Semantic Classification of liver radiological reports

STSM Scientific Report
Carry out by María del Mar Roldán Garcia form June, 22th to June 28th 2014

1 Introduction and Background

Semantic annotation of unstructured content in the form of ontologies with data as instances is a task of paramount importance for a wide range of Semantic Web and Knowledge Management applications. The majority of current ontologies are expressed in the well-known Web Ontology Language (OWL) [1] which is based on a family of logical formalisms called Description Logic (DL). Managing large amounts of OWL data, including query answering and reasoning, is a challenging technical prospect, but one which is increasingly needed in numerous real-world application domains from Health Care and Life Sciences to Financials and Government.

Using ontologies for information retrieval has certain advantages over simple keyword based access methods: An ontology provides a shared vocabulary for expressing information about the contents of documents. In addition, it includes axioms for specifying relationships between concepts. Such an ontology may then in turn be used to formulate more complex queries and to deliver exactly the information we are interested in. Furthermore, the axioms provide a means for deriving information which has been specified only implicitly. Therefore, ontologies are very suitable as the basis of a Semantic Classification System.

In the clinical world, the indexing of clinical documents and patient records is called "coding". One of the most widely used "code sets" in this context is the International Classification of Diseases (ICD), which has been used for more than a century for coding morbidity, mortality, billing and reimbursement. Another standard that is getting widely adopted, at least within the European Union, is SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms [2]). Providing a wide and comprehensive coverage for diseases, findings, procedures, pharmaceuticals etc., SNOMED-CT provides a consistent way to index, store, retrieve, and aggregate clinical data across specialties and sites of care. Similarly, the UMLS Metathesaurus has also been used in some cases for coding clinical documents [3].

The last decade has seen a marked increase in representation of "biomedical entities" in the form of ontologies. Arguably, this has led to their adoption in various aspects of biomedical informatics ranging from clinical decision support to scientific knowledge management and semantic data integration and exchange. Biomedical ontologies provide essential domain knowledge to drive data integration, information retrieval, data annotation, natural-language processing and decision support. Developing an ontology for a particular domain requires training and a background in ontology development best practices.

The proliferation of ontologies in the biomedical sciences brings about a new problem for knowledge sharing namely that concepts of the same meaning in different ontologies are not explicitly cross-linked. Many of these ontologies contain overlapping information. To be able to use multiple ontologies they have to be aligned or merged. Ontology alignment is the task of identifying such equivalence relations between concepts across ontologies.

1.1 ONLIRA

Radiologists inspect CT scans and record their observations in reports to be shared with physicians. These reports may suffer from ambiguous language and inconsistencies resulting from subjective reporting styles, which present challenges in interpretation. Standardization efforts, such as the lexicon RadLex for radiology terms, aim to address this issue by developing standard vocabularies. While such vocabularies handle consistent annotation, they fall short in sufficiently processing reports for intelligent applications. To support such applications, the semantics of the concepts as well as their relationships must be modeled, for which, ontologies are effective. They enable software to make inferences beyond what is present in the reports. The researchers of VAVLab have developed the open source ontology ONLIRA [4] (Ontology of the Liver for Radiology), the aim of which is to support such intelligent applications, such as identifying and ranking similar liver patient cases. The development of ONLIRA was partially supported by the ICT COST Action IC1302 KEYSTONE[1].

1.2 CaReRa PROJECT

Based on the fact that clinical experience plays a key role in the performance of medical professionals, it is conjectured that a Clinical Experience Sharing (CES) platform, i.e. a searchable collective clinical experience knowledge-base, accessible by a large community of medical professionals, would be of great practical value in clinical practice as well as in medical education. Such a CES would be composed of a multi-modal medical case database, would incorporate a Content Based Case Retrieval (CBCR) engine and would be specialized for different domains.

Project CaReRa[2] aims to develop such a CES for the domain of liver cases. During the course of the project, multi-modal case data is being collected, anonymized and stored in a structural database, CBCR technologies will be developed, experiments for the assessment of its impact on the clinical workflow as well as medical education will be designed and conducted.

