**Technology:** Degraded Text Search

1. **Brief Overview**

This work has to do with searching degraded text that was degraded as a result of optical character recognition and automatic speech recognition. The need for this work comes as a direct result of massive digitization efforts of old books (such as the Million Book Project in which the Library of Alexandria is a member of), old manuscripts (such as digitization efforts going on Al-Ahram newspaper), and the ubiquitous recordings from the Arabic radio and television. This technology would allow for the seamless search in such content with the need for expensive manual annotation.

1. **State of the Art (For Latin Languages)**
   1. Technology and Future Trends: much has been invested in enabling search in such media. Some of the most notable efforts were conducted by the Text Retrieval Conference (TREC (<http://trec.nist.gov/>)) as follows:

* Confusion Track at TREC-4 and TREC-5 which focused on searching OCR degraded documents
* Spoken Document Retrieval Track at TREC-9 and TRC-10 which focused on searching ASR degraded documents
* Legal Track which has been running for the past few years at TREC and uses OCR degraded documents from the US Tobacco settlement

The document collections available for English are quite large totaling hundreds of thousands of documents. Some other efforts were conducted as part of the INEX evaluation and GALE project.

* 1. Applications and Reported Performance: this has been used for a variety of uses including searching scanned enterprise documents (Taghva, K., J. Borasack, A. Condit, and J. Gilbreth. Results and implications of the noisy data projects. Technical Report 94-01, Information Science Research Institute, University of Nevada, Las Vegas, 1994.); in handling news stories and audio broadcasts (as in the GALE project) for intelligence purposes; for searching digitized books (as in INEX evaluation); for searching archives of interviews (as in the MALAKH project).

1. **State of the Art (For Arabic Language)**
   1. Technology and Future Trends: Limited work has been done on OCR text retrieval, most of it done by Kareem Darwish on a collection of 3700 documents, with a serious lack of large document collections for this work. Also, there is no publically available Arabic document collection of ASR output.
   2. Current and Envisioned Applications and Market Priorities: The are many opportunities for this work revolving around:

* Book search to support efforts from large libraries such the Library of Alexandria and university libraries that have a large archive of MSc and Ph.D. theses that are currently not easily accessible
* Archive search to support organizations that have large archives such as Al-Ahram and Al-Akhbar news papers and many magazine publishers
* Audio and video archive search to support media organization that have large archive of audio material such as newscasts, shows, and audio report.
* Printed news search to support information gathering organization such as the foreign ministry

1. **Dependency Between Technologies**

This work would require access to OCR and ASR engines and large Arabic corpora to build language models

1. **Language Resources**
   1. Available Resources (English, Arabic):

* Small dataset to evaluate OCR retrieval
* Proprietary Arabic OCR engines
* Proprietary Arabic ASR engines
* Open source search engines
  1. Needed Resources (English, Arabic):
* Publically available large collection of OCR degraded text
* Publically available large collection of ASR degraded text
* Easy access to OCR and ASR engines with prohibitively expensive licensing fees

1. **Strengths, weaknesses, opportunities and threats** 
   1. Strengths

* Good local expertise
* Broad interest from many potential stakeholders
  1. Weaknesses
* No good publically available OCR and ASR for Arabic
  1. Opportunities
* Many potential local consumers for the technology
  1. Threats
* No real threats

1. **Suggestions for Survey Questionnaire**
2. **List of people/organizations pioneers in each application area to be targeted by the Survey**

* Large libraries such the Library of Alexandria and university libraries that have a large archive of MS and Ph.D. theses that are currently not easily accessible
* Organizations that have large archives such as Al-Ahram and Al-Akhbar news papers and many magazine publishers
* Media organization that have large archive of audio material such as newscasts, shows, and audio report.
* Information gathering organization such as the foreign ministry

1. **Key persons in each application area (on technical/LR levels)**
2. **Suggestions for Language Resources (specific to the application area) if ALTEC would like to start collection immediately.**

* Large datasets of:
  + At least 250,000 OCR degraded documents and associated printed documents
  + At least 50,000 ASR degraded documents and associated audio

1. **Summary**

**Technology:** Cross Language Search

1. **Brief Overview:** This technology has to do with searching multilingual documents using a native language query. For example, searching French documents using an Arabic query
2. **State of the Art (For Latin Languages)**
   1. Technology and Future Trends: Much effort has gone into enabling search across languages. Some of the most notable efforts include:

* TREC Cross Language Information Retrieval Track that ran for many years to enable English users to search French, Spanish, Chinese, and Arabic (<http://trec.nist.gov/>)
* Cross Language Evaluation Forum which focuses on allowing users to search across European languages (<http://www.clef-campaign.org/>)
* NTCIR workshop that has been running for many years with the focus of allowing users to search between Chinese, Korean, and Japanese (<http://research.nii.ac.jp/ntcir/>)
  1. Applications and Reported Performance: This technology has widespread use in:
* Web search where users can issue queries in their native language and get automatically translated documents using machine translation
* Image and video search on the web or enterprise collections
* News search across languages

Where all these applications are useful to general audiences and the last been beneficial to general users as well as information gathering users.

