Journal of Biomedical Informatics 84 (2018) 59–74

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

From lexical regularities to axiomatic patterns for the quality assurance of biomedical terminologies and ontologies

Philip van Damme[a], Manuel Quesada-Martínez[b,c], Ronald Cornet[a], Jesualdo Tomás Fernández-Breis[b,*]

[a] Department of Medical Informatics, Amsterdam Public Health research institute, Academic Medical Center, University of Amsterdam, The Netherlands
[b] Departamento de Informática y Sistemas, Universidad de Murcia, IMIB-Arrixaca, Murcia, Spain
[c] Center of Operations Research (CIO), University Miguel Hernandez of Elche (UMH), Spain

ARTICLE INFO

Keywords:
Ontology quality assurance
Lexical regularities
Axiomatic patterns
SNOMED CT

ABSTRACT

Ontologies and terminologies have been identified as key resources for the achievement of semantic interoperability in biomedical domains. The development of ontologies is performed as a joint work by domain experts and knowledge engineers. The maintenance and auditing of these resources is also the responsibility of such experts, and this is usually a time-consuming, mostly manual task. Manual auditing is impractical and ineffective for most biomedical ontologies, especially for larger ones. An example is SNOMED CT, a key resource in many countries for codifying medical information. SNOMED CT contains more than 300 000 concepts. Consequently its auditing requires the support of automatic methods. Many biomedical ontologies contain natural language content for humans and logical axioms for machines. The 'lexically suggest, logically define' principle means that there should be a relation between what is expressed in natural language and as logical axioms, and that such a relation should be useful for auditing and quality assurance. Besides, the meaning of this principle is that the natural language content for humans could be used to generate the logical axioms for the machines. In this work, we propose a method that combines lexical analysis and clustering techniques to (1) identify regularities in the natural language content of ontologies; (2) cluster, by similarity, labels exhibiting a regularity; (3) extract relevant information from those clusters; and (4) propose logical axioms for each cluster with the support of axiom templates. These logical axioms can then be evaluated with the existing axioms in the ontology to check their correctness and completeness, which are two fundamental objectives in auditing and quality assurance. In this paper, we describe the application of the method to two SNOMED CT modules, a 'congenital' module, obtained using concepts exhibiting the attribute `Occurrence - Congenital`, and a 'chronic' module, using concepts exhibiting the attribute `Clinical course - Chronic`. We obtained a precision and a recall of respectively 75% and 28% for the 'congenital' module, and 64% and 40% for the 'chronic' one. We consider these results to be promising, so our method can contribute to the support of content editors by using automatic methods for assuring the quality of biomedical ontologies and terminologies.

1. Introduction

In recent years, biomedical ontologies and terminologies have been recognised as playing an important role in the achievement of semantic interoperability of clinical information, as reflected in the recommendations of international initiatives such as the FP7 Network of Excellence SemanticHealthNet [1]. The increasing importance of such semantic resources has also stimulated their development and organisation in publicly available repositories. BioPortal [2], which is likely to be the most popular repository of biomedical semantic resources, contains about 700 biomedical ontologies, terminologies and controlled vocabularies.

Ontologies are defined as formal, explicit specifications of shared conceptualisations [3]. The development of semantic resources is usually the result of cooperation between two types of users: domain experts, who provide the domain knowledge, and knowledge engineers, who provide the expertise for the use of semantic formalisms. Ontologies are meant to be useful and processable by both humans and machines. This objective has the implication that the ontology has to include content for both types of intended users. On the one hand,

* Corresponding author.
E-mail addresses: philip.vandamme@student.uva.nl (P. van Damme), mquesada@umh.es (M. Quesada-Martínez), r.cornet@amc.uva.nl (R. Cornet), jfernand@um.es (J.T. Fernández-Breis).

https://doi.org/10.1016/j.jbi.2018.06.008
Received 15 October 2017; Received in revised form 10 June 2018; Accepted 12 June 2018
Available online 14 June 2018

ontologies contain natural language descriptions of their concepts and properties for human consumption. On the other hand, ontologies contain logical axioms, which provide a precise meaning to their concepts and properties when they are expressed in a formal language, for machine consumption.

Generally speaking, the quality of a given product is measured by the degree of fulfilment of the design requirements for such product. The objective of Quality Assurance (QA) processes is to ensure that those requirements are met. This not only includes the identification of errors and making corrections, but also preventing them. The increasing popularity of semantic resources means that more applications are using them, so QA becomes a critical task.

There has actually been an increasing interest in QA and auditing initiatives in recent years [4]. The methodological review presented in [5] proposes a classification based on four criteria: the type of knowledge utilised in the auditing process, the type of techniques used (manual, automated systematic or automated heuristic), the terminology on which the method is focused, the attributes being audited and five quality factors: Concept-orientation; Consistency; Non-redundancy; Soundness; and Comprehensive coverage.

