Report on the Manchester Visit - 4.10 - 6.10

Marta Sabou

Vrije Universiteit Amsterdam, The Netherlands

1 Goal of the Visit

The goal of this visit was to evaluate the results of the ontology learning algorithm as implemented in the GATE framework. The corpus used for this extraction consisted of a collection of textual web service descriptions used within the ^{my}Grid project[1] - a large UK-based project for supporting biological research on the Grid infrastructure. We have compared the extracted ontology with the manually built ontology for the same collection of web services. We benefitted from the support and opinion of the domain experts who have built the manual ontology of the domain in performing this evaluation.

2 Topic of the visit

The topic was the evaluation of our ontology learning method. Evaluation of ontology learning methods is an unsolved and difficult issue. We believe that this work, that involves real life data used in a realistic scenario (that of bioinformatics infrastructures on the Grid) and was carried out with the support of domain experts, will highly benefit research on this overall topic.

3 Results of the visit

Together with Chris Wroe, the author of the ^{my}Grid ontology, we analyzed the extracted ontology and evaluated its potential usefulness for an ontology engineer. There is a report[2] containing this comparison, with detailed and commented examples of overlaps between the two ontologies. Hereby we present only the major conclusions.

The manually built ontology is very elaborate and contains several levels of abstraction. It is a broad and very deep ontology. It contains a lot of domain knowledge in the field of biology and bioinformatics that are not present in the corpus used as basis for our extraction (external knowledge). It also contains auxiliary knowledge about measuring units, which, being domain independent notions should probably be reused from other specialized ontologies. Finally, often different views on the same domain concepts are defined. For example, services are classified according to their input or output type. However, the web-service

[1] http://www.mygrid.org.uk/

[2] http://www.cs.vu.nl/~marta/MyTalks/VisitManchester_05.10/mygrid_comparison.pdf.

infrastructure which currently supports web service discovery only uses about 5% of the manual ontology. It is expected that other tasks, such as matchmaking, will require more domain knowledge, but they still do not know what knowledge will exactly be needed.

The extracted ontology significantly overlaps with the currently employed 5% of the manual ontology. However our learning method is limited in (1) learning different views, (2) determining abstraction levels over the learned domain concepts and (3) learning hierarchies of data structures that are not based on compositionality. Naturally, it can only learn the knowledge that exists in the given corpus. The current limitations can be tackled using several different techniques. View generation should be very easy since it only requires that the harvested lexical information is presented in different ways. It is interesting to investigate which views are useful and to provide methods to create them. The solutions for the other limitations probably rely on the use of existing knowledge sources such as medical ontologies (e.g. UMLS) or terminological databases (e.g. Termino).

During the evaluation exercise it turned out that:

- the ontology engineer considered most extracted concepts useful (numbers will follow). The conclusion is that recall was more important than precision in this case.
- many of the extracted concepts were **new** additions to his own ontology. Notably, several times he said "I wish I had that!", referring not only to concepts but to whole pieces of the extracted hierarchy.
- in some cases, it was impossible to decide the relevance of some concepts for the ontology. He required the possibility to inspect the sources from where the concept was learned in order to disambiguate its meaning.