January 2005

Analyst-Focused Arabic Information Retrieval

Robert N. Oddy
Syracuse University

Ann R. Diekema
Syracuse University

Jean Hannouche
Syracuse University

Elizabeth Liddy
Syracuse University

Grant Ingersoll
Syracuse University

Analyst-Focused Arabic Information Retrieval

Description/Abstract

An English-Arabic Cross-Language Information Retrieval Environment was created in which the analyst can query an Arabic database in English and retrieve a set of relevant Arabic documents. The retrieved Arabic documents are automatically translated into English to facilitate readability by the English-only analyst. Proper names of people, places, and organizations are extracted from the retrieved documents and transliterated from Arabic into English. They are presented to the analyst and serve to provide a brief summarization of the retrieved document search query in English. Cross-Language Information Retrieval (CLIR), itself a desideratum in the ARDA workshop, is a special case of Information Retrieval where retrieval is not restricted to the language of the query but queries in one language retrieve documents in other language(s) (Oard and Diekema, 1998).

The Arabic that is used in the system is called Modern Standard Arabic (MSA). MSA is the formal Arabic that is used throughout the Arab world in news and broadcast media, and the lingua franca of the Arab. MSA has an estimated 200 million speakers living in Iraq, the Arabian Peninsula, the Levant, Egypt, and Northern Africa.

Keywords

Foreign language processing, Search and retrieval, OSINT.

Disciplines

Computer and Systems Architecture

Additional Information

Analyst-Focused Arabic Information Retrieval

Anne R. Diekema, Jean Hannouche, Grant Ingersoll, Robert N. Oddy, and Elizabeth D. Liddy
Center for Natural Language Processing
School of Information Studies
Syracuse University
Syracuse, NY, 13244, USA
{diekemar,jhannouc,gsingers,liddy}@syr.edu, {bob}@robertoddy.com

Abstract

An English-Arabic Cross-Language Retrieval Environment was created in which the analyst can query an Arabic database in English and retrieve a set of relevant Arabic documents. The retrieved Arabic documents are automatically translated into English to facilitate readability by the English-only analyst. Proper names of people, places, and organizations are extracted from the retrieved documents and transliterated from Arabic into English. They are presented to the analyst and serve to provide a brief summarization of the retrieved document.

1. Introduction

Three years after the events on September 11, 2001, the Federal Bureau of Investigation is still having problems keeping up with the large amounts of foreign language material that needs to be examined (Lichtblau, 2004). There is a lack of linguists at the agency which exacerbates the difficulty in prioritizing which documents need to be translated first. This need provides the guiding principle in the Arabic Information Retrieval System (AIR), which is developed to match the workflow of the analysts.

1.1 Analysts' Requirement for Managing Topics

The design of the AIR System is based on an analysts' preferred methods of research – namely, being able to group searches and search results into what we call 'Topics'. This unique capability is based on requirements produced during a 2003 ARDA-sponsored workshop (Liddy, 2003). Amongst the list of analysts' top ten desiderata was the ability to retain 'topic' or 'task' while conducting multiple searches. We address this requirement explicitly in AIR by providing the ability for the analyst to maintain context.

1.2 Arabic Cross-Language Retrieval

The AIR system is an English-Arabic CLIR system where the analyst can search Arabic documents by typing their search query in English. Cross-Language Information Retrieval (CLIR), itself a desideratum in the ARDA workshop, is a special case of Information Retrieval where retrieval is not restricted to the language of the query but queries in one language retrieve documents in other language(s) (Oard and Diekema, 1998).

The Arabic that is used in the system is called Modern Standard Arabic (MSA). MSA is the formal Arabic that is used throughout the Arab world in news and broadcast media, and the *lingua franca* of the Arab. MSA has an estimated 200 million speakers living in Iraq, the Arabian Peninsula, the Levant, Egypt, and Northern Africa.

2. AIR System

The AIR system was designed with the working methods of analysts learned during the ARDA workshop in mind. The AIR system goes beyond simple searching by providing the analyst with the ability to categorize and track Topics over time, automatically translate documents, archive professional translations, and improve searching by garnering user feedback at critical points in the search process.

2.1 Search Topic Management

Use of the AIR System revolves around understanding the relationships among *topics*, and *enquiries*.

