Cooperation for Arabic Language Resources and Tools – The MEDAR Project

Bente Maegaard[1], M. Attia[2], K. Choukri[3], O. Hamon[3], S. Krauwer [4], M. Yaseen [5]

[1] University of Copenhagen, Denmark
E-mail: bmaegaard@hum.ku.dk
[2]The Engineering Company for the Development of Computer Systems, Egypt
[3]Evaluation and Language resources Distribution Agency, France
[4]University of Utrecht, the Netherlands
[5] Amman University, Jordan.

Abstract

The paper describes some of the work carried out within the European funded project MEDAR. The project has three streams of activity: the technical stream, the cooperation stream and the dissemination stream. MEDAR has first updated the existing surveys and BLARK for Arabic, and then the technical stream focused on machine translation. The consortium identified a number of freely available MT systems and then customized two versions of the famous MOSES package. The Consortium addressed the needs to package MOSES for English to Arabic (while the main MT stream is on Arabic to English). For performance assessment purposes, the partners produced test data that allowed carrying out an evaluation campaign with 5 different systems (including from outside the consortium) and two online ones. Both the MT baselines and the collected data will be made available via ELRA catalogue.

The cooperation stream focuses mostly on the cooperation roadmap for Human Language Technologies for Arabic. Cooperation Roadmap for the region directed towards the Arabic HLT in general. It is the purpose of the roadmap to outline areas and priorities for collaboration, in terms of collaboration between EU countries and Arabic speaking countries, as well as cooperation in general: between countries, between universities, and last but not least between universities and industry.

1. Background and Mission

The goals of the MEDAR project are the production and availability of shareable LRs and tools, the advancement of Arabic language technology, in particular multilingual resources and tools **and the collaboration between institutions from the Euro-Mediterranean countries towards these goals**. We are now almost through the project and can see how it has worked. We believe that the approach can be generalised and applied successfully in other regions.

MEDAR is structured in three overlapping 'streams': 1) the technical stream, 2) the Cooperation Roadmap stream, and 3) the dissemination stream. This paper has its main focus on the Cooperation Roadmap.

2. Technical stream

The overview of language resources in the BLARK for Arabic (Krauwer et al. 2009) has been extended with a special focus on parallel corpora, thus facilitating the multilingual activities of the project.

3. MT for Arabic

The consortium is working with open source MT systems and improving their results for Arabic through the use of dedicated language resources and tools. The project decided to focus on the English to Arabic translation rather than on Arabic to English that is widely addressed by the MT community nowadays. The MT task also helps promoting the evaluation spirit through information on evaluation programs and evaluation packages (data, metrics, reports, methodologies, best practices) suitable for Arabic and through the specific collaborative work on evaluation of the MEDAR MT prototypes.

3.1 Baseline MT systems

Taking as starting point the idea to use open source software to develop our baseline MT systems on the English to Arabic direction, we first decided to identify what was available among the community. That resulted in the identification of five open source MT systems: Apertium[1], GenPar[2], JosHUa[3], Matxin[4]and MOSES[5]. The latter has be then selected as the most suitable one for three main reasons: its wider community, some first experiments on English-to-Arabic translation which obtained interesting results (see for instance: (Schwenk et al., 2008), (Badr et al., 2008)), and most of all, because of the experience of some consortium partners with this system (Koehn et al., 2007). Using a statistical machine translation system may also allow us to obtain a baseline system with basic performance quickly, using adequate and available training data. After this identification step, it was decided that a baseline system would be developed for MEDAR, by University of Balamand (Lebanon) and an enhanced version would be supplied by a joint team from IBM-Egypt and Dublin City University. At the end of the process, both of them were able to provide translation from English to Arabic, as required by the project. Following this, each partner of the consortium installed, tested and made some suggestions to improve their usability. However, as these baseline MT systems needed some training data to provide good quality translations, a new step to collect training data has been carried out.

