Statistical Term Profiling for Query Pattern Mining

Paul Buitelaar
DFKI GmbH
Language Technology Lab
Saarbrücken, Germany
paulb@dfki.de

Pinar Oezden Wennerberg, Sonja Zillner
Siemens AG
Knowledge Management CT IC 1
Munich, Germany
pinar.wennerberg.ext@siemens.com, sonja.zillner@siemens.com

1 Introduction

Through advanced technologies in clinical care and research, especially the rapid progress in imaging technologies, more and more medical imaging data and patient text data is generated by hospitals, pharmaceutical companies, and medical research. For enabling advanced access to clinical imaging and text data, it is relevant to know what kind of knowledge the clinician wants to know or the queries that clinicians are interested in. Through intensive interviews and discussions with radiologists and clinicians, we have learned that medical imaging data is analyzed - and hence queried – from three different perspectives, i.e. the *anatomic perspective* addressing the involved body parts, the *radiology-specific spatial perspective* describing the relationships of located anatomical regions to other anatomical parts, and the *disease perspective* distinguishing between normal and abnormal imaging features. Our aim is to establish query patterns reflecting those three perspectives that would typically be used by clinicians and radiologists to find patient-specific sets of relevant images.

The context of our work is in the Theseus-MEDICO[1] project on cross-modal image and information retrieval in the medical domain. The focus of the work reported here is on setting up Wikipedia-based corpora of human anatomy and radiology and on obtaining a statistical profile of concepts from three semantic knowledge resources with these corpora: the Foundational Model of Anatomy (FMA), the radiology lexicon RadLex, and a subset of the international classification of disease codes ICD-9 CM. Using this information, we intend to extract relations that are likely to occur between statistically relevant terms and the concepts they express.

The final goal of our work is to derive potential query patterns from the extracted set of relations that can be used in the MEDICO semantic-based image retrieval application. For example when restaging head and neck lymphoma, clinicians and radiologists look for information and images that report on essential radiological patterns as *"an enlargement in the dimension of the lymph node in the neck"*. Therefore, within our approach, we aim at establishing hypotheses about possible user queries, i.e. the query patterns that reflect the three perspectives discussed above. Accordingly, an example query pattern might look like this:

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

Once an initial set of similar patterns has been established in this way, they will be evaluated by clinicians for their validity and relevance.

2 Corpora

A central aspect of the query pattern mining task is the statistical analysis of the FMA and RadLex terms in relevant text collections. In this way relevance scores can be assigned to terms that allow to investigate the most likely expressed (and hence queried) relations between them. For this purpose we need access to a representative corpus of texts that at the same time reflects the joint view of anatomy, spatial aspects of radiology and disease that we are targeting. Patient records would be our first choice, but due to strict anonymization requirements these are difficult to obtain. We therefore constructed a corpus based on the Wikipedia Categories Anatomy and Radiology. We then ran all text sections of each corpus through a part-of-speech tagger (Brants, 2000) to extract all nouns in the corpus and to compute a relevance score (chi-square) for each by comparing anatomy and radiology frequencies with those in the British Na-

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

BioNLP 2008: Current Trends in Biomedical Natural Language Processing, pages 114–115,
Columbus, Ohio, USA, June 2008. ©2008 Association for Computational Linguistics

tional Corpus. A next step will be to parse and annotate sentences with predicate-structure information, which may then be used for relation extraction along the lines of (Schutz and Buitelaar, 2005).

3 FMA Terms

The statistically most relevant FMA terms were identified on the basis of chi-square scores computed for nouns in each corpus. Single word terms in the FMA and occurring in the corpus correspond directly to the noun that the term is build up of (e.g. the noun ‘ear’ corresponds to the FMA term *ear*). In this case, the statistical relevance of the term is the chi-square score of the corresponding noun. In the case of multi-word terms occurring in the corpus, the statistical relevance is computed on the basis of the chi-square score for each constituting noun and/or adjective in the term, summed and normalized over the length of the term. Thus, the relevance value for *lymph node* is the summation of chi-square scores for ‘lymph’ and ‘node’ divided by 2. In order to take frequency in 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.

FMA Term	Freq.	Score	POS
lateral	464	338724,00	JJ
anterior	452	314721,00	JJ
artery	237	281961,00	NN
anterior spinal artery	2	219894,33	JJ JJ NN
lateral thoracic artery	2	217815,33	JJ JJ NN

Table 1: top FMA terms in anatomy corpus

FMA Term	Freq.	Score	POS
artery	65	6724,00	NN
coronary artery	17	5284,00	JJ NN
small bowel	11	4651,79	JJ NN
renal artery	3	4286,50	JJ NN
pulmonary artery	1	3974,50	JJ NN

Table 2: top FMA terms in radiology corpus

4 RADLEX Terms

Analogously, RadLex was used to identify the most relevant radiology terms. The most relevant RadLex terms are shown below. As with the FMA, the most relevant RadLex terms in the anatomy corpus are centered on “artery”. In contrast, in the radiology corpus the RadLex relevance scores indeed point to a radiology profile:

RadLex Term	Freq.	Score	POS
lateral	464	338724,00	JJ
anterior	452	314721,00	JJ
artery	237	281961,00	NN
anterior spinal artery	2	219894,33	JJ JJ NN
lateral thoracic artery	2	217815,33	JJ JJ NN

Table 3: top RadLex terms in anatomy corpus

RadLex Term	Freq.	Score	POS
x-ray	253	81901,64	NN
imaging modality	6	58682,00	NN NN
volume imaging	1	57855,09	NN NN
molecular imaging	4	57850,00	JJ NN
mr imaging	9	57850,00	JJ NN

Table 4: topRadLex terms in radiology cor pus

5 Conclusions and Future Work

Using ICD-9 lymphoma terminology, we will derive a Pubmed-based corpus on lymphoma to analyse the context of the statistically top most relevant terms from the FMA and RadLex terminologies. In this way we will be able to identify relationships and eventually query patterns across the three dimensions of anatomy, radiology and lymphoma research.

Acknowledgments

This research has been supported in part by the THESEUS- MEDICO Project, which is funded by the German Federal Ministry of Economics and Technology under the grant number 01MQ07016.

References

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

Langlotz, CP. (2006). RadLex: A New Method for Indexing Online Educational Materials In: *Radiographics* 26, pp.1595-1597.

Rosse C. and J.L.V. Mejino Jr. (2003). A reference ontology for biomedical informatics: The Foundational Model of Anatomy. *Journal of Biomedical Informatics*, 36(6), pp. 478–500.

Schutz A., and P Buitelaar. (2005). RelExt: A Tool for-Relation Extraction in Ontology Extension In: *Proc. the 4th International Semantic Web Conference*, Galway, Ireland.