Proceedings of the Semantic Integration Workshop
at the Second International Semantic Web Conference (ISWC 2003) 2003:109-114.

Knowledge Augmentation for Aligning Ontologies: An Evaluation in the Biomedical Domain

Songmao Zhang and Olivier Bodenreider

U.S. National Library of Medicine, Bethesda, Maryland
National Institutes of Health, Department of Health & Human Services
{szhang|olivier}@nlm.nih.gov

Abstract

The objective of this study is to evaluate the contribution to semantic integration of the semantic relations extracted from concept names, representing augmented knowledge. Three augmentation methods – based on linguistic phenomena – are investigated (reification, nominal modification, and prepositional attachment). The number of concepts aligned in two ontologies of anatomy before and after augmentation serves as the evaluation criterion. Among the 2353 concepts exhibiting lexical resemblance across systems, the number of concepts supported by structural evidence (i.e., shared hierarchical relations) increased from 71% before augmentation to 87% after augmentation. The relative contribution of each augmentation method to the alignment is presented. The limitations of this study and the generalization of augmentation methods are discussed.

Introduction

Ontologies are often organized into concepts (e.g., *Heart*, *Mitral valve*) and semantic relations (e.g., <*Mitral valve*, PART-OF, *Heart*>). As a first approximation, concepts represent categories, while semantic relations represent assertions about the concepts. Both concepts and relations are useful for the semantic integration of ontological resources. Lexical resemblance among concept names may indicate similarity in meaning. Likewise, from a structural perspective, concepts sharing similar relations to other concepts tend to be similar in meaning.

However, the difference between concepts and semantic relations my not be as clear-cut as it seems. Although representing categories, concepts such as *Vein of leg* and *Subdivision of heart* also embed partitive assertions in their names. For example, the relation <*Vein of leg*, PART-OF, *Leg*> can be deduced from the name *Vein of leg*. And *Subdivision of heart* is equivalent to the relation <*X*, PART-OF, *Heart*> where *X* is a placeholder for any concept subsumed by *Subdivision of heart*, including *Mitral valve*. In addition, from the name of the concept *Sweat gland*, one can derive the assertion <*Sweat gland*, IS-A, *Gland*>.

More generally, concept names often embed assertions, i.e., implicit knowledge, not always represented explicitly through semantic relations. In this paper, we examine three linguistic phenomena (reification, nominal modification, and prepositional attachment), which usually embed semantic relations. We show how semantic relations extracted from these concept names contribute to improving the semantic integration – through alignment – of two ontologies of anatomy.

The general framework of this study is that of lexical semantics and knowledge acquisition. Lexical semantics [1] studies the link between linguistic phenomena and the semantic relations they encode. As such, lexical semantics contributes to knowledge acquisition from textual resources. While originally applied to general relations (e.g., hypernymy, meronymy) from general corpora (e.g., machine-readable dictionaries [2]), the same techniques have been adapted to the acquisition of specialized relations (e.g., the molecular interaction BINDS [3]) from the biomedical literature. Terminologies have also been used as specialized corpora for acquiring knowledge [4]. In this particular con-

text (controlled vocabulary, closed subdomain), there is often less ambiguity than in larger textual resources, which may facilitate knowledge extraction. In previous work, we studied semantic relations embedded in biomedical terms through nominal modification [5] and reification [6].

Although sharing with these studies some of the methods used for knowledge acquisition, this paper specifically evaluates the contribution to semantic integration of the semantic relations extracted from concept names through various methods. We demonstrate how each linguistic phenomenon under investigation contributes to improving the alignment of two ontologies of anatomy. This study is not evaluation of the alignment itself, but rather a quantification of the contribution of augmented knowledge to the alignment.

