Capturing the Relationship between Evolving Biomedical Concepts via Background Knowledge

Cédric Pruski[1], Julio Cesar dos Reis[2], and Marcos Da Silveira[1]

[1] Luxembourg Institute of Science and Technology, 5, Avenue des Hauts-Fourneaux, L-4362, Esch/Alzette, Luxembourg
[2] Institute of Computing, University of Campinas, Av. Albert Einstein, 1251, Cidade Universitária Zeferino Vaz, 13083-852, Campinas-SP, Brazil
{cedric.pruski, marcos.dasilveira}@list.lu, julio.dosreis@ic.unicamp.br

Abstract. Semantic Web applications and knowledge-based systems heavily rely on the use of up-to-date ontologies. To support their adequate evolution, methods must detect the changes of meanings in concepts over time. This article proposes exploiting domain-specific external source of knowledge to characterize the evolution of concepts in dynamic ontologies. Our original technique analyses the evolution of values in concept attributes. The approach uses ontological properties and mappings between ontologies from online repositories to deduce the nature of the relationship between a concept and its successive version. The proposed algorithm is experimentally evaluated comparing different configurations by using various successive versions of biomedical ontologies. The obtained results reveal the benefits of considering external sources of knowledge to identify the correct semantic relation in ontology evolution.

Keywords: Biomedical ontologies, Ontology evolution, Background knowledge

1 Introduction

Software applications using Semantic Web technologies have gained interest in Life Sciences over the past years [10]. For example, clinical decision support systems use ontologies and their associated mappings for enhancing their reasoning capabilities [3]. However, the dynamic nature of medical knowledge forces ontology engineers to constantly revise the content of ontologies to keep it up-to-date. Also, these modifications might be propagated to depending artefacts (like mappings and annotations) to keep the underlying systems reliable.

Maintaining these artefacts up-to-date according to ontology evolution is a complex task, specially in highly dynamic domains. For example, the release of new versions of ontologies, like SNOMED CT, can impact a huge amount of mappings [9,11]. The development of automatic tools to assist domain experts in these maintenance tasks requires adequate characterization of ontology changes. Nevertheless, the detection of changes between different ontology versions remains an open issue. The limitations refer to understanding the evolution of

meanings for a given concept. For instance, if it has become more or less specific from the semantic point of view, or if the change does not affect its semantics.

We have shown the importance of attribute values (*e.g.,* concept label, synonyms, *etc.*) defining concepts in the establishment of mappings [8] and, in turn, for their maintenance [7]. Consider the evolution of *"poisoning central nervous system stimulants"*, which is the title of the concept '970' of ICD-9-CM version 2009, to *"central nervous system stimulants"* in ICD-9-CM version 2011. This shows that the concept became more general since the term *"poisoning"* was deleted. So, if one assumes that this concept was interrelated to another ontology with an "equivalent" mapping, the evolution previously described shall transform the "equivalent" mapping to an "is-a".

In this paper, we propose an approach to detect the semantic evolution of concept attribute values. Our method determines the relationship between the two concepts (one of each version) having the considered attributes as label or synonyms. To this end, the approach explores background knowledge (*e.g.*, external source of knowledge like *Bioportal* [15]) to characterize ontology evolution by identifying specific changes of attribute values defining the concepts in different versions. It aims at determining if the evolved information has become more generic, more specific, remains equivalent or if it is somehow related. We assessed the approach over a corpus of reference. It was built by domain experts, made up of two successive versions of concepts' attributes retrieved from SNOMED CT, ICD-9-CM and MeSH. To demonstrate the benefits of considering domain-specific background knowledge, we compare the obtained results with those from the constructed corpus.

The remainder of this article is organized as follows: Section 2 formalizes the addressed problem and describes the related work. Afterwards, we present the proposed method and the implemented algorithm (Section 3). Section 4 describes the evaluation including the obtained results. In the sequence, we present the discussion (Section 5), which is followed by the conclusion in Section 6.

2 Problem Statement and Related Work

This section introduces definitions and the addressed problem. We describe the related work to clarify the originality of our approach.