1.3 DBOWL

The Khaos Research group have developed DBOWL [5], a scalable reasoner for classifying instances of OWL ontologies with very large Aboxes (billions of instances). DBOWL is based on mature relational database technology. It stores the ontologies in a relational database, using a tableau-based OWL reasoner for obtaining the reasoned Tbox information, which is also stored in the database. Relational algebra expressions and fixed-point iterations are combined in order to create the knowledge base. DBOWL includes a SPARQL engine which allows the evaluation of SPARQL queries on DBOWL.

2 Purpose of the STSM

The purpose of the STSM was to visit the Bogazici University in Istambul for 5 working days to:

1. Study ONLIRA and its relationships with SNOMED CT and ICD10 terminologies.
2. Study the CaReRa (Case Retrieval in Radiological Databases) project in order to export the data about Livers and Lesions to RDF.
3. Provide technical support to classify and check the consistency of ONLIRA.
4. Provide technical support to model the CaReRa database's liver patient data (previous diseases, regular drugs, liver diseases, etc.) as an OWL ontology.
5. Provide technical support to develop a very first prototype for searching similar liver patients based on the previous modelled patient data using an ontology reasoner and a SPARQL query engine.

3 Main results

During the STSM several results were obtained:

1. ONLIRA ontology was classified using the Pellet [6] reasoner.

[1] http://www.cost.eu/domains_actions/ict/Actions/IC1302
[2] http://www.vavlab.ee.boun.edu.tr/pages.php?p=research/CARERA/carera.html

2. CaReRa data about livers and lesions was included as instances of ONLIRA and the consistency of data with respect to the ontology was checked. Detected inconsistences were solved.
3. After studying ONLIRA and its instances, some redundant axioms were detected. Furthermore, a new way to model liver regions and segments was proposed. This resulted in a new version of ONLIRA (version 1.0.6)[3].
4. CaReRa data about liver patients, which are not modelled by ONLIRA, i.e. previous diseases, common drugs, liver diseases, etc., was modelled as an OWL ontology. This ontology was populated by exporting CaReRa data to RDF and was classified using the Pellet reasoner.
5. Several simple SPARQL queries were created in order to search for similar patients using the previous ontology and the reasoner in order to validate the viability of the proposed search engine.

4 Future collaboration

Close collaboration with the hosts Dr. Burak Acar and Dr. Suzan Uskudarli will continue and several joint publications are planned.

This collaboration has the following objectives:

1. Implementing the liver patient search engine using DBOWL. In order to do this, some improvements and extensions of DBOWL must be developed.
2. Exporting CaReRa data on livers and lesions to RDF and implement a similar lesions search engine, that is based on concept lattice generation using also DBOWL.
3. Integrating both search engines in a semantic search engine.
4. Aligning ONLIRA and the new patient data ontology with other medical terminologies in order to improve the results of the semantic search engine.

References

[1] Grigoris Antoniou and Frank van Harmelen. Web ontology language: OWL. In Handbook on Ontologies, International Handbooks on Information Systems, pages 67–92. Springer, 2004.

[2] SNOMED-CT: Systematized Nomeclature of Medicine-Clinical Terms. https://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html.

[3] Pakhomov, S., J. Buntrock, and C. Chute, Automating the Assignment of Diagnosis Codes to Patient Encounters using Example-based and Machine Learning Techniques. Journal of Americal Medical Informatics Association, 2006. 13(5): p. 516-525.

[4] Kokciyan N., Turkay R., Uskudarli S., Yolum P., Bakır B., Acar B. Semantic Description of Liver CT Images: An Ontological Approach. IEEE Journal of Biomedical & Health Informatics (2013), in press.

[5] Maria Mar Roldan-Garcia and Jose F. Aldana-Montes. Evaluating DBOWL: A non-materializing owl reasoner based on relational database technology. In Ernesto Jimenez-Ruiz Ian Horrocks, Mikalai Yatskevich, editor, Proceedings of the 1st International Workshop on OWL Reasoner Evaluation (ORE-2012), volume 858 of CEUR Workshop Proceedings, 2012.

[6] E. Sirin, B. Parsia, B.C. Grau, A. Kalyanpur, Y. Katz. Pellet: a practical OWL-DL reasone. Journal of Web Semantics, 5 (2) (2007), pp. 51–53.

[3] http://www.vavlab.ee.boun.edu.tr/pages.php?p=research/CARERA/OnliraDownload.php