Deriving Clinical Query Patterns from Medical Corpora Using Domain Ontologies

Pinar Oezden Wennerberg †
Siemens AG, CT IC1
Otto-Hahn-Ring 6,
81739 Munich, Germany
pinar.wennerberg.ext@siemens.com

Paul Buitelaar
DERI - National University of Ireland,
IDA Business Park, Lower Dangan,
Galway, Ireland
paul.buitelaar@deri.org

Sonja Zillner
Siemens AG, CT IC1
Otto-Hahn-Ring 6,
81739 Munich, Germany
sonja.zillner@siemens.com

Abstract

For an effective search and management of large amounts of medical image and patient data, it is relevant to know the kind of information the clinicians and radiologists seek for. This information is typically represented in their queries when searching for text and medical images about patients. Statistical clinical query pattern derivation described in this paper is an approach to obtain this information semi-automatically. It is based on predicting clinical query patterns given medical ontologies, domain corpora and statistical analysis. The patterns identified in this way are then compared to a corpus of clinical questions to identify possible overlaps between them and the actual questions. Additionally, they are discussed with the clinical experts. We describe our ontology driven clinical query pattern derivation approach, the comparison results with the clinical questions corpus and the evaluation by the radiology experts.

Keywords

Medical ontology, information extraction, biomedical corpora, information management, medical imaging.

1. Introduction

Due to advanced technologies in clinical care, increasingly large amounts of medical imaging and the related textual patient data becomes available. To be able to use this data effectively, it is relevant to know the kind of information the clinicians and radiologists seek for. This information is typically represented in the search queries that demonstrate the information needs of radiologists and clinicians. Our context is the MEDICO use case, which has a focus on semantic, cross-modal image search and information retrieval in the medical domain. Our objective is to identify the kind of queries the clinicians and radiologists use to search for medical images and related textual data. As interviews with clinicians and radiologists are not always possible, alternative solutions become necessary to obtain this information. We aim to discover radiologists' and clinicians' information needs by using semi-automatic text analysis methods that are independent of expert interviews.

One MEDICO[1] scenario concentrates on image search targeting patients that suffer from lymphoma in the neck area. Lymphoma, a type of cancer occurring in lymphocytes, is a systematic disease with manifestations in multiple organs. During lymphoma diagnosis and treatment, imaging is done several times using different imaging modalities (X-Ray, MR, ultrasound etc.), which makes a scalable and flexible image search for lymphoma particularly relevant. As a result of intensive interviews with radiologists and clinicians, we learned that medical imaging data is analyzed and queried based on three different dimensions. These are the anatomical dimension, i.e. knowledge about human anatomy, the radiology dimension, i.e. the medical image specific knowledge and the disease dimension that describes the normal and the abnormal anatomical and imaging features. Therefore, our objective is to predict clinical query patterns related to these three dimensions.

Ontology based clinical query derivation approach we describe is a technique to semi-automatically predict possible clinical queries without having to depend on clinical interviews. It requires domain corpora (i.e. about disease, anatomy and radiology) and the corresponding domain ontologies to be able to process statistically most relevant terms (concepts)[2] from the ontologies and the relations that hold between them. Consequently, term-relation-term triplets are identified, for which the assumption is that the statistically most relevant triplets are more likely to occur in clinical queries. An example query of the radiologist can be *"All CT scans and MRIs of patient X with an enlarged lymph node in the neck area"*, which may have a corresponding query pattern as:

Concept	***relation***	**Concept**
[[RADIOLOGY IMAGE]Modality]	*is_about*	[ANATOMICAL STRUCTURE]
	AND	
[[RADIOLOGY IMAGE]Modality]	*shows_symptom*	[DISEASE/ SYMPTOM]

Once the statistically most relevant concepts and relations (i.e. query patterns) from the domain ontologies

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

[2] Throughout this paper, we do not semantically differentiate between 'term' and 'concept', but use these expressions interchangeably.

