**Optimizing Concept Mapping from Free-Text via Multiple Mapping Tools**

**Sal Volpe**

Introduction

Cohort definition is critical to designing randomized controlled trials and comparative effectiveness research studies. It is also a regularly rate-limiting step for researchers and clinicians because it is subject to frequent error and laborious1. In order to combat this issue, in an attempt to automate the extraction and definition process, Critera2Query was created2. The Criteria2Query is an online tool that allows researchers to parse eligibility and exclusion criteria from ClinicalTrials.gov and establish queries that connect tokenized concepts to ATLAS Concepts Sets, which facilitates the creation of standardized cohort definitions. A system like Criteria2Query is important considering the tremendous volume of EHR data, which ideally can be processed hands-free3.

An inevitable drawback, however, to an automated system is the propensity for carried errors. In particular, Criteria2Query lacks in substantially powerful entity normalization. Criteria2Query uses Usagi4, which maps terms and their domains into OMOP CDM format. Usagi does not have a highly reliable score metric for mapping queries to concepts. For example, the query of “neurological disease” maps to “urological disease” with an accuracy score of 0.9. In this situation, there is no “neurological disease” standard concept in LOINC or SNOMED, so the use of a best score leads to an improper linkage when connecting to ATLAS. After combining ClinicalTrials.gov data with criteria entered by testers, Criteria2Query achieved an F1 score of 0.514 for entity normalization, which is the initial step in query formation. This step, therefore, remains unaccounted for by the other normalization steps before the cohort definition is created.

Methods and Innovation

In consultation with Chi Yuan, one of the primary team members in creating Criteria2Query, I propose establishing a data method that looks to other mapping tools, particularly MetaMap, to filter-out semantically inappropriate results in favor better approximates. MetaMap relies on UMLS as its thesaurus to thoroughly parse biomedical text and map tokens to standard concepts5. MetaMap can be time-consuming and inefficient, so it would be best paired with a parallel process, like that of Usagi, which produced fully automatic query formulation in 1.22seconds/criterion2. In addition to sampling code from the MetaMap API, I am also interested in exploring the utility of other mapping methods that have been shown to improve concept normalization and mapping using UMLS. Recurrent neural networks have been shown to improve mapping accuracy when paired with semantic representations of one/multi-word expressions in social media posts6, which can be as haphazard as some clinicaltrials.gov criteria. Further, MetaMap results are in UMLS CUI format, so I would implement a tool to ensure that the results still adhere to the OMOP CDM. On a simpler note, standard *word2vec* and word-embedding algorithms could aid the parallel processing between Usagi and the UMLS-based methods. With all this in mind, I plan to build upon the efficiency of MetaMap and Usagi to improve mapping in the entity normalization phase and thus improve overall accuracy across the board. By alleviating the large discrepancy at this step, I hope to improve usability and reliability of Criteria2Query for autonomous cohort definition tasks in a way that can be scaled beyond this tool and used in other settings, such as with other clinical free-text.

Dataset and I/O

Dataset: UMLS, SNOMED, and ClinicalTrials.gov trials.

Input: Because Criteria2Query was trained on a number of Alzheimer’s Disease trials, I plan on working with two different sets: 100 neurological disease (indiscriminant) trials and 100 trials from randomly picked domains

Output: The output will be in the same as Criteria2Query: presentation of tokens and links to OHDSI concept sets

Evaluation Design

In order to evaluate any improvements that I have made, I would first consider the queries that were used in the initial Yuan et al paper. I will retrain Criteria2Query’s mapping on those trials as a primary training set before testing against ClinicalTrials.gov criteria and then collect a cohort of testers to input data. Then, to determine whether the Alzheimer’s Disease trial training biased Criteria2Query toward NER issues, I will use a random-number generator to gather a gold-standard of clinical trials across domains and repeat the same steps. If the improvements work locally, I will work with Chi Yuan and Cong Liu to implement the changes into the online version of Criteria2Query.

Solo Approach

As my main proficiencies are in Java programming and an undergraduate biomedical background, I plan to receive much of the support required to implement machine learning and statistical reasoning from Chi and Cong. They have a greater knowledge of these fields, especially in regards to Criteria2Query. I will also be consulting MDs in the department, as well as my father (who is a physician and informaticist) to help supplement my clinical weaknesses.

Timeline

* **Week 1 (9/22) – Preliminary Querying**
  + Establish connections with Chi and Cong for support in the project
  + Parse through Usagi implementation in Criteria2Query
  + Parse through MetaMap concept mapping algorithm
* **Week 2 (9/29) – Proposal Defense and Feedback**
  + Take feedback to improve timeline
  + Gather training data sets
  + Contact possible testers
  + Find tool for mapping UMLS CUI to OMOP CDM
* **Weeks 3–4 (10/6–13)**
  + Pinpoint overlay positions for Usagi and MetaMap
  + Evaluate improvements and deficits
  + Consider machine learning methods for training and integration of mapping tools
* **Week 5 (10/20) – Development Partially Completed**
  + Tweak use of Usagi+MetaMap (+ other methods, possible RNN or word embedding)
  + Reproducing Yuan et al. paper
  + Gather new random data set for retraining
* **Weeks 6–8 (10/27–11/10)**
  + Retrain Criteria2Query locally on randomly assembled gold-standard
  + Repeat reproduction of Yuan et al.
  + Implement new methodology for online querying
* **Week 9 (11/20) – Evaluation Completed**
* **Week 10 (11/24) – Start Paper; Prepare Slides**
* **Week 11 (12/1) – Submit Slides for Presentation**
* **Week 12 (12/8) – Presentation; Finishing Paper**
* **Week 13 (12/15) – Submit Final Paper**

**References**

1. Gulden C, Landerer I, Nassirian A, Altun FB, Andrae J. Extraction and Prevalence of Structured Data Elements in Free-Text Clinical Trial Eligibility Criteria. InEFMI-STC 2019 (pp. 226-230).
2. Yuan C, Ryan PB, Ta C, Guo Y, Li Z, Hardin J, Makadia R, Jin P, Shang N, Kang T, Weng C. Criteria2Query: a natural language interface to clinical databases for cohort definition. Journal of the American Medical Informatics Association. 2019 Feb 7;26(4):294-305.
3. Sun H, Depraetere K, De Roo J, Mels G, De Vloed B, Twagirumukiza M, Colaert D. Semantic processing of EHR data for clinical research. Journal of biomedical informatics. 2015 Dec 1;58:247-59.
4. Usagi. [https://github.com/OHDSI/Usagi](https://github.com/OHDSI/Usagi" \t "). Accessed September 22, 2019.
5. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association. 2010 May 1;17(3):229-36.
6. Tutubalina E, Miftahutdinov Z, Nikolenko S, Malykh V. Medical concept normalization in social media posts with recurrent neural networks. Journal of biomedical informatics. 2018 Aug 1;84:93-102.
7. Lu CJ, Tormey D, McCreedy L, Browne AC. Enhanced LexSynonym Acquisition for Effective UMLS Concept Mapping. InMedInfo 2017 Aug 24 (pp. 501-505).
8. Luo YF, Sun W, Rumshisky A. A Hybrid Normalization Method for Medical Concepts in Clinical Narrative using Semantic Matching. AMIA Jt Summits Transl Sci Proc. 2019;2019:732–740. Published 2019 May 6.1.