# Arabic Information Retrieval

**Kareem Darwish**
Qatar Computing Research Institute
kdarwish@qf.org.qa

**Walid Magdy**
Qatar Computing Research Institute
wmagdy@qf.org.qa

now
the essence of knowledge
Boston — Delft

**Foundations and Trends® in Information Retrieval**

*Published, sold and distributed by:*
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

*Outside North America:*
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

K. Darwish and W. Magdy. *Arabic Information Retrieval*. Foundations and Trends® in Information Retrieval, vol. 7, no. 4, pp. 239–342, 2013.

*This Foundations and Trends® issue was typeset in LATEX using a class file designed by Neal Parikh. Printed on acid-free paper.*

ISBN: 978-1-60198-777-8

**Foundations and Trends® in Information Retrieval**
Volume 7, Issue 4, 2013

# Editorial Scope

## Topics

Foundations and Trends® in Information Retrieval publishes survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation, and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

## Information for Librarians

Foundations and Trends® in Information Retrieval, 2013, Volume 7, 5 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.

Foundations and Trends® in Information Retrieval
Vol. 7, No. 4 (2013) 239–342

DOI: 10.1561/1500000031

# Arabic Information Retrieval

Kareem Darwish
Qatar Computing Research Institute
kdarwish@qf.org.qa

Walid Magdy
Qatar Computing Research Institute
wmagdy@qf.org.qa

# Contents

## Abstract

In the past several years, Arabic Information Retrieval (IR) has garnered significant attention. The main research interests have focused on retrieval of formal language, mostly in the news domain, with ad hoc retrieval, OCR document retrieval, and cross-language retrieval. The literature on other aspects of retrieval continues to be sparse or non-existent, though some of these aspects have been investigated by industry. Others aspects of Arabic retrieval that have received attention include document image retrieval, speech search, social media and web search, and filtering. However, efforts on different aspects of Arabic retrieval continue to be deficient and severely lacking behind efforts in other languages. The survey covers: 1) general properties of the Arabic language; 2) some of the aspects of Arabic that affect retrieval; 3) Arabic processing necessary for effective Arabic retrieval; 4) Arabic retrieval in public IR evaluations; 5) specialized retrieval problems, namely Arabic-English CLIR, Arabic Document Image Retrieval, Arabic Social Search, Arabic Web Search, Question Answering, Image retrieval, and Arabic Speech Search; 6) Arabic IR and NLP resources; and 7) open IR problems that require further attention.

K. Darwish and W. Magdy. *Arabic Information Retrieval*. Foundations and Trends® in Information Retrieval, vol. 7, no. 4, pp. 239–342, 2013.

DOI: 10.1561/1500000031.

# 1

# Introduction

Most early studies on Arabic IR relied on relatively small test collections containing hundreds of documents that are composed of character-coded Arabic text (7; 13; 88). Increased interest in Arabic processing and retrieval in the early 2000's led to significant work that mostly relied on a single large collection (from TREC-2001/2002) (68; 129). However, most of the work was restricted to ad hoc retrieval and cross-language retrieval. Later work focused on other aspects of Arabic retrieval including document image retrieval, speech search, social media and web search, and filtering. However, efforts on different aspects of Arabic retrieval continue to be deficient and severely lacking behind efforts in other languages. This survey reviews recent literature pertaining to different aspects of Arabic IR including different domains and applications. It also describes some of the Arabic specific challenges affecting retrieval and some of the proposed solutions to these challenges. Further, it identifies the available resources and open areas of research to aid those interested in Arabic IR research.

The remainder of this introductory section presents general interesting aspects of Arabic and outlines the content of subsequent sections in the survey.

## 1.1 The Arabic Language

Arabic is the most widely spoken Semitic language with an estimated 400 million speakers. Arabic shares many commonalities with other Semitic languages. These commonalities pertain to morphology, vocabulary, word order (subject-verb-object and verb-subject-object), use of short and long vowels, etc. For example, Arabic and Hebrew words are typically derived from roots that are composed of two, three, or four letters, with three letter (triliteral) roots being the most common. Words are constructed from roots by possibly inserting infixes, adding prefixes and suffixes, or doubling constants. Diacritics, which are often omitted in writing, help disambiguate words. Nouns can be singular, plural, or dual, and masculine or feminine.