2.1 Definitions

An ontology O is a set of concepts interrelated by semantic relationships [12]. We define a set of concepts of O at time j, such that $j \in \mathbb{N}$, as $C(O^j) = \{c_i^j | i \in \mathbb{N}\}$. Each concept is characterized by a set of attributes. The set of attributes defining a concept $c \in C(O^j)$ refers to the function $A(c^j) = \{Att_1^j, Att_2^j, ..., Att_n^j\}$ (*e.g.*, concept label, definition, synonym, *etc.*). The attributes can differ from one ontology to another, but in general each attribute describing a concept has a *name* and an associated string *value*. For example, the attribute value *"avian flu"* is a synonym of the concept *"avian influenza"* from SNOMED CT version

2014AB. We define $Att_i^j.name$ (*e.g.*, synonym) and $Att_i^j.value$ (*e.g.*, *"avian flu"*), but from now on we use Att_i^j to denote $Att_i^j.value$ to simplify. We define $Rel_i \in O^j, Rel_i = \{(a, b, r_i) | a, b \in C(O^j), r_i \in RelSymb\}$ where $RelSymb \in \{\perp, \equiv, \leq, \geq, \approx\}$.

The addressed problem consists in defining the semantic relationship that exists between two successive versions of an attribute of a given concept. Our technique must decide if the considered concept has become more or less specific or if it remains equivalent after evolution (*i.e.*, a new version of the ontology).

2.2 Related Work

Ontology matching is a research field where background knowledge has been implemented. A first significant tentative was proposed by Aleksovski *et al.* to align two ontologies that present poor lexical overlap and limited structural properties using a semantically rich knowledge source [1]. The approach consists in finding anchoring matches, using lexical heuristics, of the source and target ontology in the external one. The proposal uses its semantics to deduce the relationship that holds between the concepts.

Sabou *et al.* [16] proposed an ontology matching paradigm that could be complementary to existing classic ones. It automatically explores multiple and heterogeneous *on-line* knowledge sources to derive mappings. The approach aims at aligning ontologies' concepts by selecting the most appropriate knowledge distributed over several external ontologies. They explored formal properties of the background knowledge to infer the possible relationship that could exist between the concepts to be aligned.

TaxoMap uses *WordNet* as background knowledge [13]. The intervention of a domain expert is required to determine the best anchor in *WordNet* that corresponds to the concepts to align. This anchor delimits the usable sub-graph in *WordNet* to optimize the alignment phase. Once the appropriate sub-graph is identified, classic matching techniques interrelate the concepts of source and target ontologies with those previously defined in the sub-graphs.

Mougin *et al.* [14] proposed the use of *WordNet* to disambiguate information contained in biomedical systems with the UMLS[3]. The goal was to validate obtained mappings in cases of unambiguous matches between the information content and UMLS or, disambiguate the obtained alignment in case of several correspondences using the information provided by *WordNet*. They showed that general knowledge can improve the validation of direct mappings and help in the identification of indirect mappings of concepts to the UMLS.

Zhang and Bodenreider [19] aimed at taking advantage of domain-specific knowledge using the UMLS to improve the alignment between anatomical ontologies. They revealed that domain knowledge is a key factor behind the identification of additional mappings compared with the generic schema matching approach. The use of UMLS as an external resource was interesting for various aspects: (1) generating more mappings; (2) providing different synonyms for a given concept and (3) defining relations between concepts in a semantic network.

[3] www.nlm.nih.gov/research/umls

More recently, Arnold and Rahm [2] explored generic external resources and proposed a two-step enrichment technique to improve existing imprecise ontology mappings. They used linguistic techniques and resources like *WordNet* to refine the semantic relation between aligned concepts. Their work aims to transform equivalence between concepts into a "is-a" or "part-of" which may further reflect the real semantics of mapped concepts. It was not applied to ontology evolution.

The use of background knowledge has also been investigated for ontology evolution. The background knowledge was used to assess new statements identified as relevant and must be included in an ontology at evolution time [18]. This work, part of the EVOLVA framework [17], aims at enriching an ontology with additional relevant knowledge (*i.e.,* statements) by using background ontologies.