Workshop Biomedical Information Extraction 2009 - Borovets, Bulgaria, pages 50–56

are identified, they are compared against a corpus of actual clinical questions to discover overlaps. Additionally, they are presented to the experts for evaluation. The contribution of this paper is to describe these two tasks, i.e. the clinical query derivation approach and the comparison to the clinical questions corpus. We also report on the assessment of the clinical experts. The rest of this paper is organized as follows. Next section discusses related work. Then materials and methods used are introduced and the clinical query derivation approach is explained in detail. This is followed by the discussion of the results of comparing the query patterns with the clinical questions corpus. The clinical experts' assessment is reported followed by conclusion and future directions.

2. Related Work

Clinical query derivation can be viewed as a special case of term-relation extraction. Related approaches from the medical domain are reported by Bourigault and Jacquemin [2] and Le Moigno et al. [9] which, however, are independent of medical image semantics.

Price and Delcambre [11] propose to model the clinical queries as binary relations on query topics (e.g. relation (topic1, topic2)). The relations in the queries are then matched against relations in the documents. In their extended model [12] the 'semantic components', which are terms and expressions characteristic for certain types of documents, are used as arguments to the same query relations (e.g. relation(semantic component1, semantic component 2)). Later, the semantic components are used as mediators to map two Web-based document collections to certain generic clinical query patterns [13].

In our work, we also share the view of representing clinical queries as concept-relation patterns. The major difference is, however, the distinct goals and the techniques used. The semantic components model is developed for an improved medical information retrieval scenario, where for any given relation the goal is to identify medical text documents relevant to clinical questions with an optimal ranking. Our goal, however, is to be able to discover those relations as we assume that they will take us to the actual clinical queries of the clinicians and radiologists. To achieve this goal we use semantic sources such as ontologies and statistical analysis. Allen et al. [1] share the same goal with us in predicting some of users' information needs in the form of clinical questions, however, they do empirical research based on observing clinicians and on conducting surveys. Zeng and Cimino [14] assume that the information needs (i.e. the clinical queries) are already identified, so they develop applications within the InfoButtons [6] project that can be integrated into clinical information systems. Once the information need is identified, for example further information about a specific term like 'X-Ray' from a radiology report, it is mapped to generic question templates as well as to terminological resources such as the UMLS[3], MED[4] etc. A set of questions triggered by this term are then presented to the user to select. The user, i.e. the clinician or the radiologist, can thus explore the returned results, such as documents or Web resources, which are matched by the template of the question he selected. Again the most significant difference between this work and ours is that the former assumes that the clinical queries or at least their components are already identified, whereas our objective is first to identify the queries (or their components) based on ontologies and statistical analysis.

Related work on biomedical data sets and corpora include 'i2b2'[5] on clinical data and the GENIA[6] corpus. All these corpora have been designed to extract terms and their interrelations as described in [4]. This is the approach which we also follow with our query pattern derivation technique. These resources mainly concentrate on one domain such as genes or clinical reports. In contrast, the corpora that are established for this work i.e. the statistical analysis of ontology concepts and subsequent relation extraction, are designed to provide a common viewpoint of diseases, anatomy and radiology. Finally, there has been work on collecting clinical questions gathered from healthcare providers in clinical settings, which are available online under the Clinical Questions Collection[7.] This is also the resource we used to create the clinical questions corpus to evaluate the clinical query patterns. In our questions corpus, we additionally converted them to a special XML format and annotated them with part-of-speech information for subsequent linguistic processing.

3. Materials and Methods

The diagnostic analysis of medical images typically concentrates around three questions (a) what is the anatomy? (b) what is the name of the body part? (c) is it normal/abnormal? Therefore, when a radiologist looks for information, his search queries most likely contain terms from various information sources that provide knowledge about human anatomy, radiology and diseases. Four ontologies that address the questions above become relevant for our purposes. These are Foundational Model of Anatomy[8] (FMA), Radiology Lexicon[9] (RadLex), International Statistical Classification of Diseases and Related Health Problems (ICD)[10] and NCI Cancer

