MEDAR – collaboration between European and Mediterranean Arabic partners to support the development of language technology for Arabic

Bente Maegaard[1], M. Atiyya[2], K. Choukri[3], S. Krauwer [4], C. Mokbel [5], M. Yaseen [6]

[1] University of Copenhagen, Centre for Language Technology (CST)
Njalsgade 80, DK-2300 Copenhagen S
E-mail: bentem@hum.ku.dk
[2]The Engineering Company for the Development of Computer Systems, Egypt
E-mail: m_Atteya@RDI-eg.com
[3]Evaluation and Language resources Distribution Agency; ELDA, France
E-mail: choukri@elda.org
[4]University of Utrecht, The Netherlands
E-mail: steven.krauwer@let.uu.nl
[5]University of Balamand, Lebanon
E-mail: chafic.mokbel@balamand.edu.lb
[6] Amman University, Jordan.
E-mail: mustafa@ats-ware.com

Abstract

After the successful completion of the NEMLAR project 2003-2005, a new opportunity for a project was opened by the European Commission, and a group of largely the same partners is now executing the MEDAR project. MEDAR will be updating the surveys and BLARK for Arabic already made, and will then focus on machine translation (and other tools for translation) and information retrieval with a focus on language resources, tools and evaluation for these applications. A very important part of the MEDAR project is to reinforce and extend the NEMLAR network and to create a cooperation roadmap for Human Language Technologies for Arabic. It is expected that the cooperation roadmap will attract wide attention from other parties and that it can help create a larger platform for collaborative projects. Finally, the project will focus on dissemination of knowledge about existing resources and tools, as well as actors and activities; this will happen through newsletter, website and an international conference which will follow up on the Cairo conference of 2004. Dissemination to user communities will also be important, e.g. through participation in translators' conferences. The goal of these activities is to create a stronger and lasting collaboration between EU countries and Arabic speaking countries.

1. Background and Mission

The development of language resources and tools for the Arabic language is important for the economy in the Arab countries; but at the same time it is important for the culture. By focussing on Arabic language technology and making both the technology and content available in Arabic, the use of Arabic will grow and the request for foreign language information will decrease. At the same time language technology can help access information in foreign languages, even without a very good knowledge of these languages. And finally, it can help spread Arabic ideas and culture to non-Arabic languages.

The NEMLAR project, 2003-2005, was funded by the European Commission with the goal of supporting cooperation between European actors and actors in the Mediterranean region with the purpose of advancing the state of language technology for Arabic and other regional languages. See www.nemlar.org.

The goals of the NEMLAR project was the production and availability of shareable LRs and tools, the advancement of Arabic language technology, and the collaboration between countries towards these goals. MEDAR shares this mission**.**

2. MEDAR overview

MEDAR is structured in three overlapping 'streams': 1) the technical stream, 2) the Cooperation Roadmap stream, and 3) the dissemination stream.

The technical stream consists of a survey part and an LR and tools production and evaluation part. The Cooperation Roadmap stream also builds on the survey, so it contains the survey part, the roadmapping part and the network part. Finally the dissemination stream covers everything associated with dissemination.

3. Survey and BLARK

The project will identify the state-of-the-art of language resources (LRs) and tools in the region, and assess priority requirements through consultations with language industry and communication players. In doing so, the project will build on already existing material, e.g. the NEMLAR survey, as well as the NEMLAR BLARK (Basic Language Resource Kit) 2006 (Maegaard et al. 2006).

The survey conducted within NEMLAR led to the first directory of players, resources, projects and technology providers. It is the plan of MEDAR to update this directory. The project will also better analyze the strengths, weaknesses, opportunities and threats to the development of Arabic and other language resources in the region and establish a set of key priorities for developing tools required by applications such as machine translation, information retrieval etc.

It is a key part of this project to provide knowledge about the language technology players, projects (ongoing activities), products etc. Therefore a survey will be carried out covering all Mediterranean countries participating in

the project, plus others where possible, resulting in a knowledge base with details of all universities, research institutions and companies, as well as ongoing projects, and existing products, - with relation to tools and Language Resources (LRs), in particular for MT, information retrieval and indexing.

Questionnaires are being circulated through a large number of channels (existing mailing lists, linguistic societies, language technology communities, relevant conferences, (LREC, COLING,, INTERSPEECH, etc.), and through the partners involved in the consortium.

