Interactive Clinical Query Derivation and Evaluation

Pinar Wennerberg[1], Sonja Zillner[1], Paul Buitelaar[2]

[1]Siemens AG, Corporate Technology, Knowledge Management CT IC 1
Otto-Hahn-Ring 6, 81739, Munich Germany
[2]DERI - NLP Unit, National University Ireland Galway,
IDA Business Park, Lower Dangan, Galway, Ireland
{pinar.wennerberg.ext | sonja.zillner}@siemens.com, paul.buitelaar@deri.org

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 search for. This information is typically represented in the search queries of the clinicians, which they send to retrieve the related text and images. Collecting these queries via typical expert interviews, however, is inappropriate. The reason is that for a successful communication during the interview some medical knowledge background of the knowledge engineer becomes necessary, which is usually not available. Therefore, alternative techniques are required to obtain relevant information about clinical search queries that are independent of the expert interviews. The query pattern derivation approach described here is one technique to gain this information. It is based on the prediction of clinical query patterns given domain ontologies and corpora. The patterns identified in this way are then presented to the clinical experts via an interactive browser for knowledge elicitation and evaluation purposes. Being an interactive tool, the Clinical Query Pattern Browser also supports the communication process between the clinical expert and the knowledge engineer.

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 search for. This information is typically represented in the queries of clinicians, which they would send to a search engine to retrieve related text and medical images.

The context of our work is the Theseus-MEDICO[1] use case, which has a focus on semantic, cross-modal image search and information retrieval in the medical domain. Our focus here is to identify the kind of queries the clinicians are interested in. In our context, classical interview techniques based on so-called competency-questions were not feasible for supporting the knowledge transfer from the clinical expert to the knowledge engineer. This was due to the reason that medical knowledge is too specific and too sensitive for a knowledge engineer to come up with the appropriate interview questions and to be able to process the interview answers in an adequate manner. The clinical experts and the knowledge engineers have a quite different perception of the medical domain; the former mostly focuses on special cases and outstanding details, whereas the latter's focus is to generalize or to abstract the domain of interest.

To overcome the difficulties in the communication process between the clinical experts and knowledge engineers, our aim is to establish tools and methodologies that support the knowledge elicitation process. In particular, we are interested in identifying queries, which are used by clinicians to retrieve medical images and related patient data and which we were not able to identify during the expert interviews. To achieve this goal, we follow two steps: Firstly, we attempt to predict the search queries of the clinical experts. This is done based on our query pattern mining approach that uses domain ontologies, domain corpora and statistical techniques. Our approach is described in detail in Wennerberg et al., (2008) and in Buitelaar et al., (2008). Secondly, we present the identified set of query patterns to the clinicians for verification and for the evaluation of their relevance. In particular, we provide the clinicians an interactive visualization tool, ICQB (Interactive Clinical Query Browser), using which the clinical experts can easily verify (or refute) the accuracy and the relevance of each identified query pattern.

The contribution of this paper is the description of these two steps. Therefore, the remainder of this paper is organized as follows. In Section 2 we give a rough overview of one selected medical use case of the Theseus-MEDICO use case and in Section 3 we discuss related approaches. Section 4 provides details on our approach of query pattern mining and Section 5 illustrates the interactive clinical query browser (ICQB) and describes our practical experiences with it. The final section concludes with our plans for future work.

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

MEDICO Use Case

Increasingly large amount of medical imaging and the related textual patient data becomes available due to advanced imaging technologies. To be able to use this data effectively, it is relevant to know what kind of information the clinicians search for. This information is typically represented in the queries of clinicians that they send to a search engine to retrieve a coherent set of patient data and medical images.

One MEDICO scenario aims for improved image search in the context of patients suffering from lymphoma in the neck area. Lymphoma, a type of cancer originating 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 and discussions 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*, which is the medical image specific knowledge and finally the *disease dimension* that describes the normal and abnormal anatomical and imaging features. Hence, our objective is to predict clinical queries related to these three dimensions. An example query can be "All CT scans and MRIs of patient X with an enlarged lymph node in the neck". The query predictions are represented through generic patterns which we refer as the query patterns. Accordingly the corresponding query pattern can then be of the form:

[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 are established in this way, they are evaluated by clinicians for their validity and relevance using the Interactive Clinical Query Browser.