Arabic has a broad sphere of influence which is mostly due to: a) religious reasons, where Arabic is the language of Islamic scholarship and that of the Muslim holy book, the Qur'an; and b) Arabic was the language of science and technology during the Middle Ages, with major Arabic universities in Spain, Africa, and the Middle East being learning hubs. Consequently, Arabic is part of school curricula in most majority non-Arab Muslim countries such as Iran and Pakistan. Arabic is also an official language in other countries such as Eritrea, Chad, and Somalia. Arabic had influence, mostly in terms of vocabulary, on many other languages such as Spanish, Turkish, Persian, Urdu, Swahili, and Hausa. Further, Arabic script is used for writing many languages such as Persian, Urdu, Kurdish, Pashto, and Dari.

The Arab population is generally a young population with an average age in the Arab World slightly less than 24.[1] The Arabic language is ranked as the seventh top language on the web.[2] The Arab Internet users constitute 3.3% of the Internet users worldwide. Although Arabic is ranked seventh among languages on the web, it is the fastest growing language on the web among all other languages (Figure 1.1). The number of Arab Internet users grew from 2.5 million in 2000 to 65 million in 2011. Internet penetration in the Arab World is estimated to

[1] https://www.cia.gov/library/publications/the-world-factbook/index.html

[2] http://www.internetworldstats.com/

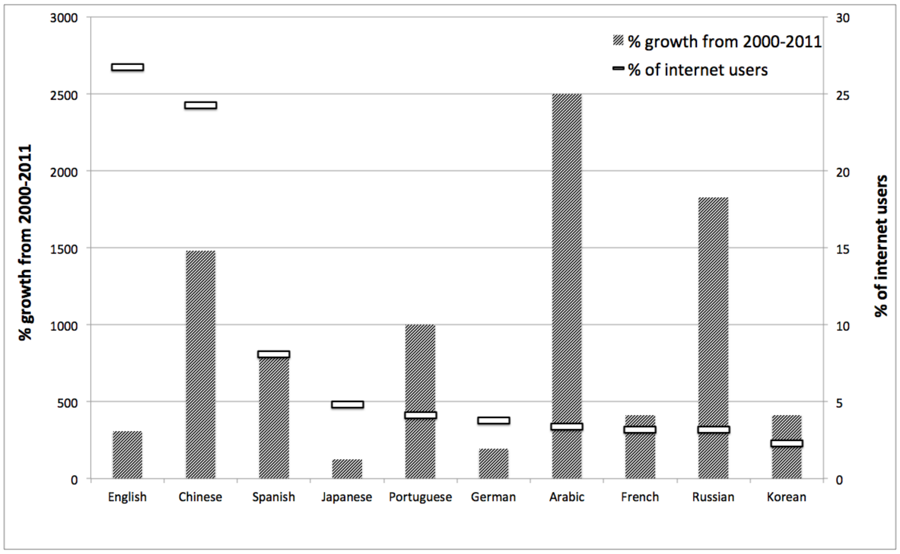
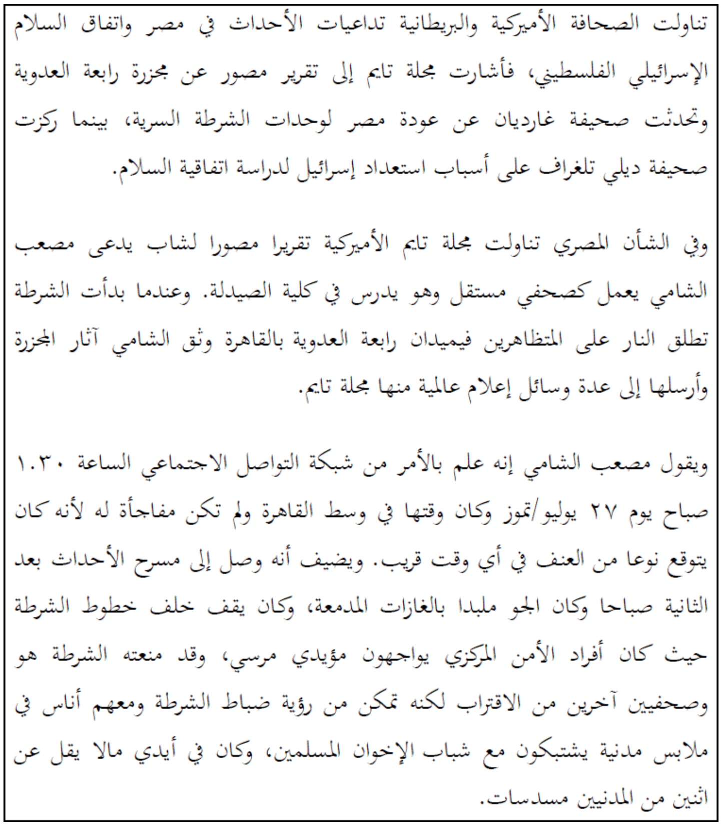


