WP7 Meeting, June 10th 2011

Agenda:

1. WP7 Description
2. Discussion
3. Tasks and Calendar
4. Other issues

Participants: Aarne Ranta (UGOT), Adam Slaski (UGOT), Mariana Damova (Ontotext), Yasen Kiprov (Ontotext), Milen Chechev (Ontotext), Lluis Màrquez (UPC), Cristina España (UPC), Meritxell Gonzàlez (UPC).

Report:

1. WP7 Description

- D7.1: Patent MT & Retrieval prototype Beta by Dec-2011

- Proposed Functionalities:
  Patent translation of titles and claims
  Patent Retrieval based on claims content

2. Discussion

- Patents Ontology.

  The patents retrieval system requires an ontology on the particular domain. Retrieval will be based on the knowledge representation of the claims. MD informs that Ontotext has an ontology on the Biomedical domain but lacks an ontology on the particular domain for the WP7. UPC will provide Ontotext patent examples from the current corpus (Domain IPC A61P: Specific therapeutic activity of chemical compounds or medicinal preparations).

  Ontotext has an ontology capturing the structure of Patent documents. It consists of several modules including some FDA terms, drugs and measurement related models.

- Indexing the patents documents.

  Patents will be indexed following the structure provided by the GF rules. For instance, compounds are described using semantic knowledge.

UGOT will provide a translation from patent translators abstract syntax to a simpler abstract syntax, which will then be translated to OWL/SPARQL by Ontotext, in a similar way to the KRI prototype we already have.

The database of patents has to be processed on batch to build the index of document and enable the retrieval process.

- Translation Process.

  The GF processes the patent claims and produces a representation in the GF abstract syntax. The GF can translate chunks and align the texts. These tasks will be used to build the hybrid MT system.

  Patents can be translated on a batch process, similarly to the retrieval pre-processing of the documents, but in a separate process.

  Online (on the fly) translation occurs somehow in the query. The user writes a query in any of the available languages (theoretically English, French and German). This query will be translated into the GF abstract syntax. Then, the SPARQL query will be build from the abstract representation provided by the GF.

- EPO Corpus

  AR informs about the negotiation with EPO regarding the terms of use of the patents corpus (namely, the non-commercial use of the applications). We will work for having a demo hosted by the EPO.

3. Tasks and Calendar

- UPC will circulate a set of examples of the patents corpus by the beginning of the next week.

- UGOT will circulate a set of examples of the abstract syntax representation by the end of the next week.

- UGOT will circulate a preliminary report (code + doc) about the GF on the patents domain by the end of June. A more detailed report will be available in August.

- Ontotext will report about the definition of the types of queries and the demo architecture, once they have analyzed the data provided, by the end of June.

4. Other issues

- AR and LM will inform the Steering Committee changes at WP5 and WP7 leaders.