Related Work

The domain corpora we established for the statistical analysis of ontology concepts seems unique in providing a common viewpoint of diseases, associated anatomy and spatial aspects of radiology. Biomedical data sets that are somewhat related to ours include 'i2b2'[2] on clinical data as well as the GENIA[3]. All these corpora have been designed to extract terms/concepts and their interrelations as described in (Ciaramita et al. 2008), the approach which we also follow with our query pattern derivation technique.

Available disease related data mostly focus on security related disease outbreaks and infectious diseases as provided by the US Center for Disease Prevention Control[4] or the MediSys Medical Information System of the European Commission[5]. Within the BioSense[6] program radiology text reports of various types from 42 acute care hospitals are collected for the purpose of early event detection, quantification and spatiotemporal visualization of public health events and risks. Further related approaches to term extraction that specifically target the medical domain are reported in Bourigault and Jacquemin (1999) and in Le Moigno et al.,(2002), which, however, are independent of the image semantics.

There is a significant volume of work on visualization with (and without) ontologies, which can be differentiated from the work presented here on the basis of various criteria, such as: what is visualized, why and how. Visualization may focus on homogeneous data from a database. Alternatively, it can be used to display data from heterogeneous sources to provide an integrative, coherent view. It can be utilized for exploratory data analysis as explained in Keim (2002). In particular cases, ontology-based visualization has been used to support queries based on temporal abstractions (Shahar and Cheng 1998); to enrich maps with additional geographic information (Ipfelkofer et al. 2007); and to assist in information mining (Castillo et al. 2003). It is popularly used to map social networks and communities of common interest (Mika 2005), (Oellinger and Wennerberg 2006).

There has been much work on ontology visualization (Bosca et al. 2005), (Noy et al. 2000), (Mutton and Golbeck 2003), (Pietriga 2002) that help the user navigate ontology concept hierarchies. For an up-to-date survey of ontology visualization methods reference must be made to (Katifori et al. 2007).

The work presented in this paper relates to visualization approaches that provide a basis for interactive knowledge engineering. Therefore, our aim is rather to establish tools, such as the browser explained here, that help involve the domain expert in the knowledge engineering process by visualizing the domain knowledge (or data), which he can evaluate and edit, than to visualize the ontology contents. In this respect, objectives of our work relate mostly to those set for the work by Sonntag and Heim (2007).

The Query Pattern Mining Approach

Our Query Pattern Mining Approach consists of two steps. First step is the set up of Wikipedia-based domain corpora,

2 https://www.i2b2.org/NLP/

3 http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA
4 http://www.cdc.gov
5 http://medusa.jrc.it/medisys/homeedition/all/home.html
6 http://www.cdc.gov/biosense/

whereas the second is the statistical processing or profiling of domain ontology concepts based on the corpora. Once the statistically most relevant ontology concepts are identified in this way, the relations that hold between the concepts are extracted. The result is a set of concept-relation-concept triplets, for which we assume that they are relevant for clinical queries. Both steps are described in the following subsections.

Wikipedia-based Corpora and Domain Ontologies

We set up corpora of texts on human anatomy, radiology and disease and we assume that these corpora contain terms and relations that are relevant for the daily practical work of the clinicians and radiologists. Patient records, despite being our first choice, are difficult to compile due to strict anonymization requirements. Therefore, we have constructed joint corpora based on the Wikipedia Categories Anatomy[7], Radiology[8] and Disease9. For the three corpora we downloaded the related web pages and generated an XML version of these using standard tools provided by Wikipedia[10]. We then ran all text sections of each corpus through the TnT part-of-speech parser (Brants 2000) to extract all nouns in the corpus and to compute a relevance score (chi-square) for each by comparing anatomy, radiology and disease frequencies respectively with those in the British National Corpus (BNC)[11]. In total there are 1410 such XML files about human anatomy, 526 about disease, and 150 about radiology.

