Additional time and resources will be required to determine what is the best tool for the nature of the data and the decision points.

Ultimately I would like to find a machine learning tool to classify a large sample size of the data to determine if incoming data matches the criteria of the large data set. If 90% of records are clearly negative/positive examples of metastatic disease then ideally the 10% could be analyzed and incorporated into the machine learning data set to come closer to 100%. The impact concerning metastatic disease is very small but would be a good proving ground for technology that could serve as a quality control measure to ensure that the codes in the record are backed up correctly. This would not fully automate the process of text to code review but could serve to focus the resources to cases that don’t match previous data. The tool would have to be robust and adaptable, but is hopefully the future of our quality control activities.