[1] http://xixona.dlsi.ua.es/apertium-www/?lang=en
[2] http://nlp.cs.nyu.edu/GenPar/GenPar.html
[3] http://www.clsp.jhu.edu/wiki2/JosHUa_-_JHU_Open_Source_Architecture
[4] http://matxin.sourceforge.net/
[5] http://www.statmt.org/moses/

3.2 Training data collection

One of MEDAR's goals is to produce a large parallel corpus and a monolingual corpus to train MT systems. The objective is to reach about 50M words that will be shared first among MEDAR partners, then with the rest of the community. This resulted in two main identification tasks: first, we established a list of existing LRs being likely to be used within MEDAR, and then we started the identification of websites that could potentially be compiled as a corpus. The identification work has been shared by all partners collecting a number of URLs, checking their content for Arabic/English data and if these data were potentially alignable.

Data have then been crawled, formatted, cleaned and finally aligned using scripts and tools such as Champollion Tool Kit [6] or Hunalign [7] regarding the alignment. In parallel to this a number of legal negotiations have been initiated to ensure that the data is usable by a wider community.

Furthermore, two monolingual corpora for both Arabic and English are currently being produced with sources coming from the Internet.

The procedure for both corpus creations is planned to finish in May 2010.

3.3 Evaluation

So as to know what kind of quality we started from, an evaluation of our two baseline systems has been carried out. Moreover, we decided to share the evaluation data with external participants and therefore make the community benefit from an evaluation campaign.

To build the evaluation corpus, 10,000 words were collected from many different websites on the climate change domain. The selected domain is rather "specific" but we assumed that its specialized jargon is becoming enough general public. The test corpus has been completed by 200,000 words used as a masking corpus, so as not to permit the identification of test data by the participants. In order to perform an automatic evaluation, four translations were produced by professional translators, using translation guidelines specific to English-to-Arabic translation. The manual translations have also been validated against validation guidelines in order to control their quality.

The evaluation campaign was held at the end of January, participants having 10 days after receiving source data to send back their automatic translations. We finally received 5 submissions from external participants, in addition to our two baseline systems' outputs and two online MT systems we used. It has been decided that the results will remain anonymous. Indeed, no training data has been provided to participants, who did not have much time to prepare and customize their systems: thus, results are not really objective and comparable, and the campaign should be seen more as a dry-run rather than as an official campaign. However, we can already mention that performance is quite low, the best MT systems getting a Word Error Rate (WER) of 65%, and our best baseline system a WER of 80%. Although this shall be considered according to the evaluation conditions, this at least shows that statistical machine translation cannot work without large training corpora. A human evaluation is also currently being carried out to check our findings.

To test the relevance of using training data, a second evaluation campaign is planned for June 2010. It will use training data for English-to-Arabic translation, using the same evaluation protocol.

The achievement of this phase of the project will be capitalized upon through several actions: a) the "baseline" systems (and any enhanced releases) developed within the project will be made publicly available to the community, assuming they adhere to the initial licensing conditions (wrt MOSES, Ghiza++, etc.), b) the evaluation resources (test suites, reference translations, metrics, etc.) will be packaged and made widely available through ELRA catalogue and its licensing schemas, c) the training data will be also made available to the community under ELRA licensing and at fair conditions.

4. Cooperation Roadmap

Given the MEDAR goals, one of the major deliverables is creating a Cooperation Roadmap for the region directed towards the Arabic HLT in general. It is the purpose of the roadmap to outline areas and priorities for collaboration, in terms of collaboration between EU countries and Arabic speaking countries, as well as cooperation in general: between countries, between universities, and last but not least between universities and industry.

The roadmap includes the usual technology and market dimensions, but we added a new and essential dimension: cooperation. There are a number of elements that all contribute to the current status and the future development of Arabic language technology, and language resources and tools are of course an important element, but cannot be the sole element. The roadmap report describes some of the most important 'instruments' or actions that can be taken in order to influence the situation and the development, and finally provides a synthesis and recommendations for actions to be taken. The target for the roadmap is 5-7 years for each element. The roadmap until year 2015 is divided into three partially overlapping phases.

