A Linguistic Approach to Aligning Representations of Human Anatomy and Radiology

Pinar Wennerberg[1], Manuel Möller [2,] Sonja Zillner[1]
[1]Siemens AG, Munich, Germany; [2]DFKI, Kaiserslautern, Germany

Abstract

In the context of medical imaging different domain ontologies are necessary that provide complementary knowledge about anatomy and radiology. This is essential for realizing applications such as medical image search. Consequently, semantic integration of these different but nevertheless related types of medical knowledge from disparate domain ontologies becomes necessary. In our work we interpret semantic integration as aligning a taxonomy on radiology and an ontology on human anatomy to find equivalent concepts that represent their shared view on medical imaging. The resulting alignments describing this common view can then be used to annotate medical images and related textual patient data. Our alignment approach has three main aspects: (a) linguistic-based, (b) corpus-based, and (c) dialogue-based. In this paper, we describe the application of the first aspect on a representation of human anatomy and a representation of radiology and report on the results.

INTRODUCTION

As the content of numerous ontologies and terminologies in the biomedical domain increases, so does the need for sharing and reusing this body of knowledge. This is especially the case in the domain of medical imaging, where different domain ontologies are required to support the heterogeneous tasks that require complementary knowledge e.g., about human anatomy and radiology. Semantic medical image search is the context of our work that lies within the THESEUS MEDICO research project.

MEDICO's proposed solution relies on ontology based semantic annotation of the image contents and the related patient data. This allows for a content mark-up with meaningful meta-information at a higher level of granularity that goes beyond simple keywords. Therefore, the data processed and stored in this way can be efficiently retrieved by a corresponding search engine as the one envisioned in MEDICO[1].

We conceive of a radiology expert as an end user who looks for, starting from a certain medical image, all related information such as patient data, lab reports, and treatment plans etc. Obtaining this kind of heterogeneous information from a single access point requires the data to have been previously integrated appropriately. The integration can be achieved by annotating the data with concepts from relevant ontologies and terminologies.

The Foundational Model of Anatomy[2] (FMA) and the Radiology Lexicon RadLex[3] are two semantic resources that can be used for this purpose. The former is a comprehensive ontology on human anatomy, whereas the latter is a lightweight terminology designed to satisfy the needs arising in annotation of radiological images and reports. Due to its complexity and scale FMA is too elaborate to use for manual annotation or for integrating in search applications. RadLex, on the other hand, is designed with practical applicability in mind but it lacks both comprehensiveness and the fine granularity of the FMA. To be able to cover all relevant information at the right level of granularity by using only one ontology as a starting point therefore requires aligning FMA and RadLex. This, additionally, allows the radiologist to use *his vocabulary* during search for convenience. The search results, i.e. the returned images and documents, thus present him a *shared view* of radiology and anatomy. For aligning medical ontologies[5] in general we follow an approach based on three main aspects: (a) on the linguistic analysis of the ontology concept labels (the *linguistic aspect*), (b) on corpus analysis (*context information aspect*) and (c) on human interaction e.g., relevance feedback (*user interaction aspect*). In this paper we report on current work on the first aspect.

The rest of this paper is organized as follows. Next section gives a brief overview of related work. Section 3 describes materials and the linguistic methods we use to align FMA and RadLex. Section 4 discusses the first results. Section 5 concludes the paper and mentions future directions.

RELATED WORK

Ontology alignment (in this paper we use the terms ontology alignment and matching interchangeably) is commonly understood as a special case of semantic integration that concerns the semi-automatic discovery of semantically equivalent (or related)

concepts across two or more ontologies. The most recent and comprehensive related work is provided by Euzenat and Shvaiko [6].

It is also becoming an increasingly active research field in the biomedical domain, especially in association with the Open Biomedical Ontologies[7] (OBO) framework. Johnson[8] et al. takes an information retrieval approach to discover relationships, between the Gene Ontology (GO) and three other OBO ontologies (ChEBI, Cell Type and BRENDA Tissue.) This approach, however, does not account for the complex linguistic structure that is typical for the medical ontology concept labels and, can therefore result in inaccurate matches.

