

On Arabic HLT Industry Status and Needs as Perceived through NEMLAR

Muhammad Atiyya; RDI¹, m_Atiyya@RDI-eg.com

Introduction:

NEMLAR www.NEMLAR.org stands for "Network for Euro-Mediterranean Language Resources and human language technology development and support". This identifies an ambitious 30-month project aiming to create a network of qualified Euro-Mediterranean partners, and to specify and support the development of high priority language resources for Arabic.

This project is supported by the INCO-MED programme under the European Commission. While the technical management of this project is conducted by ELDA "Evaluations and Language resources Distribution Agency" based in France www.ELDA.fr, the project coordination is carried out by CST "Center for Sprogteknologi" based in Denmark www.cst.dk. Including RDI; this project is shared also by other 12 academic and industrial institutions who are specialized in Arabic HLT R&D in the Arab region and Europe².

Having been launched officially in Feb. 2003, NEMLAR partners have almost completed one of its major Work Packages; namely WP3 where as much industrial and academic Arabic HLT players as possible inside and outside the Arab region are surveyed in detail. Moreover, their Arabic HLT needs of standards, tools, and resources are specified. The essential conclusions of NEMLAR's WP3 that draw an integral picture of the Arabic HLT field today are presented in the rest of this essay.

1- Data acquisition process

Over 2 waves of extensive surveys; about 50 academic and industrial, individual and institutional Arabic HLT players have been surveyed. Where 85% of them are based inside the Arab region, the other 15% are distributed over the western Europe and the US. Of those 85% based inside the Arab region, around one third are concentrated in Egypt. It is important to mention the following criteria according to which Arabic HLT players are selected to be surveyed:

- Proven effectiveness of player's participation in Arabic HLT standards, resources, tools, tuition, and/or organization & dissemination.
- Player's willingness to cooperation and information disclosure.

Before delving to the conclusions, honesty dictates stating the following facts about the process of acquiring the data filled in the surveys:

- 1- The culture of disclosing data about individual's or institution's technical and/or marketing activities was unfortunately found to be not prevalent in general esp. in the Arab region.

¹ RDI stands for (Research & Development Int'l) which is a pioneering software house based in Giza-Egypt and specialized in Arabic Human Language Technologies (HLT); see www.RDI-eg.com. RDI also carries the official name that (The Engineering Company for the Development of Computer Systems).

² A complete list of NEMLAR participants is found on www.NEMLAR.org

- 2- Some of the NEMLAR surveyors - like RDI - are recognized Arabic HLT industrials which makes the surveyed players more or less consider them as market and/or R&D competitors.
- 3- Due to (1) & (2) most of the approached contact persons for surveys were somehow conservative in disclosing information about their (institutions') activities.
- 4- However, the deep involvement of NEMLAR surveyors in their industrial and academic local community afforded a firm knowledge of what is going on inside that narrow community, which substituted to a large extent any conservation on disclosing the required information.
- 5- The data in each surveys are collected, revised, and organized at least twice per each surveyed institution.

2- Main areas of the field

The 1st and 2nd waves of surveys have revealed that the main areas of R&D in which the surveyed Arabic HLT players are involved can be categorized as follows:

a- Arabic Natural Language Processing (NLP);

Which comprises all kinds of linguistic processing of Arabic text aiming to a wide variety of applications. The most demanded target applications can be put in the following rough descending order:

- 1- Arabic<->English Machine Translation.
- 2- Arabic<->French Machine Translation.
- 3- Bilingual Arabic<->English Machine Assisted Translation tools esp. online Dictionaries.
- 4- Bilingual Arabic<->French Machine Assisted Translation tools esp. online Dictionaries.
- 5- Automatic Arabic Phonetization (i.e. Diacritization) esp. for Arabic TTS.
- 6- Online Arabic Spelling & Grammar Checkers.
- 7- Data Mining.
- 8- Arabic Full Text Search.

b- Arabic Speech Processing;

The most demanded target applications can be put in the following descending order:

- 1- ASR for Voice Commanding over distorting channels (like telephony ones) and/or through low SNR.
- 2- Arabic Speech Synthesis (for Arabic TTS).
- 3- ASR for Continuous Arabic Speech Dictation.
- 4- Very Low Bit Rate Arabic Speech Compression.
- 5- Arabic Speech Verification (esp. for the Auto Learning of Spoken Arabic in general, and learning the correct recitation of the holy Qur'an in specific).

c- Arabic OCR;

While Online Handwritten OCR (esp. on mobile devices) has the 1st priority due to the rapidly growing market of mobile devices, and Offline Typewritten OCR comes as a next important priority esp. for document management systems.

d- Hybrid Systems:

The most demanded target applications can be put in the following descending order:

- 1- Arabic<->English Speech to Speech Translation.
- 2- Arabic<->French Speech to Speech Translation.
- 3- Arabic TTS.
- 4- NLP guided Arabic OCR's.

