Journal of Biomedical Informatics 44 (2011) 155–162

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Ontology modularization to improve semantic medical image annotation

Pinar Wennerberg [a,*], Klaus Schulz [b], Paul Buitelaar [c]

[a] *Siemens AG, Corporate Technology, Global Technology Field Knowledge Management, Munich, Germany*
[b] *Ludwig-Maximilians-University, Center for Information and Language Technology, Munich, Germany*
[c] *Unit for Natural Language Processing, Digital Enterprise Research Institute, National University of Ireland, Galway, Ireland*

ARTICLE INFO

Article history:
Available online 30 December 2010

Keywords:
Medical knowledge engineering
Ontology modularization

ABSTRACT

Searching for medical images and patient reports is a significant challenge in a clinical setting. The contents of such documents are often not described in sufficient detail thus making it difficult to utilize the inherent wealth of information contained within them. Semantic image annotation addresses this problem by describing the contents of images and reports using medical ontologies. Medical images and patient reports are then linked to each other through common annotations. Subsequently, search algorithms can more effectively find related sets of documents on the basis of these semantic descriptions. A prerequisite to realizing such a semantic search engine is that the data contained within should have been previously annotated with concepts from medical ontologies. One major challenge in this regard is the size and complexity of medical ontologies as annotation sources. Manual annotation is particularly time consuming labor intensive in a clinical environment. In this article we propose an approach to reducing the size of clinical ontologies for more efficient manual image and text annotation. More precisely, our goal is to identify smaller fragments of a large anatomy ontology that are relevant for annotating medical images from patients suffering from lymphoma. Our work is in the area of *ontology modularization*, which is a recent and active field of research. We describe our approach, methods and data set in detail and we discuss our results.

1. Introduction

We conceive of a semantic medical search engine which enables radiologists to find medical images, patient reports and scientific publications more efficiently. For details please refer to [8,9]. However, a prerequisite to realizing such a semantic search engine is the annotation of medical image and patient report data.

Semantic annotations consisting of concepts and relations from domain ontologies[1], encode information contained in images and texts at a fine level of granularity, thus allowing for specific and detailed searches. This makes it possible to find similar patients based on the similarity of their semantic annotations as well as images and reports which summarize the disease history of one particular patient. The work reported in this paper is conducted towards realizing this vision.

Medical ontologies, particularly those relating to human anatomy and radiology, are important sources for semantic annotation. Moreover, use of medical ontologies requires a systematic approach. Our experiences toward this end have shown us that certain issues arise that need to be addressed before the ontologies can be effectively used in realizing applications such as a semantic medical search engine. For example, many medical ontologies are too large and complex to be utilized efficiently, so that it becomes necessary to identify smaller ontology fragments.

This issue can be addressed by techniques that fall under the research area of *ontology modularization*. In earlier work we developed and reported on a medical knowledge engineering methodology which structures and coordinates separate but related activities involving medical ontologies that address these and other related issues (see [1,3,29]).

In this article, we present the techniques that we developed in identifying smaller fragments of large medical ontologies in order to address the size and complexity problems that are typical of medical ontologies.[2]

Ontology modularization can be seen as a specialized branch of knowledge engineering that is concerned with the development and maintenance of knowledge-based systems [31]. These systems typically include an ontology that contains specific domain knowledge. Ontology modularization as a research field is increasing in

* Corresponding author.
E-mail addresses: pinar.wennerberg.ext@siemens.com (P. Wennerberg), schulz@cis.uni-muenchen.de (K. Schulz), paul.buitelaar@deri.org (P. Buitelaar).

[1] In this article we adopt a generic understanding of ontology and use the term to refer to a variety of semantic sources such as controlled vocabularies, terminologies and taxonomies.

[2] We concentrated on statistical and structural techniques and worked only with hierarchical (taxonomical) is-a relationship.

1532-0464/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2010.12.005

popularity. This is particularly evident in the life sciences due to the extensive use of large and complex ontologies in this domain. In keeping with this trend, the research we present in the current study explores corpus based methods of ontology modularization.

In the present article we explain the need for ontology modularization with the help of a use case. We then present our approach to ontology modularization and discuss our own findings.The rest of this article is structured as follows: The next section explains the need for ontology modularization and presents a use case. Section 3 on Corpus based ontology modularization describes our approach with examples and reports on the results and evaluation. This is followed by Section 4 on Related work. The article concludes in Section 5 with Conclusions and outlook.

2. Need for ontology modularization

Clinical and life science ontologies are typically difficult to navigate or process efficiently due to their size and complexity. Furthermore, much of the knowledge they contain may only be relevant in a particular application context. In most cases, there is a specific set of ontology concepts and relations that sufficiently provide the required information needed for a given application. When describing a liver carcinoma, for example, generally only those concepts and relations from a disease ontology relating to liver diseases will be relevant.

This situation has consequences for the development of new software as well as for the human experts who manually annotate and navigate large medical and life science ontologies. In particular on three specific scenarios: (a) manual annotation of medical images by human experts, (b) exploration and navigation of large (medical and life science) ontologies by humans, and (c) computation of these ontologies by algorithms for different purposes such as reasoning (for examples of this approach please see [10,11,24–26]). The first two scenarios primarily have the common goal to support humans, whereas the second scenario concerns software applications.