A Topic (i.e. Muslim Brotherhood). defines the general category under which the user is looking for information and serves as a placeholder for the specific Enquiries the user is going to ask the system concerning this general Topic. An Enquiry (i.e. the Spread of the Muslim Brotherhood from Egypt to other countries) is a specific query or question that is related to the Topic.

The AIR system allows the intel analyst to manage his or her Topics and respective Enquiries by storing, editing, deleting, sharing, and re-running existing Enquiries.

2.2 Word Sense Disambiguation

Automatic query translation is a difficult task which involves identifying the correct sense of the source word and then finding the correct translation for that sense of the source word. The AIR System enlists the user to specify the meaning of their English query, rather than having the system second guess the intended meaning of these terms. By providing the English definitions of Arabic translations, even non-Arabic speakers can select the correct translations, thus greatly improving query performance.

2.3 Relevance Feedback

The AIR System implements two types of relevance feedback: automatic relevance feedback, and user relevance feedback. The former is done automatically by the system, based on the initial set of retrieved documents, while the latter uses relevance judgments from the analyst to improve the query and subsequent retrievals.

2.4 Translation

The AIR system uses a special in-house translation lexicon for query translation and uses machine translation to translate the Arabic retrieval results back into English. The AIR System supports two third party MT systems: Systran 5.0 and Language Weaver. While the System is currently used to support searching in English and Arabic, other MT engines can easily be integrated.

2.5 Proper Name Detection

Detecting Proper Names (PNs) is quite challenging in languages like Arabic as it shares no cognates with English. The AIR PN module utilizes clue words in the document text to detect PNs in six different categories: People, Major Cities, Locations, Countries, Organizations, and Terrorist Groups.

2.6 Proper Name Transliteration

Transliteration is the representation of Arabic characters into letters of the English alphabet and vice versa. For example, we know the Palestinian president as Mahmoud Abbas, which is a transliteration of his Arabic name (محمود عباس). Unfortunately, there is no one-to-one correspondence between the different alphabets nor between the letters and their sounds, thus requiring a probabilistic transliteration model. The system incorporates transliteration models for English-Arabic as well as Arabic-English.

3. Evaluation

The AIR System underwent two evaluations: a user evaluation and a system evaluation. A user study was carried out with three analysts who found the user interface easy to use and the retrieval results easy to read. A series of "runs" of the English-Arabic system were performed in a non-interactive environment to assess the effects of the various retrieval system components. The document collection used in these tests was the TREC-10 collection of 383,872 news stories in Arabic together with 50 queries in English.

Relevance feedback to create good revised Arabic queries requires relevant documents in high ranks. We developed a metric called "Early Precision" (EP) to reflect this requirement. EP is the average of the precision values when 5, 10 and 15 documents have been retrieved. Early in the search, recall will be very low, simply because few documents have yet been retrieved. Recall is of interest later in the search, so we report recall values when 30 and 100 documents have been retrieved, averaged over all queries.

	EP	Recall @30	Recall @100
monolingual	0.583	0.130	0.271
cross-lingual	0.255	0.054	0.125

Table 1. System results.

As is typical in CLIR, the results (see Table 1) show that monolingual Arabic easily outperforms the cross-lingual English-Arabic capabilities. This shows that the translation step can be improved.

4. Conclusions and Future Research

The AIR System is a CLIR system that allows analysts to search and retrieve relevant information from Arabic language documents, even though they may not be familiar with Arabic. For future work we plan to extend our query translation capabilities, and continue work on the Topic organization and query capabilities. We also plan on expanding the Proper Name detection and transliteration modules, as well as exploring additional means for bringing CLIR performance up to the level of monolingual Arabic retrieval.

References

Lichtblau, E. 2004. FBI Said to Lag on Translating Terror Tapes. New York Times, September 28, 2004.

Liddy, E.D. (2003). Question Answering in Contexts. Invited Keynote Speaker. ARDA AQUAINT Annual Meeting. Washington, DC. Dec 2-5, 2003.

Oard, D. and Diekema, A. 1998. Cross-Language Information Retrieval. Martha Williams (Ed.), *Annual Review of Information Science (ARIST)*, Vol. 33, Information Today Inc., Medford, NJ, pp. 223-256.