In our current research, we focus on automated systematic methodologies to audit the completeness of concept definitions, which contributes to comprehensive coverage. We propose auditing ontologies by utilising the natural language descriptions associated with concepts, in line with previous studies [6]. Those studies have found that ontologies are richer in natural language content than in logical axioms. The domain knowledge expressed only in natural language is called *hidden semantics* [7]. Concepts in resources such as Gene Ontology (GO) or SNOMED CT have expressive natural language labels because developers tend to use a systematic *naming convention* for the labels of taxonomically related concepts. The use of naming conventions is a principle recommended by the Open Biological and Biomedical Ontology (OBO) Foundry for the construction of ontologies and terminologies. The lexical component of ontologies has already been used for ontology QA in [8], which exploits the semantics associated with the lexical component in ontologies to homogenise the structure of the labels in ontologies. This is done by identifying and transforming labels semantically related but expressed using a different linguistic structure. Hence, the actions taken involve the labels, not the formal concept definitions.

The comparison of what is expressed both logically and in natural language could serve the purpose of QA of biomedical ontologies and terminologies. There should be a correspondence (ideally, an equivalence) between the content expressed in natural language for humans and the content expressed in the form of logical axioms for machines. The lexical content of ontologies such as the GO has been the source of knowledge for natural language processing [9] and has driven the analysis of the compositional structure of GO concepts [10]. In terms of tooling, OBOL [11] facilitates the integration of language and meaning in bio-ontologies, by providing a grammar which permits associating axiomatic patterns with linguistic structures. It was developed for the OBO community and was used for the creation of the GO cross-products [12], and can also be applied for ontology maintenance. In [13], six main types of quality issues in SNOMED CT (see Table 1) were identified. Such issues should be targeted by QA methods. In relation to the incomplete modelling issue, previous works on SNOMED CT [14,15] have identified and illustrated situations where the formal relations are not representing the meaning associated with the natural language content.

Our work is inspired by the 'lexically suggest, logically define' (LSLD) principle [14], which states that the knowledge reflected as natural language in labels should also be represented as logical axioms. Our aim is to design an effective QA method for biomedical semantic resources, which uses resources of natural language content to propose logical axioms. This means that we will mainly address the quality issue of *incomplete modelling* described in Table 1.

In this paper, modules extracted from SNOMED CT, which is the second most audited terminology [5], are used as resources for evaluating the results of the method. Our proposal applies lexical regularities (LRs) (further defined in Section 2.1), which are groups of one or more (consecutive) tokens that appear in several concept labels in an ontology [15,16]. The assumption is that those regularities embed domain knowledge, which should be available as logical axioms. LRs function as seeds for capturing different kinds of issues, which are often concentrated on a group of concepts shared by their textual description. This can be assimilated to the idea of exploiting a 'focus concept' and its neighbourhoods presented in [17]. For example, the SNOMED CT concepts *Pseudocoarctation of aorta* and *Parallel course of aorta and pulmonary artery*, among others, exhibit the LR 'of aorta'. This LR can be used as seed for defining the axiomatic template like `X findingSite some aorta`, which could be applied for all those concepts exhibiting it. The axioms resulting from this process can then be compared with existing axioms to identify missing or incomplete axioms in the ontology. This work contributes to the QA of biomedical ontologies and terminologies by (1) proposing a pattern-based approach, which automatically analyses its lexical content and (2) proposes lexical patterns convertible into axiomatic patterns which can potentially enrich the ontology.

2. Methods

Our QA framework for the extraction of axiomatic patterns from the lexical content in ontologies is graphically described in Fig. 1. The ontology to be analysed is provided as input for the method. The output of the method is a set of axioms extracted from this ontology. The method consists of four main parts:

1. Extraction of LRs from the ontology (Section 2.1).
2. Clustering similar labels from concepts associated with each LR (Section 2.2).
3. Calculation of relevant metrics of the clusters (Section 2.3).
4. Obtaining general axiomatic patterns for each cluster (Section 2.4).

Besides, we also describe the use case (Section 2.5) and propose how to evaluate the effectiveness of the method (Section 2.6).

2.1. Extraction of LRs

The objective of this step is to find and extract the LRs existing in an ontology θ. An ontology θ contains a set of ontology concepts $OC = \{OC_1, ..., OC_n\}$, where n is the number of concepts. For each OC_i, we tokenise and lemmatise [18] its labels obtaining an ordered list of tokens $[T_1, ..., T_m]$, where m is the number of tokens obtained. Conceptually, a label refers to a natural language description associated with a concept in the ontology, which can be represented in the Web Ontology Language (OWL) using the `rdfs:label` annotation property (see the example in Fig. 2). In the case of SNOMED CT, concepts are described in natural language by means of a number of synonyms and one fully specified name, which provides an unambiguous description for a concept by concatenating a description with the name of the semantic tag in brackets, e.g., *Burn scar (morphologic abnormality)* or *Burn scar (disorder)*. In the OWL representation of SNOMED CT, this fully specified name is used for `rdfs:label` annotations. In this paper we use the term 'labels' to refer to the fully specified name of SNOMED CT concepts, without the bracketed name of the semantic tag. In the previous example, both concepts will have the 'label' *Burn scar*.

Conceptually, an LR is a single token (individual word) or a consecutive group of them (multiple words), which appear in several labels of an ontology. The formal definition of an LR is described as:

Definition 1 (*Lexical regularity (LR)*). An ordered sublist of tokens $LRT = [T_1, ..., T_i]$, where $i \in [1, max(m)]$, which is repeated in a subset