Final Report –2015-2016
Medical Informatics Postdoctoral Research Fellowship

Satyajeet Raje
satyajeet.raje@gmail.com

Lister Hill National Center for Biomedical Communications
U.S. National Library of Medicine, NIH
Mentor: Dr. Olivier Bodenreider

SUMMARY

This report provides a summary for the various projects I undertook during my one year fellowship. My primary research activity was with the Medical Ontology Group (MOR) under the guidance of Dr. Bodenreider. The research with MOR focused on interoperability of biomedical terminologies and ontologies, evaluating the semantic aspects, such as coverage, completeness and alignment. Specifically, I worked on two research projects –

1. Investigating the coverage of disease concepts across SNOMED CT and the Disease Ontology (DO).
2. Leveraging lexical features of concept names using Description Logics (DL) to identify potential missing hierarchical relations in SNOMED CT.

Apart from the projects with MOR, I also worked on a project with the Library Operations and NCBI divisions of the National Library of Medicine. The goal was to develop a method to harmonize user-defined phenotypic variables in dbGaP to promote data discoverability and reuse of the datasets. In addition, I participated as a team-lead for two NCBI hackathons during my fellowship, which resulted in two more, smaller projects. These projects focused on enabling or showcasing the reuse of publically available biomedical datasets, specifically from NLM and NIH.

INVESTIGATING COVERAGE OF DISEASE CONCEPTS ACROSS SNOMED CT AND DISEASE ONTOLOGY (DO)

Different ontologies are used to represent disease concepts in biomedical research and in clinical settings. The Disease Ontology (DO) is part of the Open Biomedical Ontologies (OBO) collection and is widely used in the research community, especially in genomic and cancer research domains. The Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) is the largest clinical terminology with over 300,000 clinical concepts, of which about 100,000 are disease concepts. In contrast to DO, SNOMED CT is primarily used in healthcare and clinical settings. Interoperability between these two important ontologies is critical for translational applications in biomedicine.

In this study, we investigate the coverage of disease concepts between DO and SNOMED CT. More specifically, we first establish a reference list of mappings of DO concepts to SNOMED CT. Next, we identify and characterize the concepts present in DO but not covered by SNOMED CT. We also analyze the differences in hierarchical structure between the two ontologies based on the mapped concepts.

We found that, overall, 4478 (65%) the 6931 DO concepts are mapped to SNOMED CT. The cancer and neoplasm subtrees of DO account for many of the unmapped concepts. Unmapped concepts usually form subtrees, and less often correspond to isolated leaf or intermediary concepts. The most frequent differentiae in unmapped concepts include morphology (for cancers and neoplasms), specific subtypes (for rare genetic disorders), and anatomical subtypes. After comparing the hierarchical organization, we

found that only about 30% of the hierarchical relations in DO and SNOMED CT are semantically consistent between the two ontologies.

The study provides a detailed analysis of the gaps in coverage and structural differences between DO and SNOMED CT contributes to the interoperability between these two ontologies and will guide further validation. It reveals critical differences in hierarchical organization and concept orientation (i.e., whether two concepts correspond to the same entity) in the two ontologies.

I presented this work as a podium abstract at the AMIA Joint Summits on Translational Science, 2017. A full paper has been selected to appear in the proceedings of the 16th World Congress on Medical and Health Informatics (MedInfo), 2017.

LEVERAGING LEXICAL FEATURES OF CONCEPT NAMES USING DESCRIPTION LOGICS (DL)

The quality assurance of large bio-ontologies is extremely critical for their effective and continued use and is an active area of research. Previous work by Dr. Bodenreider established a method to identify potentially missing hierarchical relations leveraging lexical features in SNOMED CT. It used the preferred term for each concept in SNOMED CT to create logical definitions for concepts. These definitions were used to identify potential missing hierarchical relations. The method was tested on a subset of SNOMED CT concepts, namely the namely the *Disorder of head (disorder)* and *Operative procedure on head (procedure)* hierarchies.

For this project, I expanded on the original method adding new lexical features as follows – 1) adding lexico-syntactic constraints for the concept labels based on shallow parsing to increase precision; 2) processing all synonyms in order to increase recall. I also applied the expanded method to almost all the top level sub-hierarchies of SNOMED CT. Some hierarchies were not processed as they would not be amenable to such lexical approaches (e.g.: *Pharmaceutical / biologic product (product)* hierarchy) or they did not have biomedical concepts (e.g.: “Special concept” hierarchy).