This survey and directory will also contribute to the Cooperation Roadmap that depicts the areas and themes for bilateral as well as multilateral collaboration, in particular in terms of collaboration between EU countries and Arabic speaking countries. It is essential for the project to consider the potential cooperation between private and public sectors, in particular whenever possible to get SMEs involved in this cooperation.

The BLARK for Arabic which was developed in NEMLAR (Krauwer et al. 2006) will be further elaborated based on the new surveys, and with particular emphasis on multilinguality: machine translation and other tools for translation, multilingual information retrieval etc.

4. The main part of the Technical stream

Building upon the survey, the project will address all issues related to Machine Translation and Multilingual Information Retrieval for Arabic. This covers activities like a survey of state-of-the-art with respect to the different approaches used and in particular the ones made available via open source software. It will also identify the requirements in terms of technology components to adapt and customize such software to Arabic as well as the language resources needed. MEDAR will also provide benchmarking guidelines based on best practice in major evaluation campaigns.

The survey (cf. sect. 3) will identify existing Machine Translation and Multilingual Information Retrieval tools, both at universities and provided by industry. In order to ensure that such tools are usable for Arabic, a task will focus on identifying all the obstacles that may prevent such use. This will include identification of language resources that are needed but also of basic NLP tools that may be required and that could be language-dependent (lemmatizer, POS tagger, vowelizer, etc.). The survey will cover language resources (LRs) currently available to build MT and MLIR systems and benchmark them.

The consortium will also work on the specifications and the design of cost-effective production processes for annotating and tagging resources necessary for MT and MLIR development. This task will help promote the evaluation spirit through the dissemination of information on evaluation programs and evaluation packages (data, metrics, reports, methodologies, best practices) suitable for Arabic HLT in general.

Following NEMLAR's experience on studies of this kind and in particular the ones related to the definition and the establishment of a BLARK as required by Human-Language Technologies, we aim at establishing an LR definition/roadmap programme for the field of Cross Lingual Information Retrieval (CLIR) and MT. For the CLIR tools, it is important to investigate applications similar to Cross-Language Spoken Document Retrieval (CLSDR) which combines information retrieval, cross-language translation and speech recognition. The consortium will investigate the best solution for such combinations that could be based on components that exist as open source code or as background knowledge of partners. The use of existing tools will be customised for Arabic. The baseline will be using text search requests to retrieve spoken documents.

The consortium would like also to work on enhancement of MT tools through the use of existing resources and tools. Such resources are e.g. alignment tools, lexica, named entity recognition etc. Consortium partners already have full MT systems, as well as several relevant tools. Open source components will also be used, as it is important to make sure that the results obtained will be commonly useful, not only applicable to one system.

Several methodologies exist for the evaluation and benchmarking of language technology, in particular MT. The project will build on the state-of-the-art and previous experiences of the project partners to define a baseline MT system exploiting open source software. The project will also conduct preliminary benchmarking of these tools, and any others provided by external participants willing to join.

The cost of evaluation campaigns is such that MEDAR cannot afford to conduct a full evaluation campaign. The action will adapt and customize some of the resources produced within other projects, see e.g. (Hamon et al. 2006) and go beyond the automatic metrics and investigate other approaches including the ones based on user evaluation. The tools will be assessed at the beginning and after the inclusion of resources (e.g. bilingual lexica, named entities) and comparisons will be made.

5. Cooperation Roadmap

5.1. Three interconnected roadmaps

Referring to the work done in Europe in particular under the ELSNET and other projects, we can define a roadmap as "a document that indicates directions for a planned journey and shows how and in what order goals can be reached and indicates distances". Usually one focuses either on a roadmap as "time to market" for a new product (market dimension) or on a roadmap reflecting expected "technology developments and trends" (technology roadmap). In our case we will add a new and essential dimension, which is the roadmap for cooperation between Arabic and European Union countries (cooperation roadmap) with a view to (eventually) implementing the market and technology roadmaps.

The challenge is to create strategic partnerships, i.e. long term partnerships based on mutual benefit. Such

partnerships may be used to execute research or development together, and of course to make joint proposals to open calls, be it in Europe or in Arabic countries.

For each of the three dimensions listed above we will analyze and report on the present situation in the participants' countries, we will describe the conditions that need to be fulfilled in order to arrive at key achievements and at strategic partnerships and we will describe the steps that need to be taken to get there from where we stand. Our primary focus will be on multilingual tools, in particular on machine translation and multilingual information retrieval but other areas will also be addressed (e.g. speech technologies, language processing components, etc.). For the Cooperation dimension we will make sure to place the human resources at the heart of our recommendations (staff exchange, education curricula, etc.). For the execution of this task we will build on the already existing network of key players that was created during the NEMLAR project, and that will be gradually expanded in MEDAR.

