MEDAR: Arabic language technology, state-of-the-art and a cooperation roadmap

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

[1] University of Copenhagen, Centre for Language Technology (CST)
Njalsgade 140, DK-2300 Copenhagen S
E-mail: bmaegaard@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] Al-Ahlyya Amman University, Jordan.
E-mail: mustafa@ats-ware.com

Abstract

After the successful completion of the NEMLAR project 2003-2005, a new opportunity was opened by the European Commission, and largely the same partners are now executing the MEDAR project. The project has three streams of activity: the technical stream, the cooperation stream and the dissemination stream. MEDAR has updated the existing surveys and BLARK for Arabic, and now the technical stream is focusing on machine translation and information retrieval. The cooperation stream takes as its two most important activities to reinforce and extend the NEMLAR network and to create a cooperation roadmap for Human Language Technologies for Arabic. At this conference we can now present the first version of the cooperation roadmap, and we hope it will provoke fruitful discussion at the conference and written feedback on the website, so that over time it can help create a larger platform for collaborative projects. Finally, the third stream of project focuses on dissemination; this happens through newsletter, website and in particular the present international conference. The goal of these activities is to create a lasting collaboration between EU countries and Arabic speaking countries, and also to extend this collaboration to all persons and institutions who share the goal of promoting Arabic language technology in a collaborative framework.

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 goal of the **MEDAR** project, supported by the European Commission ICT programme, is to establish a network of partner centres of best practice in Arabic dedicated to promoting Arabic HLT. The tasks of the project include surveying the state of the art on language resource needs, furthering the state-of-the-art in the field of machine translation to and from Arabic, organizing a conference, disseminating information on Arabic language technology, establishing development priorities and creating a Cooperation Roadmap for the region. As mentioned, the project has a special focus on machine translation and other multilingual tools, including information retrieval. This focus is guiding not only the technical work but also e.g. the survey and the BLARK activities, whereas the cooperation roadmap has the full scope of Arabic language technology, resources and tools.

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, and other state-of-the-art information, i.e. it contains a state-of-the-art part, a roadmapping part and the network part. Finally, the dissemination stream covers everything associated with dissemination.

3. Survey and BLARK

The project has identified the state-of-the-art of language resources (LRs) and tools in the region, in particular with a view to multilinguality: machine translation and other tools for translation, parallel aligned corpora, and multilingual information retrieval. The survey report is available on the project website, www.medar.info.

The survey conducted within NEMLAR 2003-2005 led to the first directory of players, resources, projects and technology providers. It is a key part of MEDAR to provide knowledge about the language technology players, projects (ongoing activities), products etc. Therefore a survey was 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.

The MEDAR survey was launched in April 2008 and all partners were encouraged filling the questionnaire for their institution and having it filled by their partners. As of the end of December 2008 57 questionnaires were filled in, but it is still possible to contribute and we hope to receive more responses for the next version of the survey. A number of countries are well represented (e.g. 11 responses from Egypt, 10 from Morocco).

An important part of the survey is related to the technologies our respondents feel important for Arabic and they listed a large set. Many of them consolidate our own finding listing MT, CLIR/MLIR, and ASR on the top. They also listed a number of crucial resources that should be better specified and defined by MEDAR in the framework of its updating of the BLARK.

The BLARK for Arabic which was developed in NEMLAR (Krauwer et al. 2006) has been further elaborated based on the new survey and other added information, and with particular emphasis on multilinguality: machine translation and other tools for translation, multilingual information retrieval etc. It has been a pleasure to see that for the parallel corpora, many more corpora were identified than in NEMLAR. It seems that the interest has grown.

Based on the BLARK overview of what should be there for an Arabic BLARK (the BLARK specification), and on the overview of what exists, we have made an analysis of the so-called 'gaps'. The comparison of what the survey identified as needed and the needs that were stated in the BLARK document 2004, showed total agreement.

However, the identification task carried out by ELDA showed as we said above that actually there is a certain amount of parallel corpora, so we need to check which (if any) of these can be used for MEDAR's purposes.

The analysis also showed that MEDAR will need to create an evaluation package for MT. This is in line with the foreseen work plan. Finally, for the MT tools, the analysis showed that it is advisable to focus on one of the well-known Open Source tools, such as MOSES.

