Arabic Information Retrieval Perspectives

Ahmed Abdelali (1), Jim Cowie (1) and Hamdy S. Soliman (2)
(1) Computing Research Laboratory
Box 30001/3CRL New Mexico State University
Las Cruces, NM 88003 USA
{ahmed,jcowie}@crl.nmsu.edu
(2) Computer Science Department
New Mexico Institute of Mining and Technology
801 Leroy Place
Socorro, N.M. 87801 USA
hss@nmt.edu

Abstract

Arabic IR (Information Retrieval) has recently become a focus of research and commercial development. Very few standards for evaluation of such tools are known and available. A concrete evaluation for Arabic IR systems is necessary for the advancement of this field.

In this paper we discuss available resources for testing Arabic IR systems and we propose a strategy to accelerate the development and the reliable evaluation of systems that can than be exploited by a wide range of users of varied interests.

Keywords

Arabic, Information Retrieval, Localization, Corpora, Lexicon.

1 Introduction

The fast growth of the World Wide Web (WWW) accompanied by an overwhelming explosion of multilingual resources, web spiders, indexing, and retrieving tools is driving development in multilingual IR.

Arabic one of the six official languages of the United Nations, mother tongue of more than 300 million people, has seen a very rapid growth. Statistics shows that since 1995 when the first Arabic newspaper was launched online www.asharqalawsat.com, the number of Arabic websites has been growing exponentially. By 2000 there were about 20 thousand Arabic sites on the web, about 7% of the published sites on the web.

Arabic is a Semitic language and its writing orientation is from left-to-right. The Arabic alphabet consists of 28 letters; the alphabet can be extended to ninety elements by additional shapes, marks, and vowels. Most Arabic words are derived from a root, generally composed of three consonants; occasionally the root can be also formed of two, four or rarely five consonants.

2 State of the art

Arabic IR (Information Retrieval) has become a focus of research and commercial development due to the vital necessity of such tools for people in the electronic age. The number of the Arabic-speaking Internet users in 2002 was about 4.4 million, about 1.5% of the Arab world population (Levini, 2002). But, on the other side of the picture few search engines are available to Arabic-speaking users, even though efforts are in progress to serve the increasing number of users. Recently standards for evaluation of available tools were introduced by TREC and CLEF starting in 2000.

2.1 Arabic Information Retrieval

A quick survey of Arabic IR tools shows two major categories:

- Full form based IR: Most of the commercial engines used are full form retrieval system. These include Sakhr web engine www.alidrisi.com; and www.ayna.com and other Unicode multilingual engines such as www.alltheweb.com or www.google.com.

- Morphology-based IR: The efforts that have been made in the academic environment to evaluate more sophisticated systems give an idea about the next generation of the Arabic search engines. Evaluation has been performed on systems using different approaches of incorporating morphology –stem, root based, light stem- (Larkey et al, 2002)(Gey, Oard. 2001). Other non-rule-based statistical stemmers or n-gram model have been tested for this purpose. Generally, using stemmers improves the recall as well as the precisions. (Larkey, Connell. 2002) experiments showed that the light stemmer performs better than the regular stemmer.

2.2 Arabic Resources

One of the major limitations for the Arabic IR system developer is the lack of adequate resources that could help test their system to get good evaluation of the system's performance in the real world.

2.2.1 Corpora

The only large scale resources known and available to users are the LDC collection (869 megabytes of Arabic news articles divided over of 383,872 documents from Agence France Press (AFP) used for the TREC evaluations) and the Al-Hayat newspaper collection from the European Language Resources Distribution Agency. There has been a little investigation of

these collections and their limitations in terms of richness and representativeness. This is discussed in the following section (Section 3).

2.2.2 Lexicons

Other resources such as monolingual dictionaries and bi-lingual dictionaries are needed; these types of resources can vary from Machine translation dictionaries to handcrafted dictionaries for a specific topic or usage. Available online are the Ajeeb dictionary and the Ectaco, which were used in some IR experiments (Larkey et al, 2002, 2003). Other individual efforts were carried out and deployed in different applications (Zajac, et al, 2001)

2.2.3 Tools

Arabic has a high degree of inflection; its morphology is a challenging task for IR systems. This complexity comes from issues such as the fact there is no space between words and pronouns, or that some glyphs are very ambiguous, as for example the case for Alif "ا". People tend to not to write the Hamza so the bare Alif could be either "ا" or "أ" or "إ". The prefixes and suffixes could be a conjuction of two or three or even four grammatical tokens as the case of "وسيعلمانها" or "تعلمونهن" "ليعلمانكها" "They will teach it to you", "you teach them" "to have them to teach it to you" respectively. To solve this problem two major approaches were used to build morphological analyzers, Table based morphological analyzers (Tim Backwalter) or Rule based such as that by Shereen Khoja (2001), and the Finite state Transducer of Xerox by Beesely (1996,1998).

No direct evaluation of morphology systems is available. An indirect evaluation of the performance might be obtained from the overall performance of the IR systems that the morphological analyzers were deployed in. Obviously some evaluation data and tools would be useful for Arabic computational linguistics developments.

