HLT & NLP within the Arabic world: Arabic Language and local languages processing: Status Updates and Prospects

Saturday 31st May 2008

Workshop Programme

Automatic versus interactive analysis for the massive vowelization, tagging and lemmatization of Arabic

Fathi Debili, Zied Ben Tahar,
LLACAN, INALCO, CNRS, France and Emna Souissi, ESSTT, Tunisia

Prague Arabic Dependency Treebank: A Word on the Million Words

Otakar Smrz, Viktor Bielicky, Iveta Kourilova, Jakub Kracmar, Jan Hajic, Petr Zemanek
Institute of Formal and Applied Linguistics, Charles University in Prague, Czech Republic

Arabic Named Entity Recognition using Conditional Random Fields

Yassine Benajiba and Paolo Rosso,
Natural Language Engineering Lab. Departamento de Sistemas Inform´aticos y Computaci ´on, Universidad Polit´ecnica de Valencia , Spain .

Can the building of corpus-based Arabic concordances with AraConc and DIINAR.1 tackle the issue of Arabic polyglossia?

Joseph Dichy and Ramzi Abbès,
Université Lumière-Lyon 2 and ICAR (CNRS-Lyon 2)

Amazigh Data Base

El Mehdi IAZZI, Mohamed OUTAHAJALA,
Institut Royal de la Culture Amazigh, Rabat, Morocco

Building an Arabic Morphological Analyzer as part of an Open Arabic NLP Platform

Lahsen Abouenour (*), Said El Hassani(**), Tawfiq Yazidy (**), Karim Bouzouba(*), Abdelfattah Hamdani(**)
(*) Mohammadia School of Engineers, (**) Institute for Studies and Research on Arabization, Rabat, Morocco

Morpho-syntactic tagging system for Arabic texts

A. Yousfi , A. El jihad, and L. Aouragh,
IERA (Institute for Studies and Research on Arabization) , Rabat Morocco

Guidelines for Annotation of Arabic Dialectness

Nizar Habash, Owen Rambow, Mona Diab and Reem Kanjawi-Faraj
Center for Computational Learning Systems, Columbia University, New York, NY, USA

Information retrieval in Arabic language

Malek Boualem (*), Ramzi Abbes (**)
(*) France Télécom Orange Labs, France; (**) Lyon 2 University / ICAR-CNRS, France

Memory-Based Vocalization of Arabic

Sandra Kübler, Emad Mohamed
Indiana University , Department of Linguistics, Bloomington, IN-47405, USA

Towards a human-machine spoken dialogue in Arabic

Younes Bahou, Lamia Hadrich Belguith, and Abdelmajid BEN HAMADOU
LARIS - MIRACL Laboratory, Faculty of Economic Sciences and Management of Sfax, Sfax , Tunisia

Methods for porting NL-based restricted e-commerce systems into other languages

Najeh Hajlaoui (*), Daoud Maher Daoud (**), Christian Boitet (*)
(*)GETALP, LIG,Université Joseph Fourier, Grenoble , France
(**) Amman University , Amman Jordan

Automatic Pronunciation Dictionary Toolkit for Arabic

Hussein Hiyassat(*), Mustafa Yaseen(**), Nihad Arabiat(***)
(*) e-Prucurment Project, UNDP, (**) Amman University , (***) Ministry of Education ; Amman , Jordan

Broadcast News Transcription Baseline System using the NEMLAR database

R. Bayeh (,**), C. Mokbel (**), G. Chollet (*)*
(*) TELECOM-ParisTech, CNRS-LTCI UMR-5141, Paris , France ; (**) University of Balamand , Tripoli , Lebanon

Arabic-English translation improvement by target-side neural network language modeling

Maxim Khalilov(*), José A. R. Fonollosa(*), F. Zamora-Martínez(**), María J. Castro-Bleda(**), S. España-Boquera(**)
(*) Centre de Recerca TALP, Universitat Politècnica de Catalunya Barcelona, Spain; (**) Dep. de Lenguajes y Sistemas Informáticos, Universidad Politécnica de Valencia, Valencia, Spain

Language modeling for local and Modern Standard Arabic

Ilana Heintz, Chris Brew
Department of Linguistics, Ohio State University , Columbus , USA

Towards a syntactic lexicon of Arabic Verbs

Noureddine LOUKIL, Kais HADDAR, Abdelmajid BEN HAMADOU
Institut Supérieur d'Informatique et Multimédia de Sfax, Tunisie

Automatic Morphological Rule Induction for Arabic

Ahmad Hany Hossny (*), Khaled Shaalan (**), Aly Fahmy (*)
(*) Faculty of Computers and Information, Cairo University , Egypt
(**) Faculty of Informatics , The British University in Dubai , Dubai , UAE

Motivation and Aims

This Workshop intends to add value to the issues addressed during the main conference (Human Language Technologies (HLT) & Natural Language Processing (NLP)) and enhance the work carried out at different places to process Arabic language(s) and more generally Semitic languages and other local and foreign languages spoken in the region.

It should bring together people who are actively involved in Arabic Written and Spoken language processing in a mono- or cross/multilingual context, and give them an opportunity to update the community through reports on completed and ongoing work as well as on the availability of LRs, evaluation protocols and campaigns, products and core technologies (in particular open source ones). This should enable the participants to develop a common view on where we stand with respect to these particular set of languages and to foster the discussion of the future of this research area. Particular attention will be paid to activities involving technologies such as Machine Translation, Cross-Lingual Information Retrieval/extraction, Summarization, Speech to text transcriptions, etc., and languages such as Arabic varieties, Amazigh, Amharic, Hebrew, Maltese, and other local languages. Evaluation methodologies and resources for evaluation of HLT are also a main focus.