[3] http://www.nlm.nih.gov/research/umls/

[4] http://med.dmi.columbia.edu/construc.htm

[5] https://www.i2b2.org/NLP/

[6] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA

[7] http://clinques.nlm.nih.gov/JitSearch.html

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

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

[10] ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/ICD9-CM/2007/

Thesaurus[11]. Each ontology that contains knowledge from its representative domain (i.e. anatomy, radiology or disease) is accompanied by a corresponding domain corpus. Additionally, the lymphoma corpus based on PubMed[12] abstracts on lymphoma provides more use case as well as domain specific insights. Finally, the clinical questions corpus serves as a basis for evaluating the statistically most relevant (therefore assumed to be most likely queried) concepts from the ontologies.

3.1 Terminological Sources

Foundational Model of Anatomy (FMA) 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', 'regional part of' etc.) We refer to the version available in February 2009. The FMA can be accessed online via the Foundational Model Explorer·

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 medical images. RadLex contains 11962 terms (e.g. 'Schatzki ring' and its synonym 'Lower esophageal mucosal ring') related to anatomy pathology, imaging techniques, and diagnostic image qualities. The terms are organized along several relationships hence several hierarchies. Examples of radiology specific relationships are 'thickness of projected image' or 'radiation dose'. We refer to the version available in February 2009.

The International Classification of Diseases, Ninth Revision (ICD-9 CM) is a collection of codes classifying diseases, signs, symptoms, abnormal findings and it is published by the World Health Organization (WHO)[13]. An example is 'Lymph nodes of head, face, and neck' classified under *Neoplasms* (140-249). We extracted a subset of ICD-9 CM codes that also have a corresponding term in the RadLex and in the FMA ontology, for example 'Renal artery' and 'Uterine artery'.

The National Cancer Institute Thesaurus (NCI) is a standard vocabulary 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. Every concept has one preferred name (e.g., 'Hodgkin Lymphoma') and additional 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'). We refer to the version from February 2009.

3.2 Data

The anatomy, radiology and disease corpora based on Wikipedia were constructed from the Anatomy, Radiology and Diseases sections of Wikipedia. Actual patient records would have been the first choice, but due to strict anonymization requirements they are difficult to obtain. Thus, Wikipedia corpora served as an initial step. 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 POS parser [3] using PENN Treebank Tagset 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 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 corpora with those in the British National Corpus (BNC)[14]. This follows the approach described in [7]. In total there are 1410 such XML files for anatomy, 526 for diseases, 150 for radiology.

The lymphoma corpus is based on medical publication abstracts on lymphoma from PubMed. It is set up to target the specific domain knowledge about lymphoma, as this is one major use case of MEDICO. Furthermore, medical abstracts are naturally more appropriate for our tasks as they are more domain specific. As a consequence, the PubMed corpus is larger than the other corpora. We extracted the lymphoma relevant concepts from the NCI Thesaurus and using these we identified from PubMed an initial set of most frequently reported lymphomas. These concepts were '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'. Hence, for each lymphoma type (i.e. NCI concept) we compiled a set of XML documents that are generated from PubMed abstracts and processed in the same way as the others. The resulting corpus consists of 71.973 files.

The clinical questions corpus consists of health related questions (without answers) exchanged between the medical experts. These questions (e.g., "*What drugs are folic acid antagonists?*") were collected via a scientific

[11] http://www.cancer.gov/cancertopics/terminologyresources

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

[13] http://www.who.int/en/

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

survey and are available online at Clinical Questions Collection[15]. To create the clinical questions corpus we downloaded the categories *Neoplasms*, *Hemic and Lymphatic Diseases*, *Nervous System Diseases* and finally *Neonatal Diseases and Abnormalities* from the website. For each question and its relevant information we created a corresponding XML file and processed it to include POS information as above. In the clinical questions corpus there 624 such XML files. The clinical questions collection specifies three different categories for one question, which are *General Questions*, *Short Questions* and *Original Question* and these are different formulations of the same question. Whenever present, we included all formulations of the questions. Therefore, in one XML file there can be multiple formulations of one question, which are nevertheless all semantically equivalent. The final set consists of 1248 questions in total.