3 Evaluation of the Current IR Systems

1. An Analysis of the TREC Arabic Evaluation

The performance of the systems in retrieving document from the Arabic collection seems to be comparable performance to current systems for other languages [See graphs].

Before taking a look at the published performances of some systems, we discuss the test bed used for the evaluation. As mentioned previously, the TREC evaluation uses the LDC corpus of AFP newswire articles dated from 1994 to 2000. The collection was compiled at the Middle East office of the AFP in Cypress. The language used is Modern Standard Arabic (MSA). An assessment of the corpus for its quality using Zipf's law shows that its completeness and representativeness make it an adequate collection. Some other comments could be drawn about the collection.

- It is a very small subset of actual MSA text in terms of syntax, morphology, and composition style. There are examples of style that seem unique to the AFP corpus (Figure 1). It does not take into consideration the wide variety of styles that exist from

the different Arabic countries. Abdelali (2004) details the issue of Localization in Modern Standard Arabic.

1. لكن الانفجار احدث فجوة كبيرة في الحافلة وتشاهد برك من الدم في المكان.
2. لا يوجد اليوم احد فى العالم يوءيد محاولتنا لمحو اسم مقدونيا
3. وفى حلبة سباق مونزا حيث تنظم غدا الاحد بطولة جائزة ايطاليا الكبرى (فورمولا واحد) لزم الجميع الصمت لمدة دقيقة. 45
4. احرزت اليابانية توموي ماكابي ذهبية وزن 84 كلغ بفوزها على الكورية الشمالية شا هيون هيانغ في المباراة النهائية لمسابقة الجودو ضمن دورة الالعاب الاسيوية الثالثة عشرة في بانكوك اليوم الاثنين.
5. وقام المحققون البريطانيون مساء باستجواب الموقوفين الثلاثة يساعدهم عملاء من اجهزة الاستخبارات الاميركية
6. ان شخصا قتل واصيب اربعة آخرون بجروح امس الاحد في بوروندي في هجوم بالقنابل اليدوية والسلاح الرشاشة استهدف قافلة تابعة لهذه المنظمة الفرنسية.
7. وتمكنا من العودة ليل الخميس الجمعة الى فريتاون، كما قال احدهما لوكالة فرانس برس.
8. واوضح الصحافي السيراليوني كريستو جونسون الذي يعمل لوكالة رويتر البريطانية انه افرج عنه مع عضو في بعثة الامم المتحدة في سيراليون (يونوميسيل) المعنية بمشكلات مرتبطة باحترام حقوق الانسان.

Figure 1 : Excerpts from AFP News collection

- Spelling of translated or transliterated proper names in general tends to be inconsistent in Arabic; The Figure 2 shows examples of the inconsistency. Although some of them could be considered as typos.

Word AFP	English	Occurrences	Word AFP	English	Occurrences
لوس انجليس	Los Angeles	21	نيوهامبشير	New Hampshire	15
لوس انجلوس	Los Angeles	23	نيوهامبشر	New Hampshire	9
لوس انجيلس	Los Angeles	2	جوهانس	Johannes	4
لوس انجيليس	Los Angeles	34	يوهانس	Johannes	74
انجلترا	England	2	يوهانيس	Johannes	8
انكلتر	England	1	جوهانز	Johannes	1
انكلترا	England	1	جوهانسبورغ	Johannesburg	173
كارولاينا	Carolina	26	جوهانسبرغ	Johannesburg	15
كارولينا	Carolina	14	جوهانسوبورغ	Johannesburg	1
ويسكونسين	Wisconsin	8	جوهانسورغ	Johannesburg	1
ويسكنسن	Wisconsin	2	فايمار	Weimar	3
ويسكونسن	Wisconsin	16	فيمار	Weimar	10

Figure 2 : TREC-2001 resources used by participants

This problem will affect the performance of any system and will cause systems to report incomplete results.

The performance of systems using this resource was published by the TREC conference in the first year they introduced Arabic for monolingual and cross-lingual evaluation. The published results for TREC-2001 show the achievements of the different participants in the competition. Figure 3 shows the resources and the approaches used by the participants. The graph gives a comparison of performance of the systems. Even though the performance shows satisfactory results bearing in mind that the test bed -LDC corpus- used for the evaluation is very specific. The major question raised is “Could this performance still be attained with different corpora”?

For the participants that used Stem or Root approach it wasn't clear how their approach resolved ambiguity, as we know that it is unusual to find a text marked with diacritics –short vowels- in MSA, which results in a lot of ambiguity. Some work been done by researchers to resolve the issue by restoring the vowels of the text (Beesely, 2001) (Gal, 2002).