4. The main part of the Technical stream

This section elaborates on the MEDAR findings related to Machine Translation (MT) tools and its plans to design and develop an MT system for Arabic (either as a source or a target language, the other language being English and possibly other European languages). These activities build upon the investigations described in section 3.

This MEDAR activity allowed to identify the different approaches used for MT and in particular the ones made available via open source packages. It has also identified the requirements in terms of technology/software components and Language resources. Many of these components will require an adaptation and customization to handle Arabic and the source/target languages that will be selected. It will also require the supply of adequate Language Resources (e.g. aligned corpora) to train the adopted tool. MEDAR will also provide benchmarking guidelines based on best practice in major evaluation campaigns and implement some of them. Following such evaluation, the MEDAR consortium will proceed to the development of new Language Resources for the improvement and enhancement of such a baseline MT system. The MT tools will be assessed at the beginning and after the inclusion of resources and tools developed by the project, and comparison will be made.

Following the survey and the identification of state of the art, the MEDAR consortium elaborated a detailed workplan to produce such an SMT baseline and then an enhanced version that would benefit from contributions of all partners has been drafted and adopted.

Some of the critical aspects that were to be tackled by the consortium when deciding on which path to go can be itemized as:

- Identify existing SMT systems and decide which one (ones) to consider both as a full package and as components to combine with modules of some of the partners (e.g. sentence aligners);
- Ensure that the project develops/customizes the selected tools so as to obtain a well recognized state of the art baseline.
- Identify LRs (or produce some) and tools that may improve performance
- Set up an evaluation framework and evaluate the baseline system (define an evaluation methodology, design and collect test sets, conduct an evaluation that is open to other systems, etc.)
- Perform evaluation after the use of additional material
- Ensure the largest availability of such a tool and the corresponding LRs before and after enhancement
- Work on the standardization of input and output (though we will recommend using Unicode), it is important to ensure that the tools selected can handle that.
- It is important to focus on the Modern Standard Arabic but one should not leave out colloquial Arabic(s) if possibilities are offered to do so, including for a system to translation from MSA to colloquial varieties.

In the survey we identified about 39instititions offering 71 systems: 36 systems translating from Arabic to English, 4 from Arabic to French, 2 from Arabic to Spanish, 24 translating from English to Arabic, 4 from French to Arabic and 2 from Spanish to Arabic.

The different steps that the consortium is implementing now are split into two phases. A first one (Phase 1) will ensure that a baseline is developed and evaluated, while Phase 2 will ensure that contributions from partners will provide substantial enhancements of the baseline performance. Such contributions will cover both language resources and the various tools used to optimize the translation module (e.g. multi-level alignments).

During the phase 1, several partners with the right technical background are installing and using Moses (details about SMT and Moses can be found at: http://www.statmt.org/moses/?n=Moses.Background); some others consider the use of other toolkits to compare with, such as: the GenPar (Toolkit for Research on Generalized Parsing, (including Machine Translation by Parsing) which provides also an architecture, a design, and an implementation of an integrated system for statistical machine translation by parsing (more details at: http://nlp.cs.nyu.edu/GenPar/GenPar.html)

The other tasks of Phase 1 are:

- ✓ Build a parallel corpus of Arabic/English (and may be other languages if feasible at low cost), this is being achieved by identifying data within multilingual content producers (UN, UNESCO, etc.)
- ✓ Align the corpus using the Giza++ aligner
- ✓ Collect a huge monolingual corpus for Arabic and English to train the language models
- ✓ Install and run Moses (train its decoder), exploiting its various features
- ✓ Collect a small evaluation corpus and have it translated by human translators to serve as reference times (if this ends up being too expensive, the consortium will exploit existing LRs like CESTA Corpus)
- ✓ Evaluate the whole system using BLEU (and other automatic metrics) and ensure that it is compared to systems brought by the partners (in particular SMT systems).