5.2 Technology Roadmap

The survey mentioned above will provide input to the roadmapping task by assessing priority requirements through consultations with language industry and communication players, and by establishing a protocol and a roadmap for developing a set of Language resources for all components related to MT and IR.

In the Technology Roadmap we will focus on technologies aimed at taking away language barriers (translation) or providing access to information services (text, speech or across languages). For these technologies and for each of the dimensions mentioned above we will analyze the present situation in the participants' countries, we will describe the conditions that need to be fulfilled in order to arrive at strategic partnerships and we will describe the steps that need to be taken to get there from where we stand.

The idea is to define the state-of-the-art first, then define the state-of-the-art as foreseen within the next 5 to 7 years. This will help identify the major challenges that need to be addressed to achieve the stated goals. The approaches and solutions will be associated to each challenge given our knowledge today and the emerging techniques. The ultimate goal is a common agenda based on reliable predictions to which the R&D and development community will adhere.

We will also consider the enabling technologies, in particular internet access, available within the Arabic regions and the impressive development of mobile communications and satellite broadcasts in the region.

On the basis of the-state-of-the-art and the expected application scenarios (which may be different from region to region) we will determine what the technological requirements and priorities are to create or expand the envisaged applications. This includes the identification of major obstacles that have to be addressed and resolved in order to make the scenarios come true. We anticipate that the main components that are needed are:

- Language resources for Arabic and other local languages including in their relationship to foreign languages (e.g. parallel corpora for statistical MT)
- Language technology tools for the cost-effective creation, exploration and exploitation of the resources
- Solutions, i.e. modules that serve as enabling technologies for applications.

The main output of this activity is a technology roadmap with temporally ordered priority list of specific technologies and resources needed to implement the scenarios, recommendations and priorities including costs and sources of funding.

5.3 The Market Roadmap

For a flourishing market of ICT products or services one needs customers and a delivery infrastructure. In our study we will first of all describe

- which (low cost) products and services are actually available on the market
- actual (niche) application scenarios in which they are being used
- available delivery infrastructures.

In addition we will gather (non-exhaustive) information about how similar products and services are being used in other parts of the world, preferably in countries with similar conditions (e.g. low literacy, limited knowledge of foreign languages) and on this basis we will try to sketch possible scenarios for increasing the usage (more niches, larger niches) and impact of the technologies. The report will take into account both demand and policy driven factors.

We do not envisage conducting an extensive market study, but we will try to extrapolate from what we know about what is done and what we think can be done and get some of the local government agencies involved through our local partners. The scenarios will help us to determine what is required along the other dimensions. The scenarios will also help in bringing together local actors (who know the customers and the market) and actors from EU countries (who are operating on similar markets elsewhere) who could jointly implement some of these scenarios.

The main output of this activity is an overview of what is available, and an overview of current practice examples from elsewhere, together with a number of possible evolutionary scenarios.

5.4 The Cooperation Roadmap

The importance of this new dimension of roadmapping is to boost cooperation mainly in the non-competitive areas emphasized above, such as LRs and basic HLT components, tools and solutions. For each of these topics we will analyze and report on the available skills and expertise and define scenarios and conditions for joint projects. We anticipate that human resources constitute a key factor for successful cooperation. In addition we see a number of technological enablers for cooperation, such as common interoperability and representation standards,

common benchmarks, sharing of existing resources and joint construction of new ones.

Cooperation can involve different types of partnerships: cooperation between Arabic and EU partners, cooperation between Arabic partners, cooperation between academia and industry. All three types of cooperation require mutual visibility: You cannot cooperate with someone if you are not aware of his existence, his expertise and his needs. We will use the network as one of our main instruments to create this awareness by publishing partner profiles, and by providing information about cooperation opportunities.

Staff exchange programmes are of crucial importance in this respect. This could include staff from Arabic countries to spend time in EU companies or research labs, as well as staff from EU organizations to spend time in Arabic labs. An exchange grants system should facilitate this, and the consortium will make an effort to locate funds that could be used by the partners or other interested parties. We will also analyze the horizontal possibilities of intra-Arabic countries cooperation.

As the large EC programmes give Europe a significant technological advantage we will also investigate ways to continuously disseminate and promote the results of such programmes to the Arabic countries. Recommendations for this will be included in the final Cooperation Roadmap report. For the longer term it is essential that language and speech technology find their way to higher education curricula in the Arabic speaking world. Recommendations for how to achieve this will be included in the report.

