**مركز تقنيات اللغة العربية**

**Arabic Language TEchnologies Center**

**(ALTEC)**

An Under-Establishment Egyptian Non-Profit NGO

Jan. 2009

**Executive Summary**

Arabic Language TEchnologies Center (ALTEC) is an Egyptian not-for-profit non-governmental-organization (NGO) is herein proposed as an effective mechanism to advance Arabic Human Language Technologies (HLT) with a view of creating a long-term cooperation roadmap for Egypt and the region.

HLT cover technologies that facilitate accessibility and exploitation of the *mostly unstructured* data (text, audio, image, video... etc.) produced using natural language.

ALTEC aims to:

*a)* Build a research roadmap for Arabic HLT.

*b)* Develop, improve, and augment skilled expertise in Arabic HLT

*c)* Provide accessible resources including basic training data and software tools to the Arabic HLT community

*d)* Motivate and activate the Arabic HLT community to solve problems that are relevant to our region.

*f)* Serve as an advisory board to the government, academia, and industry in matters pertaining to Arabic HLT.

*g)* Improve collaboration between relevant governmental, academic, and industrial stakeholders.

*h)* Attracting funds and sponsorships from Arab and international organizations interested in Arabic HLT.

The founders plan to dedicate ALTEC as an association that brings together the major academic and industrial stakeholders to serve and grow Arabic HLT and to play a similar role of other international institutions worldwide such as the Association of Computational Linguistics (ACL), the Linguistic Data Consortium (LDC) in the US and European Evaluations, and Language resources Distribution Agency (ELDA).

ALTEC's board will consist mainly of experts in the HLT field who represent companies and research institutions, along with some prominent national and international figures in this area.

**1- Arabic HLT: Challenges and Gaps**

Language is the prime vehicle for humans to communicate and disseminate facts, impressions, opinions, and information. With the ever increasing quantity of online text and multimedia information, there is a pressing demand for technologies that facilitate accessibility and exploitation of the *mostly unstructured* knowledge contained in natural language. Advances in Human Language Technologies (HLT) are increasing, offering nearly universal access to online information and services for more and more people with or without computer skills. HLT enable humans to communicate via varying forms of computing machinery, including PDA’s, cell phones, PC’s, and others, to participate in the information society in a totally natural way in their native language.

HLT capabilities have substantially grown in recent years, both commercially and in research. Currently, there is a wide range of applications for HLT systems, including: *automated voice response* *systems*, which exploit automatic speech recognition, text-to-speech, and question answering technologies; *web search and* *cross-language search,* which uses search and machine translation technologies; *information collection and presentation systems* where information is collected and filtered according to stated user needs and utilizes search, filtering, and summarization technologies; *read aloud system* for visually impaired persons, which uses text to speech technology; *automated question answering system*, which either finds factoids in unstructured text or identifies proper answers in an answer bank, and utilizes retrieval, question answering, and paraphrasing technologies.

The range of possibilities is almost unlimited. Most major international computer and telecoms companies currently engage in HLT research and development. These technologies will play a key role in the age of information and are cited as key capabilities for competitive advantages in global enterprises.

Native speakers of languages that are underserved by HLT likely suffer from diminished information access and higher information aggregation and authoring costs. For Arabic, although some HLT tools and resources already exist, these tools and resources continue to lag in availability and quality compared to similar tools for so-called tier-1 languages that dominate the Web such as English, French, and Japanese. This lag presents a major hurdle to effective and efficient information exchange, which adversely impacts Arab governments, businesses, and people at large.

Significant investments are required to overcome this lag, particularly because adapting tools from other language to Arabic or developing new tools than cannot be adopted necessitates handling Arabic language complexities, including:

* **Morphological richness:** Morphology refers to how words are constructed. Arabic is a morphologically rich language that can produce billions of word forms via inflectional and derivational transformations.
* **Flexible Syntax structure:** Syntax governs how a sentence is constructed and organized. Arabic syntactic structures allows for the rearrangement of syntactic elements. For example, Arabic sentences may or may not contain verbs, omissions of easy-to-guess sentence elements are routine, and verb subject and object may swap order.
* **Common lack of diacritics:** Most authors scarcely use diacritics in their writings, which increasing the ambiguity for automatic Arabic language processing, including morphological analysis.
* **Common lack of punctuation marks:** unfortunate lack of authoring standards, inconsistent use of punctuation marks, and common use of relatively long sentences without effective usage of punctuation marks significantly complicates syntactic analysis.
* **Orthographic complexity:** Orthography refers to how words are written. Arabic’s connected script and common attachment of pronouns and coordinating conjunctions complicates many HLT tasks such as search, OCR, and machine translation.