In this work we concentrate on the first two scenarios as the third scenario imposes different requirements on the modules. If the modules are designed, for example, to support reasoning then it becomes important to account for various aspects; (a) the formalism that is used for the original ontology and that will be used for the modules, (b) logical completeness, and (c) logical consistency. In the current study we concentrate on supporting human experts that are expected to annotate large volumes of medical images and patient text.[3]

2.1. Modularization use case

For our use case, we consider a human expert who is responsible for annotating large volumes of medical images manually with the help of an annotation tool and an annotation ontology. The data will usually be annotated for a specific purpose, for example, the annotation of images that show symptoms of lymphoma. This is what we refer to as the *annotation context*. Using the annotation tool a radiologist retrieves images and reports and enters his observations, such as names of organs and abnormal structures, inside the images using the concepts from the annotation ontology.

Fig. 1 below displays the interface of a typical medical image annotation tool. Using the tool, medical image contents on the left hand side will be annotated with organ names and anatomy ontology concepts that are suggested on the right hand side. The annotator's job is to go through this long list of concepts and pick the ones that relate to the contents of the images being displayed. In this particular case, the images will be annotated for lymphoma. Hence, all concepts that partially match 'Lymph' (e.g. 'Lymph node', 'Set of lymph nodes', etc.) are suggested to the radiologist. The list of suggested concepts is comprehensive due to the size of the ontology and due to the naive concept matching strategy-in this case, any concepts relating to 'Lymph'. It is important to note that only concepts which match 'Lymph' in an anatomical sense are displayed. This explains why the concept 'Lymphoma', which relates to a disease, is not present in the anatomy ontology in this particular instance.

Ontology modularization supports the radiologist during manual annotation by presenting him with a subset of the anatomy ontology that is relevant for lymphoma. He selects the context, i.e. lymphoma, prior to the annotation process and the modularization algorithm returns the pre-computed lymphoma relevant fragments of the ontology.

3. Corpus based ontology modularization

The approach we propose here represents the context by a corpus of text documents that report on a specific disease or imaging modality (e.g. X-ray). In our example the context is a text corpus consisting of documents about various types of lymphomas or imaging modalities. Additionally, corpus based ontology modularization requires the ontology fragments to have been identified previously based on the domain corpora. For each new context, e.g., breast cancer, a new representative corpus must be created. The modules are then created by applying various linguistic and statistical techniques to the corpus. Once the modules are created, they are stored in a repository and displayed to the radiologist via the user interface.

The principle behind corpus based modularization is to incorporate empirical information to the theoretical view provided by an ontology. In other words, we understand a domain ontology as a theoretical view or as a certain perspective about a domain such as radiology. Consequently, our goal is to study how much the theory is represented by real data, where the data are the corpora. As a result, our objective is to assign every concept (theoretic) a statistical value (empiric) to display its representativeness in the domain relevant data. We have explained this approach also in earlier work [4].

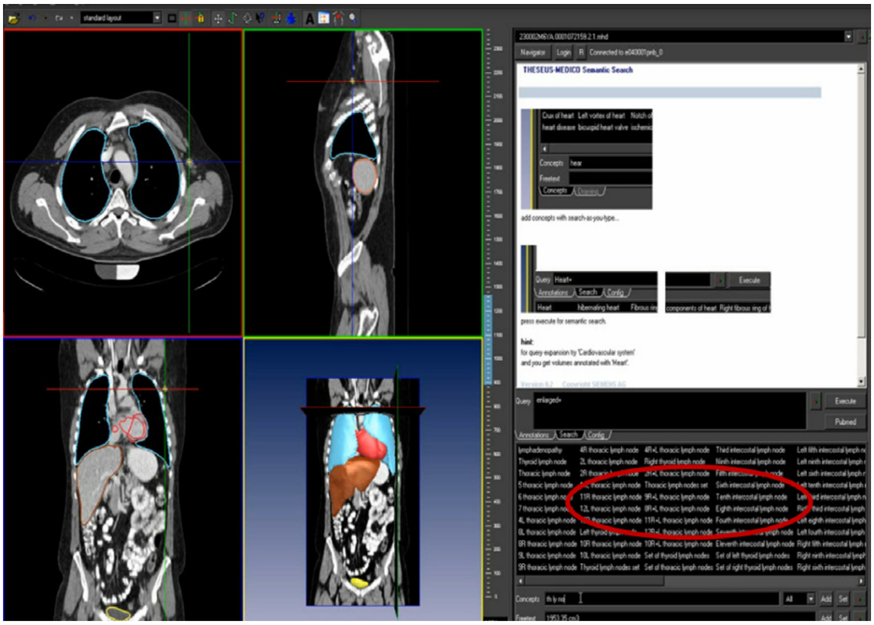
Corpus based modularization has two main components. The first component is concerned with the identification of the most context relevant concepts that compose the ontology module. Toward this end, we discuss two statistical approaches that we use in determining the most relevant concepts. The second component is concerned with the identification of actual modules from the ontology structure.

The optimal module is the smallest one that sufficiently delivers the relevant information required to accomplish a task. For example, if our task is the annotation of lymphoma images and related patient data, then the optimal module would be the smallest portion of the anatomy ontology that includes all the concepts related to lymphoma.

The core assumption common of our approach is that the application relevant ontology modules can be identified by corpus analysis. Therefore, in order to determine their context relevance, we compute corpus statistics for ontology concepts. The first approach is based on a χ^2 analysis, whereas the second approach is more complex and uses Eigenvectors to overcome the shortcomings of the first approach.