We will provide a Cooperation Roadmap, which will take into account both human, technological and market factors and which will identify the main priorities and interdependencies, and which will also provide recommendations for actions that could be undertaken to speed up the process and to pave the way for lasting strategic partnerships along each of the three dimensions.

6. Network creation

Experience shows that one of the important side-effects of many large scale research projects and coordination actions is the creation of lasting (formal or informal) transnational networks of individuals and organizations that serve as platforms that help setting up new partnerships and launching new collaborative projects.

It is on of the goals of MEDAR to create such a network that will include both players from the Arabic speaking countries and from EU countries in order to facilitate and support collaboration, both among the Arabic countries and between the EU and Arabic countries. We are building on the network that was created under the NEMLAR project and will expand it through all channels, e.g. other networks. We have a strong belief that the cooperation roadmap will be an asset in attracting new members.

The network is open to new participants and will be expanded with players from the Arabic speaking community and with parties from the EU countries who have an interest (commercial or scientific) in Arabic language and speech processing.

The MEDAR website (see below), electronic mailing lists and the electronic newsletter will ensure a continuous flow of information between the network members. Special attention will be paid to collaboration opportunities offered by the EC and national research or innovation programmes. MEDAR will publish company and organization profiles on the website in order to support the creation of new partnerships.

The project will benefit from the network in that it can be used to gather information for surveys and reports (see the sections on the survey and roadmapping above) and that it can be used for wider distribution of the results of the project. The network can also be used to (experimentally) implement recommendations aimed at creating more and better collaboration opportunities. Working groups of network members may be set up to discuss specific issues or to carry out specific tasks.

7. Dissemination

For a project like MEDAR, dissemination is a very important feature, - the third ‘stream’ as mentioned above. Dissemination is omnipresent and MEDAR uses a number of instruments for dissemination.

The dissemination targets are of different types:

- Actors at universities in Europe and the Arabic countries, i.e. researchers with an interest in Arabic language processing.
- Industrial players in Europe and the Arabic countries.
- Users and user organizations
- Funding agencies

7.1. Academic and industrial players

Obviously, the project will focus on the partner countries, but attention will be paid also to other Arabic countries than those in the project, e.g. the quite active players in the Gulf region.

The project has created the MEDAR website, www.medar.info, for collecting and disseminating global information on Arabic and local language resources, tools and technologies.

MEDAR will raise awareness of the state of play of Arabic and local language resource and tools development among all stakeholders by disseminating a regular information newsletter, ensuring information feeds to existing networks and information sources.

MEDAR also participates in relevant conferences, preferably with presentations, both nationally and internationally. MEDAR organizes workshops and panels at relevant events.

It is the hope that the network will also be an efficient dissemination channel, and that the Roadmap will attract attention.

7.2 Users and user organizations

But the dissemination should also reach other types of audiences: First of all users. Users should get information about the possibilities in new technology. This will happen e.g. at user conferences. The project will monitor relevant conferences and seek to have a participant with presentation. In particular the project will participate in the Third International Translation Conference, to be held in Beirut in 2008. The coordinator already participated with a presentation in the Second International Translation Conference, held in Amman 2007 (Maegaard 2007). This translation conference will be an excellent supplement to the MEDAR conference which will be more technological.

7.3 Dissemination to funding agencies

The project will also make funding agencies aware of the possibilities in language technology for automated translation and multilingual information retrieval.

Strategic planners, policy-makers, decision-makers and funding agencies who have an interest in promoting research & development and innovation for the Arabic language in the field of language technology will be one of the main driving forces. Such bodies may want to support the communication and cultural dialogue between e.g. Europe and the Arab countries, or they may want to attack the unemployment and poverty by enhancing the efficiency and effectiveness of the local economies and bridge the digital divide in the Arab countries.

7.4 International conference

Finally, in 2009 MEDAR will organize an international conference on current activities and future orientations in Arabic and local language resource and tools creation and management, in particular for multilingual applications, such as machine translation and information retrieval. The first International Conference on Arabic Language Resources and Tools was organized by NEMLAR in Cairo 2004, and was a big success.