To acquire the necessary knowledge about radiology, diseases and anatomy that is relevant to medical image semantics, we refer to domain ontologies and terminologies. Consequently, we used Foundational Model of Anatomy (FMA) for anatomical knowledge, the radiology lexicon RadLex for radiology specific information and a subset of the international classification of disease codes ICD-9 CM.

Relation Extraction

Using these ontologies and terminologies, we then extract relations along the lines of Schutz and Buitelaar (2005) that are likely to occur between statistically relevant terms and the concepts they express. For this purpose we implemented a simple algorithm that traverses each sentence, looking for the following pattern:

Term : Verb : Preposition : Term

This pattern enables us to identify possibly relevant relations between terms (i.e. ontology concepts). The following table presents some early results of this work. As a result, we were able to identify 1082 non-unique relations (i.e., including syntactic variants such as *analysed_by* and *analyzed_by)*. In future work we will apply further statistical measures and linguistic heuristics to identify the most salient relations within each corpus, with an emphasis on relation identification in a more specific lymphoma corpus obtained from PubMed.

Term	**Relation**	**Term**
anterior	known as	anterior scalene muscle
dentate nucleus	subdivided into	anterior
muscle	situated between	anterior
body	divided into	anterior
anterior	continued over	zygomatic arch
hand	used for	anterior
artery	supplied by	medulla
artery	released if	ulnar
vein	associated with	artery
bronchopulmonary segment	supplied by	artery

Table 1: first top 10 ranked extracted relations in anatomy corpus

Interactive Clinical Query Browser

We developed an interactive browser interface to be able to present the query patterns to the clinical experts. The purpose of our interactive browser is twofold. Firstly, it shall support the communication process between the knowledge engineer and the clinical expert. The interactive browser achieves this goal by incorporating the clinical expert in the process of the evaluation of the query patterns. By presenting the clinical expert the query patterns, we actually present him our understanding of the domain, which is his expert area. The interactive browser is a useful tool in the sense that the clinical expert can browse the patterns, can see the terms that we think are relevant for him and can either confirm or negate our understanding. He can make new suggestions and use the browser as a medium for the communication.

The second goal of the browser is to evaluate the accuracy and relevance of the semi-automatically identified query patterns. This goal is achieved naturally (i.e. as a side product) during the interaction process, when the expert gives his feedback on the patterns he explores on the browser. He can give his feedback by marking the patterns as relevant or irrelevant and by entering comments.

We presented our first ranked results of statistically concept-relation-concept triplets to the radiology expert using our browser. These triplets, despite being incomplete

7 http://en.wikipedia.org/wiki/Category:Anatomy
8 http://en.wikipedia.org/wiki/Category:Radiology
9 http://en.wikipedia.org/wiki/Category:Diseases
[10] http://en.wikipedia.org/wiki/Special:Export
[11] 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.

queries, have indeed provided a basis for our communication. Upon seeing the terms and relations, the radiologist has been able to make associations with other possible terms related to anatomy, radiology and diseases. He suggested further resources to look into and gave feedback on whether these terms and relations would ever be searched for. Below are some screenshots from the first version of our Interactive Clinical Query Browser:

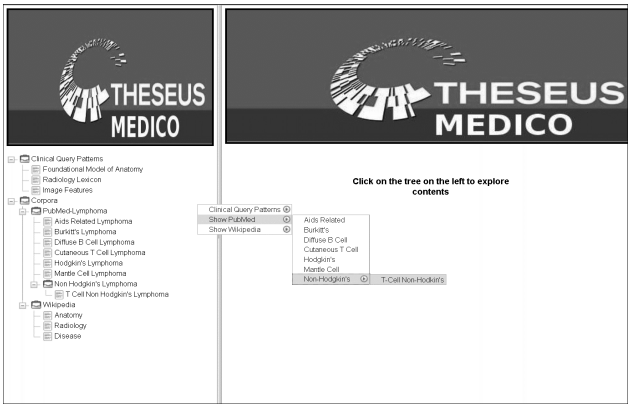
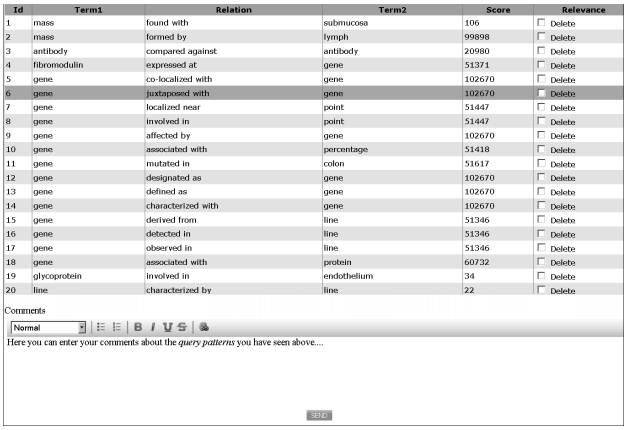


Figure 1: Initial view of the browser that shows the ontologies and corpora on which the query patterns were identified.



Id	Term1	Relation	Term2	Score	Relevance
1	mass	found with	submucosa	106	Delete
2	mass	formed by	lymph	99898	Delete
3	antibody	compared against	antibody	20980	Delete
4	fibromodulin	expressed at	gene	51371	Delete
5	gene	co-localized with	gene	102670	Delete
6	gene	juxtaposed with	gene	102670	Delete
7	gene	localized near	point	51447	Delete
8	gene	involved in	point	51447	Delete
9	gene	affected by	gene	102670	Delete
10	gene	associated with	percentage	51418	Delete
11	gene	mutated in	colon	51617	Delete
12	gene	designated as	gene	102670	Delete
13	gene	defined as	gene	102670	Delete
14	gene	characterized with	gene	102670	Delete
15	gene	derived from	line	51346	Delete
16	gene	detected in	line	51346	Delete
17	gene	observed in	line	51346	Delete
18	gene	associated with	protein	60732	Delete
19	glycoprotein	involved in	endothelium	34	Delete
20	line	characterized by	line	22	Delete

Comments

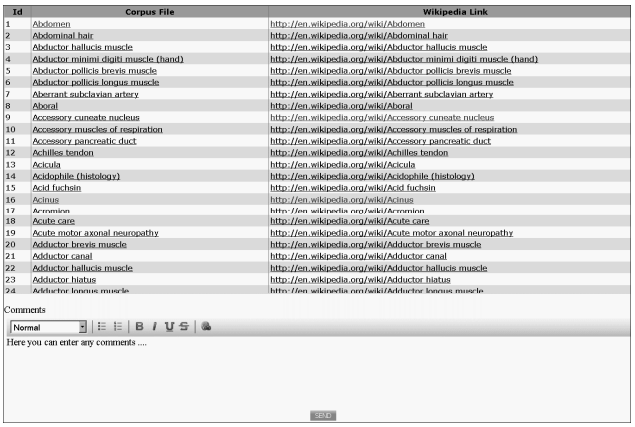
Normal

Here you can enter your comments about the *query patterns* you have seen above....

SEND

Figure 2: One view in the browser that shows the query patterns, allows the expert to delete and to enter his comments

One valuable outcome was that our radiology expert reported that most commonly occurring terms are in fact unlikely to be searched for because everyone knows and therefore writes about them in scientific literature. The most likely queries to be sent would be the ones that are least found in the literature. Hence, his hint was that most commonly searched terms will most likely be those that are rarely mentioned in text. In other words, one searches for less known and the less mentioned for the purposes of self education.