3.3 Clinical Query Pattern Derivation

The derivation of clinical query patterns consists of two steps. First step is the statistical profiling of domain ontology concepts based on corpora. Once the statistically most relevant ontology concepts are identified, the second step is to identify relations that hold between them. The result is a set of concept-relation-concept triplets to which we refer as clinical query patterns, in other words potential clinical queries. The statistical query pattern derivation process is explained in detail in Buitelaar *et al.* [4] and in Oezden Wennerberg *et al*. [10]. The resulting separate lists contain 19,337 concepts for FMA, 12,055 for RadLex and 3193 for ICD-9 CM.

Additionally, we used a list of concepts about liver lymphoma. These concepts are a set of representative image features used in the annotation of a liver image that shows symptoms of lymphoma. There are a total of 35 such image features to which we refer as image *concepts* for consistency. Some examples are 'benign', 'calcification', 'CT', 'diffuse', 'enlarged', etc. The statistically most relevant concepts are then identified on the basis of chi-square scores computed for nouns and adjectives in each corpus. 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 Wikipedia Anatomy corpus corresponds to the FMA concept 'Ear', the noun 'x-ray' from the Wikipedia Radiology corpus corresponds to the RadLex concept 'X-ray', the adjective 'respiratory' from the Wikipedia Disease corpus to 'respiratory' from the ICD, etc. 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, RadLex, ICD and Image concepts in their respective corpora are:

Table 1. 5 most relevant FMA concepts in Wikipedia anatomy corpus.

FMA Concept	Score
Lateral	338724,00
Interior	314721,00
Artery	281961,00
Anterior spinal artery	219894,33
Lateral thoracic artery	217815,33

Table 2. 5 most relevant RadLex concepts in Wikipedia radiology corpus.

RadLex Concept	Score
X-ray	81901,64
Imaging modality	58682,00
Volume imaging	57855,09
Molecular imaging	57850,00
MR imaging	57850,00

Table 3. 5 most relevant ICD concepts in Wikipedia disease corpus.

ICD Concept	Score
Acute	21609,00
Respiratory	16900,00
Fistula	8100,00
Irritable bowel syndrome	7793,68
Pulmonary hemorrhage	6038,50

Table 4. 5 most relevant concepts from an image on liver lymphoma in PubMed lymphoma corpus.

Image Concept	Score
Lymphoma	36711481,00
Tumor	183184,00
Diffuse	139129,00
Infiltration	9409,00
Neoplasm	2809,00

To obtain a more domain specific (i.e. medical) and more use case relevant (i.e. lymphoma) view, we profiled a selection of concepts from the ontologies solely on the Mantle Cell Lymphoma collection of the PubMed corpus (and we are currently extending the profiles to the rest of the lymphoma collections in the corpus). Table 5 shows the most relevant concepts from the ontologies according to their scores based on the PubMed corpus. For example, the

[15] http://clinques.nlm.nih.gov/JitSearch.html

'Lymphoma' concept, which is present in RadLex (but not in FMA) and which is also an image concept, has a relevance score of 36711481,00. This is based on its statistical analysis on the Mantle Cell Lymphoma collection of the PubMed corpus.

Table 5. 5 most relevant concepts from ontologies in PubMed lymphoma corpus. 'yes' indicates that the concept is present in the ontology, otherwise a 'no'.

Concept	FMA	Rad.	Img.	Score
Lymphoma	no	yes	yes	36711481,00
Large cell lymphoma	no	yes	no	12491501,21
Leukemia	no	yes	no	613089,00
Median	no	yes	no	305809,00
Normal cell	yes	no	no	240175,31

3.3.1 Relation Extraction

Discovering the relations between the statistically most relevant concepts is the next step for obtaining the clinical questions. Thus, we implemented a simple algorithm that traverses each sentence to find the pattern:

Noun Verb + Preposition Noun
(Concept) (Relation) (Concept)

