Towards Context Driven Modularization of Large Biomedical Ontologies

Pinar Oezden Wennerberg, Sonja Zillner
Siemens AG, Munich, Germany

Abstract

Formal knowledge about human anatomy, radiology or diseases is necessary to support clinical applications such as medical image search. This machine processable knowledge can be acquired from biomedical domain ontologies, which however, are typically very large and complex models. Thus, their straightforward incorporation into the software applications becomes difficult. In this paper we discuss first ideas on a statistical approach for modularizing large medical ontologies and we prioritize the practical applicability aspect. The underlying assumption is that the application relevant ontology fragments, i.e. modules, can be identified by the statistical analysis of the ontology concepts in the domain corpus. Accordingly, we argue that most frequently occurring concepts in the domain corpus define the application context and can therefore potentially yield the relevant ontology modules. We illustrate our approach on an example case that involves a large ontology on human anatomy and report on our first manual experiments.

INTRODUCTION

Medical research and clinical practice deal with complex and heterogeneous data, which poses challenges to realizing applications such as medical image and text search. Thus, it becomes necessary to support these applications with machine processable, explicit medical knowledge e.g., about human anatomy, radiology or diseases. This knowledge can be acquired from biomedical domain ontologies and can be used in the application, for example in semantic medical image and text search. Semantic medical image search is, indeed, the context of this work that lies within the THESEUS MEDICO research project.

Based on experience throughout the MEDICO project, we have observed that biomedical ontologies are typically very complex and comprehensive. This characteristic makes it difficult to use them straightforwardly in the target application due to efficiency reasons. At the same time, not all of the knowledge contained is relevant for the application context. In most cases, there is a specific set of ontology concepts and relations that sufficiently provide the required information. Using only parts of the ontology that are relevant for the application in mind allows for a significant improvement in the efficiency.

In MEDICO one use case focuses on patients suffering from lymphoma in the neck area. Lymphoma is a type of cancer in lymphocytes and it is a systematic disease with manifestations in multiple organs. During its diagnosis and treatment imaging is done several times based on the use of different imaging modalities (e.g. CT scan, X-Ray, MRI). This makes the lymphoma use case particularly relevant for a flexible medical image and text search application.

In this paper we describe our first ideas on an approach for modularizing large biomedical ontologies and illustrate it on the example of a large and comprehensive ontology about human anatomy. It is based on identifying statistically most relevant ontology concepts from domain corpora, in our case a corpus on lymphomas. We argue that these concepts can be used to distinguish the parts of ontologies that are most relevant for the application context. These parts can potentially yield the ontology modules that provide sufficient knowledge for the purposes of the software application. The modules identified in this way will additionally be discussed with the clinical experts for quality assessment and relevance.

The rest of this paper is structured as follows. Next section presents the related work. We then proceed with describing the relevant sources with a focus on the domain corpus and explain the statistical analysis process. In the Module Identification subsection we discuss the first (manually) identified modules supporting our ideas and assumptions. These are displayed in form of sub-hierarchies accessible through the UMLS tree browser. The paper concludes with first observations, discussions and future work.

RELATED WORK

In most cases the application scenario, the level of detail and the complexity of the medical knowledge determines the way how the modules should be identified. In other words, there is no well-defined or broadly accepted definition for the “one and only way” to modularize ontologies. On the contrary, many different approaches and techniques for ontology modularization have been implemented. [2, 3, 4] Most views agree that there is no universal way to

modularize ontologies and that the choice of a particular technique should be guided by the requirements of the considered application. Spaccapieta[5] and d'Aquin[6] provide a good overview of concepts and methods for achieving scalability through modularization of ontologies.

In general, ontology modularization can be addressed automatically or user-driven, but in both cases the modularization of the ontology is a challenging task. For example, ontology modularization approaches that guarantee logical consistency[7] may deliver too large fragments and can be slow in performance. On the other hand graph-based approaches[8] are more efficient but they do not guarantee the logical completeness. Finally, manually created ontology fragments[9] do naturally have the required level of granularity but they are expensive in terms of time and resources and are open to human errors.

The technique introduced in this paper has the objective to enable a semi-automatic identification of ontology modules and it does not prioritize completeness. Rather, we account for the practical applicability of the extracted modules to improve the efficiency of the application. Nevertheless, the extracted modules shall be discussed with clinical experts for quality assessment.

MATERIALS AND METHODS

Foundational Model of Anatomy (FMA)[10] ontology is the most comprehensive machine processable resource on human anatomy. It covers 71,202 distinct anatomical concepts (e.g., 'Neuraxis' and its synonym 'Central nervous system') and more than 1.5 million relations instances from 170 relation types. In addition to the hierarchical is-a relation, concepts are connected by seven kinds of part-of relationships (e.g., 'part of', 'constitutional part of', 'regional part of' etc.) The version we currently refer to is the version available in March 2009.