3- Distinct nature of the Arabic language:

All the surveyed Arabic HLT players (no exception) agreed to the statement that "*Arabic has distinct linguistic features!*"

While this is absolutely true at the lower linguistic layers; namely (*Phonology, Orthography, Morphology and POS tagging*), it is true to a slightly less extent at the medium ones; namely (*Syntax*), and Arabic gets common with other languages at the higher layers; namely (*Semantics, Discourse Integration, and Pragmatics*).

This means that all the aspects (standards, resources, and tools) of the lower layers including the syntax one (i.e. the infra structure) of Arabic HLT must be specially crafted to model the distinct characteristics of Arabic.

This means also that all the aspects of the higher layers of Arabic HLT can largely benefit from the advances and achievements realized in other languages on which NLP R&D have attained a great deal of maturity (e.g. Wide-spread European languages such as English, French, and Spanish). Moreover, statistical techniques necessary for disambiguation at all NLP layers are common, and Arabic like other languages can highly benefit from them as well as other languages do.

4- Market:

Most (not all) surveyed Arabic HLT players (including RDI) agreed to the statement that "*Arabic HLT market in Egypt and the Arab region is still in its infancy*".

So, in spite of being a highly demanding market of Information & Communication Technology (ICT) applications with all the more & more needed Natural User Interfaces (NUI), the still young Arabic HLT market is basically a market of HLT products and solutions not a market of HLT resources and tools yet.

This situation besides the scarceness of powerful steering, organizing, and funding Arabic HLT agencies clearly explain the lack of standards and unified specifications of Arabic HLT resources and tools.

5- Cooperation attitude, and publishing & dissemination policies:

During these surveys (and long before) we found many brilliant researchers and engineers working on Arabic/multilingual HLT R&D projects, however few of them had unfortunately realised considerable practical achievements.

This sad situation may be attributed to the lack of cooperation attitude at both the individual and the institutional level. Worse is the restrictive Arabic HLT knowledge and resources publishing and dissemination policies of the significant companies in this field. These companies are driven by the Arabic ICT market that understands only the language of final products and solutions and ignores the underlying resources and tools.

The ever worst may be the prevalent distrust in other's work due to the absence of inter-institution agreed upon validation and evaluation mechanisms.

With the absence of cooperation and evaluation, and the restriction of publishing and dissemination; no accumulation of know-how, standards, resources, and tools can ever happen. Consequently, R&D group (whether academic or industrial) starts almost from scratch on their Arabic HLT projects; i.e. "they re-invent the wheel" which has a catastrophic effect on the advance of Arabic HLT.

Many surveyed players welcomed NEMLAR as a resolution to this case. They asked for both expanding its scope of activities and elongating its duration so as to give chances for change to take place. Moreover, they also encouraged more similar projects under powerful well organized umbrellas like LDC and ELDA/ELRA.

6- Types of Arabic HLT needs:

According to the surveys, the Arabic HLT industry needs 4 types of resources, namely; Standards, Knowledge bases, Training DB's, and Tools. Listed below are those feasible needs ordered from the bottom to top NLP layers not according to importance.

- a- Standards are badly sought for
 - 1- Arabic Phonology,
 - 2- Arabic Orthography (Graphical Representation),
 - 3- Arabic Morphology (Lexical Structure of words),
 - 4- Arabic POS tagging, and
 - 5- Arabic Parsing Format.

- b- Knowledge bases of
 - 1- Formal Arabic Phonological rules (Phonetic Grammar),
 - 2- Type-written and Hand-written Arabic Orthographic rules,
 - 3- Arabic Morphological rules,
 - 4- Bilingual (esp. Arabic<->English) Machine Readable Dictionaries,
 - 5- Formal Arabic Syntactic rules (Grammar),
 - 6- Arabic Lexical Semantics, and
 - 7- Arabic Word Net are badly sought.

are highly needed.