8. Conclusion

Like its immediate predecessor NEMLAR MEDAR is not really about technology but about collaboration. Both projects share the long-term objective to create better conditions for the development of language and speech technology for Arabic. NEMLAR started out defining the essential resources needed to bootstrap language and speech technology for Arabic (the BLARK for Arabic), and surveying the Arabic resources landscape on the basis of this BLARK so that gaps could be identified. At the same time it set up a network of language and speech technology experts in the Arabic speaking world and in the EU and carried out some small scale collaborative resources creation projects.

MEDAR continues this line of activities. We are extending the survey and will improve the BLARK but we have added a new, future-oriented dimension aimed at creating conditions for future collaborative work between experts from Arabic speaking and EU countries. In order to give this activity a clear and concrete focus we will concentrate on technologies aimed at taking away language barriers. On the basis of an analysis of the present situation and possible application scenarios we will draw up a roadmap indicating future directions, technological obstacles to be addressed and opportunities for collaborative actions within the Arabic speaking world and with EU partners to take away the obstacles.

In order to provide some empirical background we will initiate a number of small-scale projects aiming at enhancing existing technologies in the light of the emerging scenarios. This is not only expected to improve these technologies but will also give us insight in possible collaboration models.

As we have seen in US and EU language and speech technology programmes that evaluation can become a driving and stimulating force behind technological advancement we will also develop models for evaluation, including user evaluation, in the development of language and speech technology tools for Arabic.

9. Acknowledgements

We want to thank the European Commission for the support to this important activity.

This paper builds on work done in NEMLAR, as well as the preparation and the first part of MEDAR. MEDAR has 15 partners, and we want to acknowledge the contribution of all of them:

- Bente Maegaard, University of Copenhagen, Denmark
- Khalid Choukri, ELDA - Evaluations and Language resources Distribution Agency, France
- Chafik Mokbel, University of Balamand, Lebanon
- Mustafa Yaseen, Al-Ahlyya Amman University, Jordan
- Steven Krauwer, Universiteit Utrecht, The Netherlands
- Stelios Piperides, ILSP - ATHENA Research Center, Greece
- Mohammad Attiyya, RDI, The Engineering company for computer systems development, Egypt
- Kanan Ali, Birzeit University, West Bank and Gaza Strip
- Abdelhak Mouradi, ENSIAS - University of Mohammed V Soussi, Morocco
- Nasredine Semmar, CEA - Laboratoire d´ingénierie de l´information multimédia multilingue, France
- Fathi Débili, CNRS - Centre Nationale de la Recherche Scientifique, France
- Anne DeRoeck, The Open University, United Kingdom
- Joseph Dichy, Université Lumière Lyon 2, France
- Ossama Emam, IBM - Human Language Technologies Group, Egypt
- Achraf Chalabi, Sakhr Software Company, Egypt

10. References

Hamon, O., **A.** Popescu-belis, K. Choukri, M. Dabbadie, A. Hartley, W. Mustafa El Hadi, M. Rajman, I. Timimi

(2006): CESTA: First Conclusions of the Technolangue MT Evaluation Campaign. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy, 2006.

B. Maegaard, S. Krauwer, K. Choukri, L. Jørgensen: The BLARK concept and BLARK for Arabic. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, 2006. p. 773-778

Krauwer, S., B. Maegaard, K. Choukri, L. D. Jørgensen: *BLARK for Arabic*, NEMLAR report, 2006, www.nemlar.org

Maegaard, B., L. Damsgaard Jørgensen, S. Krauwer, K. Choukri (2004): NEMLAR: Arabic Language Resources and Tools, In: K. Choukri and B. Maegaard (ed.): *Proceedings of Arabic Language Resources and Tools Conference*, p. 42-54, Cairo.

Bente Maegaard: Machine Translation and Multilingual Language Technology. In: Second International Translation Conference Proceedings, Amman, 2007, p. 228-238.

Maegaard, B. (2004): NEMLAR – an Arabic Language Resources project. In: *Fourth International Conference on Language Resources and Evaluation, Proceedings Vol I*, p. 109-112, Lisboa.

Maegaard, B, Choukri, K, Mokbel, S and Yaseen M. (2005) Language Technology for Arabic, University of Copenhagen, Denmark. See www.nemlar.org

M. Yaseen, M. Attia, B. Maegaard, K. Choukri, N. Paulsson, S. Haamid, S. Krauwer, C. Bendahman, H. Fersøe, M. Rashwan, B. Haddad, C. Mokbel, A. Mouradi, A. Al-kufaishi, M. Shahin, N. Chenfour, A. Ragheb (2006): Building Annotated Written and Spoken Arabic LRs in NEMLAR Project. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, 2006. p. 533-538.