It is clear from the various projects that Arabic has become a major language for HLT. During this workshop we will emphasize the need to focus on specific issues that would help citizens living in Arabic countries to have access to information and technologies in their mother tongues and therefore discuss requirements to customize existing technologies for pairs of languages e.g. English to Arabic, Amazigh, etc. A particular stress will be put on tools, technologies, resources that tackle colloquial Arabic and other local languages such as Amazigh.

We expect to identify problems of common interest, and possible mechanisms to move towards solutions, such as sharing of resources, tools, standards, sharing and dissemination of information and expertise, adoption of current best practices, setting up joint projects and technology transfer mechanisms, etc.

By bringing together players in the Arabic NLP field, we would like to follow activities discussed at similar workshops (e.g. LREC2002) but also at the NEMLAR conference on Arabic Language (2004, Cairo Egypt), the workshop on Arabic NLP (Fez, April, 2007, http://www.dsic.upv.es/~prosso/workshopAECI_ArabicNLP.pdf) as well as work carried out in projects such as NET-DC, NEMLAR (www.nemlar.org) or the LDC project on the "Less Commonly Taught Languages". The objective is also to introduce activities that will be

launched shortly within the MEDAR project (the follow-up of NEMLAR project under FP7 of the European Commission). Among the crucial issues that require particular attention is the construction/update of a broadly supported Roadmap for these languages in relationship with Multilinguality and Evaluation of HLTs.

Topics of Interest

The submissions should address some of the LREC issues that are specific and of paramount importance to the Arabic resources and evaluation; some of these issues are:

- Issues in the design, the acquisition, creation, management, access, distribution, use of Language Resources (Standard Arabic, Colloquial Arabic, other Semitic languages, Amazigh, Coptic, Maltese, English/French spoken locally, etc.)
- Impact on LR collections/processing and NLP of the crucial issues related to "code switching" between different dialects and languages
- Specific issues related to the above-mentioned languages such as role of morphology, named entities, corpus alignment, etc.)
- Multilinguality issues including relationship between Colloquial and Standard Arabic
- Exploitation of LR in different types of applications
- Industrial LR requirements and community's response;
- Benchmarking of systems and products; resources for benchmarking and evaluation for written and spoken language processing;
- Focus on some key technologies such as MT (all approaches e.g. Statistical, Example-Based, etc.), Information Retrieval, Speech Recognition, Spoken Documents Retrieval, CLIR, Question-Answering, Summarization,
- Local, regional, and international activities and projects;
- Needs, possibilities, forms, initiatives of/for regional and international cooperation.

Format of the Workshop

It will be a full-day workshop. The workshop is not intended to be a mini-conference, but as a real workshop aiming at concrete results that should clarify the situation of Arabic with respect to Language Resources and Evaluation. Sessions will include introductory speeches, invited talks, a small number of refereed presentations, etc.

Workshop chair

Khalid Choukri (ELRA/ELDA, France)

Workshop Co-chairs

Mona Diab, Columbia University, USA
Bente Maegaard (CST, University of Copenhagen, Denmark)

Paolo Rosso, Universidad Politécnica Valencia, Spain
Abdelhadi Soudi ENIM (Morocco)
Ali Farghaly, Oracle USA and Monterey Institute of International Studies,

Program and Scientific Committee (tentative)

Ken Beesley , Xerox Research Centre Europe, France
Malek Boualem , France Telecom Orange Labs (France)
Tim Buckwalter , University of Maryland, (USA)
Violetta Cavalli-Sforza, San Francisco State University (USA)
Achraf Chalabi , Sakhr (Egypt)
Khalid Choukri, ELRA/ELDA (France)
Christopher Cieri, Linguistic Data Consortium, Philadelphia, (USA)
Fathi Debili, CELLMA - ENS LSH Lyon (France)
Mona Diab, Columbia University, (USA)
Joseph Dichy, Lyon -2 university Lyon (France)
Everhard Ditters, University of Nijmegen (The Netherlands)
Khaled Elghamry, (University of Florida, USA)
Ossama Emam, IBM (Egypt)
Ali Farghaly, Oracle USA and Monterey Institute of International Studies (USA),
Abdelkader Fassi-Fehri, University of Newcastle upon Tyne, (UK)
Gregory Grefenstette, LIC2M/CEA-LIST, (France)
Ahmed Guessoum, University of Sharjah, (UAE)
Nizar Habash, Columbia University, (USA)
Mohamed Hassoun, ENSIB, Lyon (France)
Steven Krauwer, ELSNET and Utrecht University Utrecht (The Netherlands)
Mohamed Maamouri, LDC, University of Pennsylvania, (USA)
Bente Maegaard, CST, University of Copenhagen, (Denmark)
John Makhoul, BBN Technologies, GTE Corp (USA)
Chafic Mokbel, University of Balamand (Lebanon)
Abdelhak Mouradi, ENSIAS (Morocco)
Owen Rambow, Columbia University, (USA)
Mohsen Rashwan, RDI (Egypt)
Horacio Rodríguez, Universitat Politècnica Catalunya, (Spain)
Paul Roochnick, -Apptek, (USA)
Mike Rosner, University of Malta, (Malta)
Paolo Rosso, Universidad Politécnica Valencia, (Spain)
Salim Roukos, IBM T.J. Watson Research Center (USA)
Jean Senellart, SYSTRAN (France)
Abdelhadi Soudi, ENIM (Morocco)
Mustafa Yassen, Amman University Amman (Jordan)