The tasks of phase 2 target the enhancement of the baseline defined and implemented during phase 1, so phase 2 will depend heavily on the performance of phase 1 and will be planned in details afterwards. In particular two options will be considered: a) increase the size of the language resources to train the tools versus b) change the domain/genre of the data to see how robust the system is to new domain/genre. The objective is to identify which components can/should be adapted, customized and enhanced for Arabic and which resources and tools are needed to enhance the baseline. Again, the main tasks will be:

- ✓ Build a parallel corpus Arabic/English (and maybe other languages if feasible at low cost), either to enrich what has been used in Phase 1 or to cover a new domain (health, economics, etc.)
- ✓ Align the corpus using some of the partners' aligners in addition to the Giza++ aligner
- ✓ Collect a new and huge monolingual corpus for Arabic and English to train the language models (only) if we feel that the domain is so different that it requires a new language model
- ✓ Consider the possibility to use new features within Moses like exploitation of morphological analyzed corpus (alignment of Pos)
- ✓ Run Moses with the new datasets
- ✓ Conduct a second evaluation to assess the new performances and improvements

The Moses components that will have to be customized for our project are:

- Language resources and data preprocessing (e.g. the language model should be trained on a corpus that is suitable to the domain, preferably a parallel corpus)
- Language Modeling toolkit , here we will choose the SRI language modeling toolkit (but we also consider the IRST and the RandLM language modeling toolkit)
- GIZA++ for word alignments
- Tuning the translation models (minimum error rate training) but also exploit a number of features such as:
  - Reorder phrases and lexicons
  - First pass of translation using Moses generating n-best (e.g. n=1000)
  - Second pass reordering the n-best solutions with a more precise language model

The results will be evaluated using automatic metrics including BLEU and exploiting human translated texts (called «reference translated texts» with at least 4 different translations). It is also envisaged to carry out human evaluation to complement the automatic metrics.

5. Cooperation Roadmap

5.1. Three interconnected roadmaps

Referring to the work done in Europe in particular under the ELSNET project, we can define a roadmap as "a document that indicates directions for a planned journey, that shows how and in what order goals can be reached and that indicates distances".

Usually one focuses either on a roadmap as reflecting expected "technology developments and trends" (technology roadmap) or as "time to market " for a new product (market dimension). 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). So, the Cooperation Roadmap will in a sense consist of three interconnected roadmaps although we will not develop each of them independently, but rather take aspects from all of them into consideration.

At this moment we see that much ICT cooperation between EU and Arabic partners is based on third party incentives (e.g. the EC ICT programmes). It is good that such incentives exist, and we would need more of this type of support in the future, but at the same time we also face a challenge: to turn these partnerships into strategic partnerships, i.e. long term partnerships based on mutual benefit.

Taking into account the three dimensions listed above, we provide below an analysis and report on the present situation in the participants' countries, we describe the conditions that need to be fulfilled in order to arrive at particular key achievements and at some strategic partnerships and we describe the steps that need to be taken to get there from where we stand.

Our primary focus is on multilingual tools, in particular on machine translation and multilingual information retrieval but other areas are mentioned if considered relevant.

This document first describes the various elements or factors that have to be taken into account when creating a roadmap for Arabic HLT. We then describe some of the most important 'instruments' or actions that can be taken in order to influence the situation and the development, and finally we provide a synthesis and recommendations for actions to be taken. It is important to realize that at this stage most sections are still in a preliminary state, and this is in particular true for the synthesis and recommendations. We therefore invite comments and proposals for all aspects of the Roadmap document.

5.2 Elements of the roadmap

There are a number of elements that all contribute to the current status and the future development of Arabic language technology. In the following we will describe the status for each element and the target in 5-7 years.

One of the most important elements is the players and the human resources. Human resources have to be highly skilled, and therefore education is another key element in the roadmap. The next elements that are described are the technology, and the environment (or infrastructure) in which the technology operates and that can be used to disseminate it. We then discuss the market for Arabic HLT in the region, as it is important that the market has a certain size for players to find it attractive. Many of these elements are obviously interrelated, but still we find it useful to try to distinguish them and describe them separately, even if some redundancy may appear.

5.2.1 Players and human resources

In the universe of HLT many different players fulfil their tasks in the long chain from research idea to end user product. In this section we will highlight a few of these players, which can be both organisations and humans. In the roadmap report (see www.medar.info) we have brought together a large number of players (or potential players) connected to the field of Arabic HLT, with a special focus on players active in areas that are concerned with multilinguality, such as translation support, machine translation and cross-lingual information retrieval.