Modules essentially consist of concepts that are statistically the most context relevant as well as other concepts that are hierarchically related to them. In other words, if 'Left lung' is statistically relevant then 'Lung' is also as relevant as its parent concept and

[3] Manual annotation of large volumes of images is not only common in the medical domain. This scenario is also valid in large industrial settings such as automotive or construction industry where large volumes of product texts and images need to be annotated for subsequent information management.



Connected to e040001prb_0

Crux of heart Left vortex of heart Notch of
heart disease bicuspid heart valve ischemic

Concepts: hear

Freetext:

add concepts with search-as-you-type...

Query: Heart+

Heart hibernating heart Fibrous rin

components of heart Right fibrous ring of

press execute for semantic search.

hint:
for query expansion try 'Cardiovascular system'
and you get volumes annotated with 'Heart'.

Version 0.2 Copyright SIEMENS AG

Query: enlarged+

lymphadenopathy	4R thoracic lymph node	4R+L thoracic lymph node	Third intercostal lymph node	Left fifth intercostal lymph n
Thyroid lymph node	2L thoracic lymph node	Right thyroid lymph node	Ninth intercostal lymph node	Left ninth intercostal lymph
Thoracic lymph node	2R thoracic lymph node	2R+L thoracic lymph node	Fifth intercostal lymph node	Left sixth intercostal lymph
5 thoracic lymph node	[illegible] thoracic lymph node	Thoracic lymph nodes set	Sixth intercostal lymph node	Left tenth intercostal lymph
6 thoracic lymph node	11R thoracic lymph node	9R+L thoracic lymph node	Tenth intercostal lymph node	Left [illegible] intercostal lymph n
7 thoracic lymph node	12L thoracic lymph node	8R+L thoracic lymph node	Eighth intercostal lymph node	Right third intercostal lymph
4L thoracic lymph node	[illegible] thoracic lymph node	11R+L thoracic lymph node	Fourth intercostal lymph node	Left eighth intercostal lymph
8L thoracic lymph node	Left thyroid [illegible]	12R+L thoracic lymph node	Seventh intercostal lymph node	Left fourth intercostal lymph
8R thoracic lymph node	10R thoracic lymph node	10R+L thoracic lymph node	Eleventh intercostal lymph node	Right fifth intercostal lymph
9L thoracic lymph node	10L thoracic lymph node	Set of thyroid lymph nodes	Set of left thyroid lymph nodes	Right ninth intercostal lymph
9R thoracic lymph node	Thyroid lymph nodes set	Set of thoracic lymph nodes	Set of right thyroid lymph nodes	Right sixth intercostal lymph

Concepts: th ly no

Freetext: 1953.35 cm3

Fig. 1. Medico image annotation tool.

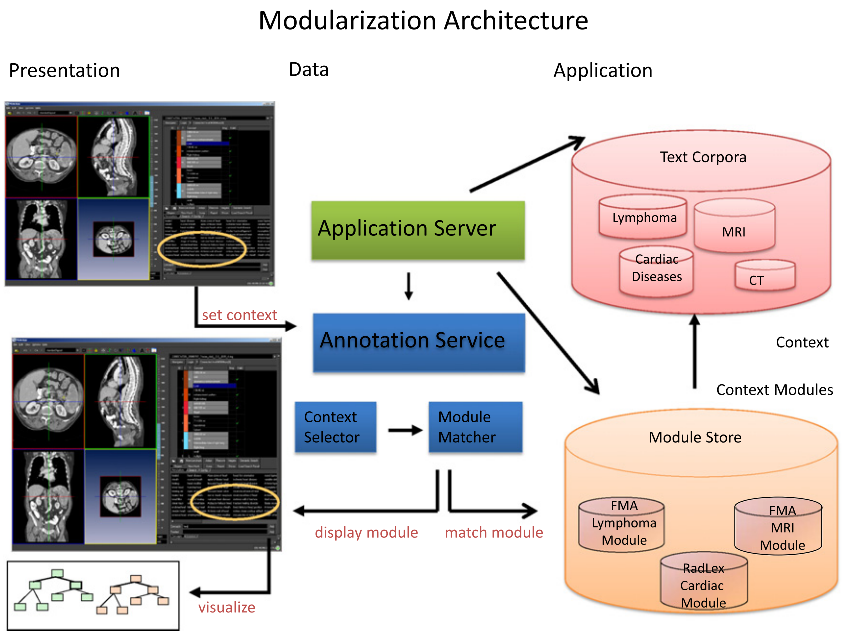


Fig. 2. Modularization architecture. The radiologist sets the annotation context, e.g. lymphoma, which retrieves the corresponding pre-computed lymphoma module from the module library and presents it to the radiology as the relevant annotation subvocabulary.

will therefore be included in the module. Likewise, if 'Lymph node' is relevant then 'Peripancreatic lymph node' is also potentially relevant as it is its direct child.

Thus, the resulting set of concepts and hierarchical relations that comprise the subset of the original ontology are presented to the radiologist as the context relevant module. In this way the radiologist has a significantly smaller number of concepts to choose from for annotation purposes. Fig. 2 displays the architecture of our corpus based modularization approach. Context modules are computed a priori based on statistical methods applied to domain corpora. When radiologist selects the context for example lymphoma, the related module is retrieved from the module library and is displayed to the annotator.

