NEMLAR – an Arabic Language Resources project

Bente Maegaard

Center for Sprogteknologi, University of Copenhagen
Njalsgade 80, 2300 Copenhagen S, Denmark
bente@cst.dk

Abstract

The NEMLAR project is a European Commission supported project with partners from the EU and from Arabic speaking countries in the Mediterranean region. The project aims at surveying the stat-of-the art- of language resources and tools for Arabic in the region, at developing a BLARK definition for Arabic, and at starting development of language resources or updating of existing language resources. The project also aims to create visibility for Arabic language technology, through a newsletter and through an international conference.

Motivation

There is abundant evidence that language technologies can only be developed using large bodies of language resources (LRs) for language modelling, as test beds, for evaluation, example bases, and terminology source. The need for LRs applies both for research and for commercial applications.

Not only raw data, but also 'derived' LRs, e.g. annotated corpora, lexica and grammars, as well as tools for manipulating data form part of the material of interest. The production of such LRs also enables the linguistic cultural heritage of a community or nation to be preserved in an age of digital access and storage.

This is the reason the NEMLAR project was started. There is a strong interest in supporting the Arabic language, in the region, in Europe and elsewhere. The project runs 2003-2005.

The NEMLAR project covers recognised European centres and recognised partners in 6 non-EU Mediterranean countries, namely Jordan, Morocco, Egypt, Lebanon, Tunisia, West Bank and Gaza Strip. (See list at the end of the paper).

NEMLAR goals

The goal of the **NEMLAR (Network for Euro-Mediterranean LAnguage Resources)** project is to create a network of qualified Euro-Mediterranean partners to specify and support the development of high priority LRs for Arabic and other local languages in a systematic, standards-driven, collaborative learning context. The project focuses on identifying the state-of-the-art of LRs in the region, assessing priority requirements through consultations with language industry and communication players, and establishing a basic LR kit for the major forms of the region's predominant language – Arabic, and other local widely spoken languages where appropriate.

Survey: Key players, LRs, industrial needs

It is a key part of this project to provide knowledge about the language technology players, projects (ongoing activities), products etc. So a 'mapping' is made covering all Mediterranean countries participating in the project, 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 Language Resources (LRs). This knowledge base has appeared in its first version (Nikkhou et al. 2004).

Institutions

It covers 35 institutional players in the region, and some 20 individual players. It will further develop during the lifetime of the project, but we believe to have identified the most important players already.

Of the 35 institutional players, 22 are based in Arabic speaking countries, obviously in particular in our partner countries. E.g. there are 8 entities in Egypt, 4 in Lebanon, 3 in Jordan and 3 in Palestine. Kuwait houses the Sakhr company with subsidiaries in many countries, incl. Egypt.

In Europe, companies such as Systran and France Télécom also take an interest in Arabic language processing.

Tools and LRs

Existing Arabic tools and LRs in the region, in Europe or elsewhere have been identified, and the first version of the survey report describes state-of-the-art of LRs for the languages of the region. It should be stressed that the survey has focussed on resources in the region, not in the whole world.

Arabic NLP technologies and tools	31
Speech processing technologies	11
Text processing technologies	11

Figure 1: Number of tools

Here NLP tools are modules that normally are parts of systems, e.g. morphological analyzer, POS tagger, language identifier, term finder etc. Also classified as NLP tools are research results that have not yet been commercialised, e.g. grammar checker, grapheme recognition for OCR.

It has been encouraging to see that e.g. POS taggers do exist, and not only at the universities, but also as products. Morphological analyzers exist at the universities and also as a component of commercial products, e.g. machine translation. It is important to make morphological analyzers available in a source format, so that researchers can further elaborate on the morphological analysis and can combine this analysis with other components in their efforts to gain new insights and develop ideas for new and better language modules.

Syntactic analyzers exist in some universities, and as an important part of e.g. MT systems. Overall, it seems at present that there is no large scale grammar and parser freely available for researchers. It is foreseeable that commercially developed syntactic analyzers cannot be made available, so we believe that some interest should go into investigating the existence of syntactic analyzers and the possibilities of developing them further.

Speech processing technologies cover Arabic text-to-speech, speech recognition, speaker recognition etc.

Arabic text-to-speech exists in several versions as products. Arabic text-to-speech is of good quality which can be easily compared to similar tools for other languages. It would be important to identify open source modules which can be used by researchers for further improvement and research.

At present we have identified the following amount of language resources:

Speech databases	22
Lexical databases	29
Text corpora	24
Multimodal resources	1

Figure 2: Number of language resources

As can be seen there are already several LRs available, 15 LRs are provided by LDC and 3 by ELRA at present. The objective is that after the NEMLAR project these figures should have raised.

