Aligning Medical Domain Ontologies for Clinical Query Extraction

Pinar Wennerberg
Siemens AG, Munich Germany
TU Darmstadt, Darmstadt Germany
pinar.wennerberg.ext@siemens.com

Abstract

Often, there is a need to use the knowledge from multiple ontologies. This is particularly the case within the context of medical imaging, where a single ontology is not enough to provide the complementary knowledge about anatomy, radiology and diseases that is required by the related applications. Consequently, semantic integration of these different but related types of medical knowledge that is present in disparate domain ontologies becomes necessary. Medical ontology alignment addresses this need by identifying the semantically equivalent concepts across multiple medical ontologies. The resulting alignments can then be used to annotate the medical images and related patient text data. A corresponding semantic search engine that operates on these annotations (i.e. alignments) instead of simple keywords can, in this way, deliver the clinical users a coherent set of medical image and patient text data.

1 Introduction

As the content of numerous ontologies in the biomedical domain increases, so does the need for sharing and reusing this body of knowledge. Often, there is a need to use the knowledge from multiple ontologies. This is particularly the case within the context of medical imaging, where a single ontology is not enough to support the necessary heterogeneous tasks that require complementary knowledge about human anatomy, radiology and diseases. Medical imaging constitutes the context of this work, which lies within the Theseus-MEDICO[1] use case.

The Theseus-MEDICO use case has the objective of building the next generation of intelligent, scalable, and robust search engine for the medical imaging domain. MEDICO's proposed solution relies on ontology based semantic annotation of the medical image contents and the related patient data.

Semantic annotation of medical image contents and patient text data allows for a mark-up with meaningful meta-information at a higher level of granularity that goes beyond simple keywords. Therefore, the data which is processed and stored in this way can be efficiently retrieved by a corresponding search engine such as the one envisioned in MEDICO.

The diagnostic analysis of medical images typically concentrates around three questions (a) what is the anatomy here? (b) what is the name of the body part here? (c) is it normal or is it abnormal? Therefore, when a radiologist looks for information, his search queries most likely contain terms from various information sources that provide this kind of knowledge.

To satisfy the radiologist's information need, this scattered knowledge has to be gathered and integrated from disparate ontologies, in particular from those about human anatomy, radiology and diseases. Subsequently, the medical image contents and the related patient data have to be annotated with this information (i.e. ontology concepts and relationships) rather than the single elements from independent ontologies.

Three ontologies that address the three questions above are relevant to gather the necessary knowledge about human anatomy, radiology and diseases. These are the Foundational Model of Anatomy[2] (FMA), Radiology Lexicon[3] (RadLex) and the Thesaurus of the National Cancer Institute[4] (NCI), respectively.

[1] http://theseus-programm.de/scenarios/en/medico

[2] http://sig.biostr.washington.edu/projects/fm/FME/index.html

[3] http://www.rsna.org/radlex

[4] http://nciterms.nci.nih.gov/NCIBrowser/Connect.do?dictionary=NCI_Thesaurus&bookmarktag=1

Proceedings of the EACL 2009 Student Research Workshop, pages 79–87,
Athens, Greece, 2 April 2009.

Given this context, the semantic integration of these ontologies as knowledge sources becomes critical. Ontology alignment addresses this requirement by identifying semantically equivalent concepts in multiple ontologies. These concepts are then made compatible with each other through meaningful relationships. Hence, our goal is to identify the correspondences between the concepts of different medical ontologies that are relevant to the medical image contents.

The rest of this paper is organized as follows. In the next section we explain the motivation behind aligning the medical ontologies. Section 3 discusses related work in ontology alignment in general and in the biomedical domain. In section 4 we introduce our approach and explain why it goes beyond existing methods. Here we also explain the application scenario, which exhibits how aligned medical ontologies can contribute to the identification of relevant clinical search queries. Section 5 introduces the materials and methods that are relevant for this work. Finally 6 and 7 discusses the planned evaluation and presents the roadmap for the remaining work, respectively

2 Motivation

The following scenario illustrates how the alignment of medical ontologies facilitates the integration of medical knowledge that is relevant to medical image contents from multiple ontologies. Suppose that we want to help a radiologist, who searches for related information about the manifestations of a certain type of lymphoma on a certain organ, e.g. liver, on medical images. As discussed earlier the three types of knowledge that serves him would be about the human anatomy (liver), the organ's location in the body (e.g. upper limb, lower limb, neighboring organs etc.) and whether what he sees is normal or abnormal (pathological observations, symptoms, and findings about lymphoma).