PubMed Lymphoma Corpus

The lymphoma corpus is based on medical publication abstracts on lymphoma from PubMed[11]scientific abstracts database. Its purpose is to provide specific domain knowledge about lymphoma, as this is one major use case of MEDICO. To establish the corpus we first extracted a set of the lymphoma relevant concepts from the NCI Thesaurus and then used these to identify from PubMed most frequently reported lymphomas, which are 'Non-Hodgkin's Lymphoma', 'Burkitt's Lymphoma', 'T-Cell Non-Hodgkin's Lymphoma', 'Follicular Lymphoma', 'Hodgkin's Lymphoma', 'Diffuse Large B-Cell Lymphoma', 'Aids Related Lymphoma', 'Extranodal Marginal Zone B-Cell Lymphoma of Mucosa-Associated Lymphoid Tissue', 'Mantle Cell Lymphoma', 'Cutenous T-Cell Lymphoma'. For each lymphoma type we compiled a set of XML documents that are generated from the PubMed abstracts. The text sections of the XML files were run through the TnT part-of-speech parser to extract all nouns and adjectives in the corpus.

The reason for including adjectives is based on our observations with the concept labels. Especially for the anatomy domain, the adjectives carry information that can be significant for medical decisions, for example, when determining whether an image is related to the *right* or to the *left* ventricle of the heart. Therefore, throughout the paper, when we talk about concepts, we refer to both adjectives and nouns. Then a relevance score (chi-square) for each noun and adjective was computed by comparing their frequencies in the domain specific corpus with those in the British National Corpus (BNC).[12] The resulting corpus consists of 71.973 files.

Statistical Analysis of Concepts

The objective of the statistical analysis is to identify a set of concepts that are most relevant for the application and for the use case. The process starts with converting the ontology into a flat list of concepts after some filtering is applied to the concept labels in the same way as explained in our previous work[13]. The statistically most relevant concepts are then identified on the basis of chi-square scores computed for nouns and adjectives. Ontology concepts that are single words and that occur in the corpus, correspond directly to the noun/adjective that the concept is build up of. For example, the noun 'ear' from the corpus corresponds to the FMA concept 'Ear'. Thus, the statistical relevance of the ontology concept is the chi-square score of the corresponding noun/adjective.

In the case of multi-word ontology concepts, the statistical relevance is computed on the basis of the chi-square score for each constituting noun and/or adjective in the concept name, summed and normalized over its length. Thus, relevance value for 'Lymph node', for example, is the summation of the chi-square scores for 'Lymph' and 'node' divided by 2. In order to take frequency into account, we further multiplied the summed relevance value by the frequency of the term. This assures that only frequently occurring terms are judged as relevant. A selection from the list of most relevant FMA concepts in the corpus is shown below (Table 1). We only focused on the Mantle Cell Lymphoma collection of the PubMed corpus, however currently we are extending the profiles to the rest.

FMA	Score
1. normal cell	240175,31
2. cell morphology	197495,31
3. stem cell	193389,88
4. plasma cell	190968,82
5. cell membrane	189984,02
6. cell surface	189981,54
7. lymphoid tissue	152765,58
8. lymph	99856,00
9. immunoglobulin	53361,00
10. inguinal lymph node	34943,38

Table 1. 10 most relevant FMA terms in the PubMed corpus

Identification of Potential Modules

Module identification starts with locating the statistically most relevant ontology concepts in the ontology hierarchy. The work reported in this paper was done manually. For the first experiments, we examined the context of the three concepts; 'Inguinal lymph node', 'Plasma cell' and 'Plasma membrane'. To locate the concepts in the hierarchy we used the UMLS Knowledge Server and selected the FMA view. We then searched for the three concepts using exact match. The hierarchical contexts are displayed below (Figure 1, Figure 2, Figure 3).

The locations of the 'Inguinal lymph node', 'Plasma cell' and 'Plasma membrane' in the FMA hierarchy display the 'Anatomical structure' as their next common parent. Therefore, it is most likely to be the root of the ontology module. The sum of the shortest paths from each concept (i.e. 'Inguinal lymph node', 'Plasma cell' and 'Plasma membrane') to 'Anatomical structure' will, in this case, be appended to it as its children. The sub-hierarchy consisting of 'Anatomical structure' as root and 'Cardinal organ part', 'Cell' and 'Cardinal cell part' as its children (and the children's descendants) may then be the potential ontology module. Consequently, the expectation from this module would be that it contains sufficient information about anatomy that relates to lymphoma.

FIRST OBSERVATIONS AND DISCUSSION

The concept labels reveal lexical overlaps. This suggests that further interrelations can be discovered by comparing the descendant concept labels at the lexical level. In this way, we expect to be able to find lexical correspondences that potentially convey further useful hierarchical information.

One drawback we have observed is that the ontology modules can be rather large. This means that it would be hard to identify the focus of the module. One possible way to avoid too large and too generic ontology modules may be by allowing only a certain number of concepts that were identified as statistically relevant and then by locating only these in the hierarchy. We currently investigate this.

