Exploring Complexity in Health: An Interdisciplinary Systems Approach
A. Hoerbst et al. (Eds.)

doi:10.3233/978-1-61499-678-1-384

Suggesting Missing Relations in Biomedical Ontologies Based on Lexical Regularities

Manuel QUESADA-MARTÍNEZ[a], Jesualdo Tomás FERNÁNDEZ-BREIS[a,1] and Daniel KARLSSON[b]
[a]Facultad de Informática, Universidad de Murcia, IMIB-Arrixaca, CP 30100 Murcia
[b]Department of Biomedical Engineering, Linköping University, Sweden

Abstract. The number of biomedical ontologies has increased significantly in recent years. Many of such ontologies are the result of efforts of communities of domain experts and ontology engineers. The development and application of quality assurance (QA) methods should help these communities to develop useful ontologies for both humans and machines. According to previous studies, biomedical ontologies are rich in natural language content, but most of them are not so rich in axiomatic terms. Here, we are interested in studying the relation between content in natural language and content in axiomatic form. The analysis of the labels of the classes permits to identify lexical regularities (LRs), which are sets of words that are shared by labels of different classes. Our assumption is that the classes exhibiting an LR should be logically related through axioms, which is used to propose an algorithm to detect missing relations in the ontology. Here, we analyse a lexical regularity of SNOMED CT, congenital stenosis, which is reported as problematic by the SNOMED CT maintenance team.

Keywords. lexical regularities, ontologies, quality assurance, SNOMED CT

1. Introduction

In recent years, the biomedical research community has increased its effort in the development of biomedical ontologies, some of which being the result of a collaborative effort of domain experts and ontology engineers teams [1]. Given the high level of activity in ontologies, ontology builders should be supported by quality assurance methods (*QA*), which would guarantee that the ontologies hold some desired features. An ontology is mainly composed by classes, properties and instances that can be logically related through the properties. Ontology components can have extra-logical relations (annotations), which do not affect the set of consequences derivable from an ontology, and which are described in natural language. Labels are one such type of annotations, and they should describe a class without ambiguity the classes.

The content of ontology labels has been exploited with the aim of increasing the axiomatisation of the ontologies. For example, [2] aimed to axiomatically enrich Gene Ontology (GO) using a Genus-Differentia construct, and [3] created knowledge enrichment patterns based on the dissection labels of a part of GO. All these works apply, in some sense, the *"lexically suggest, logically define"* principle presented in [4] in the context of the quality assurance of SNOMED-CT. In [4], the consistency

[1] Corresponding Author: jfernand@um.es

between ontology class labels and the semantic model codified as logical relations (axioms) was analysed. Recently, the compositional structure of GO classes was exploited in [5] to detect redundant and missing logical relations, and [6] proposes an algorithm to combine lexical and structural indicators to detect inconsistent modelling.

All those methods focused on a particular ontology, but there are no generic methods or tools capable of checking such principle to any ontology. The assumption is that classes that exhibit a *lexical regularity* should be logically related. In this work we extend our OntoEnrich framework [7] with a method that detects lexical relations that are not logically expressed. We also describe its application to a SNOMED CT module, and we pay special attention to the case of Congenital Stenosis, which is one of the cases for which logical relations are found missing by our method and that had previously been identified by the SNOMED CT maintenance team as a potential case of ill-defined content. We believe that this method may contribute to improve ontology quality assurance processes.

2. Methods

2.1. Basic definitions of lexical regularities

We define a *lexical regularity (LR)* as a group of consecutive *tokens* that appears in several labels of an ontology. The tokens are obtained using Natural Language Processing techniques: tokenisation, lemmatisation and nominalisation. Thus, an *LR* like *"neoplasm of phalanx of"* captures the singular and plural of *"phalanx"* in classes like *"Primary malignant neoplasm of phalanx of foot"* or *"Malignant neoplasm of phalanges of foot"*. In this work, we are interested in those *LRs* that correspond to full labels of classes. For example, *"degeneration"* is an *LR* in SNOMED CT, but there is no class labelled "degeneration". On the contrary, there are classes whose labels are *"congenital"* or *"congenital stenosis"*, which are also LRs in SNOMED CT.