Resources and Methods

Ontologies of anatomy

The **Foundational Model of Anatomy**[1] (FMA) [August 30, 2002 version] is an evolving ontology that has been under development at the University of Washington since 1994 [7]. Its objective is to conceptualize the physical objects and spaces that constitute the human body. The underlying data model for FMA is a frame-based structure implemented with Protégé-2000. With 59,422 concepts, FMA claims to cover the entire range of gross, canonical anatomy. Concept names in FMA are pre-coordinated, and, in addition to preferred terms (one per concept), 28,686 synonyms are provided (up to 6 per concept). For example, there is a concept named *Uterine tube* and its synonym is *Oviduct.*

The Generalized Architecture for Languages, Encyclopedias and Nomenclatures in medicine[2] (**GALEN**) [version 5] has been developed as a European Union AIM project led by the University of Manchester since 1991 [8]. The GALEN common reference model is a clinical terminology represented using GRAIL, a formal language based on description logics. GALEN contains 25,192 concepts and intends to represent the biomedical domain, of which canonical anatomy is only one part. Unlike FMA, GALEN is compositional and generative. Concept names in GALEN are post-coordinated, and only one name is provided for each concept.

Both FMA and GALEN are modeled by *IS-A* and *PART-OF* relationships and allow multiple inheritance. Relationships in GALEN are finer-grained than in FMA. For the purpose of this study, we considered as only one *PART-OF* relationship the various kinds of partitive relationships present in FMA (e.g., *part of, general part of*) and in GALEN (e.g., *isStructuralComponentOf, isDivisionOf*).

Extracting relations from concept names

We used three methods for extracting relations from concept names. Each method takes advantage of one particular linguistic phenomenon. The relations embedded in concept names through these phenomena sometimes coexist with equivalent semantic relations represented explicitly in the ontology. However, cases where a relation is only embedded in a concept name in one ontology and only represented explicitly in the other are likely to impair semantic integration. In order to make ontologies more easily comparable, we systematically extracted the relations embedded in concept names. In this study, we focused on taxonomic (i.e., *IS-A* and *INVERSE-IS-A*) and partitive (i.e., *PART-OF* and *HAS-PART*) relations.

The **reification of *PART-OF*** consists of using a concept named *Part of W* to subsume a concept *P* instead of using a *PART-OF* relationship between the concept *P* (the part) and *W* (the whole). From a linguistic perspective, the concept name *Part of W* reifies the *PART-OF* relationship from concept *P* to *W*. The two representations, <*P*, *IS-A*, *Part of W*> and <*P*, *PART-OF*, *W*>, are equivalent for most purposes [9]. We systematically extracted <*P*, *PART-OF*, *W*> and <*W*, *HAS-PART*, *P*> from concept names such as *Subdivision of X, Organ component of X*, and *Component of X*, where *X* is a concept present in the ontology. For example, because the concept *Component of hand* subsumes *Finger*, we generated the two relations <*Finger*, *PART-OF*, *Hand*> and <*Hand*, *HAS-PART*, *Finger*>.

Nominal modification often represents a hyponymic relation involving the head of the noun phrase. For example, a *Cranial nerve* is a kind of *Nerve* and the *Carotid artery* is a kind of *Artery*. Therefore, the relations <*X Y*, *IS-A*, *Y*> and <*Y*,

[1] http://sig.biostr.washington.edu/projects/fm/AboutFM.html

[2] http://www.opengalen.org/

INVERSE-IS-A, *X Y* > can be tentatively extracted from the term *X Y*. However, this method is not applicable when the head of the noun phrase is polysemous in the domain under investigation. For example, *Body* (human body) does not subsume *Carotid body* (a small neurovascular structure). The problem here lies in the several senses of *body*: "the material part or nature of a human being" for the former and "a mass of matter distinct from other masses" in the latter. Domain knowledge is required for identifying such cases.

In anatomical terms, **prepositional attachment** using "of" (*X of Y*) often denotes a partitive relation between *X of Y* and *Y*. For example, we generated the relations <*Bone of femur*, *PART-OF*, *Femur*> and <*Femur*, *HAS-PART*, *Bone of femur*> from the term *Bone of femur*. Because it does not fully analyze the concept names, this method is not suitable for complex anatomical terms (e.g., names containing prepositions other than "of", such as *Groove for arch of aorta*).

