NEMLAR Newsletter

News on Arabic language resources and Arabic language technologies

Issue 19

April 2010

Newsletter content:

1. News from the MEDAR project

First Evaluation Campaign within MEDAR

In MEDAR, two baseline MT systems have been developed by customizing the MOSES (http://www.statmt.org/moses) open-source MT system. The translation focused on the English-to-Arabic direction. One of the systems has been customized by IBM/Egypt in cooperation with Dublin City University (Ireland), and the other one by University of Balamand (Lebanon). Both baselines have been made available to the MEDAR partners and are used with a minimal training corpus.

So as to estimate the performance of the two baseline MT systems, an evaluation campaign has been carried out. External systems have been invited to participate in this campaign so that the community could also profit from the evaluation data.

These evaluation data consist in the following: 210,000 words have been collected from internet, focusing on the Climate Change topic. The corpus is split into two sub-corpora: the former is the test corpus, against which the systems are evaluated (10,000 words), while the latter is the masking corpus used to keep the test corpus unknown. The test corpus has been translated four times according to specific guidelines, and then validated so as to obtain high quality data.

Finally, 9 submissions have been received: two from the MEDAR baseline systems, five from external participants, and two from online MT systems. The evaluation campaign has been carried out between January and February 2010: participants received the scores provided by an automatic evaluation and a human evaluation is currently ongoing.

This campaign, considered as a dry-run for a first experimental evaluation on English-to-Arabic leads to a second evaluation campaign that will take place in June 2010. Besides estimating the performance of the MEDAR MT systems, it will focus on the improvement capability of MT systems after a training period. To that aim, training data will be provided to participants.

MEDAR at LREC 2010

The seventh international conference on Language Resources and Evaluation (LREC 2010) is this year held in Valetta, Malta 17 – 23 May.

The MEDAR project co-organising the Workshop on Language Resources (LRs) and Human Language Technologies (HLT) for Semitic Languages - Status, Updates, and Prospects on May 17.

The MEDAR project will also be present at the ***EU Project Village***.

The NEMLAR Network

The NEMLAR Network is reconstructed in order to make it easier for the members to share information and make contact. We have created The ***NEMLAR Network*** group in LinkedIn. The Network is open to all experts in the field of Arabic language technology who share our goals. It is the intention that all members can use the network to discuss new activities, find partners etc.

Therefore we invite you to join the new ***NEMLAR Network*** group in LinkedIn http://www.linkedin.com/e/vgh/2410852/ (You sign up as an individual).

You can contribute to the ***NEMLAR Networ****k* group by:

- giving a description of your fields of interest and expertise in Arabic language technology (in your LinkedIn Profile)
- linking to your activities and languages resources
- starting and participating in discussions
- starting and participating in subgroups

2. Upcoming Events

◆ LREC 2010 - The 7th edition of the Language Resources and Evaluation Conference will take place in Valetta (Malta) on May 17-23, 2010.

◆ Workshop on Language Resources (LRs) and Human Language Technologies (HLT) for Semitic Languages - Status, Updates, and Prospects
To be held in conjunction with LREC 2010
17 May 2010, Mediterranean Conference Centre, Valetta, Malta

◆ Language Resources: From Storyboard to Sustainability and LR Lifecycle Management
To be held in conjunction with LREC 2010
23 May 2010, Mediterranean Conference Centre, Valetta, Malta

◆ Call for papers:
First Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)
A NAACL-HLT 2010 Workshop , 5 or 6 June 2010, Los Angeles, CA

◆ Call for papers:
International Conference on Information Society (i-Society 2010)
Technically Co-Sponsored by IEEE UK/RI Computer Chapter
28-30 June, 2010, London, UK

◆ Call for papers:
Arabian Journal for Science and Engineering (AJSE)
Theme Issue on Arabic Computing
Deadline for submission: June 30 2010

◆ 6th International Computing Conference in Arabic
May 20-21, 2010
Hammamet, Tunisia

◆ Call for papers:
TSD 2010
Thirteenth International Conference on TEXT, SPEECH and DIALOGUE
Brno, Czech Republic, 6-10 September 2010

◆ IMCSIT 2010
International Multiconferenc on Computer Science and Information Technology
Wisla, Poland, October 18-20, 2010

◆ Workshop:
Computational Linguistics - Applications (CLA'10)
Wisla, Poland, October 18-20, 2010

◆ CALL FOR COMPETITION:
ICFHR2010 Arabic Handwriting Recognition Competition
November 16-18, 2010
Kolkata, INDIA

3. Other news

◆ Linguistic Data Consortium has created detailed guidelines for manual word alignment in Chinese-English and Arabic-English, for the DARPA GALE program. Corpora developed under these guidelines

will be published in LDC's catalog in the coming months.
See the alignment guidelines for Arabic here.

◆ **Call for Contributions for a Special Issue of MT Journal**

Special Issue on: Machine Translation for Arabic

Submission Guidelines:

Contributors must send a "Submission Intent" email message to habash@ccls.columbia.edu no later than May 1, 2010.
Contributions will be accepted starting May 1, 2010 through June 15, 2010.

Special Issue Guest Co-Editors:

Nizar Habash (Columbia University)
Hany Hassan (Microsoft Research)

◆ Tashaphyne : Arabic Light Stemming and segmentor. API for python

Features

- Arabic word Light Stemming
- Root Extraction
- Word Segmentation
- Word normalization
- Default Arabic Affixes list

◆ **Call for Participants**

Dr. John Andrew Morrow and Barbara Castleton are currently conducting research on the impact of Global English on the Arabic Language. Native-speakers of Arabic, who are also proficient in English, are kindly invited to complete the online survey.

◆ ACM Transactions on Asian Language Information Processing (TALIP)
Special Issue on Arabic Natural Language Processing (ANLP)
Volume 8 , Issue 4 (December 2009), table of contents

◆ Arabic Gigaword Fourth Edition, a comprehensive archive of Arabic newswire text that has been acquired over several years at LDC.

◆ Encyclopedia of Arabic Language and Linguistics, 5

◆ Meedan Releases the World's First Open Access Arabic/English Translation Memory.
For more info

◆ A new version of the Crescent Quran Corpus is now freely available online. The corpus contains both morphological and syntactic annotation of the Quran in Arabic. Previous releases of the corpus focused on the morphology of Classical Arabic, but this new release now includes an in-progress syntactic treebank of the Quran.

4. How to contribute to the MEDAR project

Please help by sending us information on coming events, new software, resources, books, papers and journals. We will distribute it via this newsletter and put it on the MEDAR homepage www.medar.info .

Join the new ***NEMLAR Network*** group in LinkedIn.
You can contribute by:

- giving a description of your fields of interest and expertise in Arabic language technology
- linking to your activities and languages resources
- starting and participating in discussions
- starting and participating in subgroups

To subscribe or unsubscribe to the newsletter, please send an email to: nemlar@hum.ku.dk

The newsletter is published by the MEDAR project www.medar.info and produced by Centre for Language Technology, University of Copenhagen, Denmark.

Contact: Bente Maegaard (Coordinator of MEDAR) or
Dorte Haltrup Hansen,
Centre for Language Technology,
University of Copenhagen,
Njalsgade 140,
Copenhagen S,
Denmark
Tel: +45 35 32 90 74, Fax: +45 35 32 90 89
nemlar@hum.ku.dk