- c- Training DB's of accurate
 - 1- Annotated corpus of online handwritten Arabic text patterns (from hundreds of writers) that is lexically and graphically labelled,
 - 2- Offline typewritten documents that cover the different documents layout as well as fonts and styles,
 - 3- Large Arabic Continuous and Isolated Speech Recordings covering the formal as well as different dialects across the Arab regions along with their Arabic Phonetic Transcription,
 - 4- Arabic Lexicon,
 - 5- Morphologically, Syntactically, Phonetically, and Semantically tagged large Arabic text corpora.

are critically vital to use them in building (Statistical) Language Models.

d- Tools (core engines):

- 1- Arabic Character Recognition
- 2- Small Vocabulary ASR
- 3- Arabic Morphological Analysis
- 4- Arabic Diacritization (Vowelization)
- 5- Arabic Speech Synthesis
- 6- Arabic POS tagging
- 7- Arabic Lexical Semantic tagging,
- 8- Arabic Syntactic Parsing are highly sought.
- 9- For all and before all; Statistical Language Modelling as well as Statistical Disambiguation tools are essential.

7- Industry priorities:

Due to our surveys and data collection, as well as our deep involvement at RDI in this field in Egypt and the Arab region; we list in a descending order in the table below the industry priorities of Arabic HLT systems from both a revenue generation perspective (Market Preference), and from a technological shortage point of view (R&D Shortcoming).

Market Preference	R&D Obligation
1. Limited domain Speech-to-Speech Translation (SST) systems where Arabic is a terminal language.	Limited domain Speech-to-Speech Translation (SST) systems where Arabic is a terminal language.
2. Automatic Arabic small-vocabulary Speech Recognition systems which are robust versus high noise and channel distortion. (N.B: Necessary for SST systems)	Automatic Arabic semantic analyzer.
3. Concatenative/Parametric Arabic Text-To-Speech (TTS) systems. (N.B: Necessary for SST systems)	Automatic Arabic syntactic analyzer.
4. Very low bit rate Arabic speech encoders/decoders.	Speech verification systems for the (full or computer assisted) self learning of Arabic pronunciation and the Tajweed of the holy Qur'an.
5. Automatic Arabic text diacritizer. (Necessary for Arabic TTS)	Automatic Arabic large-vocabulary (dictation) Speech Recognition systems for the office environment.
6. On-line hand-written Arabic/Latin (mainly English) OCR systems.	Very low bit rate Arabic speech encoders/decoders.
7. Off-line type-written Arabic/Latin (mainly English) OCR systems.	Bilingual French/Arabic, and Arabic/French, computer assisted MT systems.
8. Automatic Arabic large-vocabulary (dictation) Speech Recognition systems for the office environment.	Bilingual English/Arabic, and Arabic/English, computer assisted MT systems.
9. Bilingual English/Arabic, and Arabic/English, computer assisted MT systems.	Off-line type-written Arabic/Latin (mainly English) OCR systems.
10. Bilingual French/Arabic, and Arabic/French, computer assisted MT systems.	Automatic Arabic POS tagger.
11. Arabic derivative/semantic full-text search engines.	Automatic Arabic text diacritizer. (Necessary for Arabic TTS)
12. Speech verification systems for the (full or computer assisted) self learning of Arabic pronunciation esp. Tajweed of the holy Qur'an.	Concatenative/Parametric Arabic Text-To-Speech (TTS) systems.
13. Automatic Arabic morphological analyzer and disambiguator (N.B: The basis for all the other systems in this list).	On-line hand-written Arabic/Latin (mainly English) OCR systems.
14. Automatic Arabic POS tagger. (N.B: A basis for 15 & 16).	Arabic derivative/semantic full-text search engines.
15. Automatic Arabic syntactic analyzer. (N.B: A basis for 2, 9, 10 & 16)	Automatic Arabic morphological analyzer and disambiguator (N.B: The basis for all the other systems in this list).

16. Automatic Arabic semantic analyzer. (N.B: A basis for 2, 9 & 10)

Automatic Arabic small-vocabulary Speech Recognition systems which are robust versus high noise and channel distortion.

About the Author

Muhammad Atiyya has been graduated from the department of Electronics and Electrical Engineering, Faculty of Engineering, Cairo University, in 1995. He has been awarded MSc. degree from the department of Computer Engineering of the same faculty, in 2000. Currently, he is finishing his PhD in the department of Electronics and Electrical Engineering, Faculty of Engineering, Cairo University.

Muhammad has been working for RDI since about 9 years, and he now occupies the position of her NLP R&D team manager where he has been involved in developing many of its Arabic HLT products. His research interests mainly focus on software intelligent systems that aim for more natural human/machine interaction especially via written and spoken natural languages.