ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

Debugging Mappings between Biomedical Ontologies: Preliminary Results from the NCBO BioPortal Mapping Repository

Jyotishman Pathak, Christopher G. Chute
Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN, USA

Abstract

The ability to provide semantic mappings between multiple large biomedical ontologies is considered as a very important, albeit labor-intensive and error-prone task. To facilitate such a process, several approaches for collaborative ontology mapping building and sharing have been proposed in the recent past. However, despite the improvements in community-wide mappings development, more often the mapping rules are redundant, incoherent, and at times, incorrect. In this paper, we present an approach for identifying such "erroneous mappings" using Distributed Description Logics. Specifically, we illustrate how logical reasoning can be used to discover semantic inconsistencies caused by erroneous mappings, and provide preliminary results of experiments based on the National Center for Biomedical Ontology BioPortal mapping repository.

Introduction

The ability to specify semantic mappings between biomedical ontologies is an important research agenda in the medical informatics community. Several approaches have been proposed for alignment between ontologies ranging from entirely manual[1], to semi-automatic[2,3], to fully-automatic[4] mapping techniques, many of which have met with varying degrees of success. More recently, with the growing number of ontologies in the biomedical domain, and hence the increasing requirement for their alignment, community-based approaches to create mappings have been proposed that allow users and domain experts to specify semantic correspondences in a collaborative manner[5,6]. However, despite these advancements, an important limitation of the existing efforts is the lack of ability to identify, debug, and invalidate semantically inconsistent mappings (or erroneous mappings). As mentioned by Noy et al.[5], such a requirement is vital because in many cases a concept definition may change with a new version of the ontology, and thereby making an existing mapping invalid, or users may add new or delete existing mappings that result in the aligned ontologies becoming logically inconsistent.

Toward this end, we propose a technique for identifying erroneous mappings between biomedical ontologies. In particular, we exploit the underlying semantics of the mappings as well as the mapped ontologies based on Distributed Description Logics (DDL)[7] to pinpoint mappings that are logically inconsistent[8]. Our basic assumption is that a mapping that correctly states the semantic correspondences between the ontology concepts should not cause inconsistencies in the mapped ontologies. The advantage of using DDL is that it allows us to detect such inconsistencies, which can then be regarded as symptoms caused by erroneous mappings. For example, Figure 1 shows two equivalent mappings between concepts A and B in ontology O_1 (source) with concepts C and D in ontology O_2 (target), respectively. Furthermore, B is asserted as a subConcept of A in O_1, whereas both C and D are asserted as disjoint from each other in O_2. Assuming that both the mappings are valid as well as ontologies O_1 and O_2 are logically consistent, one can infer (via global interpretation) that the concept D should be a subConcept of C in O_2. However, since they are asserted as disjoint in O_2, thereby causing a logical inconsistency, implies that at least one of the mappings is erroneous—identification of which is our objective. Specifically, the main contributions of the proposed work are:

Figure 1: Inconsistent Mappings Example

- We leverage DDL[7] and ontology mapping repair techniques[8] to describe a formal framework for identifying erroneous biomedical ontology mappings.
- We illustrate the applicability of our approach by experimenting with the NCBO BioPortal mapping repository[5] and provide preliminary results.
- We provide an open-source prototype implementation of our software based on the DRAGO distributed reasoning system: http://code.google.com/p/bioontologies-mapping-debugger.

Background

Distributed Description Logics (DDL)[7] is a knowledge representation formalism for representing sets of ontologies and semantic relations between them. It provides a mechanism for referring to ontologies and for defining rules that connect "concepts" in different ontologies. This is achieved using the notion of importing and reusing concepts between ontologies and enabling reasoning with multiple ontologies interconnected by directional semantic mapping (called the bridge rules). In particular, DDL extends the notion of interpretation introduced above to fit the distributed nature of the model and to reason about concept subsumption across ontologies.