If we look at the human resources side we can observe that many of the skills required for the production of HLT products and services coincide with those required for ICT products in general (e.g. software engineering, testing, interfaces).

In the roadmap report we focus on the special skills required to build HLT for Arabic, which include both knowledge about Arabic language and linguistics, and the capability to communicate and collaborate with generic software engineers. According to a brief survey we made in the countries represented in the consortium there seems to be a definite shortage of people with these skills.

In order to increase the number of people with the required skills we have to look at the opportunities for the education of a new generation of researchers and developers with adequate skills in HLT. We will discuss this in the following section.

5.2.2. Education

The main players in the education system are universities and other institutes for higher education. In our survey (conducted in the MEDAR partner countries) we have tried to gather information about the total number of universities (excluding specialized institutions such as medical, agricultural or veterinary institutions) and about the number of those that have HLT oriented courses in their curriculum on a structural basis (i.e. not as spin-offs of the presence of individual researchers interested in HLT).

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

On the basis of our analysis it appears that there is a need for initiatives that could lead to an improvement of the education situation. As this analysis is made in the light of the creation of a collaboration roadmap we suggest a number of possible bilateral or multilateral cooperation actions between EU and Arab states, or cooperation between Arab states.

5.2.3 Technology

Successful R&D leading to the development of HLT-related products and services does not only presuppose the availability of organisations and human resources as described above, but also a wide range of other key ingredients. We list a number of them in the roadmap report and give a brief overview of the present situation and of where we see possibilities for joint actions.

From the cooperation point of view priority should be given to tools and resources (i) that are re-usable, (ii) that can be shared with other players, and (iii) that adhere to formal or de facto representation and interoperability standards, so that they can have a maximal impact. The survey shows that at this moment external standards (formal or de facto) do not play a role of significance in Arabic HLT. This is a serious obstacle for any form of collaboration.

From the specific MEDAR perspective, where the focus is on cooperation, a number of observations can be made. First of all the universe of HLT players in the Arab world is small and fragmented with a majority of small players (as far as their HLT efforts is concerned) and a very few larger ones [see *Survey of actors, projects, products*, MEDAR Report 3.1]. Many of the available HLT resources and applications originate from outside the Arab world, some from large global players such as Microsoft, Google and IBM.

The global players (some of which already have local branches in the Arab world, especially in Egypt, such as Microsoft and IBM) are in an excellent position to establish relationships with local players or even to acquire them if they feel they need them.

In this fragmented landscape it will be extremely hard for other local players to enter into competition with the big players, unless they join forces. The European experience has shown that in spite of the linguistic fragmentation in Europe many small players have entered into lasting cooperation relationships leading to a strengthening of HLT in Europe.

Cooperation, especially at the international level, is very unlikely to happen spontaneously and cannot be enforced either, but is an absolute necessity if local players want to continue to coexist and collaborate with the global players.

Cooperative projects could take many different shapes, such as medium size and large RTD projects (STREPs or IPs in EC jargon), participation in Networks of Excellence, Coordination and Support actions, but also staff exchanges involving both academic and commercial organisations leading to better transfer of knowledge between academia and industry end between EU and Arab partners.

As already stated above representation and interoperability standards are an important instrument to ensure re-usability and sharability of resources and to create opportunities for cooperation aimed at integration of tools and services with a view to offering more advanced products and services. It is therefore recommended that players from the Arab states be offered ways to participate in EU-wide actions aimed at the creation of infrastructures to share resources and technologies, and at providing recommendations for policies regarding language resources and technologies (e.g. CLARIN and FlaReNet), where standards play a crucial role. Without that they would isolate themselves from what is going on in mainstream HLT in Europe.

To summarize and conclude we list a number of typical examples of possible actions aimed at strengthening the advancement of Arabic HLT and at creating lasting partnerships, and that should be included in the HLT Cooperation Roadmap for Arabic HLT.

This includes Coordination and Support actions aimed at:

- Strengthening of participation of players from the Arab states in Networks of Excellence in fields related to multilingual HLT or HLT in general
- Participation from the Arab HLT community in HLT related actions concerned with language resources infrastructures, standards and policies
- Developing schemes for staff exchange
- 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

And it would include joint RTD actions involving EU and Arab partners aimed at

- Creation of basic resources (development of the BLARK)
- Creation of application or domain specific resources
- Creation of HLT products, applications and services, especially in the field of multilingual technologies

5.2.4 E-infrastructure

An important factor for the further development and use of Arabic language technology is the availability and penetration of enabling technologies, in particular internet access, available within the Arab regions, as these technologies are a prerequisite for the development of the market.