Zhang and Bodenreider[9] report on their experience with aligning two anatomy ontologies. They showed that concept labels contain implicit relationships and with the help of linguistic methods these can be discovered to find more correspondences. Mungall's[10] research objective is to enable discovering such relations systematically with the help of a formal language. Our work is inspired by these findings.

Only recently, there has been work on creating an application ontology from RadLex and FMA[10]. This is done by incorporating subsets of the FMA into the organizational structure of RadLex. In contrast to this approach, we only align RadLex to obtain an additional view to the FMA. This allows us to preserve the entire information from the FMA for automatic image/text annotation whenever necessary.

MATERIALS AND METHODS

Foundational Model of Anatomy (FMA)

FMA is our primary source of anatomical knowledge. It is developed and maintained by the Structural Informatics Group at the University of Washington. Besides the specification of anatomy taxonomy, the FMA provides definitions for conceptual attributes, part-whole, location, and other spatial associations of anatomical entities. FMA also provides synonym information (up to 6 per concept), for example one synonym for *'Neuraxis'* is the *'Central nervous system'*. It is currently the largest ontology about human anatomy. The version we currently refer to is the version available in February 2009.

RadLex (Radiological Lexicon)

RadLex is a radiology taxonomy developed and maintained by the American College of Radiology (ACR). It covers terms from different sub-domains of medicine and clinical practice such as imaging modality and observations, pathologies, anatomy etc. Its purpose is to provide a standardized terminology for radiological practice. Synonym information is given whenever it is present as in *'Schatzki ring'* and *'lower esophageal mucosal ring'*. The version we currently refer to is the version available in February 2009. For this work we focus on its anatomy subset.

General Approach to Aligning Medical Ontologies

As a result of our experience with the medical ontologies throughout the MEDICO use case we have identified a set of common characteristics that are relevant for the alignment process. The most significant observations are that (a)they generally are very large models, (b) they have extensive *is-a* (*part-of*) hierarchies organized according to different views, (c) they contain complex relationships, (d) their terminologies are rather stable (especially for anatomy), i.e. they should not differ too much in the different models.

Based on these characteristics and the general MEDICO expectations, we derived a set of requirements[11] for aligning medical ontologies, which eventually led to our combined approach based on the three aspects described earlier. Accordingly, the *linguistic aspect* suggests exploiting the information-rich concept labels in the medical ontologies to discover further relations. The *context information aspect* based on corpus analysis assumes that ontology concepts from different ontologies with similar meaning will have similar contexts in the corpus (i.e. documents, sentences and surrounding words, in which the concept label appears) and suggests to compare the contexts to be able to determine the concept similarity. Finally, the *user interaction aspect* conceives of a dynamic model, where the alignment happens during an interactive dialogue between the user and the system. In this way, clarifications and questions from the user's feedback support the ontology matching process.[12]

Linguistic Alignment of FMA and RadLex

The linguistic alignment proposes to use rules for detecting the syntactic variants of the ontology concept labels to discover semantic relations e.g., equivalence, hyponymy, hyperonymy. The assumption is these relations can potentially support the alignment process. For example, take the concept label *'blood in aorta'* from the FMA and its lexical pattern (*noun preposition noun*). We can apply the transformation rule: [noun1 preposition:in noun2 à noun2 noun1] and generate a syntactic variant for this concept label that nevertheless has equivalent semantics, i.e., *'blood in aorta'* == *'aorta blood'*. Indeed, in FMA such syntactic variants (or semantic equivalents) are partially present. With the help of this rule we thus augment FMA with additional syntactic variants.

More interesting is the case with detecting hyponyms and hyperonyms (sub-, superconcept relations). For example, the concept '*superficial femoral artery*' from RadLex does not have a correspondence in the FMA. However, its hyperonym (superconcept) '*femoral artery*' does. Thus, by adding it to RadLex we can match the corresponding FMA concept.