In this pattern Verb+Preposition is the relation we look for. Subsequently, we identified relations, e.g. 'recommended for' and obtained a set of term-relation-term triplets e.g., "lymphoma recommended for therapy". Eventually, we were able to identify 1082 non-unique relations (i.e. including syntactic variants such as analy*s*ed_by and analy*z*ed_by) from the PubMed lymphoma corpus (so far only from the Mantle Cell Lymphoma collection). The triplets thus demonstrate how concepts from different ontologies relate to each other specifically within the medical imaging context. Some patterns are:

Table 6. Relations between the statistically most relevant concepts based on PubMed corpus, where R is for RadLex, F for FMA and I for Image concepts.

Concept	*Relation*	Concept
Lymphoma (R, I)	*associated with*	Adenocarcinoma (R)
Leukemia (R)	*compared with*	Lymphoma (R, I)
Normal cell (F)	*micro-dissected from*	Tonsil (F, R)
Cell membrane (F)	*detected by*	Flow (R)
Tumor (R, I)	*found in*	Gastrointestinal tract (R, F)

4. Results

We compared the clinical query patterns with actual clinical questions from the clinical questions corpus to identify overlaps. In the first place, we concentrated on comparing the ontology concepts and in this paper we focus on reporting their results. We additionally discussed the patterns and results with clinical experts.

4.1 Results on Clinical Questions Corpus

For space reasons, we only display the detailed results from matching against *Neoplasms* questions. The concepts being matched are those from the ontologies that were identified as most relevant based on corresponding corpora. Table 7 shows the comparison results in detail (up to first 10, frequencies in paranthesis and number of different concept types in *italics* and paranthesis.) For example, 2653 most relevant FMA concepts (according to anatomy corpus), 827 RadLex, 95 ICD and 8 image concepts were compared against 358 questions about *Neoplasms*. In case of FMA there are 196 matches, for RadLex 303, for ICD 68 and for image concepts 25, where the same question might have been matched multiple times by different concepts. Table 8 shows a summary for the rest of the question categories. Finally, Table 9 displays a selection of the most and the least relevant ontology concepts (based on their corpus profiles) concepts and their occurrences in the questions.

4.2 Analysis

According to comparison results, more than half of the 358 *Neoplasm* questions, (%54,7) were matched by the FMA concepts. For RadLex the results were higher, %88,5 percent of the questions had correspondences among the RadLex concepts. Another clear observation was the high number of matches for the few (8) image concepts in the *Neoplasms* category, which was not the case in the other question categories. We believe the reason for this is that the image concepts come from a lymphoma image, they are profiled on the basis of the PubMed lymphoma corpus, which is a highly domain specific and use case relevant corpus and they are matched against questions about neoplasms also known as tumors related to cancers.

A parallel observation is that from a rather large set of FMA concepts (2653), only 33 different types of FMA concepts were found in the *Neoplasm* questions. From the smaller RadLex set, however, 76 different types were found. These profiles remained similar across question categories. So, we can say when the anatomy concepts occur in the questions then this is more of a small and focused set. It is not possible to say this for radiology concepts. Also for ICD or image concepts, the input set of concepts proved to be not large enough to be able to make statements. Acknowledging this as background information, the most significant observation for us, however, is the correlation between the relevance scores of the concepts and their occurrences in the questions. In contrast to our expectations, the concepts with the highest relevance scores did not occur more often in the questions, regardless of the category. The results showed rather the opposite; those concepts that showed up most often in the questions did in fact have lower scores. This means the following; for predicting potential clinical queries our assumption that the most frequently occurring concepts would also be the most relevant ones shall be reversed. In other words, those concepts that have rather lower relevance profiles (because they occur too less i.e. too specific) are much more relevant for predicting clinical queries. This can be explained by the fact that, when the clinicians and radiologists search for information, they are mostly after a specific piece of information. That is, they have a special case at hand, for example a medical image (e.g., of liver) which show abnormal symptoms (e.g., of lymphoma), and they need to find targeted, specific information. First observations on comparing relations to clinical questions reveal caused by ("Is this anemia caused by iron deficiency?") and affected by ("Is platelet function affected by nonsteroidal anti-inflammatory drugs?") to be most frequent.

