Artificial Intelligence in Medicine 65 (2015) 35–48

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Approaching the axiomatic enrichment of the Gene Ontology from a lexical perspective

Manuel Quesada-Martínez [a,*], Eleni Mikroyannidi [b], Jesualdo Tomás Fernández-Breis [a], Robert Stevens [b]

[a] Facultad de Informática, Campus de Espinardo, Universidad de Murcia, 30100 Murcia, Spain
[b] School of Computer Science, The University of Manchester, Oxford Road, Manchester M13 9PL, United Kingdom

ARTICLE INFO

Keywords:
Ontology engineering
Axiomatic enrichment
Biomedical ontologies
Gene Ontology

ABSTRACT

Objective: The main goal of this work is to measure how lexical regularities in biomedical ontology labels can be used for the automatic creation of formal relationships between classes, and to evaluate the results of applying our approach to the Gene Ontology (GO).
Methods: In recent years, we have developed a method for the lexical analysis of regularities in biomedical ontology labels, and we showed that the labels can present a high degree of regularity. In this work, we extend our method with a cross-products extension (CPE) metric, which estimates the potential interest of a specific regularity for axiomatic enrichment in the lexical analysis, using information on exact matches in external ontologies. The GO consortium recently enriched the GO by using so-called cross-product extensions. Cross-products are generated by establishing axioms that relate a given GO class with classes from the GO or other biomedical ontologies. We apply our method to the GO and study how its lexical analysis can identify and reconstruct the cross-products that are defined by the GO consortium.
Results: The label of the classes of the GO are highly regular in lexical terms, and the exact matches with labels of external ontologies affect 80% of the GO classes. The CPE metric reveals that 31.48% of the classes that exhibit regularities have fragments that are classes into two external ontologies that are selected for our experiment, namely, the Cell Ontology and the Chemical Entities of Biological Interest ontology, and 18.90% of them are fully decomposable into smaller parts. Our results show that the CPE metric permits our method to detect GO cross-product extensions with a mean recall of 62% and a mean precision of 28%. The study is completed with an analysis of false positives to explain this precision value.
Conclusions: We think that our results support the claim that our lexical approach can contribute to the axiomatic enrichment of biomedical ontologies and that it can provide new insights into the engineering of biomedical ontologies.

1. Introduction

Many biomedical ontologies have been developed in recent years, and their development has been stimulated by their increasing importance in the scientific community [1]. An indicator of such an increasing importance is that ontologies are considered to be a key technology for semantic interoperability in healthcare; see the semantic health project [2] and SemanticHealthNet[1] for examples. According to [3], an *ontology* is a set of *logical axioms* that are designed to account for the intended meaning of a vocabulary; in other words, it is a representation that captures the categories of *objects* in a field of interest and the relationships that those objects have to each other in such a way that it is possible to recognise category membership. For example, the Gene Ontology (GO) [4] has the aim of standardising the representation of gene product attributes across species and thus across databases. The *objects* of an ontology encompass different *components*, such as classes, individuals and object properties/relationships [5]. For human readability, ontology

* Corresponding author.
E-mail addresses: manuel.quesada@um.es (M. Quesada-Martínez), eleni.mikroyannidi@manchester.ac.uk (E. Mikroyannidi), jfernand@um.es (J.T. Fernández-Breis), robert.stevens@manchester.ac.uk (R. Stevens).

[1] http://www.semantichealthnet.eu accessed September 2014.

http://dx.doi.org/10.1016/j.artmed.2014.09.003
0933-3657/

authors include strings of characters as labels that describe an ontology component. However, machines need *logical axioms* that are expressed in a *formal language* with which to reason.

The Open Biomedical Ontologies (OBO) Foundry [6] contributes to the development of an orthogonal collection of biomedical ontologies and defines criteria[2] to be followed by biomedical ontology authors who contribute to the OBO Foundry. Ideally, different contributors would model an *ontology* that focuses on a specific sub-domain, but they would re-use *components* from other ontologies, where appropriate. However, the high level of activity in biomedical ontologies [7] makes reaching this goal a complex task. Moreover, the largest repository of biomedical ontologies is the National Center for Biomedical Ontology's BioPortal [1], which has 372 ontologies at the time of this writing.

According to [8], the labels in biomedical ontologies can embed a meaning that is not always represented as *logical axioms* in the ontology. Such hidden semantics constitute not only implicit references to *components* within an ontology but also implicit references to other ontologies. For example, the GO class '*oocyte differentiation*' is a type of '*cell differentiation*' that implicitly references the class '*oocyte*' from the Cell Ontology [9]. The goal of axiomatic enrichment is to make explicit such implicit relationships.