All the surveyed approaches fail in considering the impact of ontology evolution on dependent artifacts as ultimate focus, so several gaps remain open. First, in most of the approaches, the used background knowledge is usually of general nature (the knowledge is described at a higher level of abstraction). This might be interesting for disambiguating the context, but not to characterize the evolution of concepts, especially if their definition become more finely described. Second, a single source of knowledge (*i.e.*, only one ontology) is usually implemented, which limits its coverage. This is true mainly for domain specific ontologies, which require a very precise description of the external knowledge. The use of multiple connected ontologies might optimize the coverage of the domain through a more precise description of the domain.

3 Determination of Semantic Relationship between Changing Concepts

Our method aims to reach an accurate characterization of the semantic evolution of concepts by analysing multiple and domain specific interconnected ontologies contained in *Bioportal*. We assume that performing a match between different domain-specific ontologies might provide necessary and sufficient facts to determine the relationship between evolving concepts. Mappings that link them might allow reasoning over several ontologies. This aspect, combined with the richness of the content of ontologies, provides a good support for ontology evolution, and in particular, the characterization of the modifications that affect entities.

We define the Algorithm 1 that exploits search modules, the structure and properties of ontologies and mappings stored in *Bioportal*. Concept attributes play a key role in the definition of mappings and for their maintenance [4]. In consequence, we assume that the modifications in attribute values of concepts directly impact their semantics, which in turn, might influence dependent mappings. For this reason, given a changed concept, our algorithm takes as input the value of the attribute before evolution and its value after evolution. In the following, we explain the algorithm.

1. **Search for concepts** (statement 1 to 3). The goal is to find the attribute values Att_1 and Att_2 in the description of concepts in ontologies different

Algorithm 1: Background knowledge-based semantic relationship identification between evolving concepts

Require: $Att_1 \subset O^j; Att_2 \subset O^{j+1}$ - Two attributes of concepts
Ensure: $r \in \{\perp, \equiv, <, >, \approx\}$ - The link between concepts defined by Att_1 and Att_2

```
1: r ← ⊥;
2: C_Att1 ← findConcepts(Att1)
3: C_Att2 ← findConcepts(Att2)
4: O_common ← findCommonOntologies(C_Att1, C_Att2, O^j)
5: if O_common = ∅ then
6:    M ← getMappings(C_Att1, C_Att2)
7:    if M = ∅ then
8:       return r
9:    end if
10: end if
11: if areEquivalent(Att1, Att2, O_common) = true then
12:    r ←≡
13: else if isMoreSpecific(Att1, Att2, O_common) = true then
14:    r ←<
15: else if isMoreSpecific(Att2, Att1, O_common) = true then
16:    r ←>
17: else if areSiblings(Att1, Att2, O_common) = true then
18:    r ←≈
19: end if
20: return r
```

from the one of input O. We use Bioportal's search module to find exact match between Att_1 and the preferred terms and synonyms of existing concepts. The same procedure is taken to search concepts for Att_2.

2. **Identify a set of common ontologies** (statement 4 to 10). To detect the semantic relationship that exists between the concepts found at the previous step and Att_1 and Att_2. Two cases can be distinguished:
   (a) **Direct method**. Both attribute values are found together in concepts belonging to the same group of ontologies excluding the original O, *i.e.*, there is at least one common ontology that contains both attributes.
   (b) **Indirect method**. In this case, we use existing mappings defined in the background knowledge to retrieve equivalent concepts than the ones found in the search phrase.
3. **Characterize the relationship that links the identified concepts in background knowledge** (statement 11 to 19). The algorithm first selects the more detailed ontology. We assume that the more concepts an ontology contains the most precise the relationship we are looking for will be. The algorithm calculates the path that exists between the two anchored concepts using the hierarchy. The relation is then determined as follows:
   (a) Concepts are considered equivalent if those identified in the search step are the same or an explicit equivalent mapping exist between them.

(b) The path only contains subsumed concepts then the concept located at the highest level of the hierarchy is considered as less specific.
(c) If the two concepts are siblings then it returns ≈ *"partially related to"*. This relation is imprecise, but of interest for mapping maintenance [6].
(d) Otherwise, no relationship can be precisely determined and the *"undefined"* value is returned.

Consider the example of Fig. 1. We observe that the concept code 'M0006899' whose label is "*Pituitary dwarfism*" in MeSH evolved from version released in 2012 to "*Pituitary dwarfism II*" ('M0452907') in version 2013. Algorithm 1 identifies that these concepts are siblings in SNOMED CT, by searching *Bioportal*, therefore the relationship symbol "partially related to" is returned.