More formally, let I be a set of non-empty indices, such that $\{O_i\}_{i\in I}$ is a set of ontologies. Concepts and axioms are represented with the index of the ontology they belong to such that i:C denotes a concept in ontology O_i and j:C⊆D represents that concept C is a sub-concept of D in ontology Oj , where i:C and j:C are different concepts. Semantic relations between concepts in different ontologies are represented via axioms, called bridge rules that are of the following form: (1) i:C → j:D (into-rule); and (2) i:C ← j:D (onto-rule); where, C and D are concepts in ontologies O_i and O_j , respectively. Furthermore, the derived bridge rule i:C≡j : D can be defined as a conjunction of the into- and onto-bridge rules. These rules do not represent the semantic relations stated from an external observation point of view such as the Web. Instead, a rule i to j expresses relations between i and j viewed from j-th subjective point of view. Specifically, an into-bridge rule i:C→j:D states that, from j-th point of view, the concept C in i is less general that its "local" concept D. Equivalently, the onto-relation i:C←j:D expresses the more generality relation. In general, note that the into-rule (i:C→j:D) is not necessarily an inverse of the onto-rule (i:C←j:D) since these rules reflect a subjective point of view. Thus, a "distributed ontology" D_{OR} can now be defined as a tuple, $(\{O_i\}_{i\in I}, \{R_{ij}\}_{i\neq j\in I})$, where $\{O_i\}_{i\in I}$ is the set of ontologies, and $\{R_{ij}\}_{i\neq j\in I}$ is the set of bridge rules between those ontologies.

An important aspect of DDL is that for the fundamental reasoning services of verification of consistency and concept satisfiability, in addition to the ontology itself, the reasoning depends on other ontologies to which it has semantic mappings. This is due to the ability of the bridge rules to transitively propagate knowledge across ontologies in the form of subsumption axioms as illustrated in Figure 2.

$$\frac{i:A\subseteq B, i:A\longrightarrow j:G, i:B\longrightarrow j:H}{j:G\subseteq H}$$

Figure 2: Subsumption Propagation of DDL Bridge Rules

The main objective of our work is to leverage DDL[7] and existing techniques for repairing ontology mappings[8] to provide a formal framework for identifying erroneous mappings between biomedical ontologies. In what follows, we formalize ontology mappings (with respect to DDL) and outline steps for identifying erroneous mappings.

Mappings and Correspondences: At an abstract level, a mapping between ontologies O_i (source) and O_j (target) can be defined via a set of correspondences, where each correspondence represents a semantic relation between concepts in O_i and O_j.

Definition 1 (Semantic Correspondence): Given ontologies O_i and O_j , a semantic correspondence can be represented (minimally) by a 3-tuple <C,C′,r>, such that C∈F(O_i), C′∈F(O_j), and r is a semantic relation, where F is a function that identifies elements in O_i and O_j. Furthermore, in this work, we restrict r to the set {≡, ⊆, ⊇}, essentially limiting to equivalence and subsumption. Given a set of semantic correspondences, we can define the notion of a mapping as a collection of such correspondences.

Definition 2 (Ontology Mapping): Given ontologies O_i and O_j, M is a mapping between O_i and O_j , iff for all correspondences <C,C′,r>∈M, we have C∈F(O_i), and C′∈F(O_j).

To formalize ontology mappings in terms of DDL presented earlier, we encode the semantic correspondences as bridge rules. In particular, each correspondence <C,C′,r> between a pair of ontologies O_i and O_j is translated into a bridge rule via a translation function T as follows:

$$T(\langle C,C',\subseteq\rangle) \equiv i:C\longrightarrow j:C' \wedge j:C'\longrightarrow i:C$$
$$T(\langle C,C',\supseteq\rangle) \equiv i:C\longrightarrow j:C' \wedge j:C'\longrightarrow i:C$$

Inconsistent Mappings. A mapping M of a distributed ontology ℑ can be defined as inconsistent with respect to a particular concept i:C if it becomes unsatisfiable modulo the mappings

Definition 3 (Mapping Consistency): Given a distributed ontology ℑ, the mapping M between ontologies Oi,Oj∈ℑ is consistent with respect to a concept i:C iff concept C is unsatisfiable in O_i implies that i:C is also unsatisfiable in ℑ. Otherwise, M is inconsistent with respect to i:C. By extrapolation, M is consistent with respect to O_i iff for all i:C, M is consistent with respect to i:C; otherwise M is inconsistent with respect to O_i.

For example, based on Figure 1, M = {O_1:A ≡ O_2:C, O_1:B ≡ O_2:D}. Furthermore, by applying distributed reasoning it can be inferred that O_2:D⊆C should hold. However, at the same time both C and D are defined as disjoint concepts in O_2, thereby making M inconsistent with respect to D since it cannot be

satisfied in the global interpretation. Algorithm 1 follows directly from Definition 3 which also states that the inconsistency of one ontology, or some sub-group of connected ontologies, does not automatically render the entire distributed ontology inconsistent. Arguably, the goal is to determine an erroneous mapping set and identify which of the semantic correspondences involved can be removed to maintain consistency. In particular, we want to determine a "minimal erroneous mapping set" which has the property that none of its subset is an erroneous mapping set.