Going through the various resources it turns out that e.g. the lexical databases are strong in morphology and morpho-syntax, but fewer of them contain valency information and very few semantic information. No semantic lexica like e.g. Euro-Wordnet exist.

Which language resources and tools are available

As several companies produce products in the field of Arabic language technology, e.g MT, speech technology etc., LRs do exist within these companies. Such resources are e.g. large corpora, lexica, morphological components, speech corpora etc. However, such basic resources are normally not available to others. It is a well-known fact, and not specific for these companies that the LRs developed have been expensive and constitute a competitive advantage that companies can usually not share with others.

According to the present version of the survey, the situation wrt. lexica seems to be positive, in particular lexica with morpho-syntactic information.

Industrial needs

Industry in the Mediterranean countries and other industry working with Arabic has been consulted with respect to needs for LRs for the Arabic language and/or bilingual LRs for the development of language technology. This is detailed in a second survey report which provides a record of the LR needs of industry and an analysis of missing LRs in the current situation ('LR gaps').

At the time of writing, this report is work in progress, and the results will probably not be that significant, as the number of industrial players identified is actually quite small. However, it is important to note that the most important industries in the region do take part in the survey, so the survey will be representative for the industry in the region.

Industry in other areas, e.g. European industry is also consulted, and it is encouraging that there is a growing interest in Arabic language technology also outside the Arabic speaking countries.

It is to be hoped that the focus NEMLAR places on Arabic language resources will give more companies a possibility to join the market.

Significant commercial players

There are some significant companies in the region, e.g. Sakhr, Kuwait (with subsidiaries in several other countries of the region), IBM Egypt and RDI in Egypt. Additionally, there is a number of smaller companies. Sakhr's list of products includes i.a. MT English-Arabic, Arabic English, keyword extraction, categorisation, summarisation, grammar checking. IBM Egypt concentrates on speech, using the IBM ViaVoice methodology. RDI provides e.g. Arabic text to speech, POS tagger, diacritizer, morphological analyzer, etc.

Also outside the region, companies have an interest in the development of language technology for Arabic, e.g. Systran, France, will market an MT system in 2004.

A few of these companies are partners in the NEMLAR project; and in order to establish a broader basis, the project is establishing relationships with external actors, including companies as the above.

Validation of language resources

In order to make sure LRs are of the intended quality, it is important that they are validated.

The majority (67%) of the interviewed experts and institutions validate their data internally only. 19% of the interviewees reported that they commission the validation to external organizations. Finally 13% reported they do not validate them at all.

Validation references

A variety of validation references are mentioned. Magma' Allugha Al-Arabiyya (The Institute of Standard Arabic) also called "Magma' Al-Khalideen" in Egypt, and its counter part in Damascus, Syria, are important institutes. Here, the ELRA validation efforts will also become useful. Validation manuals exist for spoken as well as written language (Van den Heuvel et al. 2000; Fersøe 2004). These manuals may be used for the Arabic resources, possibly with slight modifications.

One of the project's goals is to contribute to bringing existing language resources up to a standard level, so that they may be reused by others. The validation methodology developed in ELRA may contribute to this task. Therefore we mention its main characteristics below.

Validation of language resources consists of formal validation and content validation. Basically the validation consists in checking that the resource obeys the principles laid down in its own documentation. This puts an importance on the documentation: It is important that documentation exists, and that it describes e.g. how attributes and values are assigned. This is important not only from a formalistic validation point of view, but also for the possible buyer of a resource.

In formal validation i.a. technical aspects are checked, and it is checked that all and only the attributes and values mentioned in the documentation are used.

The content validation can be performed only by native speakers of the language in question, or by persons with a very good knowledge of the language. It checks e.g. for a lexicon that the entries are coded correctly.

Distribution of LRs

Over half of the interviewed institutions and experts wish to make their resources available to others according to a negotiated standardised distribution agreement.

Only 19% said they do not want to distribute their resources and this due to legal (9%), commercial (6%) and strategic (4%) reasons. This is encouraging, and we believe that the efforts spent in the project to identify the resources are very well spent. Some of the resources may be fully ready for distribution, others after a slight updating. The list is made available at the NEMLAR web site, www.nemlar.org.

BLARK

In parallel with the survey, work is ongoing to specify the Basic Language Resource Kit for Arabic. The BLARK constitutes what is seen as the minimum requirements with respect to language resources. The BLARK concept has not been developed for Arabic before, and it is interesting to note the differences in comparison with other languages. E.g. for Dutch for which the BLARK concept was first developed, a diacritizer tool (for inserting vowels) is not necessary, but for Arabic it is.