Less than ten percent of people in the Arab world are internet users: According to a study (March 2008) conducted by Internet World Stats (www.internetworldstats.com) the internet penetration percentage in the Arab region is only equal to 9.4%. This average comes from figures as low as 0.1% for Iraq and 38.4% for UAE. If the Arab region is compared to other parts of the world, we see .e.g. Europe with 47.7% in average and North America with 73.1%. But it is also important to consider the growth rate, and here several Arab countries show a very good growth rate for the period 2000-2008.

We have also investigated the penetration of mobile phones. The reports and statistics were taken from: http://www.itu.int/ITU-D/ict/newslog/CategoryView,category,Arab%2BStates.aspx dates are shown for each report.

A new report from Arab Advisors Group[1] analyzes and ranks 30 fixed services operators and 50 cellular operators in nineteen Arab countries. STC's Al Jawwal, Egypt's Mobinil and Vodafone Egypt are the largest Arab cellular operators in terms of subscribers.

With the advent of new operators and increased competition in 2008, cellular subscribers in 19 examined Arab countries reached 194.533 million. ALJAWAL and MobiNil sustained their top rankings by H1 2008, with 17.8 million and 16.328 million subscribers respectively. Vodafone Egypt ended the first six months of 2008 with 15.202 million subscribers, settling as the third largest mobile operator in the region.

Looking at the future for internet and the mobile market, there are several ways to support the growth in Internet penetration. One way is to reduce the rates that users have to pay; e.g. the Internet rates in Jordan have recently been reduced by 30%, and the same was done in Saudi Arabia and other countries.

Another way of boosting the Internet penetration is for the governments to increase their services on the Internet, i.e. developing e-Government. This is a well-known method, which has proven its efficiency in other parts of the world. However, e-Government can only take off when a reasonable amount of users are online, so maybe only some of the Arab countries are ready for this development.

By 2015 we expect an Internet penetration of 25% in average for the region, but it may be wise to strive for a

[1] 2/12/2008

higher penetration.

From the reports and statistics for some of the Arab countries listed above, it is noticeable that the mobile phones usages are growing at a very high speed, and many companies are looking at more penetration through enhancing services and providing clients with an edge and added value for subscribing with their networks. One good opportunity will be to utilize language technology into mobile added services, and to approach the mobile companies to support R&D in this direction.

5.2.5 Market

This section will briefly elaborate on the existing market and its different components (suppliers, consumers, impacting factors, business models). The market foreseen here comprises the Arabic domestic market that could be addressed by commercial offers of products and services but also foreign markets that could attract exports from players located within the Arabic countries. We also consider the different suppliers and users: those located in Arabic countries and supplying HLT products/services to the markets mentioned above, but also those located outside the Arabic region and supplying products/services to Arabic market as well as international ones. The idea of the project is, based on the market structure today, propose scenarios to boost the offers of Arabic players (for domestic and export markets) but also the international offers of Arabic HLT for the Arabic markets.

The first aspect to take into consideration is the profiles of our consumers/users of HLT products and services. We distinguish at least three categories: individual consumers, professional/business users, and institutional users. Such categories have to be located within Arabic countries and outside.

It is clear that products for individual users (e.g. spell/grammar checkers, MT, TTS, ASR, etc.) will be impacted by the software business in general and here piracy is a serious issue.

The annual survey of the Business Software Alliance (BSA) about the software piracy (the 2007 BSA and IDC Global Software Piracy Study) covers piracy of all packaged software running on personal computers (PC). The worldwide PC software piracy rate increased three percentage points to 38% from 2006 to 2007 (meaning that 38% of the packages are not "paid"). The average piracy rate for the middle-east and Africa region is 60%. Let us share some data about some of the Arabic countries: Yemen is the "first" Arab country (ranked 7 with 89%), Libya (ranked 8 with 88%), Algeria ranked 14 with 84%, UAE is within the low rates (with 35%). Egypt's government made a deal with right holders to provide software packages for government and educational use, and piracy rate dropped to 60%. In terms of revenues, BAS estimates that the whole market is 4,000M$ with losses of 2,446M$ due to piracy.