3.1. Semantic sources and corpora

As our modularization approach is based on corpus analysis, we constructed a domain corpus for our experiments. Additionally, we used a publicly available, domain-independent corpus that reflects general English language use. We applied two statistical analysis techniques to these corpora to determine the most relevant concepts for annotation.

3.1.1. PubMed lymphoma corpus (BNC)

This corpus is based on medical publication abstracts on lymphoma research from the PubMed[4] scientific abstracts database. Its purpose is to provide domain-specific information about lymphoma. We concentrated on the five most commonly reported lymphoma types: 'Non-Hodgkin's Lymphoma', 'Burkitt's Lymphoma', 'T-Cell Non-Hodgkin's Lymphoma', 'Hodgkin's Lymphoma' and 'Diffuse Large B-Cell Lymphoma'. For each lymphoma type we compiled a set of XML documents generated from the PubMed abstracts. Thus, the resulting corpus consists of 12.865 XML files in total.

3.1.2. The British national corpus

This is a 100 million word collection of samples of written and spoken British English.[5] The latest edition is the BNC XML Edition released in 2007. We used the written part of the BNC that includes samples from regional and national newspapers, academic books, specialist periodicals and journals for all ages and interests.

3.1.3. Foundational model of anatomy (FMA)

This is the most comprehensive machine processable resource on human anatomy. The FMA is developed by the University of Washington and the US National Library of Medicine. It covers 71,202 distinct anatomical concepts and more than 1.5 million relations instances from 170 relation types [5].

3.2. Identifying relevant concepts based on χ^2 statistics

We applied a χ^2 test to our corpora to identify the domain specific, i.e., lymphoma related, relevance of FMA concepts, relative to distributions in the lymphoma corpus. The motivation behind our use of a χ^2 analysis is to determine the statistical significance of each FMA concept based on its frequency in a domain corpus (PubMed Lymphoma) relative to that in a general corpus (BNC). This is in line with the approach taken by Kastrin and Hristovski [32]. The underlying assumption is that domain relevant concepts occur more often in a domain corpus than in a general corpus.

Assumption 1. Domain relevant concepts occur more often in a domain corpus than in a general corpus.

The most statistically relevant concepts are identified on the basis of χ^2 scores computed for nouns and adjectives as these lexical items are the most information bearing. More precisely, the χ^2 scores for nouns and adjectives were computed by comparing their frequencies in the domain specific corpus with those in the general corpus. Formulas (1) and (2) show the computation of the χ^2 scores,

$$\chi^2 = \frac{(O_G - E_G)^2}{E_G} + \frac{(O_C - E_C)^2}{E_C} \tag{1}$$

where E_G and E_C are expected frequencies and O_G and O_C are observed frequencies, respectively. The expected frequencies are calculated as follows:

$$E_G = \frac{N_G(O_G + O_C)^2}{N_G + N_C} E_C = \frac{N_C(O_G + O_C)^2}{N_G + N_C} \tag{2}$$

where N_G and N_C are the total frequency of FMA concepts in a generic (i.e. BNC) and context (PubMed lymphoma) corpus, respectively.

Ontology concepts that consist of single words and that occur in the corpus correspond directly to the noun or adjective of which the concept is built. For example, the noun 'ventricle' from the corpus corresponds to the FMA concept 'Ventricle'. So, the statistical relevance of the ontology concept is the χ^2 score of the corresponding noun/adjective. For multi-word ontology concepts, the statistical relevance is computed on the basis of the χ^2 score for each constituting noun and/or adjective in the concept name, summed and normalized over its length. Thus the relevance value for 'Left ventricle', for example, is the sum of the respective χ^2 scores for 'Left' and 'ventricle' divided by 2.

Table 1 displays a selection of FMA concepts with their frequencies in the Lymphoma vs. BNC corpus. Table 2 lists the most lymphoma relevant FMA concepts identified as a result of this process. In Table 1 we can see that the concept 'differentiation' occurs more often in the Lymphoma corpus than in the BNC corpus.

3.2.1. Results and discussion

One drawback we have observed with χ^2 ranking is that the identified concepts tend to be generic high level concepts. This eventually leads to rather large and less focused ontology modules. 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. However, this approach may be too simplistic because it does not select the concepts based on their information content. So, it would be possible to exclude concepts that would have been relevant.

We investigated another statistical approach that allows for a more focused and strategic selection of ontology concepts. The more specific concepts are again located in the ontology hierarchy. In the next section we report on our experiments with a statistical approach based on Eigenvector values to select more domain focused ontology concepts that will be used to identify the potential modules.

3.3. Relevant concepts based on HITS algorithm

Our goal is to reduce the module size by using a more strictly selected and therefore more domain-specific set of ontology concepts. We expect that modules resulting from this set of concepts will be more domain focused. We adopt a statistical approach based on Eigenvectors along the lines of Seidel [34], which is an

Table 1
FMA concept frequencies (number of occurrences) in the PubMed Lymphoma vs. BNC corpus.

FMA concept	BNC	PubMed lymphoma
cage	1100	5
zone	3223	1591
basis	14420	663
differentiation	912	1148
border	5028	30

Table 2
Ten most lymphoma relevant FMA concepts in the PubMed Lymphoma corpus according to their χ^2 scores.