Furthermore, a deeper analysis of the multi-word concept labels has shown that noun+preposition pairs almost always specify relations between anatomical entities. Take, for example, the RadLex concept '*articular cartilage of distal medial cuboid*'. Here, '*articular cartilage*' is related to '*medial cuboid*' through the relationship *to*. Similarly, in '*fossa for right fifth costal cartilage*' (FMA concept) '*fossa*' and '*right fifth costal cartilage*' are related to each other through the relation *for*. A common observation is that in most cases nouns (or noun phrases) that occur in the left or right hand side of the prepositions are anatomical entities. Furthermore, these anatomical entities participate in relations (often spatial) that are indeed specified by the prepositions. Both RadLex and FMA are rich with prepositions, where the most frequent five are: *of* (119886 FMA, 2180 RadLex), *to* (119886 FMA, 2180 RadLex), *for* (3167 FMA, 58 RadLex), *with* (438 FMA, 28 RadLex) and *in* (145 FMA, 21 RadLex) for both ontologies. Consequently, we can expect to discover a considerable number of relationships that are implicitly present. We currently investigate how these relationships can possibly support the alignment process.

Generating variants requires assigning concept labels their lexical categories in order to apply the transformation rules. We started with creating a flat text file from the OWL version of FMA that included, for each concept, its id, preferred name, all synonyms and its superconcept. For RadLex the resulting text file contained for each concept its preferred name, one or more synonyms, superconcept and optionally its definition with a total of 81800 FMA and 5026 RadLex concepts. We then annotated all information in both files with part-of-speech (POS) information i.e. we assigned the words their lexical categories. Eventually, for the most frequent prepositions we generated 924 FMA and 135 RadLex variants (i.e. semantic equivalents) using the previous rule.

For generating hyperonyms we, in the first place, concentrated on adjective+noun sequences that do not contain prepositions. As discussed earlier, we observed that the prepositions in multi-word concept labels relate different anatomical entities to each other meaning that one multi-word concept label may contain multiple *anatomical* entities. Therefore, the hyperonym generation in this case should be handled with care. This is necessary to avoid situations where, for example, '*thoracic wall*' is a hyperonym for '*neural network of posterior thoracic wall*' leading to an overgeneralization.

To generate the hyperonyms we identified all those multi-word concept labels from FMA and RadLex, where the last noun in the concept label is preceded by at least one or more successive adjectives. Then, for each such case we repeatedly omitted an adjective from the beginning of the multi-word concept label until we were left with one adjective+noun combination. Each newly generated concept label was added to the original as its hyperonym. For example, the RadLex concept '*superficial femoral artery*' is assigned '*femoral artery*' as its hyperonym. Eventually, we generated for FMA 1504 and for RadLex 902 hyperonyms that we incorporated in the alignment process.

Lexical Alignment uses the concept labels from the two ontologies as well as the generated variants and hyperonyms to determine their string similarity. We applied simple string matching after normalization to the concept labels i.e. to the preferred names and to the synonyms. We followed a strict matching strategy and considered only the exact matches. In the future we will relax it to accord for string overlaps and we will include other similarity measures (particularly those available in the Simpack[14] software library) to expand this initial set.

RESULTS

As a result of the exact string matching we identified 1147 common concepts. The generated variants (i.e. semantic equivalents) matched additional 62 concepts. Using the hyperonyms we further found a total of 846 correspondences. With the hyperonyms we found 448 additional matches in the FMA and 398 in RadLex. The reason for the higher number of matches in RadLex is that RadLex included more synonyms than FMA. The results are shown in the table below (Table 1).

Matching Mode	**F**	**R**
exact string matches	1.4%	22,8%
matches with generated variants	1.4%	24%
matches with generated hyperonyms	1.9%	28%

Table 1. The percentage of FMA/ RadLex that was matched according to the matching technique.