International players may be discouraged to invest in such markets while local software industries can be crippled by competition from pirating actors. These analyses also are in agreement with the trend to offer most of these "applications" as web based services. This is also the big trend in software industry

The other dimension mentioned above is the illiteracy combined with the poor reading and publishing habits (publication of less than 900 books/year in Morocco compared to 200,000 in UK, 6,500 in Turkey, 12,000 in all Arab countries together). This has an impact on the global Arabic economy and the PC/Internet penetration. This dimension can be also taken as a serious growth factor if national agencies consider the use of HLT products and services in the literacy programs and offer boosters to a positive consumption of web-based content.

The other consumer category is "businesses" (SME and large companies) which could adopt such technologies if they are useful economic growth instruments. In that scenario, some well designed technologies could be very useful: Automatic speech recognition technologies for dictation, language learning, text to speech synthesis in local colloquial varieties of Arabic coupled with news and media content suppliers, MT both desktop and on mobile for tourism industry, better Arabic search engines, etc. In addition we can also imagine supplying HLT tools for the content production players (media, news, publishers, etc.): tools such as bilingual editing software with grammar checkers, translation memories, etc.

The third consumer category is the Institutional Agencies (e-government, e-health, e-education, etc.) that could adopt the HLT products and provide web-services for citizens. This requires more R&D to come up with e.g. advanced search engines, speech enabled services, etc.

An important aspect that we considered is the degree of ICT friendliness of the Arabic countries and their capacity to enter into the ICT and digital world as contributors as well as consumers. The International Telecommunications Union (ITU) in its 2007 report "Measuring the Information society" introduced what it calls the ICT development index to "captures the level of advancement of information and communication technologies (ICTs)" [2] and compares progress made between 2002 and 2007. In this report, the top three are Sweden, Korea and Denmark. Sweden ranked number 1 with an ICT Development Index (IDI) of 7.50 in 2007. Korea is ranked number 2 and finally Denmark ranked number 3 (IDI of 7.22).

The first Arabic country is United Arab Emirates, ranked 32 with a 2007 IDI of 5.29, the next one is Bahrain (ranked 42 (IDI 4.69)), followed by Qatar (ranked 44 with IDI 4.44). Other countries: Saudi Arabia (55th, 3.62), Kuwait (57th), Lebanon (64th), Jordan (76th), Oman (77th), Palestine (79th), Libya (81th), Tunisia (83th), Syria (89th), Egypt (94th), Algeria (97th), Morocco (101st, 2.34), etc.

The market vision for 2015 will be impacted by a large number of factors. Among the negative impacting factors

[2] The index considers features like Fixed telephone lines, and Mobile cellular telephone subscriptions, Internet bandwidth and Internet users, Fixed/Mobile broadband subscribers, households with a computer, Adult literacy rate, etc.

we can quote the poverty, illiteracy rate and IP piracy and more drastically the ICT low friendly indexes. Among the positive impacting factors we can consider the very impressive mobile penetration, the Internet growth, etc. Other factors that will have an impact but that are hard to assess are all the driving factors like external R&D investments i.e. for security purposes (both within and outside Arabic countries), the relocation of big players that outsource part of their activities (off-shore) e.g. Orange Labs, Microsoft and IBM in Egypt, Alcatel and SMT-Microelectronics in Morocco, etc. The role of local players will be very sensitive. We do not expect them to grow per se but rather surf on the open-source trends to offer high value services to local administrations consolidate and merge with big players or simply disappear.

5.2.6 Roadmap recommendations

For all the other roadmap elements treated above, we can conclude that as a result of the spread of ICT in general, more penetration, more trends, and consequently more needs will attract the international major players to this area and start planning for dominating or getting more share in this emerging, promising market. Major companies such as the mobile and telecommunication companies will have needs to enhance their services by facilitating the utilization of the services in the Arabic language. Major international applications and software providers and manufacturers need to add tools and utilities to their products, such as Microsoft and Google; who are basically dealing with language aspects and always in need to enhance their services with professional, not superficial tools and utilities. Their focus is directed towards utilities for the enhancement of Arabic processing (spell checkers, morphological analyzers, syntactic analyzers, lexicons, search engines based on Arabic main features, etc).