Once we know what the radiologist is looking for we can support him in his search in that we present him an integrated view of only the liver lymphoma relevant portions of the patient health records (or of *that* patient's record), PubMed abstracts as reference resource, drug databases, experience reports from other colleagues, treatment plans, notes of other radiologists or even discussions from clinical web discussion boards.

From the NCI Thesaurus we can obtain the information that '*liver lymphoma*' is the synonym for *'hepatic lymphoma'*, for which holds:

'hepatic lymphoma'
'disease_has_primary_anatomic_site'
'liver'
'hematopoietic and lymphatic system'
'gastrointestinal system'

With this information we can now move on to the FMA to find out that *'hepatic artery'* is a *part of* the *'liver'* (such that any finding that indicates lymphoma at the *hepatic artery* would also imply the lymphoma at the *liver*). RadLex on the other hand informs that *'liver surgery'* is a '*treatment*' '*procedure'*. Various types of this '*treatment*' '*procedure'* are '*hepatectomy'*, '*hepatic lobectomy'*, '*hepatic segmentectomy'*, '*hepatic subsegmentectomy'*, '*hepatic trisegmentectomy'* or '*hepatic wedge excision'*, which can be used for disease treatment.

Consequently, the radiologist who searches for information about liver lymphoma is presented with a set of patient health records, PubMed abstracts, radiology images etc. that are annotated using the terminology above. In this way, the radiologist's search space is reduced to a significantly small portion of the overdose of information available in multiple data stores. Moreover, he receives coherent data, i.e. images and patient text data that are related to each other, from a single access point without having to login to several different data stores at different locations.

3 Related Work

Ontology alignment is commonly understood as a special case of semantic integration that concerns the semi-automatic discovery of semantically equivalent concepts (sometimes also relations) across two or more ontologies.

There are two commonly adopted approaches to ontology alignment; schema-based and instance-based, where most systems use both. Accordingly, the input of the former approach is the ontology schema only, whereas the input of the latter is the instance data i.e. the data that have been annotated with the ontology schema. Both approaches take advantage of linguistic and graph-based methods to help identify the correspondences. The most recent and comprehensive overview of work ontology alignment in general is reported by Euzenat and Shvaiko (2007).

Ontology alignment is an increasingly active research field in the biomedical domain, especially in association with the Open Biomedical

Ontologies (OBO)[5] framework. The OBO consortium establishes a set of principles to which the biomedical ontologies shall conform to for purposes of interoperability. The OBO conformant ontologies, such as the FMA, are available at the National Center for Biomedical Ontology (NCBO) BioPortal[6].

Johnson *et al.* (2006) take an information retrieval approach to discover relationships between the Gene Ontology (GO) and three other OBO ontologies (ChEBI[7], Cell Type[8] and BRENDA Tissue[9]). Here, GO ontology concepts are treated as documents, they are indexed using Lucene[10] and are matched against the search queries, which are the concepts from the other three ontologies. Whenever a match is found, it is taken as an evidence of a correspondence. This approach is efficient and easy to implement and can therefore be successful with large medical ontologies. However, it does not account for the complex linguistic structure typically observed at the concept labels of the medical ontologies, which may result in inaccurate matches.

The focus of the work reported by Zhang *et al.* (2004) is to compare two different alignment approaches that are applied to two different ontologies about human anatomy. The subject ontologies are the FMA and the Generalized Architecture for Languages, Encyclopedias and Nomenclatures for Medicine[11] (GALEN). Both approaches use a combination of lexical and structural matching techniques, however one of them additionally employs an external resource (the Unified Medical Lexicon UMLS[12]) to obtain domain knowledge. In this work the authors point to the fact that medical ontologies contain implicit relationships, especially in the multiword concept names that can be exploited to discover more correspondences. This thesis builds on this finding and investigates further methods, e.g. the use of transformation grammars, to discover the implicit information observed at concept labels of the medical ontologies.

On the medical imaging side, there are activities that concentrate around ImageClef[13] campaign, which concerns the cross-language image retrieval and which runs as a part of the Cross-Language Evaluation Forum (CLEF)[14] on multilingual information access. Here, the Medical Annotation and the Medical Retrieval tasks benchmark systems on efficient annotation and retrieval of medical images. However, these activities are organized taking an information retrieval and image parsing perspective and do not focus on semantic information integration. Nevertheless, the campaign releases valuable imaging and text data that can be used.

