A Hybrid Natural Language Approach to Manage Semantic Interoperability for Public Health Analytics

Maxime Lavigne, Arash Shaban-Nejad, Anya Okhmatovskaia, Luke Mondor, David L. Buckeridge

McGill Clinical & Health Informatics, Department of Epidemiology and Biostatistics, McGill University, Montreal, Quebec, H3A 1A3, Canada
maxime.lavigne@mail.mcgill.ca
(arash.shaban-nejad|anya.okhmatovskaia|luke.mondor|david.buckeridge)@mcgill.ca

Abstract. This paper discusses the integration of an ontology with a natural language query engine to calculate and interpret epidemiological indicators for population health assessment. In this paper, we discuss the application of this approach to one type of possible query, which retrieves health determinants, causally associated with diabetes mellitus.

Keywords: Ontology, Natural language interface, Causal inference, Epidemiology

1 Introduction

The Population Health Record (PopHR) platform [1][2] aims to improve population health decision-making. It calculates and presents measures or indicators of health determinants and health outcomes in a manner that, unlike most current web portals, is intuitive to access and provides up-to-date indicators that are contextualized by public health knowledge. In this paper, we describe our approach to querying the PopHR knowledge base using an natural language interface (NLI).

Early in its development, it became apparent that even though we were restraining the language of recognized queries, the breadth of pre- and post-conditions made implementation difficult. We therefore partitioned the space of possible queries and called these, query types. By partitioning intents of user inputs into collectively exhaustive and mutually exclusive query types, we were able to overcome the difficulty of designing a single data processing pathway for all queries. Linking a query's concepts with our domain ontology is simplified and it allows us, for example, to disambiguate concepts, which could have different interpretation in different query types.Partitioning restricts the software contract of our system when processing a query. Finally, this approach allows us to make assumptions about the domains of concepts, such as statistics and geography, which are relevant in our context.

In this paper, we use a representative query as an example: *What determinants increase the risk of diabetes?* The following sections introduce the relevant parts of our domain ontology and then describe the strategy used to answer the query.

2 Ontological Representation

PopHR uses its own domain ontology representing knowledge relevant to population health, including a taxonomy of human diseases, various groups of health determinants and public health interventions, measures of disease occurrence and other epidemiological concepts. In addition to the hierarchy of concepts, the ontology encodes associative relations to allow for meaningful inference. One specific type of associative relation represents a causal link between two entities (i.e., cause and effect). For example, body mass index (BMI) has a positive effect on an individual's disposition towards developing type 2 diabetes mellitus (see Figure 1). More generally, this relationship is an example of a probabilistic causal link from a health determinant to a process of developing a disease. We can also describe a causal relation between a health determinant and a process that modifies another health determinant.

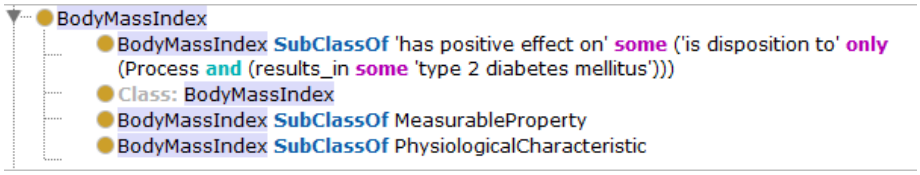

Fig. 1. Example of encoding BMI in the ontology, as seen with the Protégé editor.

3 Processing Pipeline

In PopHR, the natural language interface is the preferred method for querying information. All queries must respect a proper subset of the English language that is formally defined to be context free. The subset is built around question answering and was conceived with the intent to provide all the expressivity needed. This design decision implied that we needed an intuitive, consistent user experience. For the system to succeed at providing proper guidance, it needed to suit both the needs of the inexperienced users and experts.

3.1 Lexical and Syntactic Analysis

We used our formally defined grammar in conjunction with the ANTLR framework[3] for langage processing. The first step of the process is to break the input down into lexemes. The token stream produced by this step from the example input is:

```
QUESTION, ID, VERB, ARTICLE, ID, QUALIFIER_START, ID, QUESTION_MARK
```

Once the individual components of the question are separated, an LL(*) parser uses production rules to generate a syntactic tree. If the creation of such a tree is impossible, then we know that the input text was not part of our language and proper guidance will be given on how to correct the issue. This syntactic tree (Figure 2) is the artifact that will be used by the rest of the system.