Evaluation

The two ontologies of anatomy, FMA and GALEN, were aligned using a combination of lexical techniques (resemblance among concept names) and structural techniques (similarity and conflicts based on the semantic relations) [10]. In order to evaluate the role of the relations generated by augmentation, the alignment based on the explicit knowledge alone was compared to the alignment based on both explicit and augmented knowledge. In practice, structural techniques were used to refine the alignment of lexically related concepts, called anchors. Structural similarity, used as positive evidence, is defined by the presence of common hierarchical relations among anchors across systems. Conflicts, on the other hand, used as negative evidence, are defined by the existence of opposite hierarchical relationships (e.g., *PART-OF* and *HAS-PART*) between the same two anchors across systems.

Based on such evidence, the anchors (i.e., pairs of lexically related concepts *X* and *X'*) can be classified into three main groups:

1. anchors with no structural evidence (i.e., *X* and *X'* do not share any hierarchical relationships to other anchors),
2. anchors with positive evidence, (i.e., *X* and *X'* share similar relationships to other anchors), and
3. anchors with negative evidence (i.e., *X* and *X'* share opposite relationships to other anchors).

In order to quantify the contribution of augmented knowledge to the alignment of two ontologies, we compared the number of anchors in each group before and after augmentation. Since the augmentation methods applied to the two ontologies generate additional relations, it is expected that some of these new relations provide additional structural evidence to some anchors, thus reducing the number of anchors with no structural evidence.

Results

Number of relations generated

The number of concept names exhibiting the three linguistic phenomena under investigation (reification of *PART-OF*, nominal modification, and prepositional attachment) is presented in Table 1.With the exception of nominal modification, the lexical phenomena of interest in this study were more often present in FMA than in GALEN. This is especially true for prepositional attachment. Most names in FMA are anatomical terms and a majority FMA names contain the preposition "of" (e.g., *Muscle of pelvis*, *Nail of third toe*, *Cruciate ligament of atlas*, *Base of phalanx of middle finger*, etc.). In contrast, only part of GALEN concepts are related to the anatomical domain, which may explain the lexical differences observed between the two ontologies. Because a given name may exhibit more than one lexical phenomenon, the sum of the numbers of names for each phenomenon is greater than the total number of names.

The number of relations generated by the three augmentation methods described earlier is shown in Table 2. Note that a method may extract more than two relations (direct and inverse) from a concept name. This happens when the same linguistic phenomenon is present more than once in a name (e.g., from *Base of phalanx of middle finger* (*BoPoMF*), we generate both <*BoPoMF*, *PART-OF*, *Phalanx of middle finger*> and <*BoPoMF*, *PART-OF*, *Middle finger*>, as well as their inverses). A majority of relations extracted from the concept names are also explicitly represented in GALEN, but not in FMA. Because of some redundancy between explicit and extracted relations (and, to a lesser degree, among extracted relations), the total number of relations after augmentation is less than the sum of the numbers of explicit and extracted relations.

Additional structural evidence acquired

The alignment consists of identifying equivalent concept in FMA and GALEN. These anchors are concepts present in the two ontologies exhibiting the following two properties: lexical similarity (their names are lexically equivalent) and structural similarity (they share relationships to other anchors). 2353 lexically equivalent concepts, called anchors, were identified, of which 1668 (71%) also exhibited structural similarity before augmentation techniques were applied to FMA and GALEN. This proportion rose to 87% when relationships generated through augmentations were used.

The details of the alignment before and after augmentation are presented in Table 3. The relations generated by augmentation enable 388 anchors (+16%) to acquire positive evidence. Before augmentation, there was no support for these concepts to be considered either aligned (positive evidence) or distinct (negative evidence). Anchors acquiring positive evidence after augmentation include *Ciliary gland* (the sweat gland of eyelid), which acquired through augmentation *ISA* relation to *Gland* and *PART-OF* relation to *Head*, themselves anchors.