4 Approach and Contributions

Here, we describe our approach for the alignment of medical ontologies and outline the contributions of this thesis. In this respect, we first specify the general requirements for medical ontology alignment, which are then addressed by our approach. These are followed by the statement of the hypotheses of this work. Secondly, the materials that are relevant for this work are introduced. In particular, we describe the semantic resources and our domain corpora. Finally, an application scenario is described that exhibits the benefits of aligning medical ontologies. We describe this scenario as '*Clinical Query Extraction*' and explain the idea behind.

4.1 Requirements for medical ontology alignment

Drawing upon our experiences with the medical ontologies along the MEDICO use case we have identified some of their common characteristics that are relevant for the alignment process. These can be summarized as:

1. Generally, they are very large models.
2. They have extensive *is-a* hierarchies up to ten thousands of classes, which are organized according to different views.
3. They have complex relationships, where classes are connected by a number of different relations.
4. Their terminologies are rather stable (especially for anatomy) in that they should not differ much in the different models.
5. The modeling principles for them are well defined and documented.

Based on these characteristics and the general requirements of the MEDICO use case, we de-

[5] http://www.obofoundry.org/
[6] http://www.bioontology.org/ncbo/faces/index.xhtml
[7] www.obofoundry.org/cgi-bin/detail.cgi?id=chebi
[8] www.obofoundry.org/cgi-bin/detail.cgi?id=cell
[9] www.obofoundry.org/cgi-bin/detail.cgi?id=brenda
[10] http://lucene.apache.org/java/docs/
[11] http://www.opengalen.org
[12] http://www.nlm.nih.gov/research/umls/
[13] http://imageclef.org
[14] http://www.clef-campaign.org/

rived the following requirements specifically for aligning medical ontologies:

Linguistic processing: Medical ontologies are typically linguistically rich. For example, the FMA contains concept names as long as '*Anastomotic branch of right anterior inferior cerebellar artery with right superior cerebellar artery*'. Such long multi-word terms are usually rich with implicit semantic relations. This characteristic shall be exploited by an intensive use of linguistic alignment methods.

Use of external resources: As we are in a specific domain (medicine) and as we are not domain experts, we are in lack of domain knowledge. This missing domain knowledge shall be acquired from external resources, for example UMLS. Synonymy information in this resource and in other terminological resources is of particular interest.

Non-machine learning approach: We do not have access to much instance data. This is partly because we are domain dependent. A more important reason, however, is that the special resource, the patient health records, which would provide a large amount of relevant instance data is very difficult to obtain due to legal issues. Therefore, machine learning approaches, which require large portions of training data are not the optimal approach for our purposes.

Structural matching: Medical ontologies typically come with rich structures that go beyond the basic *is-a* hierarchy. Most of them include a hierarchical ordering along the *part-of* hierarchies. Ontologies such as FMA additionally have part-of classification with higher granularity that include relations such as **'***constitutional part-of*', '*systemic part-of*' *etc.* This rich structure of the medical ontologies shall be used to validate (or improve) the alignments that have been obtained as a result of the linguistic processing and the lexical matching.

Sequential matching: Medical ontologies are complex, so that their automatic processing is usually expensive. Therefore, a target concept will be identified (this target concept/term will be in practice the search query of the clinician. More details are explained under section 6.2) First lexical matching techniques shall be applied to identify the search query relevant parts of the ontologies. In other words, those concepts that lexically match the query shall be aligned as first. In this way, the lexical match acts as a filter on the medical ontology and decreases the amount of the computation necessary.

4.2 Assumptions

Given this context, we focus on the evaluation of the following hypotheses:

1. Valid relationships (equivalence or other) exist between concepts from FMA, RadLex and from NCI.
2. Relationships between non-identical concept labels from the three ontologies can be discovered if these have common reference in a more general medical ontology.
3. Concept labels in these ontologies are most often in the form of long natural language phrases with regular grammars. Meaningful relationships (e.g. synonymy) across the three ontologies can be derived by processing these labels using transformation grammars.
4. Identification of medical image related query patterns (i.e. a certain combination of concept labels and relations) from corpora is more efficient when it is done based on the alignments.