Table 7. Comparison to Neoplasms questions.

Neoplasms: # Questions: 358	
FMA Total # of concepts: 2653 # of matches: 196 (%54,7) *(# of different types of concepts: 33)*	Anterior(2), Artery(4), Carotid artery(2), Coronary artery(2), Internal(2), Basal(7), Throcacic vertebra(2), Basal cell(7), Renal cell(2), Bone(2), …
RadLex Total # of concepts: 827 # of matches: 303 (%84,6) *(# of different types of concepts: 76)*	X-ray(2), Magnetic resonance Imaging(1), Dual energy x-ray absorbtiometry(1), Ultrasound(9), Small(5), First(3), Artery(4), Tissue(5), Brain(8), Soft tissue(1)….
ICD Total # of concepts: 95 # of matches: 68 (%22,4) *(# of different types of concepts: 12)*	Lung(8), Soft tissue(1), Renal failure(2), Vagina(1), Brain(8), Stomach(2), Tongue(2), Colon(14), Prostate(22), Neck(2),...
Image Concepts Total # of concepts: 8 # of matches: 25 (%6,9) *(# of different types of concepts:2)*	Tumor(15), Mass(10)…

Table 8. Comparison to rest of the questions (concept frequencies in paranthesiss and number of different concept types in *italics* and paranthesiss).

Hemic & Lymphatic Diseases # of Questions 296	# of matches and *(# of different types of concepts:85)*	FMA	67(%22,6)
		Rad.	181(%63,1)
		ICD	11 (%3,7)
		Img.	(%1,6)
Neonatal Diseases & Abnormalities # of Questions 294	# of matches *(# of different types of concepts:90)*	FMA	197(%67)
		Rad.	201 (%68,3)
		ICD	11 (%3,74)
		Img.	5(%1,7)
Nervous System Diseases # of Questions 300	# of matches and *(# of different types of concepts:78)*	FMA	38(%12,6)
		Rad.	194 (%64)
		ICD	13 (%4,3)
		Img	(%1,6)

Table 9. 5 *most* and *least* relevant concepts from the ontologies. F = FMA, R = RadLex, I = Image concepts.

Concept	Relevance	Freq. in Questions
Lymphoma (R, I)	36711481,00	2
Large cell lymphoma(R)	12491501,21	0
Leukemia (R, I)	613089,00	0
Median (R)	305809,00	0
Normal cell (F)	240175,31	0
Prostate (F,R)	441,00	66
Blood (F,R)	3133,52	52
Iron (F, R)	1,25	36
Hemoglobin (F,R)	1521,00	30
Platelet(R)	25,00	26

So far we have compared the concepts and the relations to the questions independent of each other, to be able to obtain maximum information. However, we conducted first experiments to compare them in combination (e.g. lymphoma recommended for therapy), which naturally, returned less matches. The most probable two reasons for this can be that the clinical questions corpus is not sufficiently domain conformant as it is compiled based on the questions asked among the family physicians.

Therefore, it is not sufficiently radiology specific. A possible second reason is due to the natural characteristic of the questions: they are fairly short. Therefore, it becomes less probable to match longer chunks of patterns against short questions. However, we continue extending the questions corpus to continue with the experiments.

4.3 Discussions with Clinical Experts

We discussed the query patterns with the clinicians and radiology experts, who also confirmed our observations and agreed with the explanations. In their daily tasks, when the healthcare experts search for information they have a specific case at hand, so their information need is very much focused. As a result, the search queries are accordingly specific. The more generic concepts belong to commonly known and shared facts, so there is no need to investigate. Otherwise, attempting to predict typical clinical query patterns had another useful side effect; they served as a basis medical vocabulary for us when communicating with the medical experts.