DISCUSSION

The strict exact matching strategy naturally returned fewer matches. Strict matches are useful to identify

the exact view of the two sources on the anatomy. Nevertheless, it will be expanded to include the results of a more relaxed matching strategy e.g., string overlaps. At the same time we expect such a relaxed strategy to return a very large set of correspondences (first experiments indicate this), therefore a threshold value will be required. This will disallow those matches that are not 'similar enough' to be included in the correspondences set. Using the generated variants (i.e. generated semantic equivalents) we were able to identify additional correspondences. In the case of FMA this was not a significant addition therefore it did not change the proportion that was matched. In the case of RadLex was a 2% increase. We assume that relaxing the match strategy will also yield more correspondences via the syntactic variants in the future. Finally, with the generated hyperonyms we identified a considerably larger amount of matches than those with the variants.

FUTURE WORK

As next we will incorporate the hyperonyms generated as described for structural alignment to exploit the hierarchical information in the ontologies for deriving more correspondences. This can be done by traversing the sub-hierarchy around the hyperonyms. In parallel, we will investigate potentials of processing the information conveyed by the prepositions acting as relations to support the alignment process. A helpful next step would be to identify the semantically equivalent or subsuming relations for the prepositions from the OBO Relations Ontology (RO) to assign them explicit and stable semantics. In this way, not only the RadLex or FMA concepts but also the relations of the RO can be used for annotating the medical images. Finally, a succeeding natural step will be to enhance the alignments obtained from the linguistic aspect of our general alignment approach with those obtained from the corpus based aspect that accounts for context similarity.

Acknowledgements

This research has been supported in part by the THESEUS Program in the MEDICO Project, which is funded by the German Federal Ministry of Economics and Technology under the grant number 01MQ07016. The responsibility for this publication lies with the authors. Special thanks to Kamal Najib of Siemens AG for implementing the tests. We are also thankful to our clinical partner Dr. Alexander Cavallaro of the University Hospital Erlangen for his expert contribution.

References

1. Möller, M., Regel, S., and Sintek, M. "RadSem: Semantic Annotation and Retrieval for Medical Images", In Proc. of the 6th Annual European Semantic Web Conference 2009 (to appear).
2. Rosse C. and Mejino J.L.V. Anatomy Ontologies for Bioinformatics: Principles and Practice. The Foundational Model of Anatomy Ontology 2007;6: 59-117
3. Langlotz C.P. Radlex: A new method for indexing online educational materials. RadioGraphics, 2006;26:1595-1597.
4. Möller M. and Mukherjee S. Context-driven ontological annotations in DICOM images: Towards semantic PACS. In: Proc. of BIOSTEC 2009. (to appear)
5. Wennerberg P. Aligning Medical Ontologies for Clinical Query Extraction. In: Proc. of EACL 2009, PhD Symposium, Athens, Greece (to appear)
6. Euzenat J, Shvaiko P., Ontology Matching. Juni 2007, Springer-Verlag
7. The Open Biomedical Ontologies. Online available: http://www.obofoundry.org/
8. Johnson H.L, Cohen K.B., Baumgartner W.A. Jr., Lu Z, Bada M, Kester T, Kim H, Hunter L, 2006: Evaluation of lexical methods for detecting relationships between concepts from multiple ontologies. In: Proc. of Pac. Symp Biocomput, 2006;28-39, Washington, DC
9. Songmao Z., Bodenreider O. Aligning representations of anatomy using lexical and structural methods. In Proc AMIA Symp 2003: 753–757
10. Mungall C.J, 2004: Obol: integrating language and meaning in bio-ontologies Comparative and Functional Genomics, vol.5, no. 6-7, pp. 509+, August 2004
11. J. L. Mejino, D. L. Rubin, and J. F. Brinkley. FMA-RadLex: An application ontology of radiological anatomy derived from the foundational model of anatomy reference ontology. In Proc. of AMIA Symposium, pages 465-469, 2008.
12. Sonntag D, 2008. Towards dialogue-based interactive semantic mediation in the medical domain In Third International Workshop on Ontology Matching at ISWC, 2008
13. The OBO Relation Ontology. Online available: http://www.obofoundry.org/ro/
14. SimPack. Online available: http://www.ifi.uzh.ch/ddis/simpack.html