FMA concept	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

[4] http://www.ncbi.nlm.nih.gov/pubmed/

[5] http://www.natcorp.ox.ac.uk/.

adaptation of the Hypertext Induced Topic Search (HITS) algorithm by Kleinberg [33].

Seidel's goal was to identify a domain specific vocabulary on the basis of compound words (e.g., 'Iliac lymph node') as these are good representatives of a given domain. We on the other hand view an ontology as a domain specific vocabulary. Proceeding on this assumption, we argue that all ontology concepts are good domain representatives as they are specific to topics such as anatomy or radiology. Subsequently, our goal is to identify the most specific ontology concepts according to a given context such as lymphoma.

While χ^2 ranking processed the PubMed lymphoma corpus as a whole, here we break it down to its sub-corpora. Thus, the PubMed lymphoma corpus becomes a collection of its five parts containing text about five different lymphoma types. Consequently, each PubMed file is classified under one sub-corpus. No file is left unclassified. In this way we achieve a finer level of granularity in terms of categorization so that the resulting modules become more specific yielding one module for each lymphoma type.

Each text in each sub-corpus is represented in terms of the ontology concept labels it contains. Also here, we work with ontology concept labels[6] that contain up to four words. Multiple occurrences of one concept are kept. For example, 1513068.xml in the Burkitt's Lymphoma sub-corpus is represented in terms of FMA concept labels:

Cell
Maxillary sinus
Maxillary sinus
Orbit

In other words, 1513068.xml contains theFMA concepts 'Cell', 'Maxillary sinus' and 'Orbit', whereby 'Maxillary sinus' occurs two times. Subsequently, each file in each corpus is represented in this way. Hence, Burkitt's Lymphoma sub-corpus for example, consists of 1258 concepts × 19169 texts including the multiple occurrences of a concept label.

The adapted version of the HITS algorithm starts with building an adjacency matrix over PubMed abstracts (documents) from sub-corpora and ontology concept labels (terms). If a concept c occurs in document d its entry value a_{cd} in the matrix is 1. Otherwise, it is 0. For example, matrix A below shows the representation of concepts 'Cell', 'Skin' and 'Bone marrow' in texts '1.xml', '2.xml' and '3.xml'.

'Cell' ⇒ '1.xml', '2.xml'
'Skin' ⇒ '2.xml', '3.xml'
'Bone marrow' ⇒ '3.xml'

$$\underset{\downarrow}{\text{documents}}$$

$$\underset{\rightarrow}{\text{terms}}$$

$$A:\begin{pmatrix}1&1&0\\0&1&1\\0&0&1\end{pmatrix}$$

Subsequently, each concept label that occurs in a document receives a weight i.e., a_{wd} determined after a number of iterations until the algorithm reaches the optimal point. This weight eventually represents the domain relevance of its related concept. For a more detailed explanation of the algorithm please refer to Seidel [34]. Each file in each sub-corpus (e.g. Burkitt's Lymphoma) is subject to the algorithm in this way.

Table 3
Ten highest ranking FMA concepts for Burkitt's Lymphoma according to adapted HITS.

FMA concept	Score
1. Cell	1,000000
2. Line	0,287337
3. Gene	0,181194
4. Surface	0,131989
5. Protein	0,131047
6. Immunoglobulin	0,101872
7. Blood	0,090374
8. Genome	0,087076
9. Membrane	0,085013
10. Bone	0,059269

The algorithm results in lists containing new rankings for FMA concepts. More precisely, there is a list of newly ranked FMA concepts for example for sub-corpus 'Burkitt's Lymphoma'. There are five lymphoma ranked lists for FMA for the five lymphoma types: 'Burkitt's Lymphoma', 'Hodgkin's Lymphoma', 'Non-Hodgkin's Lymphoma', 'Diffuse Large B-Cell Lymphoma', 'T Cell Non-Hodgkin Lymphoma'. Each list consists of the concept labels from the respective ontology with an associated numeric ranking value. Table 3 displays the list with 10 FMA concepts that are highly relevant for Burkitt's Lymphoma.

3.3.1. Results and discussion

Having examined all the lists that were output by the algorithm, we observed that a specific set of concepts with high scores occurred in every list. These are the so called *stop-concepts*, which are not discriminative enough, where the discrimination criterion is the spatial aspect. For example, such concepts as 'Cell', 'Gene', 'Protein', 'Tissue', 'Artery', 'Body', 'Blood' occur almost non-exclusively in every list and they are everywhere in the body. This makes it difficult to conclude that marking a region in an image and annotating it as a tissue implies annotating the thorax or abdomen. In contrast, annotating the heart in the same way would indeed imply marking a location in the thorax as the heart is located in the thorax. A stop-concept list for FMA with the lymphoma corpus contains $StopConceptsFMA = \{Cell, Gene, Protein, Line, Median, Surface, \ldots\}$, etc.

Further, the resulting ranked lists have been presented to a clinical expert for evaluation and additional selection. We have explained the methods we used to collect expert feedback in earlier work [2]. According to the expert's assessment, all concepts that have been identified by the algorithm as relevant for lymphoma have actually been relevant. Additionally, he confirmed that concepts with higher ranks were more relevant than those with lower ranks.

One problem that the expert reported concerned the selection of the lymphoma types. These lymphomas were closely related to each other so that it was difficult for the expert to find discriminative criteria that significantly distinguish one type from the other. As a consequence, the resulting five ranked lists for five lymphoma types contained overlapping ontology concepts. This problem can however be overcome by including other types of diseases with more discriminative features. Table 4 displays the list with 10 FMA concepts as highly relevant for Burkitt's and Hodgkin lymphomas after stop-concept elimination and expert assessment.