As mentioned, the work with the BLARK concept takes the original BLARK ideas as the point of departure. For the notion of Availability, the NEMLAR project has adopted a more fine-grained description than (Binnenpoorte 2002). Additionally the notion of quality has been introduced. The reason for this is that the simple availability of a language resource is not sufficient: its quality, including its adherence to standards is also an important feature.

The surveys mentioned above provide input about what is already available, and where there are gaps, or resources that have to be updated and improved in order to fit the specifications. Consequently, we have the necessary basis for detailed work on updating or creating languages resources for the Arabic language.

Work on language resources, update and production (future work)

Following the surveys of existing LRs and LR needs, as well as the BLARK specifications, the project will decide on priority needs for LR update or development, and develop a work plan for this work. The plan will also take into account the available human resources. The work plan will specify projects/pilot projects for project partners. Such pilot projects may concern the updating of existing resources (e.g. change of format, change of standards, validation and updating of existing LRs etc.). They may also concern collaboration with ongoing projects in order to ensure that the specifications are met. The development of LRs from scratch can hardly be considered in this project, given the amount of resources this requires.

Dissemination

The objectives of this project are on the one hand the technical work with the surveys, the specifications and the resources.

On the other hand, dissemination of the knowledge that has been acquired, and awareness about Human Language Technology in general is also a key objective. The project web site, www.nemlar.org, and the quarterly newsletter contribute to the general awareness raising.

Additionally, an international conference will be held in September 2004 in order to disseminate the surveys and the specifications, the industrial needs, and research in the field of LRs and tools for Arabic.

Acknowledgements

This project is supported by the European Commission. The author is co-ordinator of the project and would like to acknowledge the contribution of the partners, in particular Khalid Choukri, ELDA, France, Steven Krauwer, Utrecht University, Netherlands, and Muhammad Atiyya, RDI, Egypt.

The NEMLAR partners also contributed through their work:

Lise Damsgaard Jørgensen. Center for Sprogteknologi, University of Copenhagen, Denmark

Mahtab Nikkhou, ELDA European Language Resources Distribution Agency, France

Joseph Dichy, Université Lumière Lyon 2 - Faculté des Langues, France

Christian Fluhr, CEA -LIST/DTSI/SRSI/Laboratoire d´ingénierie de l´information multimédia multi-lingue, France

Fahti Débili, CNRS - Délégation Rhône-Alpes, Site Vallée du Rhône, France

Stelios Piperidis, Institute for Language and Speech Processing, Greece

Mustafa Yaseen, Amman University, Faculty of Information Technology, Jordan

Chafik Mokbel, University of Balamand, Lebanon

Abdelhak Mouradi, ENSIAS, University of Mohammed V Soussi, Ecole Nationale Supérieur d´informatique et d´analyse des Systèmes, Morocco

Salem Ghazali, SOTETEL-IT - Société Tunisienne d´Entreprises de Télécommunications – Information Technology, Tunisia

Anne DeRoeck The Open University, Computing Department, Maths & Computing Faculty, United Kingdom

Mahdi Arar, Birzeit University – Birzeit Information technology UNIT (BIT) & Arabic Department, West Bank and Gaza Strip

Kanan Al-Ali, Birzeit University – Birzeit Information technology UNIT (BIT) & Arabic Department, West Bank and Gaza Strip

References

Atiyya, M., *A Large-Scale Computational Processor of The Arabic Morphology, and Applications*, MSc. thesis, Dept. of Computer Engineering, Faculty of Engineering, Cairo University, 2000.

Binnenpoorte, D., F. De Friend, J. Sturm, W. Daelemans, H. Strik, C. Cucchinari, *A Field Survey for Establishing Priorities in the Development of HLT Resources for Dutch,* In: Proceedings LREC 2002, (Third International Conference on Language Resources and Evaluation), Las Palmas de Gran Canaria, Spain 2002.

Fersøe, H.: *Validation Manual for Lexica*, ELRA, Paris, 2004

Hamza, W.M., *A Large Database Concatenative Approach for Arabic Speech Synthesis*, PhD. thesis, Dept. of Electronics and Electrical Communications, Faculty of Engineering, Cairo University, 2000.

Krauwer, S., M. Atiyya, K. Choukri, B. Maegaard: *BLARK for Arabic*, NEMLAR report, 2004, www.nemlar.org

Monachini, M., F. Bertagna, N. Calzolari, N. Underwood, C. Navarretta: Towards a Standard for the Creation of Lexica, ELRA, Paris, 2003

Nikkhou, M., K. Choukri: Survey on the existing institutions and Language Resource using or developing Arabic, NEMLAR report, 2004. www.nemlar.org.

Van den Heuvel, H., Louis Boves, Eric Sanders: *Validation of Content and Quality of Existing SLR: Overview and Methodology*, ELRA, Paris, 2000