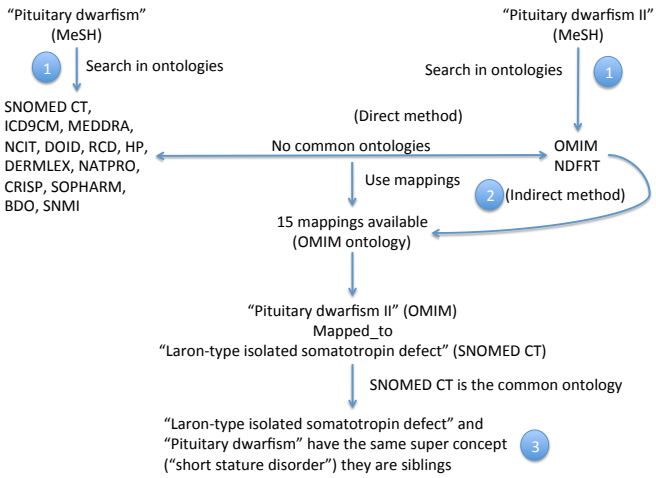

Fig. 1. Illustrative example in applying the method

4 Experimental Evaluation

We assessed the effectiveness of our approach on realistic case studies of the biomedical domain. The goal is to show the ability of the algorithm to infer the semantic relations for characterizing ontology evolution in modifications of concept attribute values.

4.1 Materials and Procedure

Experiments relied on successive versions of ICD-9-CM, SNOMED CT and MeSH. The evaluation consists in characterizing the evolution of concepts of

these ontologies by analysing the evolution of their attributes value. However, as no Gold Standard for such an evaluation exists, we needed to construct our own corpus of reference to compare the obtained results. We conducted the following three steps method to obtain our corpus of reference:

1. We selected 1.000 couples of attributes from SNOMED CT, ICD-9-CM and MeSH from concepts affected by change operations. One attribute of each couple comes from a concept at time j and the other attribute comes from the context of the concept at time $j+1$. We chose these couples based on the similarity score between the attribute values (*i.e.*, we excluded attributes with very low similarity and unchanged attributes at time $j+1$). The context of a concept denotes all its subconcepts, superconcepts and siblings and the similarity refers to syntactic and word-based distances [5].
2. Three experts evaluated the selected couples of attributes to determine "equivalent", "more specific", "less specific", "partially matched to" or "no relation" as the relationship that link them.
3. They performed one round of evaluation and we merged the answers that are the same for all reviewers. The experts collaborated and re-evaluated a second time those attribute couples for which no agreement was found. We achieved an average agreement rate of 86% for the concerned attributes. Finally, we retained 675 pairs of attributes that have the consensus of opinion about their semantic relationship.

4.2 Results

Fig. 2 presents two sets of results. The graphic on the left side depicts all the cases tested with our algorithm including those for which the algorithm did not detect a relationship (*e.g.*, cases where the pairs of attribute were not found in external ontologies). Aiming to better evaluate the precision of the algorithm, the right side of Fig. 2 presents the results excluding all cases of disagreement, where the algorithm returns *no relation*. The main reason for this disagreement happens when the analyzed attributes do not exist in external ontologies.

The obtained results are convincing since a general precision of 77% is obtained (CR=BK in Fig.2, *i.e.*, experts and algorithm agree on the semantic relationship between the concept attributes). The most significant results are obtained for "no relation" (unmappable in Fig. 2, left side), "equivalent" and "partial match" relations, reaching an average of 88% of precision. However, this number decreases to 53% when considering the subsumption relationship (More specif. and Less specif.).

The graphic on right side of Fig. 2 reveals the general efficiency of our algorithm in a subset of the cases (by excluding disagreement on the "no relation"). Globally, the results are significant since 96% of precision is reached. We even obtain 100% precision for the identification of "equivalent" relation (*i.e.*, all the cases returned by our algorithm are correct according to domain experts). The ability of the algorithm to identify the subsumption relationships is less convincing as shown by the results (13% of the cases were not correctly identified).