Algorithm 1 Identification of Mapping Inconsistency

```
1: procedure ISCONSISTENT(ℑ = ({O_i}_{i∈I}, {R_ij}_{i≠j∈I})), i)
2:   for all concepts i : C ∈ T_i do
3:     if (T_i ⊭ C ⊑ ⊥) and (ℑ ⊨ i : C ⊑ ⊥) then
4:       return false
5:     end if
6:   end for
7:   return true
8: end procedure
```

Evaluation

Materials. We evaluated our methods proposed above using the NCBO BioPortal mappings repository. As stated in Noy and Musen[5], the inability to impose any quality control on the mappings that the users submit is a limitation of the existing BioPortal infrastructure, and our work provides preliminary steps in addressing this requirement.

At the time of our evaluation, the repository contained approximately 30,000 mappings between various biomedical ontologies, and a majority of these mappings were between the Open Biomedical Ontologies (OBO) and Web Ontology Language (OWL 1.0) ontologies. Since our technique for inconsistency detection has been implemented on top of the DRAGO distributed reasoning system, which is an OWL-DL based reasoner, we transformed all the mapped OBO ontologies into OWL ontologies via the OBO-in-OWL Protege plugin. Furthermore, the mappings in the BioPortal repository do not use "true" logical equivalence (e.g., owl:equivalentClass), but rather the notion of "similarity"[5]. Since such a weaker definition of equivalence is not modeled in DDL, we transformed each "similar" mapping into an equivalence (≡), into (⊆), and onto (⊇) bridge rules for experimentation. All data can be accessed at http://code.google.com/p/bioontologies-mapping-debugger.

Results

Table 1 shows the results of our evaluation. From the mapping repository, we chose only those mapped ontologies which had at least 2 or more mappings specified between them. We also did not include mappings involving the Foundational Model of Anatomy (FMA) and International Classification of Diseases (ICD-9) because the current release of DRAGO (version 2.1) does not support nominals (e.g., owl:oneOf, owl:hasValue constructs) present in FMA, and there is no ClaML (Classification Markup Language used to represent ICD-9) to OWL transformer available, respectively. Furthermore, the columns L-Satisfiable and D-Satisfiable in Table 1 represent the total number of classes found satisfiable in the target ontology that are determined by the local axioms of the ontology (localized reasoning) and by propagation of the axioms via mappings (distributed reasoning), respectively.

Discussion

Result Analysis. For mappings between OBO ontologies, no inconsistencies were found. We believe this can be attributed to the fact that none of the evaluated OBO ontologies had disjoint class axioms, and hence none of the mappings were conflicting. Similarly, for mappings between OBO and OWL ontologies, no inconsistencies were observed even though the two original OWL ontologies that were evaluated, Nano Particle Ontology (NPO) and NCI-Thesaurus (NCI-T), had 12,265 and 171 disjoint class axioms, respectively. We believe that the lack of mapping inconsistency can be attributed to: (i) for many mappings, the classes from the disjoint class axioms were not involved, and (ii) for those mappings where such classes were involved, the mappings were logically correct. For example, NPO and ChEBI had the mappings npo:Gold≡chebi:CHEBI 29287 and npo:Carbon≡chebi:CHEBI 27594, such that npo:Gold is disjointWith npo:Carbon, and the classes CHEBI 29287 and CHEBI 27594 (with labels gold and carbon, respectively) had no associations between them. Consequently, there was no conflict in the mappings as well. Finally, due to performance issues, we were not able to evaluate mappings between original OWL ontologies (namely, Galen and NCI-T).

Limitations and Further Work. As mentioned earlier, in this work we limited our scope to one-to-one concept mappings, and further considered only equivalence and subsumption mappings. However, in reality, it is possible to specify arbitrary mappings (e.g., disjoint) between any ontological entities (e.g., relationships) and the ability to consider such

Source Ontology	Target Ontology	Mapping Type	# Mappings	# L-Satisfiable	# D-Satisfiable
Cereal Plant Trait (OBO)	Plant Environmental Conditions (OBO)	≡, ⊑, ⊒	3	506 (n=506)	506 (n=506)
Phenotypic Quality (OBO)	NCI-Thesaurus (OWL)	≡, ⊑, ⊒	4	66726 (n=66726)	66726 (n=66726)
Nano Particle Ontology (OWL)	ChEBI (OBO)	≡, ⊑, ⊒	4†	21377 (n=21377)	21377 (n=21377)
Cell Type (OBO)	Fungal Gross Anatomy (OBO)	≡, ⊑, ⊒	10	71 (n=71)	71 (n=71)
Molecule Role (OBO)	ChEBI (OBO)	≡, ⊑, ⊒	21	21377 (n=21377)	21377 (n=21377)
Zebrafish (OBO)	Mouse Adult Gross Anatomy (OBO)	≡, ⊑, ⊒	145	2877 (n=2877)	2877 (n=2877)
Galen (OWL)	NCI-Thesaurus (OWL)	≡, ⊑, ⊒	271	N/A	N/A
Mouse Adult Gross Anatomy (OBO)	NCI-Thesaurus (OWL)	≡, ⊑, ⊒	2870	66726 (n=66726)	66726 (n=66726)
Human Disease (OBO)	NCI-Thesaurus (OWL)	≡, ⊑, ⊒	6883	66726 (n=66726)	66726 (n=66726)