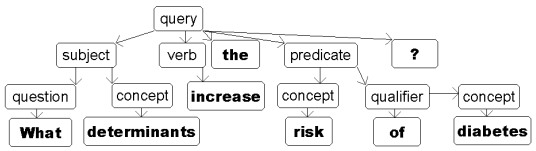

Fig. 2. Syntactic Tree Produced by our Example

A formal representation of the question is a necessary but not a sufficient step to understand the intent of the user. Although it is trivial for a human, performing this step programmatically requires the ability to match the query to some known patterns. This role is played by the oracle: all known patterns are manually entered in the system and take the general form: *What (To Be) ID? is a description query*. With this mapping, we are making the assumption that a question that starts with the question word *What* and uses a derivate of the verb *To Be* that has a final concept *ID* is asking the system for a description of this concept. Applying the Oracle to our example would classify it as a *CausalityEnumerationQueryWithConcept*. We can intuitively concur that we did want an enumeration of all determinants that have some causal relationship to diabetes mellitus.

3.2 Semantics

At this point in the process, we have gathered information regarding the domain and general intent of the query. Nevertheless, we still have no information on which concepts are used and what they mean. It is at this point that we query the ontology for concept such as determinants, increase, risk, and diabetes. Fetching these concepts by their textual representation, searching labels, synonyms and other annotations, we obtain the following:

Table 1. Association Between Query Terms and Ontology

Input	Concept or Relation in Ontology
determinants	health determinants
increase	'has positive effect on' some _
risk	'is disposition of' only (Process and 'results in' only _)
diabetes	diabetes mellitus

It is noteworthy that misspelings, difference in case and such are handled outside of the ontology. We then check for special markings that define processing triggers to activate. In our example, *'has positive effect on'* requires transitively walking upstream to identify additional causal factors. At this point, all of the information needed to understand what the user requested has been gathered. We would then reformulate the question into a format that can be answered by a description logic (DL) reasoner, such as Fact++[4]. From our example, we need *SubClassOf+* of *'Health Determinants'* that are described by: *" 'has positive effect on' some ('is disposition to' only (results_in some 'diabetes mellitus') "*

The results, will be a list of health determinants that directly influence the risk of the event in a positive way. From our processing trigger associated with *'has positive effect on'*, we know that the answer should also include any health determinants that positively affects health determinant having a direct influence on the risk event. We know, for example, that BMI is one of those direct factors and that it is a measurable property. Therefore, we look for other health determinants that have a positive effect on a disposition to increase the level of BMI. If the result is a measurable property we would repeat the same step.

4 Discussion

Developing a system that is accessible via a natural language interface is challenging. To address this challenge, we make use of all the contextual information we can learn about the intent of the query, we restrict ourselves to a proper context-free subset of the English language, and we use a domain ontology. The resulting system gives useful and correct answers to practical questions. We are looking forward validating our solution in user testing. It will enable us to broaden our scope from a prototype state to that of day to day use.

Acknowledgments

The Canadian Foundation for Innovation (CFI) and the Canadian Institutes of Health Research (CIHR) provide funding for this research.

References

1. Buckeridge DL, Izadi MT, Shaban-Nejad A, Mondor L, Jauvin C, Dubé L, Jang Y, and Tamblyn R. An infrastructure for real-time population health assessment and monitoring. IBM Journal of Research and Development, 2012, 56(5): 2
2. Izadi M, Shaban-Nejad A, Okhmatovskaia A, Mondor L, Buckeridge DL (2013). Population Health Record: An Informatics Infrastructure for Management, Integration, and Analysis of Large Scale Population Health Data. In Proc. of AAAI (HIAI 2013), Belleveue, Washington, USA July 14-18.
3. Parr T., Fisher KS, LL(*): The Foundation of the ANTLR Parser Generator, PLDI 2011
4. Tsarkov, D, and Horrocks, I. FaCT++ description logic reasoner: system description. In proc. of IJCAR 2006, Springer, pp. 292-297.