3.4. Identifying modules

A list of ranked ontology concepts is the first step towards identifying the modules. However, we are also interested in the information that is conveyed by the ontology structure. Hence, we proceed with identifying the ontology subgraphs(or subtrees) on the basis of the selected concepts. We start by locating the two highest ranking concepts from the list in the ontology hierarchy.

[6] We use the terms concepts and concept labels interchangeably, while statistical corpus analysis essentially works with concept labels.

Table 4
Ten expert selected FMA concepts for Burkitt's and Hodgkin lymphomas.

Burkitt's Lymphoma	Score	Hodgkin Lymphoma	Score
1. Immunoglobulin	0,101872	Lymphocyte	0,107332
2. Chromosome	0,100269	Lymph node	0,057859
3. Genome	0,087076	Solid	0,023740
4. Lymphocyte	0,058284	Spleen	0,014441
5. Serum	0,056749	Cell phenotype	0,014242
6. Bone marrow	0,055911	Lymphoid tissue	0,012644
7. Leukocyte	0,021819	Neck	0,010120
8. B lymphocyte	0,015393	Mediastinum	0,009339
9. Cell phenotype	0,014848	Abdomen	0,005737
10. Lymph node	0,012465	B lymphocyte	0,003783

When entering the ontology hierarchy, our first assumption is that the parents and ancestors of these concepts are also domain relevant.

Assumption 2. Parents and ancestors of the domain relevant concepts are also domain relevant.

Their next common parent, i.e. the common denominator, is treated as the module root, while the sum of the shortest paths from each concept to the common parent is appended to it as its children. Further, those paths to the root concept are excluded that are too generic to be sufficiently informative. For example, such concepts as 'Anatomical entity' or 'Anatomical structure' are in almost every concept's ancestor hierarchy and they do not have the sufficient level of granularity.

In this way the following concepts have been excluded from almost every concept ancestor hierarchy: 'Non-physical anatomical entity', 'Physical anatomical entity', 'Immaterial anatomical entity', 'Material anatomical entity', 'Anatomical set', 'Anatomical structure', 'Acellular anatomical structure', 'Anatomical cluster'. After the removal, some solitary concepts remain that have no ancestors such as 'Genome' or 'Serum', or that have only one like 'Head'. These are also eliminated.

For example, as shown in Table 5 below, the three concepts 'Lymphocyte', 'Leukocyte' and 'B Lymphocyte' share the first common denominator concept 'Differentiated hemal cell'. This is marked as the root of the module. We then move upwards in the hierarchy until we find the most commonly occurring denominator. In this case, 'Nucleated cell' occurs three times as ancestors of three different concepts, while the rest of the ancestors occur only once. Consequently, we reduce the module to those three concepts sharing the ancestors, in this case 'Lymphocyte', 'Leukocyte' and 'B Lymphocyte' together with their ancestors. Solitary concepts ('Genome', 'Serum', etc.) as seen in the table are also removed.

Table 5
FMA concepts with their ancestors after too generic concepts have been removed. Solitary concepts are removed as next.

FMA concept	Ancestors
Immunoglobulin	Biological macromolecule, Protein
Chromosome	Cardinal cell part, Cell component
Genome	–
Lymphocyte	Cell, Nucleated cell, Somatic cell, Hemal cell, Differentiated hemal cell, Leukocyte, Nongranular leukocyte
Serum	–
Bone marrow	Cardinal organ part, Organ region, Organ zone, Zone of bone organ
Leukocyte	Cell, Nucleated cell, Somatic cell, Hemal cell, Differentiated hemal cell
B lymphocyte	Cell, Nucleated cell, Somatic cell, Hemal cell, Differentiated hemal cell, Leukocyte, Nongranular leukocyte, Lymphocyte
Cell phenotype	–
Lymph node	Cardinal organ part, Organ component
Spleen	Organ, Solid organ, Parenchymatous organ, Corticomedullary organ
Head	Cardinal body part
Abdomen	Subdivision of cardinal body part, Subdivision of body proper, Subdivision of trunk, Subdivision of trunk proper
Cervical lymph node	Cardinal organ part, Organ component, Lymph node

Having worked with the ancestors, we proceed with the children and descendants. As lower levels of the hierarchy include more specific concepts, our expectation is to increase the context specificity of the modules and thus provide more focus.

Assumption 3. Children and descendants of the domain relevant ontology concepts are also domain relevant.

The list of concepts for which we would like to see the children and descendants include only the three concepts 'Lymphocyte', 'Leukocyte' and 'B Lymphocyte' as explained earlier. Most of the time the number of descendants that one concept has is several times higher than that of the ancestors as they may have numerous direct children (fan out), each of which may have a long descendant path before arriving at the leaf. Starting with a reduced number of concepts from the beginning helps to address this problem. Table 6 shows a portion of the resulting FMA module for Burkitt's Lymphoma.