3.1 Phase 1 (2010-2012): Laying the foundations: Education and Basic Language Resources

The MEDAR Survey shows that the number of Arabic HLT professionals is very low, and not sufficient to build and maintain strong HLT industry in the Arab states. This means that we will have to create opportunities for the education of a new generation of researchers and developers with adequate skills in HLT.

The education system should aim at providing HLT training both to students who want to graduate and to professionals who are already working in the ICT field but who lack specific knowledge about HLT and language in

[6] http://champollion.sourceforge.net
[7] http://mokk.bme.hu/resources/hunalign

general. In addition to that there is also a need to train people to become HLT educators, as this is necessary for a sustainable supply of *HLT-enabled* professionals. We use the term *HLT-enabled* since we do not believe that the main goal of the education system should be to create a completely new HLT discipline with its own professionals, but rather to provide people who have a firm basis in one of the fields relevant for the advancement of HLT with an additional component of knowledge and skills that allow them to contribute to the development of HLT related products or services without the risk that their profile would be too narrow for a future in ICT at large.

As for digital material we have already described in the *BLARK for Arabic* document [MEDAR Report 3.2] what the basic set of language resources required for education should contain. In the same report we also gave an account of what is already available (and accessible) and what is still missing.

Possible bilateral or multilateral cooperation actions between EU and Arab states or cooperation between Arab states include joint training of teachers, curriculum or course development, schemes for industrial placements or traineeships for students, development of teaching material and e-learning courses, and creation of (essential parts of) the BLARK, where many tools might be ported from other languages in collaborative actions.

In this phase, the Internet rates will be reduced, mobile phones will spread, e-government projects will start in many countries and the ICT infrastructure in most countries will be more appropriate and affordable. Those are essential elements for pushing the Arabic HLT industry forward. Products will help in handling the literacy problems; moreover, IP laws and copyright need to be strengthened to help the industry and overcome the high rate of piracy.

3.2 Phase 2 (2012-2014): Moving forward

This phase will witness the implementation of the first HLT-enabling curricula in some Arab states, as well as teaching staff and student exchanges in both directions, and development of some essential BLARK components will be realized. It also foresees developing schemes to bring players from the Arab HLT community together in order to make them more competitive vis-à-vis the global players on the Arab HLT market More focus on e-Government and content creation for the general public, and linguistic support via mobile phones for the illiterate will boom. Industries will start utilizing Arabic HLT, there will be infotainment and entertainment services, access to educational services in particular for language learning will grow and support applications for mobiles in the Arab world. At least one application made in collaboration with universities in this phase. Development of at least one product from each country on the market in the areas of: MT systems, text analysis tools, LRs.

3.3 Phase 3 (2013-2015): First consolidation

The main expectations are:

- Implementation of improved curricula on the basis of experience gained and new technological developments,
- Regular student and staff exchanges between Arab states and EU, and between academia and industry,
- Joint projects and training activities across Arab states
- Joint projects between EU and Arab players to build new resources, applications and services.

Further development of the Arabic BLARK and creation of application or domain specific resources and tools for priority areas will be also on the consolidation agenda. By 2015 an Internet penetration of 25% in average for the region is expected, developing more e-Government projects, and the number of mobile telephones will be at least three times as is currently, at least one project should have resulted in an educational product for illiterates on the Arab market. As a result of fighting piracy, international players will be encouraged to invest in Arab markets. Local software industries can be freed from competition from pirating actors; and the local players can penetrate export markets.