† One of the mappings between Nano Particle Ontology and ChEBI was deemed invalid because the class in the source ontology did not exist.

Table 1: `BioPortal` Mapping Evaluation Results

mappings to find inconsistencies becomes vital. Furthermore, in the current evaluation, we took a snapshot of the mapping repository, thereby not considering how different versions of an ontology will affect the associated mappings. In future, we plan to evaluate how mapping consistency and satisfiability results vary with the evolution of the ontologies. Another limitation of our work is the complexity of the reasoning procedure. DDL subsumption reasoning has been shown to be NEXPTIME[7], thereby significantly impacting the efficiency of the consistency checking process. For example, evaluating the mappings between GALEN and NCI-Thesaurus was not feasible as the program runs out of memory (with a maximum Java heap space of 4GB). Hence, our objective is to leverage approximate reasoning services that apply correct but incomplete heuristics for performance gain[9].

Complementary to our work, the problem of identifying erroneous mappings has been addressed using the notion of a "global ontology"[10]. Consequently, reasoning is done with respect to the global ontology which, in certain cases, can result in increased complexity compared to distributed reasoning that exploits the structure provided by semantic relations for the propagation of reasoning through the local ontologies. However, there are no studies verifying this hypothesis, and our goal is to adapt our approach for such an investigation. Finally, our work raises the issue of evaluating "similarity" mappings between simple ontologies because, for example, in the absence of disjoint class axioms in both source and target ontologies, the mappings, although logically consistent, may still represent incorrect knowledge. We believe this can be partially addressed by leveraging the subsumption propagation of DDL (Figure 2) to create a distributed hierarchy which can be evaluated for correctness and accuracy, although such a proposal warrants further research.

References

1. Vikstrom A, Aner YS, Strender LE, Nilsson GH. Mapping the Categories of the Swedish Primary Health Care Version of ICD-10 to SNOMED CT Concepts: Rule Development and Intercoder Reliability in a Mapping Trial. BMC Medical Informatics and Decision Making. 2007;7(9).
2. Noy NF, Musen MA. PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment. In: 17th National Conference on Artificial Intelligence. AAAI Press; 2000. p. 450–455.
3. Fung KW, Bodenreider O, Aronson AR, Hole WT, Srinivasan S. Combining Lexical and Semantic Methods of Inter-terminology Mapping Using the UMLS. In: 12th World Congress on Health (Medical) Informatics. vol. 129. IOS Press; 2007. p. 605–609.
4. Doan A, Madhavan J, Dhamankar R, Domingos P, Halevy A. Learning to Match Ontologies on the Semantic Web. VLDB Journal. 2003;12(4):303–319.
5. Noy NF, Griffith N, Musen MA. Collecting Community-Based Mappings in an Ontology Repository. In: 7th International Semantic Web Conference. Springer-Verlag, LNCS 5318; 2008. p. 371–386.
6. Correndo G, Alani H, Smart PR. A Community based Approach for Managing Ontology Alignments. In: 3rd International Workshop on Ontology Matching. vol. 431. CEURWorkshop Proceedings; 2008. p. 61–72.
7. Borgida A, Serafini L. Distributed Description Logics: Assimilating Information from Peer Sources. J. of Data Semantics. 2003;1:153–184.
8. Meilicke C, Stuckenschmidt H, Tamilin A. Repairing Ontology Mappings. In: 22nd AAAI Conference on Artificial Intelligence. AAAI Press; 2007. p. 1408–1413.
9. Meilicke C, Stuckenschmidt H. Applying Logical Constraints to Ontology Matching. In: 30th Annual German Conference on AI. Springer-Verlag, LNCS 4667; 2007. p. 99–113.
10. Cardillo E, Eccher C, Serafini L, Tamilin A. Logical Analysis of Mappings between Medical Classification Systems. In: Artificial Intelligence: Methodology, Systems, and Applications. 2008. p. 311–321.