Evaluation of a Study Cohort Query Formulation Tool: Criteria2Query

Paul M. Heider, PhD¹, Chunhua Weng, PhD², Stéphane M. Meystre, MD, PhD¹
¹ Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC
² Department of Biomedical Informatics, Columbia University, New York, NY

Introduction: Evaluation of natural language processing performance should not be limited to precision, recall and F-measure as these do not capture the intricate failings that appear when evaluating performance in one domain for a system developed and trained on a separate domain. In an effort to generate more actionable insights into our domain adaptation work, we have put together a guide to help classify types of system errors we have encountered. First, we will introduce criteria2query¹, a tool for semi-automatically transforming criteria text into executable standards-based cohort database queries. Then we will analyze our evaluation of generalizing this tool from a neurology domain to an oncology domain and discuss how this analysis can serve as a guide in other ontology-rich annotation tasks.

Methods: To evaluate criteria2query's off-the-shelf generalized performance for the oncology domain, we used criteria2query, as developed for neurology trials, to parse three clinical trial eligibility criteria listings for real cancer trials. A reference annotation of these listings, done in collaboration with an oncologist, involved linking concepts in the clinical trial eligibility criteria description with their counterparts in an ontology². The concepts were composed with numerical and logical operators to match the semantics of the trial's criteria. For example, all three trials required a minimal ECOG Performance Status, which is often used in oncology and measures a patient's level of functioning. We linked mentions of ECOG Performance Status with UMLS CUI C1520224. In tandem, we created a rule limiting values associated with code C1520224 to zero or one for trials that required participants to have "ECOG PS<=1".

Results: Table 1 highlights a basic evaluation of the criteria2query extraction against the manual annotation. Let us walk through the first trial's results from left-to-right to understand this table. A manual annotation yielded six concepts and four criteria rules. The criteria2query extraction included two relevant concepts. One of those two concepts was included in a manually extracted rule (Ref. % in Table 1). The other concepts are important to trial eligibility but were not the focus of this experiment and, thus, were not annotated. Criteria2query derived two rules. Zero of them matched the four manually annotated reference rules. One of those two rules was a correct criterion (but did not count towards the previous column because it contained concepts not in the manual annotation).

Discussion: Manually extracted and criteria2query's results differed in several aspects. We will discuss reasons for missing a concept extraction and then will discuss reasons for incorrectly composing a rule. As a sample, the most common failure was extraction at the wrong ontological level or extraction of inconsistent granularity levels. For instance, "neoplasm" was extracted rather than "neoplasm of the breast", a child concept of the former term. Across all of these errors, it is useful to calculate how many are domain issues versus general vocabulary issues. Because the manual annotation only focused on domain concepts, the "matching concepts" ratio approximates this value. Some notable composition errors arise because a criterion can be encoded in multiple ways. For instance, an ECOG Performance Status <= 1 is the same as an ECOG Performance Status of 0 or 1, since zero is the lowest valid value. We will need to include special logic around evaluation of any rules including numerical components. A less tractable variant of this same problem is for non-numerical ordinal concepts. Tumor stages are often defined as ranges (e.g., T1-T3) but ontologies treat these as equal siblings with no annotation that T0 < T1 < ... < T4.

Acknowledgements: Research supported by R01LM009886-10 and the SmartState Endowment for Translational Biomedical Informatics.

References