Ontology customization for biomedical informatics

Author	**Pablo López-García**
Supervisors	Arantza Illarramendi
Studies/Stage	PhD in Web Information Systems, expected Fall 2012
Affiliation	University of the Basque Country UPV/EHU (Spain)
E-Mail	pablo.lopez@ehu.es

Aims and Objectives of the Research

The overall aim of the research is to explore to which extent existing semantic technologies, particularly biomedical ontologies and ontology mappings, can be customized to better serve biomedical applications.

Justification for the Research Topic

Semantic technologies, such as comprehensive biomedical ontologies (e.g., SNOMED CT [1]), and ontology mappings linking several ontologies (e.g., SNOMED CT, NDF-RT [2], and RxNorm [3]), are offered as valuable tools to the biomedical community. However, due to the high degree of specialization in health care and life sciences, two undesirable scenarios might occur: (i) the whole breadth of an extended biomedical ontology might be excessive for some applications and users or, on the other hand, (ii) information of interest might be missed because it is scattered among several ontologies.

Research Questions

1. How useful are ontology modularization techniques to extract subsets from a large biomedical ontology to annotate medical records? Can the size of those subsets be reduced by using term frequency information from an authoritative corpus? How does size reduction affect coverage?

2. Once a module from a reference ontology has been adopted for a domain, how useful are ontology mappings and ontology repositories as an approach to add new, related information from another domain to the module?

Research Methodology

1. *Case study: Annotating discharge summaries in cardiology using SNOMED CT.* Four graph-traversal heuristics [4], and one logic-based technique [5] were explored to extract subsets from SNOMED CT. Extracted subsets were subsequently filtered with term frequency information from MEDLINE. As signatures and test sets, twenty manually coded discharge summaries from cardiology patients were used. The coverage, size, and precision of extracted subsets were measured.

2. *Case study: Using disorders in SNOMED CORE to extract drugs from RxNorm.* The CORE problem list subset of SNOMED CT [6] was used as signature in the domain of diseases to find related drugs in RxNorm, using NDF-RT as a cross-ontology, and the UMLS Metathesaurus to provide the mappings.

Research Results to Date

1. *Published results [7]:* Graph-traversal heuristics provided high coverage (71-96% of terms in the test sets of discharge summaries) at the expense of subset size (17-51% of the size of SNOMED CT). Logic-based techniques extracted small subsets (1%), but coverage was limited (24-55%). Filtering reduced the size of large subsets to 10% while still providing 80% coverage. The study shows that the requirements of preserving ontology entailments (key to ontology modularization), and achieving high precision (key to annotation), necessarily conflict. Attempting to satisfy the two aims simultaneously can lead to unsatisfactory results.

2. *Preliminary results:* Using the CORE subset's 5,814 SNOMED CT concepts as input, we were able to extract 7,499 related drugs in RxNorm. The CORE subset is roughly 6%, and 9% the size of SNOMED CT's Disease (ID 64572001), and Clinical finding (ID 404684003) hierarchies, respectively. The extracted RxNorm subset was 3.87% the size of the full RxNorm. The current research is focused on measuring the coverage of the new subset (CORE plus the extracted RxNorm) when annotating terms from a large, real dataset.

References

1. IHTSDO (2012), International Health Terminology Standards Development Organization, http://www.ihtsdo.org/.
2. Brown SH, Elkin PL, Rosenbloom ST, Husser C, Bauer BA, Lincoln MJ, Carter J, Erlbaum M, Tuttle MS (2004), VA National Drug File Reference Terminology: a cross-institutional content coverage study, in: Studies in Health Technology and Informatics 107(Pt 1), 477-481.
3. Liu S, Ma W, Moore R, Ganesan V, Nelson SJ (2005), RxNorm: prescription for electronic drug information exchange, in: IT Professional 7(5), 17-23.
4. Seidenberg J, Rector A (2006), Web ontology segmentation: analysis, classification and use, in: Proceedings of the 15th International Conference on World Wide Web, Edinburgh, Scotland, UK: Association for Computing Machinery, 13-22.
5. Cuenca Grau B, Horrocks I, Kazakov Y, Sattler U (2008), Modular reuse of ontologies: theory and practice, in: Journal of Artificial Intelligence Research 31, 273-318.
6. Fung KW, McDonald C, Srinivasan S (2010), The UMLS-CORE project: a study of the problem list terminologies used in large healthcare institutions, in: Journal of the American Medical Informatics Association 17(6), 675-680.
7. López-García P, Boeker M, Illarramendi A, Schulz S (2012), Usability-driven pruning of large ontologies: the case of SNOMED CT, in: Journal of the American Medical Informatics Association 19(e1), e102-e109.