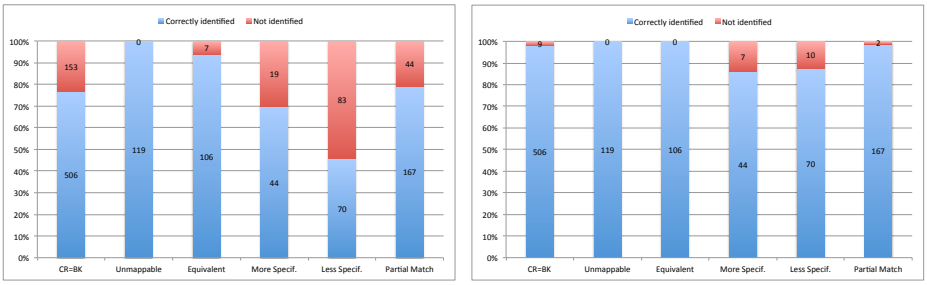

Fig. 2. Experimental results. CR denotes the experts data, and BK is the algorithm outcome. The left side shows the overall results, the right side excludes the results where the algorithm predicts *no relation* and the experts disagree with it.

5 Discussion

The proposed method is able to correctly detect 95% of the semantic relationship in concept evolution when the concepts exist in ontologies stored in *Bioportal*. The remaining 6% differs mainly for two reasons:

- The level of granularity to describe one concept can differ from one ontology to another. This situation impacts on the outcome of our method when one ontology includes as synonyms a list of terms that have subsumption relations in another ontology. For instance, the term *"Chondrosarcoma of bone"* is defined in SNOMED CT as a sub-concept of *"Chondrosarcoma"*, but in CTV3 these two terms are synonyms. While building the corpus of reference, the experts adopted the definition of SNOMED CT for these terms (*i.e.*, more specific than) so the algorithm found a different result.
- Domain experts were more precise than existing ontologies. This was, for instance, observed when using the terms *"autonomic peripheral nervous system diseases"* and *"autonomic central nervous system diseases"*. MeSH, National Drug File – Reference Terminology, and Neuroscience Information Framework Standard Ontology define these two terms as synonyms of *"autonomic nervous system diseases"*. Our method detects that these terms are equivalent, but domain experts considered these two terms as siblings. Consequently, the outcome of the algorithm differs from the corpus of reference.

The use of background knowledge shows the possibility of automatically capturing the impact of changes in concepts from a semantic viewpoint. However, this approach contains some limitations:

- In our method, semantic changes can only be measured if the considered attribute is the label (or synonym) of a concept in another external ontology. This situation was observed in 83% of the concepts in the corpus.
- Mappings between ontologies in the repository must be correct and up-to-date. Our method uses these mappings to select the ontologies that are

analysed, since we assumed that unmapped ontologies do not describe the same domain (*i.e.*, they do not have overlapping concepts).
- Non-equivalent relation need to be inferred by our method. *Bioportal* only contains equivalent mappings, *i.e.*, if there is a mapping between two concepts from different ontologies, the interrelation between these concepts is always an equivalence. This can lead to cases where subsumed concepts (in ontology A) are mapped to the same concept (in ontology B). For instance, *"left bundle branch blocks"* and *"right bundle branch blocks"* are sibling concepts in ICD9 and CTV3, but they are interrelated to the same concept of MeSH (where these two terms are described as synonyms). The outcome of our method could be even more accurate if other types of relationships exist in available mappings from the repository.

To complement the proposed technique, we are currently working on an algorithm to determine non-equivalent relations between concepts. We are studying the use of several mappings to navigate from one ontology to another relying on domain-specific background knowledge. This can allow collecting more information about the attribute to select the appropriated relation according to the level of granularity used to describe the concept in the original ontology.

6 Conclusion

Dealing with the evolution of ontologies and of their dependant artefacts relies on appropriate ontology changes identification. It demands characterization of the semantic relation between evolving concepts. In this paper, we proposed an approach exploiting domain-specific background knowledge to determine the semantic evolution of concepts. Our algorithm analyzed the modifications in the values of concept attributes and the experiments showed the effectiveness of the technique with large life sciences ontologies from *Bioportal*. Future work involves the refinement of the algorithm and further experiments with additional datasets.