3.4 Summary of the Roadmap activity

Development and growth will continue throughout the duration of the period specified (till 2015). As a result, cooperation will be binding and linking all players together, by encouraging the universities and research centers to provide HLT-enablement programs and human resources; international major players to keep and maintain their interest and to support and build the local industry. Governments and funding agencies should facilitate, support and help companies and universities to initiate and sustain their products. Governments should launch services/applications for citizens (e.g. e-government) that will be accessed and navigated in Arabic language. Local mobile companies, internet service providers and telephone companies should provide the support and encourage the local companies and universities to direct their efforts towards producing tools and utilities that could be integrated and added to the provided services.

5. Network creation and extension

One of the important side-effects of 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 one of the goals of MEDAR that its network will facilitate and support collaboration, in the Arabic countries and between the EU and Arabic countries. The network is also being used to (experimentally) implement some recommendations of the Cooperation Roadmap.

5. Dissemination

Dissemination is an important and omnipresent feature of this project.

MEDAR uses a number of instruments for dissemination that target different types of actors with interest in Arabic language processing such as:

1. Researchers and students in Europe and the Arabic countries.
2. Industrial players in Europe and the Arabic countries.
3. Users and user organizations.
4. Funding agencies.

While the project focuses on the partner countries, attention is being paid also to the rest of Arabic countries e.g. 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, technologies, and scientific literature. MEDAR also raises awareness by disseminating a regular information newsletter, ensuring information feeds to existing MEDAR network members and others.

MEDAR also participates actively in relevant regional and international conferences and events with papers, presentations, and workshops. The network is being persistently expanded through the website and the participation in events (both scientific and user related) aiming to attract attention to the project and the Roadmap.

5.1 Dissemination to funding agencies

Making funding agencies aware of the wide possibilities in language technology multilingual applications as well as other HLT enabled technologies is a particular priority for the success and continuity of the MEDAR initiative.

A main driving force in this regard is the strategic planners, policy-makers, decision-makers and funding agencies - who have an interest in promoting effective ICT applications based on Arabic language processing via innovative R&D and robust education in this field. Such bodies may want on one side support the communication and intercultural dialogue between the Arab countries and the rest of the world, and on the other side may want to attack the unemployment and poverty by enhancing the efficiency and effectiveness of the local economies via bridging the digital gap between the Arab countries and the more developed parts in the world.

5.2 International conference

MEDAR has been well exposed to both the specialized community and mass media through the Second International Conference on Arabic LRs & Tools, Cairo, Egypt, 2009. The project organized this conference with the support, sponsorships, and auspices of local and international players. Current activities and future orientations in Arabic and local LRs and tools creation and management were discussed, and the ground was laid for future cooperation, following the lines of the cooperation roadmap.

6. Conclusions

As we have seen, MEDAR is working with MT technology, but with a clear focus on language resources, tools and evaluation, and in particular on **collaboration**.

On the basis of an analysis of the present situation and possible application scenarios we have drawn up a roadmap indicating future directions, technological obstacles to be addressed, language resources to be built, collaborative education programmes to be developed, and other opportunities for collaborative actions within the Arabic speaking world and with EU partners to take away the obstacles. We believe that the approach can be generalised and applied successfully in other regions.

7. Acknowledgements

MEDAR has 15 partners, and we want to acknowledge the contribution of all of them. Below we only give one name per group, and only for those groups which do not appear as authors:

- University of Copenhagen, Denmark
- ELDA - Evaluations and Language resources Distribution Agency, France
- Chafik Mokbel, University of Balamand, Lebanon
- Al-Ahlyya Amman University, Jordan
- Universiteit Utrecht, The Netherlands
- Stelios Piperides, ILSP - ATHENA Research Center, Greece
- 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
- Michael Ghali, Sakhr Software Company, Egypt

7. References

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 (2009), *BLARK for Arabic*, www.medar.info

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

Maegaaard, Yaseen, Krauwer, Choukri (2009): Cooperation Roadmap, http://www.medar.info/MEDAR-roadmap.pdf

Second International Conference on Arabic Language Resource and Tools www.MEDAR.info/conference_all/2009/index.php.

M. Yaseen, et al. (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.