Not surprisingly, augmented knowledge also revealed a few more conflicts across systems. For example, the two anchors *Dorsum of Foot* and *Dorsal Region of Foot* originally received positive evidence through some shared hierarchical relations. After augmentation, they acquire negative evidence because the extracted relation <*Surface of dorsum of foot*, *PART-OF*, *Dorsum of foot*> in FMA conflicts with the explicit relation <*Dorsal Region of Foot*, *HAS-PART*, *Dorsum of Foot*> in GALEN (*Surface of dorsum of foot* and *Dorsal Region of Foot* are synonymous in FMA).

Relative contribution of each method

Before augmentation, the number of anchors not supported by structural evidence was 665, i.e., 28% of the 2353 anchors. If only one method were applied, this number would decrease by about 9%, since about 200 anchors acquire evidence through reification of *PART-OF* (203) and nominal modification (201), and by 7% with propositional attachment (158). This shows the relative contribution of the three augmentation methods in providing evidence for anchors.

Finally, Table 4 simulates what would happen if augmentation methods were applied only to one system and not to the other. The alignment mostly benefited from augmenting relations in FMA. The relations required for concepts to acquire evidence were generated from concept names in FMA in 364 cases out of 388 (94%).

Discussion

Generalization. Knowledge augmentation can be applied to other subdomains of biomedicine than anatomy and can be applied beyond the biomedical domain. Because of the prominence of hierarchical relations in anatomy, this study focused on *IS-A* and *PART-OF* relations. However, associative relations could benefit from the same approach. Roles and functions are often reified (e.g., *Iron transporter*, *Calcium channel blocker*). New rules would have to be developed to target specific relations.

Likewise, depending on the context, prepositions other than "of" could be used to identify relations (e.g., *Urine test for glucose*, where the preposition "for" expresses the relationship *analyzes*). Possibly, other linguistic phenomena such as appositives could be used as well. Finally, by increasing the number of relations available, knowledge augmentation should benefit not only semantic integration, but also other approaches relying on semantic relations such as semantic interpretation.

Limitations. One obvious limitation of this study is that no validation of the 2353 anchors has been performed yet. In the absence of a gold standard resulting from such a validation, it may be difficult to evaluate the actual benefit of any method generating the relations used as structural evidence in the identification of equivalent concepts across ontologies. Since the validation of 2353 anchors represents a significant effort involving domain experts, we elected to maximize the amount of structural evidence first (e.g., through augmentation) so that it could be used by the experts in a validation environment. Nevertheless, the results of this study suggest that relations generated by augmentation only provided structural evidence to a significant number of anchors (16%). An informal evaluation conducted on a limited number of anchors showed that, in most cases, anchors supported by structural evidence denote equivalent concepts across ontologies.

Alternative approaches. Our approach to aligning ontologies relies on lexical and structural similarity. In this regard, it is close to approaches such as PROMPT [11]. However, the augmentation techniques presented here are typically not used in their alignment algorithm. A different approach to aligning FMA and GALEN has been reported by Mork & al. [12]. These authors use a generic schema matching technique. While their approach is essentially generic, and therefore virtually domain-independent, ours takes advantage of domain knowledge. The augmentation techniques described in this paper are in many cases specific to anatomy. However, we believe that this study may be an illustration of the importance of domain knowledge in alignment techniques.

Conclusions

Knowledge augmentation based on semantic relations embedded in concept names through various linguistic phenomena has proved a powerful technique, generating as many relations as are represented explicitly in FMA. Moreover, knowledge augmentation also clearly benefited the alignment of FMA and GALEN, enabling 16% more anchors to acquire evidence (mostly positive, but also negative), compared to the use of explicit relations alone.

Acknowledgements

The research was supported in part by an appointment to the National Library of Medicine Research Participation Program administrated by the Oak Ridge Institute of Science and Education through an interagency agreement between the U.S. Department of Energy and the National Library of Medicine.

Thanks for their support and encouragement to Cornelius Rosse, José Mejino, and Kurt Richards for FMA and Alan Rector, Jeremy Rogers, and Angus Roberts for GALEN. Thanks also to Pieter Zanstra at Kermanog for providing us with an extended license for the GALEN server.