4.3 Approach

The ontology alignment approach proposed in this thesis has three main aspects. It suggests a combinatory strategy that is based on (a) the linguistic analysis of the ontology concept labels (the linguistic aspect), (b) on corpus analysis (context information aspect) and (c) on human-computer interaction e.g. relevance feedback (user interaction aspect).

The linguistic aspect draws on the observation that concept labels in medical ontologies (especially those about human anatomy) often contain implicit semantic relations as discussed by Mungall (2004), e.g. equivalence. By observing common patterns in the multi-word terms that are typical for the concept labels of the medical ontologies these relations can be made explicit.

Transformation grammars can help here to detect the syntactic variants of the ontology concept labels. In other words, with the help of rules, the concept labels can be transformed into semantically equivalent but syntactically different word forms. For example, one concept label from the FMA and its corresponding commonly observed pattern (in brackets) is:

'Blood in aorta' (noun preposition noun)

Using a transformation rule of the form,

noun1 preposition:'in' noun2 => noun2 noun1

we can generate a variant as below with the equivalent semantics:

'aorta blood' (noun noun)

This is profitable for at least two reasons. Firstly, it can help resolve possible semantic ambiguities (if one variant is ambiguous the other one can be preferred). Secondly, identified variants can be used to compare linguistic (textual) contexts of ontology concepts in corpora leading to the second aspect of our approach.

Subsequently, the second aspect, the corpus analysis, builds on comparing linguistic (textual) contexts of ontology concepts in corpora and it assumes that concepts with similar meaning (originating from different ontologies) will appear in similar linguistic contexts. Here, the linguistic context of an ontology class (e.g. *'terminal ileum'* from the FMA as in the example below) can be defined as the document in which it appears, the sentence in which it appears and a window of size N in which it appears. For example, a window size -5, +5 for the FMA concept *"terminal ileum"* would be:

*'Focal lymphoid hyperplasia of the **terminal ileum** presenting mantle zone hyperplasia with clear cytoplasm'*

can be represented as a vector in form of:

<token -5, token -4, … , token +4, token +5>
<focal, lymphoid, hyperplasia, of, the, presenting, mantle, zone, hyperplasia, with>

These vectors can then be pairwise compared, where most similar vectors indicate similar meaning of corresponding ontology concepts and alignment between ontology concepts follows from this.

Finally, with the user interaction aspect we understand dynamic models of the ontology integration process. Within this dynamic process the ontology alignment happens during an interactive dialogue between the user and the system. In this way, clarifications and questions that elicit user's feedback support the ontology alignment process. An example interactive dialogue can be:

(1) **Radiologist**: Show me the images of Ms. Jane Doe, she has "Amyotrophic Lateral Sclerosis" (*NCI Cancer Thesaurus concept*)

(2) **System**: Ms. Doe doesn't have any images of "Amyotrophic Lateral Sclerosis". Is it equivalent to "Lou Gehrig Disease" (*equivalent NCI Cancer Thesaurus concept*) or to "ALS" (*equivalent RadLex concept*)? That attacks the neurons i.e. the nerve cells (*FMA concept*) Stephan Hawkins has it.

(3) **Radiologist**: Yes, that is true.

(4) **System** Ok. ALS is a kind of "Neuro Degenerative Disorder" (*super-concept from RadLex*) Do you want to see other images on Neuro Degenerative Disorders?

This dialogue illustrates a real life question answering dialogue; where the utterances (2) and (4) contain the system questions, and utterance (3) is the user's interactive mapping feedback. This aspect is based on the approach explained in more detail in (Sonntag, 2008).

5 Materials and Methods

5.1 Terminological resources

Foundational Model of Anatomy (FMA) is the most comprehensive machine processable resource on human anatomy. It covers 71,202 distinct anatomical concepts and more than 1.5 million relations instances from 170 relation types. The FMA can be accessed via the Foundational Model Explorer[15].

FMA also provides synonym information (up to 6 per concept), for example one synonym for *'Neuraxis'* is the *'Central nervous system'*. Because single inheritance is one of the modeling principles used in the FMA, every concept (except for the root) stands in a unique *is-a* relation to other concepts. Additionally, concepts are connected by seven kinds of part-of relationships (e.g., *part of, constitutional part of, regional part of*). The version we currently refer to is the version available in August 2008.