Two scenarios could be seen in this case:

1) The major international players will dominate the market, develop technologies and enforce their vision, methodologies, procedures, and as a result monopolize the whole industry.
2) Local efforts should be undertaken, governments and major funding agents should encourage and incubate such activities, giving the local companies opportunities to grow and develop their own tools and services; build the capacity in local forces; and build and maintain national industry that could be competitive on the international level.

What we recommend is a hybrid solution between the two scenarios; the Arabic HLT should get the benefit out of the huge interest, huge allocation of investments, and immediate need for the services and applications in the ICT in general, and in related applications to Arabic HLT in particular (search engines, mobile services, e-commerce, e-government, e-learning, etc).

As a result, cooperation should be sought binding and linking all players together, that will result in cooperation between all the players to build an Arabic HLT industry; that is encourage the international major players to keep and maintain their interest and to build the local industry and workforce needed to sustain it. The following could be thought of as different components and directions that could lead into success of the strategy:

- Specialized companies should play a significant role in this area and should build and enhance tools, utilities and applications for Arabic HLT,
- Universities and research centres should provide the basic and applied research in cooperation with industry to produce solid products,
- 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 sub-projects) that will be accessed and navigated in the Arabic language
- Universities and other educational institutions should create the proper training, and re-training (rehabilitation program for personnel from other disciplines who could be re-trained to fit the new requirements)
- International companies specialized and interested in the HLT (and Arabic HLT) in particular should be encouraged:
    - to maintain the interest in Arabic,
    - to providing services to the region, and should be given facilities to make this attractive to them
    - to maintain relationships with local companies and task forces, and
    - to utilize what is available locally
- 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.

6. Networking

It is on of the goals of MEDAR to continue the formation of 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 and other countries who have an interest (commercial or scientific) in Arabic language and speech processing, and in particular multilingual aspects of this.

It is the hope that the network will not only be instrumental in creating partnerships, but also be an efficient dissemination channel, and that the Roadmap will attract attention, which will support the network.

7. International conference

Dissemination is one of the very important activities of

the MEDAR project, cf. above the third stream: the dissemination stream. Dissemination has a number of different target groups and uses a number of different channels.

Here, we just want to mention the Second International Conference of Arabic Language Resources and Tools that is being held in Cairo 2009. The first International Conference on Arabic Language Resources and Tools was organized by NEMLAR in Cairo 2004, and was a big success.

The intention is to have a broad spectrum of participants, from all over the world, and from academia as well as industry, with the advancement of Arabic language technology as a common goal. The conference will have presentation of new approaches for Arabic HLT, and will have an important panel discussion on the cooperation roadmap, as mentioned above.

8. Summary and conclusions

In this paper we have given an overview of the MEDAR goals and activities, and we have concentrated on those aspects that we are focussing on at present, in particular the Cooperation Roadmap, which we hope will have a lasting impact for Arabic HLT.

9. Acknowledgements

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

This paper builds on work done by the MEDAR partners. 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 Attia, 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 - Commissariat à l'Energie Atomique - Laboratoire d´Ingénierie de la Connaissance Multimédia Multilingue, France
- Fathi Débili, CNRS, Centre National de la Recherche Scientifique, Laboratoire LLACAN - UMR 8135 du CNRS, Langage, langues et cultures d'Afrique Noire, France
- Anne DeRoeck, The Open University, UK
- Joseph Dichy, Université Lumière Lyon2: Groupe SILAT, France
- Ossama Emam, IBM - Human Language Technologies Group, Egypt
- Michael Ghali, Sakhr Software Company, Egypt

10. References

Choukri, C., M. Diab, B. Maegaard, P. Rosso, A. Soudi, A. Farghaly (ed.): *HLT & NLP within the Arabic world: Arabic Language and local languages processing: Status Updates and Prospects*, LREC Workshop Proceedings, Marrakech, Morocco, 2008, 121 pages.

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

Bente Maegaard, M. Atiyya, K. Choukri, S. Krauwer, C. Mokbel, M. Yaseen: MEDAR – collaboration between European and Mediterranean Arabic partners to support the development of language technology for Arabic. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, 2008.

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.