References

1. Cruse DA. *Lexical semantics*: Cambridge University Press; 1986.
2. Dolan W, Vanderwende L, Richardson SD. Automatically Deriving Structured Knowledge Bases From On-Line Dictionaries. Redmond, WA: Microsoft Corporation; 1993.
3. Rindflesch TC, Hunter L, Aronson AR. Mining molecular binding terminology from biomedical text. *Proc AMIA Symp* 1999:127-31.
4. Aussenac-Gilles N, Bourigault D, Condamines A, Gros C. How can Knowledge Acquisition benefit from Terminology? *Proceedings of the 9th Knowledge Acquisition Workshop* 1995;1:11-16.
5. Bodenreider O, Burgun A, Rindflesch TC. Lexically-suggested hyponymic relations among medical terms and their representation in the UMLS. *Proceedings of TIA'2001 "Terminology and Artificial Intelligence"* 2001:11-21.
6. Burgun A, Bodenreider O, Le Duff F, Mounssouni F, Loréal O. Representation of roles in biomedical ontologies: a case study in functional genomics. *Proc AMIA Symp* 2002:86-90.
7. Rosse C, Mejino JL, Modayur BR, Jakobovits R, Hinshaw KP, Brinkley JF. Motivation and organizational principles for anatomical knowledge representation: the digital anatomist symbolic knowledge base. *J Am Med Inform Assoc* 1998;5(1):17-40.
8. Rector AL, Bechhofer S, Goble CA, Horrocks I, Nowlan WA, Solomon WD. The GRAIL concept modelling language for medical terminology. *Artif Intell Med* 1997;9(2):139-71.
9. Schulz S. Bidirectional mereological reasoning in anatomical knowledge bases. *Proc AMIA Symp* 2001:607-11.
10. Zhang S, Bodenreider O. Aligning representations of anatomy using lexical and structural methods. *Proc AMIA Symp* 2003:(to appear).
11. Noy NF, Musen MA. PROMPT: algorithm and tool for automated ontology merging and alignment. *Proc of AAAI* 2000:450-455.
12. Mork P, Pottinger R, Bernstein PA. Challenges in precisely aligning models of human anatomy using generic schema matching. *Personal Communication* 2003.

Table 1. *Number of concept names exhibiting the three linguistic phenomena under investigation (a given name may exhibit more than one lexical phenomenon)*

	FMA		GALEN	
Reification of *PART-OF*	1,618	(2%)	227	(1%)
Nominal modification	19,395	(22%)	8,282	(33%)
Prepositional attachment	53,103	(60%)	1,886	(7%)
None	23,049	(26%)	15,353	(61%)
Total (unique names)	88,108		25,192	

Table 2. *Number of relations generated by the three augmentation methods (In parentheses is the percentage of relations not present before augmentation for each linguistic phenomenon and, on the last line, the percentage of relations only generated by augmentation techniques)*

	FMA		GALEN	
Before augmentation	342,889		322,092	
Reification of *PART-OF*	215,300	(93%)	58,358	(38%)
Nominal modification	55,328	(37%)	19,732	(21%)
Prepositional attachment	145,960	(74%)	3,886	(27%)
Total (unique relations)	658,749		349,366	
From augmentation only	*315,860*	*(48%)*	*27,274*	*(8%)*

Table 3. *Repartition of the 2353 anchors by type of evidence, before and after augmentation*

Type of evidence	Before		After		Difference	
None	665	(28%)	277	(12%)	-388	(-16%)
Positive	1668	(71%)	2054	(87%)	+386	(+16%)
Negative	20	(1%)	22	(1%)	+2	(+0%)

Table 4. *Number of anchors acquiring structural evidence (positive or negative) after augmentation, by method (first applied to each ontology separately, then applied to both)*

	FMA	GALEN	Both
Reification of *PART-OF*	193	13	203
Nominal modification	183	8	201
Prepositional attachment	137	10	158
All three combined	364	26	388