The resulting module for Burkitt's Lymphoma contains 'Cell' as its root, and is followed by 'Nucleated cell', 'Somatic cell', 'Hemal cell' as the first, second and third level descendants, respectively. The module additionally includes all the descendants of 'Lymphocyte', 'Leukocyte' and 'B Lymphocyte' and terminates with leaves such as 'T helper cell type 1', 'T helper cell type 1'. 'B Lymphocyte' which has no descendant concepts itself remains as a leaf. In total there are 37 unique concepts.Five modules for five lymphoma types have been identified in this way and have been discussed with the clinical experts. According to their assessment the modules included all the concepts that were related to the context. Nevertheless, the level of granularity was coarse, meaning that the modules required a higher level of specificity.

We will turn back to this problem and discuss it in more detail in the final section of this article. However, at this point we would like to mention that the problem arises more because of the textual characteristics of PubMed articles and is less related to the methods used.

4. Related work

Various groups from industry and academia have been promoting ontology engineering methodologies [13]. Some of these are based on experiences from projects [12,14–16,30], whereas others

Table 6
A partial list of FMA concepts with their descendants.

Series FMA concept	Series descendants
Lymphocyte	Null lymphocyte, Medium sized lymphocyte, Thymocyte B lymphocyte, Small lymphocyte, Large lymphocyte, Plasma cell T lymphocyte, Delayed type hypersensitivity-related T lymphocyte, Suppressor T lymphocyte, . . . , T helper cell type 1, T helper cell type 2
Leukocyte	Granular leukocyte, Basophil, Eosinophil, Neutrophil, Nongranular leukocyte, Peripheral blood mononuclear cell, Monocyte Lymphoblast, T lymphoblast, B lymphoblast, Lymphocyte, Null lymphocyte, Medium sized lymphocyte, Thymocyte B lymphocyte, Small lymphocyte, Large lymphocyte, Plasma cell, Natural killer cell, Lymphokine-activated natural killer cell, . . .
B Lymphocyte	–

are stand-alone cross-domain engineering approaches [16,17], that use different formalisms, e.g. [28] and tools [27].

Regarding ontology modularization, current approaches may be generalized as being semantics-driven and structure-driven. The former centers on a strategy for identifying modules that fulfill a specific subset of the application requirements and can meaningfully stand on their own (e.g.,[18–20]). The latter concentrates on the ontology as a graph and uses graph partitioning algorithms to extract the most tightly interconnected nodes (i.e. concepts) irrespective of their semantics (e.g., [21]). For a recent and comprehensive overview of this field refer to [22]. Other related research concerns approaches towards discovering relationships across different clinical domain ontologies, which are discussed and evaluated by Johnson and colleagues [6]. These approaches, or others towards achieving the same goal [7,23] can potentially be applied to ontology modules. The modularization technique introduced in this paper aims for semi-automatic identification of ontology modules on the basis of the analysis of a domain corpus as context. Further work relating to medical image annotation and mark-up is explained by Rubin and colleagues.

5. Conclusions and outlook

As we have mentioned earlier one problem with corpus (context) based modularization is the coarse granularity level of the resulting modules. We have been able to increase the specificity of the modules and therefore improve the focus by adapting a more selective statistical approach to identify relevant concepts. Even though this helped improve the results over those obtained by the χ^2 approach, this problem still remains.

We attribute this problem to the nature of the PubMed scientific abstracts, which mainly report on research issues instead of practical clinical knowledge such as observations, diagnoses and pathology as found in actual patient reports. Therefore, the concepts that occur in the PubMed abstracts are significantly different, or more superficial, from the perspective of medical image annotation.

Another step towards increasing the specificity, and therefore the focus of the modules is to use other type of data such as radiology reports or discharge summaries that include more practical, real-life related information. Toward this end, we obtained a large data set on clinical patient reports (radiology, discharge summaries, cardiology reports, etc.) from the University of Pittsburgh (BlueLab)[7] and conducted first experiments. Our expectation is to identify more specific ontology concepts (and their variants) occurring in this kind of text. In parallel, we are working on adapting the PageRank (i.e., Google) algorithm (which HITS underlies) to restrict the statistical concept selection further.

Finally, an important but long term 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 one must usually find a compromise between optimal size, logical completeness and consistency.

Acknowledgments

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 grant number 01MQ07016 and by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2). The responsibility for this publication lies with the authors.

