Title: LexMapr - A rule-based biomedical text-mining tool for entity recognition and ontology-driven classification

Authors: Gurinder Gosal, Emma Griffiths, Damion Dooley, Ivan Gill, William Hsiao.

Abstract: LexMapr is a rule-based text-mining tool for entity recognition that has been developed in 2018 to extract biosample entities from short free-text phrases and map these to standard ontology terms. The tool was initially motivated to fulfill biosample metadata harmonization objectives of public health surveillance networks like the US FDA's GenomeTrakr system and the US National Antimicrobial Resistance Monitoring System (NARMS). Their objective is to harmonize short phrases describing food pathogen source data using standard vocabularies for reporting of transmission dynamics in public health foodborne pathogen surveillance and investigation realms. The initial focus of LexMapr development has been on providing a text-mining tool to clean up the short free-text biosample metadata that contained a lot many inconsistent punctuation, abbreviations and typos, and map the identified entities to standard terms from ontologies. Two key food biosample domain ontologies, FoodOn and GenEpiO that cover clinical, epidemiological and food semantics, have been selected as the target ontologies for standardizing biosamples. These ontologies also include a subset of relevant terms and relationships considered vital for the breadth of biosample domain borrowed from a few other key ontologies for anatomy, environment and taxonomy such as UBERON, ENVO and NCBITaxon.

Because the problem space of short biosample phrases has a very focused semantic domain of text and very specific challenges to deal with, we have employed a **rule-based approach** that draws upon wide-ranging lexical resources. LexMapr combines basic lexicographic transformation with light Natural Language Processing (NLP), synonymy, ontology and other resource lookup to produce a tokenized equivalent description suitable for keyword and ontology-driven search of biosample database contents. LexMapr pipeline addresses many challenges in the processing of biosample phrases such as grammatical incorrectness, misspellings, plurality, abbreviations and acronyms, synonymy, non-English usage, term context ambiguity, detection of overlapping entity boundaries and missing ontology vocabulary. LexMapr implements different rules for pre-processing, normalization, entity recognition and ontology term mapping tasks and makes use of domain-specific lookup-tables for abbreviation and acronyms normalization, non-English food terms translation and spelling correction that have been created locally for the biosample domain. To demonstrate and test the entity linking system of LexMapr amongst the biomedical ontology community, we conducted a workshop at the International Conference on Biological Ontology 2018.

Once the primary task of linking free text to standard ontology terms has been accomplished, LexMapr provides a mapping and reporting platform for many potential ontology-driven applications. LexMapr has functionality to classify biosamples as per institution-specific classification schemes. LexMapr performs its ontology-driven classification of biosample metadata provided by GenomeTrakr and NARMS using the IFSAC epidemiology-focused food classification scheme for categorizing foods implicated in outbreaks. LexMapr uses predefined nodes of ontologies as buckets (containers) to characterise specific third-party classes. The LexMapr pipeline classification component provides functionality for biosamples to be linked to these buckets (ontology ids) and hence to be categorized according to IFSAC classes. To support specific requirements of IFSAC classification schema, LexMapr uses many classification rules to further refine the preliminary classification results. For example, a post-refinement rule classifies an input phrase as "Multi-ingredient" (an IFSAC class) in case it contains more than one food ingredient combined together.

Applied to real-world GenomeTrakr and NARMS biosample metadata, LexMapr exposed and reported the incompleteness of the existing scheme in describing a variety of biosamples. Subsequent deliberations with GenomeTrakr and NARMS has helped to come up with an enhanced and improved classification scheme "IFSAC+" to effectively represent the biosamples. LexMapr has enabled the reorganization of classes and introduction of many new classes such as dietary supplement, engineered seafood, flavoring or seasoning, animal feed, veterinary clinical/research and others in the schema. The US FDA is implementing FoodOn facets in a new Minimum Information about any (x) Sequence) Food Environmental Metadata Standard (MIxS) extension for food related data using LexMapr as the enabling tool. LexMapr has been recently equipped with certain additional functionalities to cater specific user needs such as embedding scientific names of food source plant and animal organisms, reporting the normalizations (e.g. from non-English to English food names) to provide more visibility in the linking and classification process. LexMapr has also helped in reporting many new candidate terms as potential terms for curation and inclusion in the ontologies and has helped ontologies, especially FoodOn to add new terms and their synonyms in order to capture GenomeTrakr food descriptions.

Although LexMapr was initially developed to serve the biosample domain, we believe that our general approach of cleaning and harmonization of data can be used to address other content domains by adding selected domain specific ontologies and rules. LexMapr recently has been configured to allow users to select a set of ontologies and to substitute their own set of lexical resources (default lexical resources are more pitched towards biosample domain), if any, to support wider coverage and better accuracy. Using this approach, there is an ongoing initiative to

use generalized LexMapr to link COVID-19 metadata to a selected set of ontologies covering COVID-19 domain. Work is also in progress to equip LexMapr with a mechanism for performing ontology-driven classification configured to any sort of institution-specific classification schemas provided by the user. The LexMapr source code is publicly available at https://github.com/Public-Health-Bioinformatics/LexMapr. LexMapr is available both as a locally installable command-line tool and via a Django-based website providing a simple graphical interface (http://watson.bccdc.med.ubc.ca:8000/lexmapr/) that is being enhanced in usability and functionality.