5. Conclusions and Future Work

We reported on our work towards predicting typical clinical queries for retrieving medical images and textual patient data. Subsequently, we described the clinical query pattern derivation approach for achieving this goal. It is based on statistical profiling of concepts from medical ontologies on a special set of domain corpora. The query pattern derivation approach takes as input the concepts from the ontologies and assigns them relevance scores to indicate their specificity based on frequencies in domain vs. generic corpora. For the statistically most relevant concepts we additionally extracted relations from the domain corpora.

The comparison results with a corpus of clinical questions showed that the statistically less relevant concepts have more potential to be parts of clinical search queries. This was also confirmed by the clinical experts. We will take this finding as a basis for our future concept/relation profiling and for deciding for a most representative set of clinical query candidates. We further plan to extend this work to map the selected concepts/relations to a set of generic medical question templates, e.g. ‘What is the drug of choice for condition X?’ [8]. In this way we expect to obtain full question patterns for a selection of most interesting concepts and relations. Consequently, we can investigate methods to determine the most radiology specific full question patterns.

Another potential future work is based on the observation of a characteristic of the clinical questions; that they are usually short. Thus, questions, like news headlines, contain highly interrelated concepts (like symptoms, diseases, drugs, anatomical parts) that are in the immediate context of each other. This provides a good basis for term/relation extraction from the clinical questions corpus.

6. References

[1] Allen, M., Currie, L.M., Graham, M., Bakken, S., Patel, V.L., Cimino, J.J. 2003. The classification of clinicians' information needs while using a clinical information system. AMIA Annual Symposium Proc. 2003:26–30.

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

[3] Brants T. 2000. TnT - A Statistical Part-of-Speech Tagger. In: Proc. of the 6th ANLP Conference, Seattle, WA.

[4] Buitelaar P., Oezden Wennerberg P., Zillner S. 2008. Statistical Term Profiling for Query Pattern Mining. In: Proc. of ACL 2008 BioNLP Workshop. Columbus, Ohio, 2008.

[5] Ciaramita, M., Gangemi, A., Ratsch, E., Saric, J., Rojas,I. 2008. Unsupervised Learning of Semantic Relations for Molecular Biology Ontologies. In Paul Buitelaar, Philipp Cimiano (Eds.) Ontology Learning and Population: Bridging the Gap between Text and Knowledge. Frontiers in Artificial Intelligence and Applications Series, Vol. 167, IOS Press.

[6] Cimino, J., Elhanan, G., Zeng, Q. 1997. Supporting Infobuttons with Terminologic Knowledge. In Proc. of the AMIA Annual Symposium. 1997 pp. 528–532.

[7] Drouin, P, 2003. Term extraction using non-technical corpora as a point of leverage, In Terminology, 9, 99-117.

[8] Ely, J.W. A Osheroff J., Gorman, N. P., Ebell, M. H., Chambliss, M.L., Pifer, E. A., Stavri, P. Z. 1999. A Taxonomy of Generic Clinical Questions: Classification Study. BMJ, 321(7258), 1999, pp. 358-361.

[9] 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, pp. 9-13. USA.

[10] 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: Workshop on Building and Evaluating Resources for Biomedical Text Mining at LREC, 2008.

[11] Price S.L. and Delcambre, L. M. 2005. Using concept relations to improve ranking in information retrieval. In Proc. of the AMIA 2005, Washington DC.

[12] Price, S., Delcambre, L., Nielsen, M. L., Tolle, T., Luk, V., and Weaver, M. 2006. Using semantic components to facilitate access to domain-specific documents in government settings. In Proc. of the 2006 International Conference on Digital Government Research, vol. 151. ACM NewYork, NY, pp. 25-26.

[13] Price, S. L., Nielsen, M. L., Delcambre, L. M., and Vedsted, P. 2007. Semantic components enhance retrieval of domain-specific documents. In Proc. of the Sixteenth ACM Conference on Conference on information and Knowledge Management. ACM New York, pp. 429-438.

[14] Zeng Q. and Cimino J. 1997. Linking a Clinical System to Heterogeneous Information Sources. In Proc. of the AMIA 1997 Annual Fall Symposium, pp. 553-5