Team	Arabic Terms Indexed				Query	Translation Resources Used		
	Word	Stem	Root	*n*-gram	Lang	MT	Lexicon	Corpus
BBN		X			A,E	X	X	X
Hummingbird		X			A			
IIT	X	X	X		A,E	X	X	
JHU-APL	X			X	A,E,F	X		
NMSU	X	X			A,E		X	
Queens	X			X	A,E	X		
UC Berkeley		X			A,E	X	X	
U Maryland	X	X	X	X	A,E	X		
U Mass	X	X			A,E	X	X	
U Sheffield	X				A,E,F	X		

Figure 3 : Arabic TREC-2001 resources used by participants

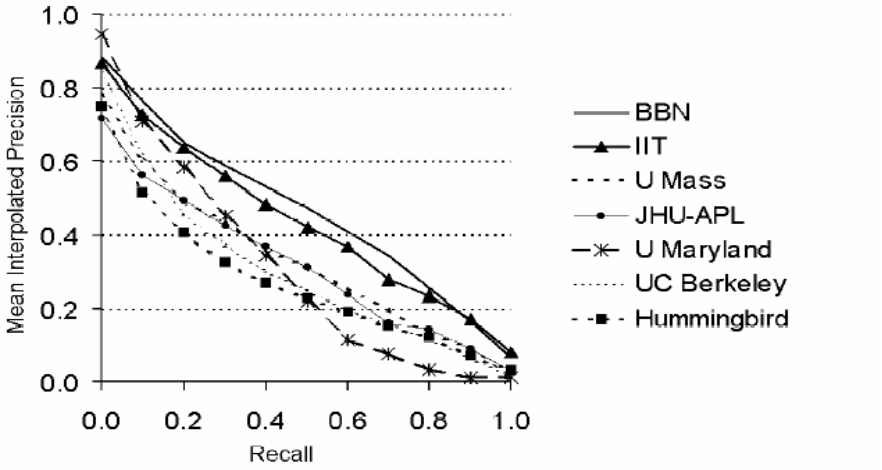

Figure 4 : Arabic TREC-2001 effectiveness

4 Conclusion: Perspectives for Arabic IR

As was shown in the previous section, so many issues hinder current systems from demonstrating real effectiveness. The next generation of Arabic IR systems will need to consider working with richer data. Current data found on the Internet could be a starting point for building a considerable resource adequate to challenge systems and reflect their ability to handle real data and text circulated by the different media resources. Initial work done by (Goweder and De Roeck, 2001) and (Abdelali, Cowie, and Soliman, 2003) demonstrated the feasibility of considering resources from daily newspapers published throughout the Arab world to construct well balanced corpora that could be used for the task of improving Arabic IR and many other tasks.

It is very important to get a concrete evaluation for the morphological analyzers used as the backbone in many of the IR systems. This will lead to a real assessment of the efficacy of analyzers and help improve them.

References

Abdelali, Ahmed. 2004. Localization in Modern Standard Arabic. Journal of the American Society for Information Science and Technology (JASIST), Vol. 55, N. 1, 2004. pp. 23-28.

Abdelali, Ahmed. Cowie, Jim. and Soliman, Hamdy S. 2003. Building Modern Standard Arabic Corpus. Unpublished manuscript.

Beesley, Ken. 1996. Arabic finite-state morphological analysis and generation. In Proceedings of COLING-96, the16th International Conference on Computational Linguistics, Copenhagen.

Beesley, Ken. 1998. Arabic morphological analysis on the internet. In Proceedings of the 6th International Confer-ence and Exhibition on Multi-lingual Computing, Cambridge, April.

Beesley, Kenneth R. 2001. Finite-state morphological analysis and generation of Arabic at Xerox Research: Status and plans in 2001. In ACL Workshop on Arabic Language Processing: Status and Perspective, pp. 1–8, Toulouse, France, July 6th.

Gey, Fredric C. Oard, Douglas. 2001. The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic Using English, French or Arabic Queries. TREC 2001. pp. 16-25

Gal, Ya'akov. 2002. An HMM Approach to Vowel Restoration in Hebrew and Arabic ACL 02 Semitic Language Workshop. pp. 27-33

Goweder, A. and De Roeck, A. Assessment of a significant Arabic corpus. Presented at the Arabic NLP Workshop at ACL/EACL 2001, Toulouse, France, 2001.

Larkey, Leah S. and Connell, Margaret. 2002. Arabic Information Retrieval at UMass in TREC-10. The Tenth Text Retrieval Conference. NIST, pp. 562-570.

Larkey, Leah S., James Allan, Margaret E. Connell, Alvaro Bolivar, and Courtney Wade. 2003. UMass at TREC 2002: Cross Language and Novelty Tracks. Ellen M. Voorhees and Lori P. Buckland (Eds.) TREC 2002, NIST Special Publication 500-251, pp 721-732.

Levini, Jenifer. 2002. The Internet Minute: Languages on the Net September. http://sonomabusiness.com/archives/2002-09-column-levini.html Retrieved January 5, 2004

Shereen, Roger Garside and Gerry Knowles. 2001. An Arabic Tagset for the Morphosyntactic Tagging of Arabic. Corpus Linguistics 2001, Lancaster University, Lancaster, UK.

Zajac, Rémi, Malki, Ahmed, Abdelali, Ahmed, Cowie, James, Ogden William C. 2001. Arabic-English NLP at CRL, Proceedings of the Arabic NLP Workshop ACL/EACL 2001.