**Figure 1.1:** Top 10 languages in the Internet by 31 Dec 2011 (www.internetworldstats.com)

be 24%, which is lower than the global average of 32.7%. There are an estimated 45 million Arab Facebook users constituting roughly 5.6% of Facebook users globally. Though no exact estimates are available, Arabic online content is believed to constitute less than 1.5% of the global content. The relative size of Arabic forum content is disproportionately larger compared to the English forum content. English forum content is often considered of lower quality. However, such content is often of high quality in Arabic.

Modern Standard Arabic (MSA) is the lingua franca for the so-called Arab world, which includes northern Africa, the Arabian Peninsula, and Mesopotamia. Figure 1.2 shows a sample document written in MSA, which is an article from the Aljazeera.net news website. The article is written in MSA and would generally be understood by most Arabic speakers. However, Arabic speakers generally use dramatically different languages (or dialects) in daily interactions. There are six dominant dialects, which are Egyptian (85+ million speakers), Maghrebi (75+ million), Levantine (35+ million), Iraqi (25 million), Gulf (25+



تناولت الصحافة الأميركية والبريطانية تداعيات الأحداث في مصر واتفاق السلام الإسرائيلي الفلسطيني، فأشارت مجلة تايم إلى تقرير مصور عن مجزرة رابعة العدوية وتحدثت صحيفة غارديان عن عودة مصر لوحدات الشرطة السرية، بينما ركزت صحيفة ديلي تلغراف على أسباب استعداد إسرائيل لدراسة اتفاقية السلام.

وفي الشأن المصري تناولت مجلة تايم الأميركية تقريرا مصورا لشاب يدعى مصعب الشامي يعمل كصحفي مستقل وهو يدرس في كلية الصيدلة. وعندما بدأت الشرطة تطلق النار على المتظاهرين فيميدان رابعة العدوية بالقاهرة وثق الشامي آثار المجزرة وأرسلها إلى عدة وسائل إعلام عالمية منها مجلة تايم.

ويقول مصعب الشامي إنه علم بالأمر من شبكة التواصل الاجتماعي الساعة ١.٣٠ صباح يوم ٢٧ يوليو/تموز وكان وقتها في وسط القاهرة ولم تكن مفاجأة له لأنه كان يتوقع نوعا من العنف في أي وقت قريب. ويضيف أنه وصل إلى مسرح الأحداث بعد الثانية صباحا وكان الجو ملبدا بالغازات المدمعة، وكان يقف خلف خطوط الشرطة حيث كان أفراد الأمن المركزي يواجهون مؤيدي مرسي، وقد منعته الشرطة هو وصحفيين آخرين من الاقتراب لكنه تمكن من رؤية ضباط الشرطة ومعهم أناس في ملابس مدنية يشتبكون مع شباب الإخوان المسلمين، وكان في أيدي مالا يقل عن اثنين من المدنيين مسدسات.

**Figure 1.2:** Sample Arabic document from Aljazeera news website

million), and Yemeni (20+ million).[3] Aside from those, there are tens of different Arabic dialects along with variations within the dialects. Due to the spread of social media, users are increasingly using Arabic dialects online. These dialects may differ in vocabulary, morphology, and spelling from MSA and most do not have standard spellings.

The fast growth of Arabic content on the web and the large variations between MSA and different dialects make it essential to develop effective IR systems. Figure 1.1 shows the number of users for some of the languages and their growth trends. In this survey, the efforts exerted for developing these systems are explored for different IR applications.

