Extracting social and knowledge networks from the LREC registration data

Thierry Declerck, DFKI GmbH

As a possible additional contribution to the "LREC2010 Map of Language Resources, Technologies and Evaluation" (http://www.lrec-conf.org/lrec2010/?LREC2010-Map-of-Language-Resources), we investigated the kind of information that can be extracted from the LREC submission database (the 2008 edition), and if it can be organized in social and knowledge networks. This concerns basically information about authors of scientific papers, their affiliation, the country they are belonging to, the topics they address and the language(s) they cover in their work. This should then be combined with the information about language resources and technologies that will be available in the submission data of LREC 2010 (the LREC 2010 Map).

Some details of the approach, as applied to LREC 2008

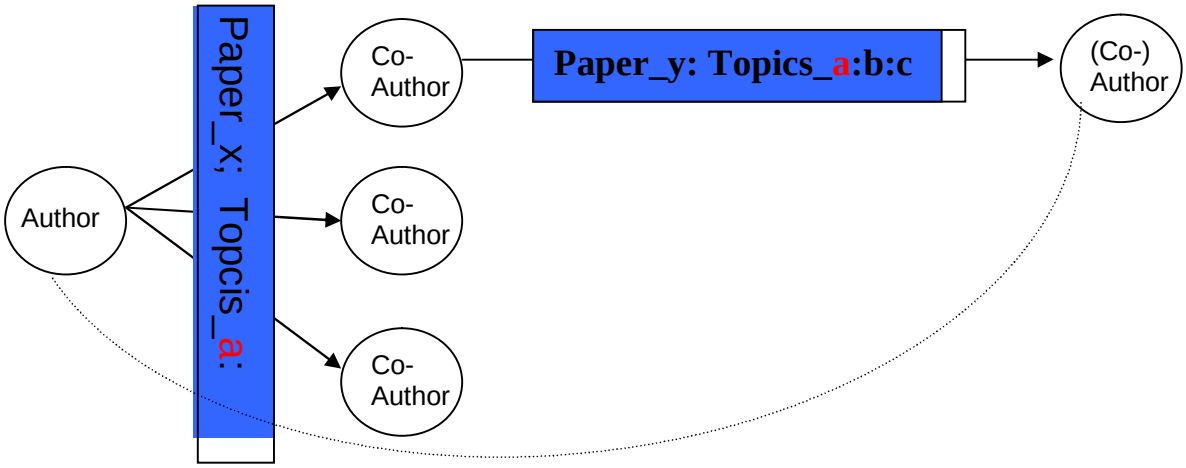
The on-line proceedings of LREC 2008 (http://www.lrec-conf.org/proceedings/lrec2008/), are displayed various interlinked lists, as shown below (with few additional comments from us):

- "Sessions" contains introductory messages, keynote speeches, panels information, and titles of papers organized by session.
- "Papers" contains a list of the papers organized by their title. Additionally the list contains the abstract, all the authors, the topics (as selected by the authors on the base of LREC suggestion), the PDF version of the paper, possibly the presentation slides, and the bibtex file.
- "Authors" contains a list of all authors and their related papers and their affiliation
- "Workshops" contains a list of all workshops and tutorials.
- "Topics" contains a list of all topics and their related papers.
- "Affiliations" contains a list of all affiliations (including name of country) and their related papers.

We map part of this information onto an XML representation, which for the time being has the following format

```
<authors>
        <author>firstname="Akinori" lastname="Abe" affiliation="ATR Knowledge
Science Labs." affiliation_location="JAPAN"
<papers>
        <paper>
        <title>Relationships between Nursing Converstaions and Activities</title>
        <coauthors>
        <coauthor>firstname="Hiromi" lastname="Itoh Ozaku"</coauthor>
        <coauthor>firstname="Kaoru" lastname="Sagara"</coauthor>
        <coauthor>firstname="Kiyoshi" lastname="Kogure"</coauthor>
        </coauthors>
        <language>Multiple languages</language>
        <topics>
        <topic>Corpus (creation, annotation, etc.)</topic>
        <topic>Information Extraction, Information Retrieval</topic>
        <topic>Acquisition, Machine Learning</topic>
        </topics>
        </paper>
</papers>
</authors>
```

On the base of this data, we can for example extract a network of authors working on similar topics, as sketched below



We are here thus extracting direct and indirect relations between authors, constrained by topics. Similar word is applied to affiliations and countries of affiliation, also using the covered language(s) as restriction for the detected relations.

Related work and resources

We aim at cooperation, integration and interaction with related initiatives, like for example LT-World (http://www.lt-world.org/), mainly in using and extending its background ontology, or with the "Take" project (http://www.dfki.de/lt/project.php?id=Project_539&l=en), especially on the topic of Information Extraction applied to scientific/technological literature in the field of language technology, which is among other dealing with the ACL Anthology initiative (http://aclweb.org/anthology-new/) or with an approach on mining expertise from scientific publications (see Paul Buitelaar and Thomas Eigner. *Mining Expertise Topics from Scientific Literature, SAAKM 2009*) . In doing this, we will aggregate information from a supervised information portal, information extracted from the content of scientific papers and information extracted from submission databases of big conferences.

Possible future work

- Use LT-World to gain additional information (for example related projects, etc) and to eliminate duplicates in the naming of persons, institutions, etc.
- Use the semantic storage developed in the European R&D project "MUSING" for supporting reasoning and (semantic) querying. Integrate the extracted networks to an ontology. Include the temporal dimension (between editions of LREC and other conferences)
- Aggregation with results of content analysis of papers of conferences (for example project TAKE), or expertise mining (work by Paul Buitelaar and Thomas Eigner).
- Identify possible cooperation links (are there countries that are working isolated on a topic, etc.)

Acknowledgements: Thanks to the CNR-ILC team for fruitful discussions, to Brigitte Joerg, DFKI, for support on LT-World, to Andreas Weber, research assistant at DFKI, for implementation support and the CLARIN and D-SPIN projects.