Id	Corpus File	Wikipedia Link
1	Abdomen	http://en.wikipedia.org/wiki/Abdomen
2	Abdominal hair	http://en.wikipedia.org/wiki/Abdominal_hair
3	Abductor hallucis muscle	http://en.wikipedia.org/wiki/Abductor_hallucis_muscle
4	Abductor minimi digiti muscle (hand)	http://en.wikipedia.org/wiki/Abductor_minimi_digiti_muscle_(hand)
5	Abductor pollicis brevis muscle	http://en.wikipedia.org/wiki/Abductor_pollicis_brevis_muscle
6	Abductor pollicis longus muscle	http://en.wikipedia.org/wiki/Abductor_pollicis_longus_muscle
7	Aberrant subclavian artery	http://en.wikipedia.org/wiki/Aberrant_subclavian_artery
8	Aboral	http://en.wikipedia.org/wiki/Aboral
9	Accessory cuneate nucleus	http://en.wikipedia.org/wiki/Accessory_cuneate_nucleus
10	Accessory muscles of respiration	http://en.wikipedia.org/wiki/Accessory_muscles_of_respiration
11	Accessory pancreatic duct	http://en.wikipedia.org/wiki/Accessory_pancreatic_duct
12	Achilles tendon	http://en.wikipedia.org/wiki/Achilles_tendon
13	Acicula	http://en.wikipedia.org/wiki/Acicula
14	Acidophile (histology)	http://en.wikipedia.org/wiki/Acidophile_(histology)
15	Acid fuchsin	http://en.wikipedia.org/wiki/Acid_fuchsin
16	Acinus	http://en.wikipedia.org/wiki/Acinus
17	Acromion	http://en.wikipedia.org/wiki/Acromion
18	Acute care	http://en.wikipedia.org/wiki/Acute_care
19	Acute motor axonal neuropathy	http://en.wikipedia.org/wiki/Acute_motor_axonal_neuropathy
20	Adductor brevis muscle	http://en.wikipedia.org/wiki/Adductor_brevis_muscle
21	Adductor canal	http://en.wikipedia.org/wiki/Adductor_canal
22	Adductor hallucis muscle	http://en.wikipedia.org/wiki/Adductor_hallucis_muscle
23	Adductor hiatus	http://en.wikipedia.org/wiki/Adductor_hiatus
24	Adductor longus muscle	http://en.wikipedia.org/wiki/Adductor_longus_muscle

Comments

Normal

Here you can enter any comments ....

SEND

Figure 3: One view in the browser that shows the domain corpora, so that the clinical expert can explore the context of the patterns..

Conclusion

We reported on our work to extract clinical query patterns using medical ontologies and domain corpora. We identified the need for clinical query extraction based on the fact that it is not easy to elicit this knowledge from clinicians via traditional expert interviews. However, information about clinical search queries to retrieve patient images and text data is an essential step for building a semantic search engine as MEDICO. Subsequently we introduced our Interactive Clinical Query Browser that is developed as a tool to support the communication and the knowledge elicitation process between the knowledge engineer and the clinical expert. We additionally reported on our first experiences in using the tool. Our query pattern derivation approach and the results presented using our browser has had positive assessment from our radiology expert and it has provided a basis for successful communication at the right level of granularity.

Future Work

We are currently applying query pattern mining approach to the scientific abstracts in PubMed. As one major use case of MEDICO is lymphoma we are specifically targeting the lymphoma related abstracts. The relevant sub-terminology from the NCI Thesaurus of the National Cancer Institute[12] was extracted for this purpose. Based on this sub-terminology and the initial version of the PubMed lymphoma corpus, a first set of approximately 500 query patterns (i.e. concept-relation-concept triplets) related to Mantle Cell Lymphoma were extracted. The NCI sub-terminology, which includes information about lymphoma types, their relevant thesaurus codes, synonyms, hyperonyms and the corresponding thesaurus definitions, was also converted to OWL format. This allows for

[12] http://www.nci.nih.gov/cancerinfo/terminologyresources

integrating the lymphoma information in the current work of MEDICO on semantic image annotation. The Interactive Clinical Query Browser will be extended to allow for more user interaction. As next it is planned add functionality that enables the clinical expert to enter new domain related concepts and relations i.e. to enter his own query patterns.