2.2. Detecting lexical regularities not supported by logical relations

Given an *LR* and its set of tokens, the classes of interest of the LR includes all the classes of the ontology whose full label corresponds to one or more consecutive tokens of the LR. Figure 1 a) shows the 4 *classes of interest* for the LR "congenital stenosis". We use the cardinality of the *classes of interest* as a metric, so we are interested in those with value greater than 0.

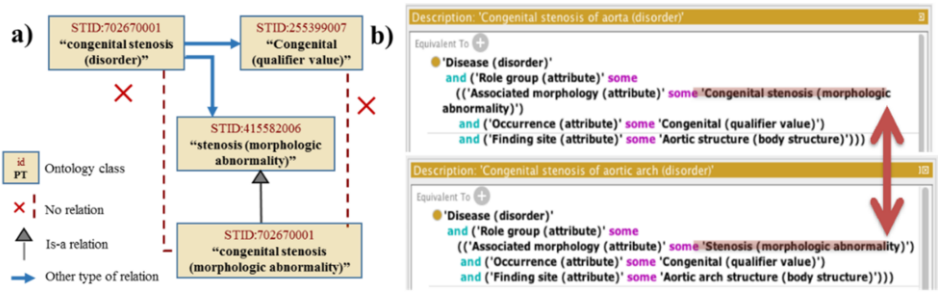

Figure 1. a) Classes that contain *"congenital stenosis"* or any of its substrings. b) Axiomatic description of two classes that contain "congenital stenosis".

Algorithm 1 computes the logical relations between a pair of classes, thus the logical relation between those classes that exhibit an *LR* and their *classes of interest* can be addressed. Let us explain it using the SNOMED CT classes *Congenital stenosis of aorta (disorder)* (TERM1) and *Stenosis (morphologic abnormality)* (TERM2). Lines 1-3 return false because they are not taxonomically related in the asserted model neither in the inferred one. Line 5 retrieves the axioms associated with TERM1, which permits line 7 to obtain that TERM1 has logical relations to *Disease (disorder*), Congenital stenosis (morphologic abnormality), *Congenital (qualifier value)* and *Aortic structure (body structure)*. Despite none of these four classes matches with TERM2, a reasoner infers that *Stenosis (morphologic abnormality)* subsumes *Congenital stenosis (morphologic abnormality)*, so TERM1 and TERM2 are logically related. This does not occur between *Congenital stenosis of aortic arch (disorder)* and *Congenital Stenosis (morphologic abnormality)*.

```
1.  IF (TERM2 Subsumes* TERM1) OR (TERM1 Subsumes* TERM2)
2.    Return true;
3.  END IF
4.
5.  AXIOMS = getAllAxioms(TERM1)
6.  FOR each AXIOM in AXIOMS
7.    ENTITIES = getEntities*(AXIOM)
8.    FOR each ENTITY in ENTITIES
9.      IF (ENTITY is TERM2) or (TERM2 Subsumes* ENTITY)
10.         Return true;
11.     END IF
12.   END FOR
13. END FOR
14.
15. RETURN false
```

```
Input:
    (1) REASONED ONTOLOGY: OWL or OBO ontology
        file reasoned by a reasoner
    (2) TERM1: source term
    (3) TERM2: destiny term
```

```
Output:
    (1) BOOL: true if there is a logical
        relation between TERM1 and TERM2
```

Algorithm 1. Detection of logical relations between two terms of the ontology.

