LREC 2012 Workshop on:

Language Resource Merging

22 May 2012 – Afternoon Session

CONTEXT

The availability of adequate language resources has been a well-known bottleneck for most high-level language technology applications, e.g. Machine Translation, parsing, and Information Extraction, for at least 15 years, and the impact of the bottleneck is becoming all the more apparent with the availability of higher computational power and massive storage, since modern language technologies are capable of using far more resources than the community produces. The present landscape is characterized by the existence of numerous scattered resources, many of which have differing levels of coverage, types of information and granularity. Taken singularly, existing resources do not have sufficient coverage, quality or richness for robust large-scale applications, and yet they contain valuable information (Monachini et al. 2004 and 2006; Soria et al. 2006; Molinero, Sagot and Nicolas 2009; Necsulescu et al. 2011). Differing technology or application requirements, ignorance of the existence of certain resources, and difficulties in accessing and using them, has led to the proliferation of multiple, unconnected resources that, if merged, could constitute a much richer repository of information augmenting either coverage or granularity, or both, and consequently multiplying the number of potential language technology applications. Merging, combining and/or compiling larger resources from existing ones thus appears to be a promising direction to take.

The re-use and merging of existing resources is not altogether unknown. For example, WordNet (Fellbaum, 1998) has been successfully reused in a variety of applications. But this is the exception rather than the rule; in fact, merging, and enhancing existing resources is uncommon, probably because it is by no means a trivial task given the profound differences in formats, formalisms, metadata, and linguistic assumptions.

The language resource landscape is on the brink of a large change, however. With the proliferation of accessible metadata catalogues, and resource repositories (such as the new META-SHARE (http://www.meta-net.eu/meta-share) infrastructure), a potentially large number of existing resources will be more easily located, accessed and downloaded. Also, with the advent of distributed platforms for the automatic production of language resources, such as PANACEA (http://www.panacea-lr.eu/), new language resources and linguistic information capable of being integrated into those resources will be produced more easily and at a lower cost. Thus, it is likely that researchers and application developers will seek out resources already available before developing new, costly ones, and will require methods for merging/combining various resources and adapting them to their specific needs.

Up to the present day, most resource merging has been done manually, with only a small number of attempts reported in the literature towards (semi-)automatic merging of resources (Crouch & King 2005; Pustejovsky et al. 2005; Molinero, Sagot and Nicolas 2009; Necsulescu et al. 2011). In order to take a further step towards the scenario depicted above, in which resource merging and enhancing is a reliable and accessible first step for researchers and application developers, experience and best practices must be shared and discussed, as this will help the whole community avoid any waste of time and resources.

AIMS OF THE WORKSHOP

This half-day workshop is meant to be part of a series of meetings constituting an ongoing forum for sharing and evaluating the results of different methods and systems for the automatic production of language resources (the first one was the LREC 2010 Workshop on Methods for the Automatic Production of Language Resources and their Evaluation Methods). The main focus of this workshop is on (semi-)automatic means of merging language resources, such as lexicons, corpora and grammars. Merging makes it possible to re-use, adapt, and enhance existing resources, alongside new, automatically created ones, with the goal of reducing the manual intervention required in language resource production, and thus ultimately production costs.

WORKSHOP TOPICS

The topics of the workshop are related to best practices, methods, techniques and experimental results regarding the merging of various types of language resources, such as lexicons and corpora, especially in support of language technology applications. In particular, new methods for automatic merging with a view towards reducing human intervention will be most welcome.

Topics for submission include, but are not limited to:

- Experiments on (semi-)automatic merging of automatically produced resources
- Experiments on the merging of two or more existing resources containing the same or different levels of linguistic information
- Studies or experiments on merging resources at different levels of granularity (corpora, lexicons, grammars)
- Studies or experiments on unifying, mapping or converting encoding formats
- Comparison between different resources and mapping algorithms to provide desired merging
- Use of linguistic information from different sources in high-level language applications
- Use of new, merged language resources in language technology applications

WORKSHOP WEBSITE:

http://panacea-lr.eu/en/news/project/2011/12/19/lrec-2012-merging-lr-workshop/

SUBMISSIONS

Interested participants must submit a preliminary paper of about 4-6 pages including references (between 2000-2500 words). For the submission please use the online form on START LREC Conference Manager at: https://www.softconf.com/lrec2012/MergingLR2012/

When submitting a paper from the START page, authors will be asked to provide essential information about resources (in a broad sense, i.e. also technologies, standards, evaluation kits, etc.) that have been used for the work described in the paper or are a new result of your research.

For further information on this new initiative, please refer to http://www.lrec-conf.org/lrec2012/?LRE-Map-2012

Papers will be peer-reviewed by the workshop Program Committee.

IMPORTANT DATES

- Deadline for paper submission: 22 February 2012 (23:59 CET +1) ****EXTENDED****
- Notification of acceptance: 15 March 2012
- Submission of camera-ready version of papers: 31 March 2012
- Workshop date: 22 May 2012 – Afternoon Session

ORGANIZING COMMITTEE

Núria Bel, UPF, Barcelona, Spain

Maria Gavrilidou, ILSP-"Athena", Athens, Greece,

Monica Monachini, CNR-ILC, Pisa, Italy

Valeria Quochi, CNR-ILC, Pisa, Italy

Laura Rimell, University of Cambridge, UK

Contacts

lrec12_workshop_merging@ilc.cnr.it

PROGRAMME COMMITTEE:

Victoria Arranz, ELDA, Paris, France

Paul Buitelaaar, National University of Ireland, Galway, Ireland

Nicoletta Calzolari, CNR-ILC, Pisa, Italy

Olivier Hamon, ELDA, Paris, France

Aleš Horák, Masaryk University, Brno, Czech Republic

Nancy Ide, Vassar College, Mass. USA

Bernardo Magnini, FBK, Trento, Italy

Paola Monachesi, Utrecht University, Utrecht, The Netherlands

Jan Odijk, , Utrecht University, Utrecht, The Netherlands

Muntsa Padró, IULA, Barcellona, Spain

Karel Pala, Masaryk University, Brno, Czech Republic

Pavel Pecina, Charles University, Prague, Czech Republic.

Thierry Poibeau University of Cambridge, UK and CNRS, Paris, France

Benoît Sagot, INRIA, Paris, France

Kiril Simov, Bulgarian Academy of Sciences, Sofia, Bulgaria

Claudia Soria, CNR-ILC, Pisa, Italy

Antonio Toral, DCU, Dublin, Ireland

Maurizio Tesconi, CNR-IIT, Pisa