ONTOLOGY EVOLUTION WITH TEXT RESOURCES

- A study on SNOMED CT clinical terminology

Chunnan She (Duran)
School of IT

Supervisor: Prof. Joseph Davis

FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

Introduction

What is Ontology?

•Ontology is sourced from philosophy research. Gruber defines the term ontology as "specification of a conceptualization". In some cases, the term 'ontology' is also used to refer to controlled terminologies.

Figure: Beer ontology – an ontology example

•Wikipedia definition of Ontology:
"An ontology is a formal representation of the knowledge by a set of concepts within a domain and the relationships between those concepts."

And why should we care?

•Benefits of ontology[1]

- Communication between systems, between humans, and between humans and systems.
- Computational inference.
- Reuse and organization of knowledge.

•Ontology in action

- SNOMED CT - Medical
- Cyc – Top Level

Motivation and Aims

Lessons learnt from SNOMED CT

•Ontology should evolve to reflect the changes in its knowledge domain

•Evolving (maintaining) ontology could be difficult

- Size is large (>1 million relations)
- Proper tools not available
- High cost[2]

Ontology evolution approaches

•Crowd sourcing

- Adapted by open projects (e.g. Freebase, OpenCYC)

•Experts

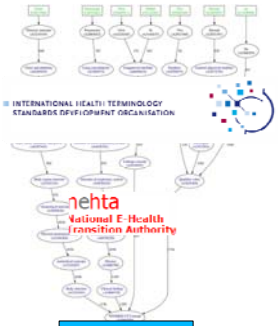
- Reliable, Controllable, Auditable

Project Aim

•An integrated system supporting the ontology evolution from textual resources by Experts

- Evolve an existing ontology without rebuilding it from the ground up.
- Automate the evolution process to which demands minimal human intervention.
- Keep the trail of proposed changes from input documents.

Proposed Method and System Framework



Relation Extraction

Hearst Rules method in NLP

•Label text with phrase chunking tool

•Apply Hearst rules[3]

#	Pattern	Relation
1	NP_0 such as {NP_1, NP_2 … , (and|or)} NP_n	hyponym(NP_i, NP_0)
2	such NP_0 as {NP_1 ,}*{(or|and)} NP_n	hyponym(NP_i, NP_0)
3	NP_1 {, NP_n}*{,} or other NP_0	hyponym(NP_i, NP_0)
4	NP_1 {, NP_n}*{,} and other NP_0	hyponym(NP_i, NP_0)
5	NP_0 {,} including {NP_1 ,}*{or|and} NP_n	hyponym(NP_i, NP_0)
6	NP_0 {,} especially {NP_1,}*{or|and} NP_n	hyponym(NP_i, NP_0)

Measuring the Equivalence of Concepts

Multi-Level approach

•Level 3 and above are considered as equivalent. Inspired by [4].

Level	Description
1	Highest – two nodes are literally identical
2	High – two nodes have same set of terms, but in different order
3	Normal – same as level 2, but terms are processed after stemming
4	Partial – two nodes have only some terms in common, after stemming

Indexing

•Direct reading SNOMED CT concepts and relations from zipped release.

•Fast search and retrieval of all SCT concepts

•Word stemming support

- Example: bodies -> body

Graph-based Algorithm

Matching sub-graphs/triples to existing ontology

•Existing ontology G=<V,E>, sub-graph/triple obtained from text C=<W,F>

•Evolved ontology candidate O≈<VUW,EUF>

•Irrelevant sub-graph/triple detection

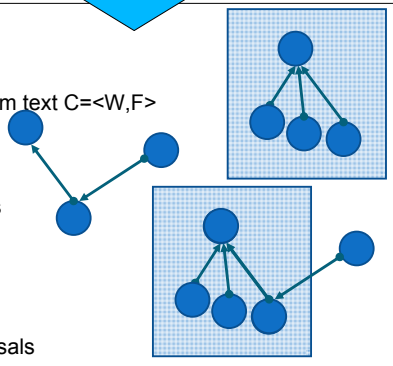
•Advanced path matching for SCT defining characteristics

Evaluating and proposing changes

•Lexico and structural based evaluation

•With references to text

•Experts are able to choose and approve based on proposals



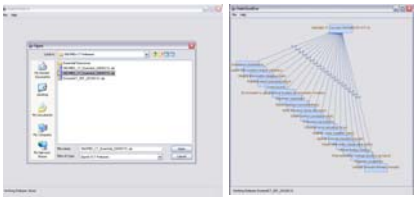
Prototyping the System (Working in Progress)

•Index new SCT releases

•Choosing working release and input documents

•Show and review evolution proposal

•Graphical User Interface



Evaluation and Expected Results

Datasets

- Two different releases of SNOMED CT (January 2009 and January 2010)
- Journal articles from MEDLINE database

Process

- Specify a subject (e.g. cancer)
- Select the subject concept as root node in both releases
- Compare the two SCT subset and identify new concepts in latter release.
- Use earlier release as working release, MEDLINE articles with new concepts keyword as input documents. System propose changes to working release.
- We expect the system to propose the same new concepts added to the working release in a hierarchy level similar to the latter release.

References

[1] T. Bürger and E. Simperl, "Measuring the Benefits of Ontologies," in On the Move to Meaningful Internet Systems: OTM 2008 Workshops, ed, 2008, pp. 584-594.
[2] F. Bakhshi-Raiez, et al., "Development and application of a framework for maintenance of medical terminological systems," Journal of the American Medical Informatics Association, vol. 15, pp. 687-700, Sep-Oct 2008.
[3] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," presented at the Proceedings of the 14th conference on Computational linguistics - Volume 2, Nantes, France, 1992.
[4] B. Cacayan, et al., "Automated Ontology Mapping: A Multi-level Approach," in HIC 2006 and HINZ 2006, Brunswick East, Vic, Australia, 2006, pp. 334-339.