Algorithm 1 is applied to all the pairs of classes <A,B>, where A belongs to the set of classes that exhibit an *LR* and B belongs to the set of *classes of interest* for the LR. We define the *logical relation* of an *LR* as the percentage of classes that exhibit it and are logically related to its *classes of interest*. *LRs* with a *logical relation* different from 1 can be reported for further analysis, since there might be missing logical relations. For example, we can see in Figure 1 b) that the class *Congenital stenosis of aortic arch (disorder)* is related to the morphologic abnormality *Stenosis*. However, in that figure, the class *Congenital stenosis of aorta (disorder)* is related to the *Congenital stenosis (morphologic abnormality)*. The *logical relation* of "congenital stenosis" for the simplified example of Figure 1 is 87,5% because the classes have logical relations with all the *classes of interest* except for *"congenital stenosis (morphologic abnormality)"*, so in Figure 1b) there is a lexical suggestion that is not logically supported for the class *Congenital stenosis of aortic arch (disorder)*: it should be linked to *Congenital stenosis (morphologic abnormality)* instead of to *Stenosis (morphologic abnormality)*. The absence of this link might be a problem because the results of the query for finding diseases with associated morphology *Congenital stenosis* would not include *Congenital stenosis of aortic arch (disorder)*.

3. Results

The algorithm has been implemented as part of the OntoEnrich tool (http://sele.inf.um.es/ontoenrich). We have applied the method to a SNOMED CT module (version July 2015) containing 18,440 classes and 18,443 logical relations.

Next, we describe the major results. More information, including the source files can be found at http://miuras.inf.um.es/mie2016.

3.1. Characterisation of lexical regularities and missing logical relations (deviations)

We obtained 19,774 *LRs.* This value is higher than the 18,440 classes in the module, and this is a consequence of the systematic naming followed in the development of the ontology. 11.17% of these LRs can be decomposed in *classes of interest* and the method found that 585 LR included situations in which lexical relations did not have the corresponding *logical one.*

Figure 2 a) shows the average number of *classes of interest* by *LR;* its value is 2 but there are some LR with more than three classes. Figure 2 b) shows that roughly 50% of the classes that exhibit an *LR* with classes of interest have missing relations. Figure 2 c) describes the *logical relation* in absolute terms; its median 1, so more than 50% of the LRs capture just 1 *missing logical relation*, which is a good indicator. Anyway, the inspection of no typical cases is worthy because a unique *LR* may capture many candidate missing logical relations.

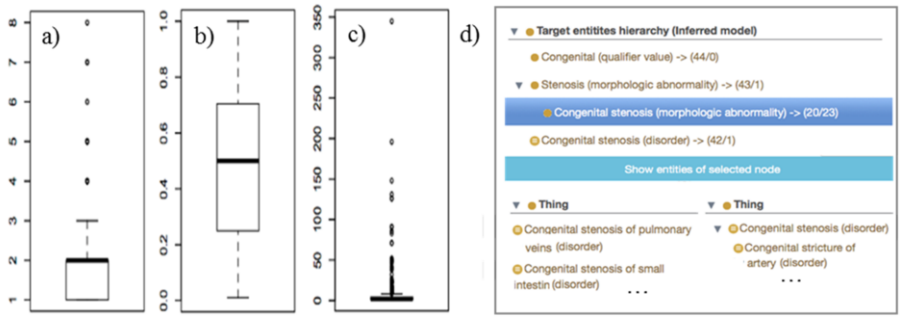

Figure 2. a), b) and c) represent graphical descriptions using boxplots of LR metrics. d) shows an example of the form implemented in OntoEnrich for exploring potentially missing relations.