1. **State of the Art (For Arabic Language)**
   1. Technology and Future Trends: the most successful approaches have focused on using machine translation directly, machine translation aligned data (either as phrases or individual words), dictionaries, and transliteration engines.
   2. Current and Envisioned Applications and Market Priorities: As mentioned previously, this technology can be applicable in web, enterprise, and news search.
2. **Dependency Between Technologies**

May depend on machine translation output if dictionaries are not available

1. **Language Resources**
   1. Available Resources (English, Arabic):

* HMM and IBM aligners
* Dictionaries and parallel data between Arabic and English
* Open source search engines
* Many collections are available in a variety of languages and translating the queries would not be prohibitive to test Arabic 🡺 Other\_language search
  1. Needed Resources (English, Arabic)
* Parallel data between Arabic and other languages, which can obtained from existing sources or automatically mined
* Dictionaries between Arabic and other languages

1. **Strengths, weaknesses, opportunities and threats** 
   1. Strengths

* Little competition
* Interest from local stakeholders (CMIC)
* Good local expertise in the area
  1. Weaknesses
* Requires significant investment in building translation resources
* May require copyright clearance for documents
  1. Opportunities
* Large demand for cross language search
* Will directly affect Arabic users
  1. Threats
* Up until this time, there is little work in this area, with the most notable being from Google, which focuses entirely on web search

1. **Suggestions for Survey Questionnaire**
2. **List of people/organizations pioneers in each application area to be targeted by the Survey**

* Web users
* Information gathering organizations

1. **Key persons in each application area (on technical/LR levels)**
2. **Suggestions for Language Resources (specific to the application area) if ALTEC would like to start collection immediately.**

* Parallel data between Arabic and other languages of interest to users
* Parallel named entity data to train transliterators
* Dictionaries between Arabic and other languages of interest to users

1. **Summary**

**Technology:** Domain Specific Search

1. **Brief Overview:** This track deals with search for specific domains of text, such as news, medical, web, forum, blog, and legal search
2. **State of the Art (For Latin Languages)**
   1. Technology and Future Trends: Much of the work has been focused on standard collection building, good feature extraction for the different domains, and good ranking algorithms including machine learning based algorithms. A lot effort has gone into making specialized engines for particular kinds of documents. The Text Retrieval Conference (TREC) has long tried to address domain specific search. Since the start of TREC in 1992, TREC has had tracks for:

* News search, which was the first and longest running of all due to the funding of DARPA
* Legal search which addresses recall oriented search in legal documents including printed documents and emails
* Medical search which ran for a few years starting 2004 and focused on biomedical abstracts
* Blog search which as the name indicates is concerned with searching through blogs
* Web search has been running for many years and focuses on search in very large heterogeneous web documents
* Enterprise search which focused on searching for documents in intranets

Another evaluation forum such INEX has focused on search in XML documents and Wikipedia, and CLEF has had tracks that deal with image search

* 1. Applications and Reported Performance: Searching in different domains has many applications in many areas. For example, forum search is critical for the Arab market where forums continue to dominate the Arabic web.

1. **State of the Art (For Arabic Language)**
   1. Technology and Future Trends: unfortunately the only serious effort on domain specific search in Arabic has been on news search in the context of the TREC 2002 and 2003 cross language search track. Otherwise, there are hardly any domain specific collections and little or no publications addressing such domains.
   2. Current and Envisioned Applications and Market Priorities: The applications are too many to enumerate here, but just to name a few examples:

* Legal search to support lawyers who want to search in legislations and case logs
* Forum and blog search on the web for general users
* News search for Arabic news aggregators or news providers
* Image search on the web
* Enterprise search

1. **Dependency Between Technologies**

No dependencies

1. **Language Resources**
   1. Available Resources (English, Arabic): the only known collection is the TREC 2002/2003 cross language search Arabic collection with 383,000 news collection.
   2. Needed Resources (English, Arabic): many collections are needed, but to be prioritized according to user needs. Suggest collections include:

* A blog/forum collection
* A large news collection
* A collection of legal documents
* A snapshot of the Arabic web

For all these collections, a set of topic, typically between 50 and 100 topics, along with their relevance judgments that map the topics to the documents, are needed. Typically, building relevance judgments is the most expensive part of building a test collection.

1. **Strengths, weaknesses, opportunities and threats** 
   1. Strengths

* Little competition
  1. Weaknesses
* Requires significant investment in collection building
* May require copyright clearance for documents
  1. Opportunities
* Large demand for domain specific search
* Will directly affect Arabic users
  1. Threats
* Up until this time, there is little interest in our market to pose a threat

1. **Suggestions for Survey Questionnaire**
2. **List of people/organizations pioneers in each application area to be targeted by the Survey**
3. **Key persons in each application area (on technical/LR levels)**
4. **Suggestions for Language Resources (specific to the application area) if ALTEC would like to start collection immediately.**

Identifying which domains are of highest priority, then collect data for the documents and topics, and build relevance judgments.

1. **Summary**