The enrichment of ontologies should establish new formal relationships between existing ontologies, increasing the potential and usefulness of the biomedical applications that are supported by such ontologies [10]. In recent years, different approaches have been proposed within this research area:

- Reference [11] defined the "lexically suggested logical closure" metric for medical terminology maturity. This metric was based on the evaluation of relationships that were proposed by lexical processing programs.
- The Gene Ontology Next Generation project aimed to provide a method for the migration of biological ontologies to *formal languages* such as the Web Ontology Language (OWL) and to explore issues that are related to the maintenance of large biological ontologies [12,13].
- The Open Bio-Ontology Language (OBOL) project [14] generated formal relationships for existing OBO ontologies using reverse engineering. Later, reference [15] described a frame-based integration of the GO and two other ontologies for improving the *logical axioms* between classes of biological concepts.
- Additionally, [16] proposed a method for the enrichment of ontologies by defining ontology design patterns [17] and their corresponding implementation in the Ontology Pre-Processor Language.[3]
- Reference [18] addressed the normalisation of GO by explicitly stating the labels of the compositional classes and partitioning them into mutually exclusive cross-product sets; they used a combination of OBOL and manual curation to generate *logical axioms*, which they called logical definitions, for selected parts of GO.
- Reference [19] detected hidden semantics, which were named underspecification, in classes from the Systematised Nomenclature of Medicine (SNOMED) that were without *logical axioms*; the authors used natural language processing, which associated each class with a set of equivalence classes that grouped lexical variants (based on their labels), synonyms and translations.
- Reference [10] represented the Foundational Model of Anatomy ontology in OWL2, exploiting the naming conventions in its labels to make explicit some hidden semantics. For example, the pattern *A_of_B* was used to enrich the class '*Lobe_of_Lung*'. In most cases, the name *A of B* is a contraction that is formed from *A* and *B* that omits some *logical axiom p* that relates the two entities, *A* and *B*. The missing *p* was recovered from scanning the list of property restrictions that are attached to the *class*. For example, '*regional_part_of*' is the *p* for '*Lobe_of_Lung*'.

Our approach [16] used a manual analysis of lexical regularities; the results were used for detecting linguistic patterns from a GO sub-hierarchy such as the following: (1) '*X binding*': the selective, non-covalent, often stoichiometric interaction of a molecule with one or more specific sites on another molecule; or (2) '*translation X factor activity*': any molecular function that is involved in the initiation, activation, perpetuation, repression or termination of polypeptide synthesis at the ribosome. These linguistic patterns inspired the core concept of this work. In the previous examples, the lexical regularities are the fixed part of the patterns (e.g., *binding*, *translation* or *factor activity*). Another example of a lexical regularity is '*negative regulation*', which in general stands for the prevention or reduction of a biological process. This linguistic expression appears in several biomedical ontologies, but it is not usually represented with *logical axioms*. The '*negative regulation of transcription*' and the '*negative regulation of translation*' in the Gene Regulation Ontology or the '*negative regulation*' in the Phenotypic Quality Ontology are similar examples.

Our initial hypothesis was that classes that exhibit lexical regularity encode the meaning of a domain object, and there should be a relation between this class and other classes that exhibit that regularity. In previous work, our method demonstrated its ability to retrieve a large set of classes that exhibited regularities, but not all of them are relevant for enriching the ontology. Hence, we identified the need for methods that select which sets are relevant for such a purpose. Therefore, in this paper, we extend our method with a new metric that analyses the relation between the lexical regularities exhibited by the labels of the classes and the labels of the classes that are defined in the ontologies and used for enrichment, namely, the cross product extension (CPE). This metric can be understood to be an estimation of the enrichment of those classes that exhibit such regularities. Moreover, we propose three different conditions of the CPE metric that define different types of matches. Our hypothesis here is that such conditions provide information about the degree and type of enrichment that can be expected. For example, in GO, if '*translation*' is a lexical regularity that can be generalised as the pattern '*X translation*', then the usefulness of the pattern can be estimated by the percentage of classes that exhibit the regularity and that are decomposable as cross-products.

Here, we focus on the GO for several reasons. First, the GO provides a controlled vocabulary for the functional annotation of gene products. To date, GO classes have been used to produce millions of annotations, which are available in resources such as the GO annotation database [20]. Its enrichment would have an impact on the exploitation possibilities of the GO. Such enrichment would enable machines to not only exploit the GO labels but also manage and exploit more fine-grained *objects*, such as biochemical substances or links between molecular functions, biological processes and cellular components. Consequently, enrichment provides additional dimensions for analysis, in this case, functional biomedical data. Second, our analysis of BioPortal ontologies revealed the *prima facie* suitability of the GO for its enrichment: 100% of the classes have labels, 92% of the words of the labels are repeated, and 85% of the ontology labels exhibited 67 lexical regularities [21]. Finally, the GO consortium and other scientists have already identified the necessity of increasing the axiomatic richness of GO, and they have recently developed a partially enriched version, the GO cross-product extensions [18]. In this work, we will compare our results with these GO cross-product extensions. Although each

[2] http://obofoundry.org/crit.shtml accessed September 2014.
[3] http://oppl2.sourceforge.net/ accessed September 2014.