[7] http://www.dbmi.pitt.edu/blulab/nlprepository.htm

References

[1] Wennerberg P, Zillner S, Moeller M, Buitelaar P, Sintek M. KEMM: a knowledge engineering methodology in the medical domain. In: Proceedings of the 5th international conference on formal ontology in information systems (FOIS); 2008.
[2] Wennerberg P, Zillner S, Buitelaar P. Interactive clinical query derivation and evaluation. In: AAAI 2009 spring symposium on technosocial predictive analytics (TPAI), Stanford, San Francisco USA; 2009.
[3] Buitelaar P, Wennerberg P, Zillner S. Statistical term profiling for query pattern mining. In: Proceedings of ACL 2008 BioNLP workshop; 2008.
[4] Wennerberg P, Buitelaar P, Zillner S. Towards a human anatomy data set for query pattern mining based on Wikipedia and domain semantic resources. In: Proceedings of a workshop on building and evaluating resources for biomedical text mining (LREC); 2008.
[5] Rosse C, Mejino J. A reference ontology for biomedical informatics: the foundational model of anatomy. J Biomed Inform 2003;36(6):478–500.
[6] Johnson HL, Cohen K, Baumgartner JW, Lu Z, Bada M, Kester T. Evaluation of lexical methods for detecting relationships between concepts from multiple ontologies. In: Pacific symposium on biocomputing; 2006. p. 28–39.
[7] Pinto H, Martins J. Ontology integration: how to perform the process. In: Proceedings of international joint conference on artificial intelligence; 2001.
[8] Moeller M, Sintek M, Buitelaar P, Mukherjee S, Zhou X, Freund J. Medical image understanding through the integration of cross-modal object recognition with formal domain knowledge. In: Proceedings of Healthinf; 2008.
[9] Moeller M, Regel S, Sintek M. Radsem: semantic annotation and retrieval for medical images. In: Proceedings of the 6th annual European semantic web conference (ESWC), Hersonissos, Crete; 2009.
[10] Wittekind C, Meyer H, Bootz F. TNM Klassifikation maligner Tumoren. Heidelberg: Springer Medizin Verlag; 2005.
[11] Dameron O, Roques E, Rubin D, Marquet G, Burgun A. Grading lung tumors using OWL-DL based reasoning. In: Proceedings of 9th international protege conference; 2006.
[12] Uschold M. Building ontologies: towards a unified methodology. In: 16th annual conference of the British Computer Society Specialist Group on expert systems, Cambridge, UK; 1996.
[13] Jones D, Bench-Capon T, Visser P. Methodologies for ontology development. In: Proceedings of IT& KNOWS conference, XV IFIP world computer congress, Budapest; August 1998.
[14] Lopez MF, Gomez-Perez A, Juristo N. Methontology: from ontological art towards ontological engineering. In: Proceedings of AAAI97 spring symposium series, workshop on ontological engineering; 1997. p. 3340.
[15] Gruninger M, Fox MS. The design and evaluation of ontologies for enterprise engineering. In: Proceedings of the workshop on implemented ontologies, European conference on artificial intelligence (ECAI); 1994.
[16] Sure Y, Studer R. To-knowledge methodology – final version. Institute AIFB, University of Karlsruhe, On-to-knowledge deliverable; 18, 2002.
[17] Schreiber G, Akkermans H, Anjewierden A, Dehoog R, Shadbolt N, Vandevelde W, et al. Knowledge engineering and management: the CommonKADS methodology. The MIT Press; 1999.
[18] Rector A. Modularisation of domain ontologies implemented in description logics and related formalisims including OWL. In: Proceedings of the K-Cap 2003 conference; 2003. p. 121–8.
[19] Parent C, Spaccapietra S. An overview of modularity. In: Stuckenschmidt H, Parent C, Spaccapietra S, editors. Modular ontologies. Springer; 2009. p. 5–23.
[20] Cuenca G, Horrocks I, Kazakov Y, Sattler U. Just the right amount: extracting modules from ontologies. In: Proceedings of the 16th international conference on WWW; 2007. p. 717–26.
[21] Stuckenschmidt H, Schlicht A. Structure-based partitioning of large ontologies. In: Stuckenschmidt H, Parent C, Spaccapietra S, editors. Modular ontologies. Springer; 2009. p. 5–23.
[22] Stuckenschmidt H, Parent C, Spaccapietra S. Modular ontologies concepts, theories and techniques for knowledge modularization. Springer; 2009.
[23] Mungall C. Obol: integrating language and meaning in bio-ontologies. Comp Functional Genomics 2004;5:509.
[24] Marquet G, Dameron O, Saikali S, Mosser J, Burgun A. Grading glioma tumors using OWL-DL and NCI-Thesaurus. In: Proceedings of the American Medical Informatics Association conference AMIA; 2007.
[25] C. Goldbreich, O. Dameron, O. Bierlaire, B. Gibaud, What reasoning support for ontology and rules? The brain anatomy case study. In: Proceedings of the Workshop on OWL experiences and directions; 2005.
[26] Zillner S, Hauer T, Rogulin D, Tsymbal A, Huber M, Solomonides T. Semantic visualization of patient information. In: Proceedings of the 21th IEEE international symposium on computer-based medical systems (CBMS); 2008.
[27] Horridge M, Bechhofer S. The OWL API: a Java API for working with OWL 2 ontologies. In: Proceedings of the 6th OWL experienced and directions workshop (OWLED2009); 2009.
[28] Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider PF. The description logic handbook: theory, implementation, and applications. In: Description logic handbook. Cambridge University Press; 2003.
[29] Wennerberg P, Zillner S, Schulz K. Context driven modularization of a representation of human anatomy. In: Proceedings of the AMIA 2009 summit on translational bioinformatics (AMIA), San Francisco, USA; 2010.

[30] Schreiber G, Akkermans H, Anjewierden A, Dehoog R, Shadbolt N, Vandevelde W, et al. Knowledge engineering and management: the CommonKADS methodology. The MIT Press; 1999.

[31] Kendal SL, Creen M. An introduction to knowledge engineering. Springer; 2007.

[32] Kastrin A, Hristovski D. A fast document classification algorithm for gene symbol disambiguation in the BITOLA literature-based discovery support system. In: AMIA 2008 annual symposium. Washington (DC): American Medical Informatics Association; 2008.

[33] Kleinberg J. Authoritative sources in a hyperlinked environment. J ACM 1999;46(5):604–32.

[34] Seidel C. Christian Seidel – System zur dynamischen Erfassung domaenenspezifischer Texteinheiten und Texte mit Hilfe von Eigenvektormethoden. Muenchen: Verlag Dr. Hut; 2010.