The **Radiology Lexicon (RadLex)** is a controlled vocabulary developed and maintained by the Radiological Society of North America (RSNA) for the purpose of uniform indexing and retrieval of radiology information, including images. RadLex contains 11962 terms related to anatomy pathology, imaging techniques, and diagnostic image qualities. RadLex terms are organized along several relationships hence several hierarchies. Each term will participate in one of the relationships with its parent. Synonym information is given whenever it is present such as

[15] http://fme.biostr.washington.edu:8089/FME/

in *'Schatzki ring'* and *'lower esophageal mucosal ring'*. Examples of radiology specific relationships are *'thickness of projected image'* or *'radiation dose'*.

The **National Cancer Institute Thesaurus (NCI)** provides standard vocabularies for cancer research. It covers around 34.000 concepts from which 10521 are related to Disease, Abnormality, Finding, 5901 are related to Neoplasm, 4320 to Anatomy and the rest are related to various other categories such as Gene, Protein, etc. The ontology model is structured around three components i.e. Concepts, Kinds and Roles. Concepts are represented as nodes in an acyclic graph, Roles are directed edges between the nodes and they represent the relationships between them. Kinds on the other hand are disjoint sets of concepts and they constrain the domain and the range of the relationships. Each concept belongs to only one Kind. Except for the root concept, every other concept has at least one *is-a* relationship to another concept.

Every concept has one preferred name (e.g., *'Hodgkin Lymphoma'*). Additionally, 1,207 concepts have a total of 2,371 synonyms (e.g., *Hodgkin Lymphoma* has synonym *'Hodgkin's Lymphoma'*, *'Hodgkin's disease'* and *'Hodgkin's Disease'*). The version we currently refer to is the version in June 2008 (08.06d).

5.2 Data

The **Wikipedia anatomy, radiology** and **disease corpora** have been constructed based on the Anatomy[16], Radiology[17] and Diseases[18.] sections of the Wikipedia. Patient records would be the first choice, but due to strict anonymization requirements they are difficult to compile. Therefore, as an initial resource we constructed the corpora based on the Wikipedia.

To set up the three corpora the related web pages were downloaded and a specific XML version for them was generated. The text sections of the XML files were run through the TnT part-of-speech parser (Brants, 2000) to extract all nouns in the corpus. Then a relevance score (chi-square) for each noun was computed by comparing anatomy, radiology and disease frequencies respectively with those in the British National Corpus (BNC)[19.] In total there are 1410 such XML files about human anatomy, 526 about disease, and 150 about radiology.

The **PubMed lymphoma corpus** is set up to target the specific domain knowledge about lymphoma, a special type of cancer (one major use case of MEDICO is lymphoma). Thus, the lymphoma relevant subterminology from the NCI Thesaurus was extracted. This subterminology includes information about lymphoma types, their relevant thesaurus codes, synonyms, hyperonyms (or parent terms) and the corresponding thesaurus definitions.

Using the lymphoma terminology, we identified from PubMed an initial set of most frequently reported lymphomas, e.g. the top five is *'Non-Hodgkin's Lymphoma'*, *'Burkitt's Lymphoma'*, *'T-Cell Non-Hodgkin's Lymphoma'*, *'Follicular Lymphoma'*, and *'Hodgkin's Lymphoma'* in that order. The lymphoma corpus currently consists of XML files about two main lymphoma types i.e. *'Mantle Cell Lymphoma'* and for *'Diffuse Large B-Cell Lymphoma'*. The former includes 1721 files and the latter 111.

The **clinical questions corpus** consists of health related questions asked among the medical experts and that were collected during a scientific survey. These questions (without answers) are available through the Clinical Questions Collection[20] online repository. It can either be searched or browsed, for example, by a specific disease category. An example question from the Clinical Questions Collection is *"What drugs are folic acid antagonists?"* For each question, additional information about the expert asking the question, e.g. time, purpose etc. are encoded.

To create the clinical questions corpus we downloaded the categories Neoplasms as well as Menic and Lymphatic Diseases from the Clinical Questions Collection website. For each existing HTML page that reports on a question, we created a corresponding XML file. Currently there are 796 questions our questions corpus.

The **clinical discussions corpus** is ongoing work and it will be a corpus, whose contents will be compiled from the various clinical discussion boards across the Web. These discussion boards usually contain questions and answers between and among the medical experts and patients. We expect the language to be less technical because of the user profile. The purpose of this corpus is to have a resource of clinical questions together

[16] http://en.wikipedia.org/wiki/Category:Anatomy
[17] http://en.wikipedia.org/wiki/Category:Radiology
[18] http://en.wikipedia.org/wiki/Category:Diseases
[19] The BNC (http://www.natcorp.ox.ac.uk/) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English.
[20] http://clinques.nlm.nih.gov/JitSearch.html

with their answers as well as experience reports, links to other useful resources in a less technical language. We have already identified a set of relevant clinical discussion boards and analyzed their contents and structure.