Acknowledgements

This work is supported by the National Research Fund (FNR) of Luxembourg and São Paulo Research Foundation (FAPESP) (Grant #2014/14890-0).

References

1. Aleksovski, Z., Klein, M., ten Kate, W., van Harmelen, F.: Matching unstructured vocabularies using a background ontology. In: Proceedings of the 15th International Conference on Managing Knowledge in a World of Networks. pp. 182–197. EKAW'06, Springer-Verlag, Berlin, Heidelberg (2006)
2. Arnold, P., Rahm, E.: Semantic enrichment of ontology mappings: a linguistic-based approach. In: Advances in Databases and Information Systems. pp. 42–55. Springer (2013)

3. Bodenreider, O.: Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearbook of medical informatics p. 67 (2008)
4. Dinh, D., Dos Reis, J.C., Pruski, C., Da Silveira, M., Reynaud-Delaître, C.: Identifying relevant concept attributes to support mapping maintenance under ontology evolution. Web Semantics: Science, Services and Agents on the World Wide Web 29, 53–66 (2014)
5. Dos Reis, J.C., Dinh, D., Da Silveira, M., Pruski, C., Reynaud-Delaître, C.: Recognizing lexical and semantic change patterns in evolving life science ontologies to inform mapping adaptation. Artificial Intelligence in Medicine 63(3), 153 – 170 (2015)
6. Dos Reis, J.C., Dinh, D., Pruski, C., Da Silveira, M., Reynaud-Delaître, C.: Mapping adaptation rules for the automatic reconciliation of dynamic ontologies. In: Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM) (2013)
7. Dos Reis, J.C., Pruski, C., Da Silveira, M., Reynaud-Delaître, C.: Dykosmap: A framework for mapping adaptation between biomedical knowledge organization systems. Journal of biomedical informatics 55, 153–173 (2015)
8. Dos Reis, J.C., Pruski, C., Da Silveira, M., Reynaud-Delaître, C.: Characterizing semantic mappings adaptation via biomedical kos evolution: A case study investigating snomed ct and icd. In: Proceedings of the Annual AMIA Symposium (2013)
9. Dos Reis, J.C., Pruski, C., Da Silveira, M., Reynaud-Delaître, C.: Understanding semantic mapping evolution by observing changes in biomedical ontologies. Journal of Biomedical Informatics (2013)
10. Feigenbaum, L., Herman, I., Hongsermeier, T., Neumann, E., Stephens, S.: The semantic web in action. Scientific American 297(6), 90–97 (2007)
11. Gross, A., Hartung, M., Thor, A., Rahm, E.: How do computed ontology mappings evolve?-a case study for life science ontologies. In: Joint Workshop on Knowledge Evolution and Ontology Dynamics (2012)
12. Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge Acquisition 5(2), 199–220 (Jun 1993)
13. Hamdi, F., Safar, B., Reynaud, C., Niraula, N.: TaxoMap alignment and refinement modules: Results for OAEI 2010. In: The Fifth International Workshop on Ontology Matching. pp. 212–219 (2010)
14. Mougin, F., Burgun, A., Bodenreider, O.: Using wordnet to improve the mapping of data elements to umls for data sources integration. In: Proceedings of the AMIA annual Symposium. pp. 574–578 (2006)
15. Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M.A., Chute, C.G., Musen, M.: Bioportal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research (37) (2009)
16. Sabou, M., d'Aquin, M., Motta, E.: Exploring the semantic web as background knowledge for ontology matching. Journal on data semantics XI pp. 156–190 (2008)
17. Zablith, F.: Evolva: A comprehensive approach to ontology evolution. In: The Semantic Web: Research and Applications, the 6th European Semantic Web Conference, ESWC, pp. 944–948. Springer, Heraklion, Crete, Greece (2009)
18. Zablith, F., d'Aquin, M., Sabou, M., Motta, E.: Using ontological contexts to assess the relevance of statements in ontology evolution. In: Proceedings of EKAW 2010 - Knowledge Engineering and Knowledge Management by the Masses (2010)
19. Zhang, S., Bodenreider, O.: Experience in aligning anatomical ontologies. International journal on Semantic Web and information systems 3(2), 1 (2007)