The methodology can be summarized as follows. We create logical definitions for concepts, using DL, leveraging all concept labels (preferred terms and synonyms) provided by SNOMED CT and their lexico-syntactic analysis. We use the ELK semantic reasoner to automatically infer hierarchical relations based on the logical definitions of the concepts. We filter out those already present in the original hierarchy of SNOMED CT. Finally, we validate our method by reviewing the remaining set of relations as being potentially missing relations.

The novel aspect of this work is to use a DL approach to lexical similarity. In practice, it means that no ad hoc programming is required for identifying partial ordering relations among sets of words for terms in an ontology reflecting hierarchical relations among the corresponding concepts. Instead, logical definitions created from lexical features can simply be represented in DL formalism and run through a reasoner to infer the relevant hierarchical relations. This work is a contribution to quality assurance in SNOMED CT. Importantly, the missing hierarchical relations identified by our methods may only indicate underlying issues in SNOMED CT’s concept definitions, which will require domain expertise to address.

A preliminary study comparing the effects of adding synonyms and lexico-syntactic constraints on the identification of potentially missing hierarchical relations has been submitted to the proceedings of AMIA Annual Symposium 2017 and currently under review. We are awaiting results from on-going evaluation and plan to submit a full paper to a suitable journal soon.

HARMONIZING USER-DEFINED PHENOTYPIC VARIABLES TO IMPROVE DATA DISCOVERABILITY IN DBGAP

The NLM Database of Genotypes and Phenotypes (dbGaP) provides a rich source of studies with genotypes, phenotypes and associations between them. As part of the submission process, dbGaP requires researchers to submit a data dictionary of study variables including their short descriptions. It is critical to harmonize these user submitted variables to allow researchers to effectively find studies that share the same or similar variables in dbGaP. A recent study documents a manual effort to map variables in dbGaP to a standardized set of measures in the PhenX toolkit. However, the rapidly expanding size of dbGaP variable dictionary (200,000+ variables by early 2017) renders sustaining such manual post-coordinated efforts infeasible. In this project, I attempt to (semi-)automate the harmonization of variables in dbGaP using the user-submitted variables.

We reuse the data from the study mentioned above for this work. It provides PhenX IDs (by RTI International) and LOINC codes (by NCBI staff) for 20635 variables in dbGaP. We use Latent Semantic Analysis (LSA) to compute a pairwise similarity score between these variables based on the user-provided variable descriptions. We then cluster together variables based on the calculated similarity score (threshold of 0.8 was empirically chosen). Thus, each variable has a list of "duplicates" identified by LSA. To validate the output of the LSA method, three evaluators manually reviewed 10% (~2000) randomly sampled duplicate variables independently. We also compare the results of our LSA based method with the existing PhenX and LOINC mappings.

Manual evaluation showed that the LSA method can identify "duplicate" variables with very high accuracy (96% precision when all 3 evaluators agree that a result is valid). The inter-rater agreement among evaluators was also high (kappa = 0.7). However, we observed poor conformance between the LSA derived clusters with existing PhenX ids and LOINC codes for some concepts. This means that the LSA is able to cluster variables that were not originally clustered by the LOINC or PhenX IDs. Further analysis revealed this is mainly due to discrepancies between semantic and syntactic characteristics. For instance, implicit synonymy among terms in a terminology (e.g. "sex" and "gender" have same LOINC code) cannot be clustered by LSA. Further refinement would be needed to account for such conditions.

This work has been submitted as a poster to the AMIA Annual Symposium 2017. We are currently finishing the full paper and expect to submit soon.

MESHGRAM: AN OPEN SOURCE TOOL TO VISUALLY BROWSE CO-OCCURRENCE OF MESH TERMS IN PUBMED

MeSHgram was developed as part of the January NCBI Hackathon. It is a tool for convenient, visual and interactive exploration of the co-occurrence of MeSH terms over the entire PubMed corpus. The tool can assist in the quantification of known research patterns as well as potentially aiding novel hypotheses generation.