6 Evaluation Strategies

We distinguish between two kinds of evaluation techniques that can be applied to assess the quality of the alignments.

Direct evaluation methods compare the results relative to human judgments as explained by Pedersen *et al.* (2007), which in our case would be the assessment and the resulting feedback of the clinical experts. This kind of evaluation, however, is not very realistic in our context due to the unavailability of a representative number of clinical experts.

Indirect evaluation methods, on the other hand, consider the performance of an application that uses the alignments. Hence, any improvement in the performance of the application when it uses the alignments can be attributed to the quality of the alignments. In the following two subsections we first describe the baseline and then explain the planned application that shall use the alignments. The performance of this application, with and without the alignments, will be taken as a measure on the quality of these alignments.

6.1 Baseline and Comparison to Other Systems

Our baseline for comparison is string matching after normalization on the concept labels from the input ontologies. Survey results (van Hage and Aleksovski, 2007) suggest that this method is currently the simplest and the most intuitive method being used for ontology alignment (or similar) tasks. Thus, the results of our matching approach will be in the first place compared with the results of this simple matching strategy.

The Ontology Alignment Evaluation Initiative[21] (OAEI) offers a service evaluate the alignment results for its participant matching systems. The competing systems are evaluated on consensus test cases at four different tracks. The evaluation at the anatomy track, which is the most relevant one for us, has been done either by comparing the systems' resulting alignments to reference alignments (absolute comparison) or to each other (relative comparison).

[21] http://oaei.ontologymatching.org

6.2 Clinical Query Extraction

We conceive of the clinical query extraction process as a use case that shows the benefits of semantic integration by means of ontology alignments.

Clinical query extraction, (Oezden Wennerberg *et al*., 2008; Buitelaar *et al*., 2008) is the process of predicting patterns for typical clinical queries given domain ontologies and corpora. It is motivated by the fact that when developing search systems for healthcare professionals, it is necessary to know what kind of information they search for in their daily working tasks. As interviews with clinicians are not always possible, alternative solutions become necessary to obtain this information.

Clinical query extraction is a technique to semi-automatically predict possible clinical queries without having to depend on clinical interviews. It requires domain corpora (i.e. disease, anatomy and radiology) and domain ontologies to be able to process statistically most relevant concepts in the ontologies and the relations that hold between them. Consequently, concept-relation-concept triplets are identified, for which the assumption is that the statistically most relevant triplets are more likely to occur in clinical queries.

Clinical query extraction can be viewed as a special case of term/relation extraction. Related approaches from the medical domain are reported by Bourigault and Jacquemin (1999) and Le Moigno *et al*. (2002).

The identification of query patterns (i.e. the concept-relation-concept triplets) starts with the construction of domain corpora from related Web resources such as Wikipedia[22] and PubMed[23]. As next, use case relevant parts from domain ontologies are extracted. The frequency of the concepts from the extracted sub-ontologies in the domain corpora versus the frequencies in a domain independent corpus determines the domain specificity of the concepts.

This statistical term/concept profiling can be viewed as a function that takes the domain (sub)ontologies and the corpora as input and returns the partially weighted domain ontologies as output, where the terms/concepts are ranked according to their weights. An example query pattern can look like:

[22] http://www.wikipedia.org/

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

[ANATOMICAL STRUCTURE]	located_in	[ANATOMICAL STRUCTURE]
	AND	
[[RADIOLOGY IMAGE]Modality]	is_about	[ANATOMICAL STRUCTURE]
	AND	
[[RADIOLOGY IMAGE]Modality]	shows_ symptom	[DISEASE SYMPTOM]

The clinical query extraction approach, as illustrated so far, builds on using domain ontologies, however on using them independently. That is, the entire statistical term profiling is based on processing the use case relevant terms (i.e. concepts) of the ontologies in isolation. In this respect the clinical query pattern extraction is a good potential application that can be used to evaluate the quality of the ontology alignments.

As the current process is based on single concepts, the natural extension will be to perform the extraction based on aligned concepts. Any improvement in the identification of the query patterns from corpora can then be attributed to the quality alignments.