In parallel, we are working on our indirect evaluation approach to be able to verify the validity of our patterns semi-automatically. The Interactive Clinical Query Browser offers functionalities to enter feedback from the clinical expert, which can then be stored in a database for future automatic processing.

Finally, multilinguality is one of the topics to be addressed within the next steps. This will become most relevant, when the patient records are available as we assume to obtain at least one dataset in German. Using a multilingual knowledge resource such as Wikipedia therefore proves to be an additional advantage.

Acknowledgements

This research has been supported in part by the THESEUS Program in the MEDICO Project, which is funded by the German Federal Ministry of Economics and Technology under the grant number 01MQ07016. The responsibility for this publication lies with the authors. Paul Buitelaar contributed most of the work reported here while at DFKI. We are especially thankful to our clinical partner Dr. Alexander Cavallaro from the University Hospital Erlangen, Germany.

References

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

Bosca A. and Bonio D, 2005. *Ontosphere: More Than a 3D Ontology Visualization Tool*. In SWAP The 2nd Italian Semantic Web Workshop.

Buitelaar P., Wennerberg P., Zillner S. *Statistical Term Profiling for Query Pattern Mining* BioNLP, ACL, 2008.

Castillo J.A.R, 2003. *Information Extraction and Integration from Heterogeneous, Distributed, Autonomous Information Sources - A Federated Ontology-Driven Querycentric Approach*. In Conference on Information Reuse and Integration, pp. 183–191.

Ciaramita, M., Gangemi, A., Ratsch, E., Saric and 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

Ipfelkofer F., 2007. *Ontology Driven Visualisation of Maps with SVG - An Example for Semantic Programming.* In Conference on Information Visualization, pp. 424–429

Katifori A, 2007. *Ontology Visualization Methods: A Survey*. In ACM Comput. Surv., (39) pp.10, New York, NY, USA, ACM.

Keim, D. A, 2002. *Information Visualization and Visual Data Mining*. In Transactions on Visualization and Computer Graphic, volume 8, pp. 1–8, 2002

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.

Mika P, 2005. *Ontologies Are Us: A Unified Model of Social Networks and Semantics*. In The Semantic Web - ISWC 2005, volume 3729/2005, pp. 522–536.

Mutton, P. and Golbeck J, 2003. *Visualization of Semantic Metadata and Ontologies*. In IV '03: Proceedings of the Seventh International Conference on Information Visualization, pp. 300, Washington, DC, USA .

Noy N., Fergerson R.W., and Musen M.A, 2000. *The Knowledge Model of Protege-2000: Combining Interoperability and Flexibility* In Journal of Knowledge Engineering and Knowledge Management Methods, Models, and Tools, pp.69-82.

Oellinger T, Wennerberg P.O.: *Ontology Based Modeling and Visualization of Social Networks for the Web*. GI Jahrestagung (2) 2006: 489-497

Pietriga E, 2002. *Isaviz, a visual environment for browsing and authoring RDF models*. In Eleventh International World Wide Web Conference Developer's Day.

Shahar Y. and Cheng C, 1998. *Ontology-driven Visualization of Temporal Abstractions*. In EKAW'98 Eleventh Workshop on Knowledge Acquisition, Modeling and Management.

Sonntag D. and Heim P., 2007. *Semantic Graph Visualisation for Mobile Semantic Web Interfaces*. In: B. Falciendo, M. Spagnuolo, Y. Avrithis, I. Kompatsiaris, Paul Buitelaar (eds.): *Semantic Multimedia.* Proceedings of the 2nd International Conference on Semantic and Digital Media Technologies (SAMT-2007)

Wennerberg, P. Buitelaar P., Zillner S, 2007. *Towards a Human Anatomy Data Set for Query Pattern Mining based on Wikipedia and Domain Semantic Resources* Building and Evaluating Resources for Biomedical Text Mining, ELDA, 2008