We used the NLM PubMed XML corpus. As of Jan 2017, the corpus contained approximately 24.5 million publications from 1809 to 2016. Publications indexed in PubMed have human curated Medical Subject Heading (MeSH) terms associated with them. We leveraged these MeSH terms in MeSHgram. We parsed the entire PubMed corpus and extracted the ID, year of publication and MeSH terms associated with each document. We excluded duplicate items such as revision entries and those with no MeSH terms, resulting in approximately 23 million publications in our database

The tool allows searching with multiple MeSH terms via an auto-complete search box. It displays article counts by year for the searched terms along with the number of co-occurrences. It also displays a word cloud for other MeSH terms that appear alongside the searched MeSH terms to provide contextual

visualization of other associated MeSH terms. Visually selecting specific year ranges in the graph updates the graph and the word cloud.

This work was submitted as a poster to the AMIA Annual Symposium 2017. The source code is open and available on GitHub under MIT license.

EXTENDING TCGA QUERIES TO AUTOMATICALLY IDENTIFY ANALOGOUS GENOMIC DATA FROM DBGAP

A major obstacle to data discovery is the disconnectedness of various data sharing resources. Automated tools that can connect these databases and reduce the time that researchers spend on data discovery are critically needed. Such tools will promote reproducibility, increase the efficiency of research, and aid in solving the problem of small sample sizes. These issues are especially relevant to genomic data, which is typically expensive to gather.

Here, we focus on connecting two popular genomic data repositories, the Database of Phenotypes and Genotypes (dbGaP) and The Cancer Genome Atlas (TCGA), hosted by the Genomic Data Commons (GDC). TCGA is a popular resource for individual-level genotype-phenotype cancer related data. DbGaP contains many datasets similar to those in TCGA. In this project, we developed a software toolkit that will allow researchers to discover relevant genomic data from dbGaP, based on matching TCGA metadata. The resulting research provides an easy to use tool to connect these two data sources.

We developed an easy-to-use tool that can be used to find additional data from dbGaP (and SRA) by expanding TCGA queries automatically. The first part of the pipeline allows researchers to query either repository by TCGA Project ID, File ID, Case ID, disease type, or experimental strategy via a metadata mapping dictionary. It returns not only a list of TCGA IDs, but also a list of related dbGaP study IDs. For dbGaP studies with NCBI SRA data, the second part of the pipeline will return the .sam files that contains reads aligned to a genomic region of interest to be used with the SRA Toolkit.

To our knowledge, this is the first easy-to-use tool for harmonizing TCGA and dbGaP study metadata for the purpose of data discovery and consolidated querying. This work was also part of a NCBI hackathon (August 2016). An article describing the project has been published in the F1000 open research publication platform. The source code is open and available on GitHub under MIT license.

PUBLICATIONS

Raje, S., Bodenreider, O. (2017). Investigating the Coverage of Diseases across Biomedical Research and Clinical Ontologies. *[abstract] Proc. of AMIA Summits on Translational Science Proceedings.*

Raje S., Bodenreider O. (2017) Interoperability of Disease Concepts in Clinical and Research Ontologies – Contrasting Coverage and Structure in the Disease Ontology and SNOMED CT. *Studies in health technology and informatics (Proc. of MedInfo 2017) (accepted)*

Raje S, Bodenreider O. (2017). Identifying potentially missing hierarchical relations in SNOMED CT based on lexical features – Impact of synonyms and lexico-syntactic constraints. *[abstract] AMIA Annu Symp Proc 2017 (submitted)*

Amos L, **Raje S**, Kimura M, et. al. (2017) Harmonizing User-defined Phenotypic Variables using Latent Semantic Analysis (LSA) to Improve Data Discoverability in dbGaP *[poster] AMIA Annu Symp Proc 2017 (submitted)*

Raje S, Bhupatiraju RT, Hosny A, Busby B. (2017) Investigating the coverage of diseases across biomedical research and clinical ontologies. *[poster] AMIA Annu Symp Proc 2017 (submitted)*

Wagner E. K., **Raje S.**, Amos L., Kurata J., Badve A., Li Y. & Busby B. (2017). Extending TCGA queries to automatically identify analogous genomic data from dbGaP [version 1; referees: awaiting peer review]. *F1000Research* 2017, **6**:319 (doi: 10.12688/f1000research.9837.1)