All these factors lead to many challenges for HLT applications.

Fortunately, Arabic HLT has gained much interest in the past few years, and currently there are many active teams worldwide in academy and industry working on such technologies. However, Arabic HLT continues to suffer from:

* **The absence of a clear national** **roadmap:** Some Western countries, particularly the US, have strongly impacted Arabic HLT through large grants to academia and industry. Although Western investments had a positive impact on Arabic HLT, they have not fulfilled our national needs, as their underlying requirements for the technologies are fundamentally different than ours. For example, Western investments have focused on Arabic to English machine translation, while most of our needs are towards English to Arabic machine translation. The lack of a national roadmap has resulted in fragmented work that often benefits Western roadmaps and has left many gaps in available Arabic HLT.
* **Insufficient Language Resources (LRs):** Arabic HLT requires significant language resources and tools that are either not available or accessible. Large companies often have sufficient business interest to produce their own resources, but small companies usually cannot afford such a luxury. Additionally, resources that are built by companies are normally not accessible by external parties as they are seen as a competitive advantage over their rivals.

A widely available *Basic LAnguage Resources and Toolkits* (BLARK) to academia and industry can significantly accelerate the development of next generation language processing tools. Basic BLART components such as Arabic machine readable dictionaries and bilingual dictionaries continue to be insufficient or proprietary.

* **Inadaptable technologies:** Although, some language tools can be adapted to Arabic, many tools require new processing paradigms. For example, Arabic syntactic analysis is sufficiently different from a language such as English, necessitating extensive R&D to develop Arabic tag sets and perhaps new parsing paradigms.

The availability of Arabic HLT would have positive cultural impact on Arabic users [Arab Human Development Report 2004, UNDP] as more information will be easily accessible by more people in Arabic. Although a number of initiatives to facilitate the use of Arabic HLT are at hand, most are not mature with insufficient critical mass.

For all the aforementioned reasons there is a strong need for a dedicated association that brings together the major academic and industrial stakeholders to serve Arabic HLT, and for developing Arabic language technologies to a level that parallel the level achieved in other major languages. The prospected association will lead to a more widespread awareness and adoption of Arabic language technologies. It will also set and accredit standards for Arabic HLT tools and language resources. Most importantly, the association will set a clear roadmap for HLT. This association should be able to serve the entire Arab region and to be able to get support and fund from all the relevant associations and researchers. Our proposed association is to be called Arabic Language TEchnologies Center (ALTEC).

**2- Similar International Efforts and Available Tools**

There are numerous efforts worldwide that mirror the proposed activities of ALTEC. International efforts related to language resources and benchmarking include:

1. The Linguistic Data Consortium (LDC) at the University of Pennsylvania <http://www.ldc.upenn.edu> receives significant governmental funding to avail different language and speech resources to both industry and academia, and to participate in most state-of-the-art HLT projects.
2. The National Institute of Standards and Technology (NIST) in the US defines benchmarks and performance metrics for numerous natural language processing (NLP) tasks including automatic speech recognition (ASR), machine translation (MT), and information retrieval (IR). LDC and NIST significantly contributed towards many recent advances in HLT.
3. In Europe, Evaluations and Language resources Distribution Agency (ELDA) pursues similar efforts by availing language resources to academia and industry and participating in evaluation campaigns.
4. In the Arab world, Ecole Nationale d'Ingénieurs de Tunis (ENIT) in Tunisia released an OCR database that became despite it’s relatively small size a reference for researchers working on OCR.

The last few years witnessed the development of many open source tools which facilitate creating benchmark systems and thus focusing on core research and development problems. These systems include: the hidden Markov model toolkit (HTK) from Cambridge University (htk.eng.cam.ac.uk), the SRILM statistical language modeling toolkit from SRI (www.speech.sri.com/projects/srilm), Giza++ parallel text corpora aligner (www.fjoch.com/GIZA++.html), the Pharaoh and Moses machine translation systems ([www.isi.edu/publications/licensed-sw/pharaoh](http://www.isi.edu/publications/licensed-sw/pharaoh)).

**3- The Structure of ALTEC**

We aim to establish an effective entity that can *lead* and *harmonize* the research and development of Arabic HLT in the region.