7 Future Directions

Regarding the linguistic aspect of the ontology alignment approach, the next step will be to concentrate on the definition of the transformation grammar to generate the semantic equivalent concepts.

A further consideration is to explore whether other relations beyond synonymy such as hyponymy or hyperonymy can also be generated and whether this is profitable. To accord for the second aspect, the most suitable vector model will be determined and tested and applied on the current corpora. As required by the third, user interaction aspect, a dialogue that is most representative of a real life use case will be modeled.

Currently, some of the existing alignment frameworks, e.g. COMA++[24] or PhaseLibs[25] are being tested for their performance with FMA, RadLex and NCI. The observations on the strengths and the weaknesses of these systems will give more insights for the requirements for our system.

Other tasks that are relevant for achieving the goal of this thesis concentrate on two main topics; the collection and the preparation of data and the evaluation of the alignment approach. Subsequently, the clinical questions corpus will be expanded and will be used to evaluate the clinical query patterns. As explained earlier, the efficient identification of the clinical query patterns based on the alignments will be regarded as one means to assess the performance of the alignment approach. Parallels, a complementary corpus compiled from relevant clinical discussion boards will be prepared for the same purpose.

As required by the linguistic aspect of our approach an initial grammar will be set up and be continuously improved to detect the variants of the ontology concepts labels from the three ontologies mentioned earlier. Transformation rules will be used for this purpose.

The open question about whether the ontology relations shall also be aligned will be investigated to determine the trade-offs of including vs. excluding them from the process. We consider using an external resource such as UMLS to obtain background knowledge that can help resolve possible semantic ambiguities. The appropriateness and adoptability of this resource will be assessed. Finally, the evaluation the overall ontology alignment approach will be carried out, whereby a possible participation the OAEI may also be considered.

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 the grant number 01MQ07016. The responsibility for this publication lies with the authors. Special thanks to Prof. Dr. Iryna Gurevych of TU Darmstadt, to Daniel Sonntag and Paul Buitelaar of DFKI Saarbrücken, and to Sonja Zillner of Siemens AG for fruitful discussions. Additionally, we are thankful to our clinical partner Dr. Alexander Cavallaro of the University Hospital Erlangen.

References

Bourigault D and Jacquemin C, 1999: *Term extraction + term clustering: An integrated platform for computer-aided terminology*, in Proceedings EACL-99.

Buitelaar P., Oezden Wennerberg P., Zillner S., 2008: *Statistical Term Profiling for Query Pattern Mining*. In:Proc. of ACL 2008 BioNLP Workshop (ACL'2008). Columbus, Ohio, USA, 19 June 2008.

[24] http://dbs.uni-leipzig.de/Research/coma.html
[25] http://phaselibs.opendfki.de/

Euzenat J, Shvaiko P., 2007: *Ontology Matching*. Springer-Verlag; Juni 2007

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.* Pac. Symp Biocomput, pp. 28-39, 2006 American Psychological Association. 1983. Publications Manual. American Psychological Association, Washington, DC

Le Moigno S., Charlet J., Bourigault D., Degoulet P., and Jaulent M-C, 2002: *Terminology extraction from text to build an ontology in surgical intensive care*. AMIA, Annual Symposium, 2002. 9-13. USA

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

Oezden Wennerberg P, Buitelaar P, Zillner S, 2008: *Towards a Human Anatomy Data Set for Query Pattern Mining based on Wikipedia and Domain Semantic Resources*. In:Proc. of a Workshop on Building and Evaluating Resources for Biomedical Text Mining (LREC'2008). Marrakech, Marocco, 26 May 2008.

Pedersen T, Pakhomov S.V., Patwardhan S and C.G. Chute, (2007): *Measures of semantic similarity and relatedness in the biomedical domain*, Journal of Biomedical Informatics, vol. In Press, Corrected Proof.

Sonntag D, 2008. *Towards dialogue-based interactive semantic mediation in the medical domain* In Third International Workshop on Ontology Matching at ISWC, 2008

van Hage W.R, Isaac A, Aleksovski A (2007): *Sample Evaluation of Ontology-Matching Systems*. EON 2007: 41-50

Zhang S, Mork P, Bodenreider O, 2004: *Lessons learned from aligning two representations of anatomy* In: Hahn U, Schulz S, Cornet R, editors. Proceedings of the First International Workshop on Formal Biomedical Knowledge Representation (KR-MED 2004); 2004. p. 102-108