3.2. Congenital Occurrence vs. Congenital Morphology

We drive our attention now to the *LR* "congenital stenosis", for which missing relations where suggested. The tracker artefact *"artf229197: Congenital Occurrence vs. Congenital Morphology"* published by the IHTSDO, describes content changes and other known identified issues in the July 2015 release of SNOMED-CT International Release. This artefact mentions issues related to the congenital stenosis, which means that our method has been capable of identifying this potential issue. Figure 2d shows the form available in OntoEnrich for inspecting the results of our method, which shows the hierarchy of classes of the inferred model of the ontology. The figure shows the information for the LR *"congenital stenosis"*. The top panel shows the *classes of interest.* Each *class of interest* has two numbers associated, shown in brackets: classes exhibiting the *LR* with logical relations (left) and classes exhibiting the *LR* without logical relations (right). For example, 42 classes exhibit the LR *"congenital stenosis"*, 20 are related with *congenital stenosis (morphologic abnormality)*, but 22 are not. The bottom panel permits to explore these 42 classes: not related (left) and related (right). Our manual inspection of the 22 classes make us think that the suggested missing logical relations are needed, but has not been validated by the SNOMED CT team yet.

4. Conclusions

Our method detects *LRs* that capture lexical relations that might not be logically defined. Our method is complementary to the one presented in [4], where some rules were built from the analysis necessary and sufficient conditions used to define terms that exhibit the most frequent bi-grams. The authors assumed that if there is a logical pattern it should be followed by the classes that share the bi-gram, regardless it is a term or not. Our approach focuses on decomposable terms that match with other classes and studies the relation between those classes that exhibit them.

The application of the method to a SNOMED CT module of 18,440 classes has identified 585 cases of the potential existence of missing logical relations between classes. We have analysed the *"congenital stenosis"* case, which has been reported as an issue by the SNOMED CT team. This makes us think that the analysis of the reminding 584 cases might contribute to the SNOMED-CT Quality Assurance. It should be noted that SNOMED-CT allows three sub-types of labels: fully specified name (FSN), preferred term (PT) and synonym (S). We have processed PTs, but the inclusion of S for detecting *LRs* could increase the number of missing logical reglations detected.

The next steps also include the application of the method to other relevant biomedical ontologies. Finally, we will develop templates for the addition of logical relations that would cover the missing relations found by the method.

Acknowledgements

This work has been possible thanks to the funding of Spanish Ministry of Science and Innovation, the Fundación Séneca and the FEDER Programme through grants EEBB-I-15-10466, TIN2014-53749-C2-2-R, 15295/PI/10 and and 19371/PI/14.

References

[1] Hoehndorf, R., Haendel, M., Stevens, R., & Rebholz-Schuhmann, D. (2014). Thematic series on biomedical ontologies in JBMS: challenges and new directions. *Journal of Biomedical Semantics*, **5**, 15.

[2] Mungall, C. J., Bada, M., Berardini, T. Z., Deegan, J., Ireland, A., Harris, M. a., Lomax, J. (2011). Cross-product extensions of the Gene Ontology. *J Biomed Inform*, **44**(1), 80–86.

[3] Fernandez-Breis, J. T., Iannone, L., Palmisano, I., Rector, A. L., & Stevens, R. (2010). Enriching the Gene Ontology via the Dissection of Labels Using the Ontology Pre-processor Language. In P. Cimiano & H. S. Pinto (Eds.), *Knowledge Engineering and Management by the Masses* (pp. 59–73). Springer.

[4] Rector, A., & Iannone, L. (2012). Lexically suggest, logically define: Quality assurance of the use of qualifiers and expected results of post-coordination in SNOMED CT. *J Biomed Inform*, **45**(2), 199–209.

[5] Agrawal, A., Perl, Y., Ochs, C., & Elhanan, G. (2015). Algorithmic detection of inconsistent modeling among SNOMED CT concepts by combining lexical and structural indicators. In *2015 IEEE International Conference on (BIBM)* (pp. 476–483). IEEE.

[6] Mougin, F. (2015). Identifying Redundant and Missing Relations in the Gene Ontology. In *Digital Healthcare Empowering Europeans* (pp. 195–199).

[7] Quesada-Martínez, M., Mikroyannidi, E., Fernández-Breis, J. T., & Stevens, R. (2015). Approaching the axiomatic enrichment of the Gene Ontology from a lexical perspective. *Artificial Intelligence in Medicine* **6**, 35-48.