An Egyptian not-for-profit non-governmental-organization (NGO) is herein proposed as an effective mechanism to undertake the aforementioned mission. Our proposed association, ALTEC, is expected to play a multitude of roles including:

* Providing a roadmap for Arabic HLT.
* Facilitating the contribution and exchange of expertise among all the stakeholders namely; industry, academia, and potential vertical and horizontal (public) clients.
* Dealing smoothly and credibly with big bodies; e.g. governmental institutions like ITIDA, Arab League … etc.
* Attracting funds and sponsorships from Arab and international organizations interested in Arabic HLT and affording it for Arabic HLT infrastructure projects that may seem less appealing to the industry and beyond the capabilities of universities.
* Recommending technical standards and best industry practices to Arabic HLT communities.
* Setting benchmarks for the different aspects of Arabic HLT, and persistently raising the bar in order to elevate the status quo in the field.
* Disseminating Arabic HLT knowledge and awareness through attending/organizing major scientific, and/or industrial events locally, regionally, and internationally.
* Fostering collaboration among all significant Arabic HLT parties towards achieving mega projects.

**ALTEC will have the following structure and legal form:**

* 1. Founding members who represent the major Egyptian business entities that work in the area of HLT along with selected national and international key persons in the field.
  2. A group of regular members who make up the mass of the organization.
  3. The NGO would have a General Assembly made of the founding members and active members. The General Assembly elects the board of directors (BoD).
  4. The BoD elects the *executive committee*.
  5. Two kinds of memberships are proposed, as exists in the Egyptian law governing civil associations, namely *active voting member* and *associate non-voting member*. The active affiliation in this association is limited to persons who represent companies and academic entities only.

**The founders**:

In order for ALTEC to undertake its prospected roles, its board consists mainly of experts in the HLT field who represent companies and research institutions:

Industrial founders:

1. Dr Ossama Emam (IBM-Egypt)

2. Dr. Kareem Darwish (CMIC-Microsoft-Egypt)

3. Dr. Mohamed Afify (Orange Labs-Egypt)

4. Prof. Mohsen Rashwan (RDI-Egypt)

5. Dr. Khalid Wahba (TayaIT)

6. Dr. Hazem AbdelAzeem (ITIDA)

7. Mr. Hamdy Solyman Mobarak (Sakhr)

Academic founders:

8. Prof. Aly Fahmy (Faculty of Computers & Informatics-Cairo University)

9. Prof. Ahmed Rafea (AUC)

10. Prof. M. Walid T. Fakhr (Arab Academy for Science & Technologies and Maritime

Transport, AAST-MT)

11. Prof. Samia Mashary (Electronics Research Institute “ERI”)

12. Dr. Sameh Alansary (Bibliotheca Alexandria)

13. Dr. Sherif Abdou (Cairo University, FCI)

14. Dr. Sameh Alansary (Faculty of Arts, University of Alex.)

With clear conditions for the active membership, other entities and individuals are also welcomed to share in ALTEC. ALTEC is about to form committees to share all the community stakeholders. Strategic clients of Arabic HLT will also be invited as associate members. Such entities will include: telecom companies, translation/localization companies and news and content management companies/agencies, ...etc.

**4- SWOT Analysis**

|  |  |
| --- | --- |
| **Strengths**  - Human Capital  - Governmental Support  - Industrial Products/Services  - Research activities | **Weaknesses**  - Lack of National Strategy.  - Lack of infrastructure; LR’s & baseline systems.  - Lack of qualified computational linguists  - Lack of focused fund. |
| **Opportunities**  - Increasing market need  - Available local/regional/int’l fund | **Threats**  - Adverse effects to poor maintenance.  - International and regional competition. |

**5- Mission, Strategy, Goals, Objectives, and Realization Mechanisms**

|  |
| --- |
| The mission of ALTEC is to help in keeping Arabic a well served language in the digital era to achieve the following major goals:  - Create a national Arabic HLT roadmap  - Develop the Arabic community.  - Improve communications with other culture.  - Elevate the Arabic HLT to the world standards. |

The prime strategy of ALTEC is to establish a network of partner centers of best practice in Arabic written and spoken HLT with a view of creating a long-term cooperation roadmap for Egypt and the region.

**6- Conclusion**

Arabic Language TEchnologies Center (ALTEC) has been established as a non-profit NGO to work on the advance of Arabic human language technologies (HLT) to the current level of HLT in the other major languages of the world. This document has first defined the challenges and current gaps, surveyed similar organizations elsewhere in the world, proposed an initial and future structure of ALTEC along with all its stake holders, and presented a SWOT analysis of the Arabic HLT. The document has then proceeded to define ALTEC mission, strategy, goals and structure.