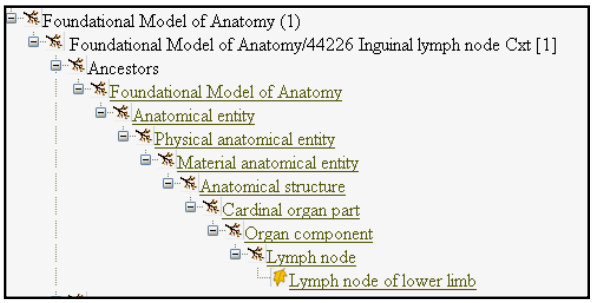

Figure 1. Hierarchical context of 'Inguinal lymph node' in the FMA view of the UMLS tree browser.

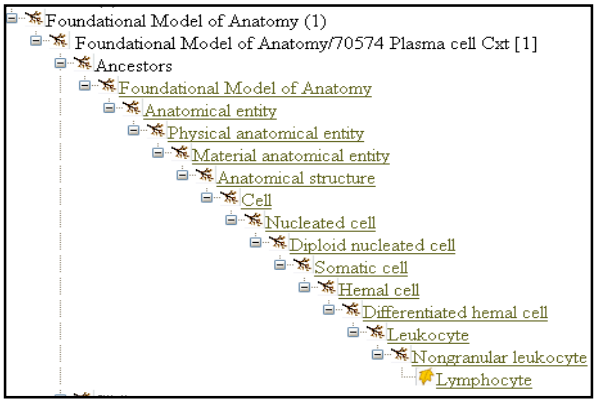

Figure 2. Hierarchical context of 'Plasma cell' in the FMA view of the UMLS tree browser.

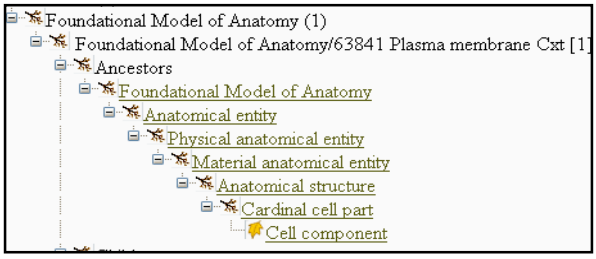

Figure 3. Hierarchical context of 'Plasma membrane' in the FMA view of the UMLS tree browser.

FUTURE WORK

As next the semi-automatic identification of the ontology modules will be realized. UMLS Knowledge Source Server and other tools from the National Library of Medicine[14] can support this process. Once this is achieved, it becomes relevant to identify the correspondences between the sub-

hierarchies. Lexical methods e.g., string similarity and overlap detection, can be used to discover correspondences between the concept labels. A long term but an important research question concerns finding an effective strategy to identify the optimal size for each module. It is essential to be able to determine when to terminate appending children to the module hierarchy. This is a challenging task, as optimal size, logical completeness and consistency usually require compromising.

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. We are also thankful to our clinical partner Dr. Alexander Cavallaro of the University Hospital Erlangen for his expert contribution.

References

1. UMLS Knowledge Source Server. Available online: http://umlsks.nlm.nih.gov/uPortal/
2. d'Aquin M. Schlicht A., Stuckenschmidt J. and Sabou M.: Ontology Modularization for Knowledge Selection: Experiments and Evaluations. 2007: In Proc. of the 18th Int. Conference.
3. Noy, N., Musen, M.: Specifying Ontology Views by Traversal. 2004: In: Proc. of the Int. Semantic Web Conference.
4. Stuckenschmidt, J., Klein, M.: Structure-Based Partitioning of Large Concept Hierarchies. 2004: In Proc. of the Int. Semantic Web Conference.
5. Spaccapieta, S.: Report on Modularization of Ontologies. 2005: Knowledge Web Deliverable 2.1.3.1.
6. d'Aquin, M., Sabou, M., Motta, E.: Modularization: a Key for the Dynamic Selection of Relevant Knowledge Components. 2006: In Proc. of the Workshop on Modular Ontologies.
7. Cuenca G.B., Horrocks I., Kazakov Y., Sattler U. Just the right amount: extracting modules from ontologies. In Proc. of the 16th international conference on WWW, 2007:717-726 NY, ACM.
8. Jimenez-Ruiz E., Berlanga R, Nebot V., Sanz I. Ontopath: A language for retrieving ontology fragments. In Meersman R. and Tari Z., (eds.), OTM Conferences (1) 2007;4803: 897-914.
9. S. Zillner, T. Hauer, D. Rogulin, A. Tsymbal, M. Huber, and T. Solomonides: "Semantic Visualization of Patient Information." 2008: In Proceedings of the 21th IEEE Int. Symposium on Computer-Based Medical Systems.
10. 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. 2008. Proc. of AMIA Symp:465-469.
11. PubMed Central. Available online: http://www.ncbi.nlm.nih.gov/pubmed/
12. British National Corpus. Available online: http://www.natcorp.ox.ac.uk/
13. Oezden Wennerberg, P., Buitelaar P., & Zillner S. *Towards a Human Anatomy Data Set for Query Pattern Mining Based on Wikipedia and Domain Semantic Resources*. 2008: In Workshop on Building and Evaluating Resources for Biomedical Text Mining at LREC.
14. National Library of Medicine. http://www.nlm.nih.gov/