[3] `http://en.wikipedia.org/wiki/Varieties_of_Arabic`

## 1.2 The Remainder of the Survey

The subsequent sections cover the following topics:

Section 2, entitled "Arabic Features Affecting Retrieval", presents key aspects of Arabic that affect retrieval. These key aspects include Arabic orthography and morphology; the use of MSA vs. dialects; the differences between formal and informal text; the use of non-standard textual representations; and Arabic properties affecting the retrieval of content in different modalities, namely print and speech.

Section 3, entitled "Arabic Preprocessing and Indexing", presents the core preprocessing steps that are required to prepare Arabic text for effective IR. The preprocessing steps including handling different encodings of Arabic, orthography, morphology, lexical and spelling variations, and stopwords. It also introduces effective index terms for Arabic.

Section 4, entitled "Arabic IR in Shared-Task Evaluations", explores the presence of the Arabic language in different IR evaluation campaigns such as TREC, TDT, BOLT, and CLEF. It also presents the different IR tasks at the campaigns, namely ad hoc retrieval, filtering, cross-language retrieval, topic detection and tracking, and question answering.

Section 5, entitled "Domain-specific IR", surveys work on different IR applications. These applications include cross-language IR, document image retrieval, general web search, social search, question answering, image retrieval, and speech search. The section addresses some of the challenges associated with different applications and some of the solutions that are reported in the literature.

Section 6, entitled "Open Research Areas in Arabic IR", explores open areas of research that require more work. These areas include ad hoc IR, question answering, social search, and web search.

Section 7 concludes the survey.

Appendix A focuses on listing and providing links to Arabic resources that can be useful for IR such as test collections, stemmers, index tools, and translation tools.

# 2

# Arabic Features Affecting Retrieval

Arabic poses many challenges for retrieval. When dealing with formal electronic texts such as news articles, which are primarily written in MSA, most of these challenges are due to orthography and morphology. However, dealing with social media, such as microblogs, and informal web content, such as forums and discussions, introduces further complications. These complications relate to the use of dialects, text decorations, abbreviations, word elongations, code switching between languages, and a Romanized version of Arabic commonly referred to as Arabizi. The retrieval of Arabic content in other modalities such as speech and printed text is affected by the orthographic and phonological properties of Arabic. In this section, we expound on these items in greater detail and describe their effect on retrieval.

## 2.1 Arabic Orthography and Print

Arabic has a right-to-left connected script that uses 28 basic letters, which change shape depending on their positions in words. There are eight other letters (or letter forms), namely different forms of hamza – إ (}) أ (&) آ (’) ء (|) ؤ (>) ئ (<), ta marbouta – ة (p), and alef maqsoura – ى (Y). Buckwalter encoding is used to Romanize Arabic text in this survey (76). The mappings are provided in section A.6 of the Appendix. Fifteen of the letters contain dots to differentiate them from other letters (78). Letters may or may not have diacritics (mostly short vowels), depending on the discretion of the producer of the document. Ligatures, which are special forms for some character sequences, and kashida, which are symbols that extend the length of words, are often employed in printed text. Figure 2.1 demonstrates some of these orthographic features.

Further, some letters are often used in place of others due to varying orthographic conventions, common spelling mistakes, and morphological transformations. Though varying forms of letters are important orthographically, morphologically, and syntactically, distinguishing between them may actually hurt retrieval effectiveness. These include:

- ي (y - ya) and ى (Y - alef maqsoura)
- ه (h - ha) and ة (p - ta marbouta)
- ا (A - alef), آ (| - alef maad), أ (> - alef with hamza on top), and إ (< - alef with hamza on the bottom)
- ء (’ - hamza), ؤ (& - hamza on w), and ئ (} - hamza on ya)

Optional diacritics, the use of kashidas, and inconsistent spellings all complicate retrieval. There are eight different diacritic marks that are commonly used in Arabic. Others, like the dagger, are less frequently used.

Concerning numerals, Arabic-Indic numerals are commonly used in Arabic writing instead of Arabic numerals. The Arabic numerals (0 1 2 3 4 5 6 7 8 9) have different codepage entries than Arabic-Indic numerals (٠ ١ ٢ ٣ ٤ ٥ ٦ ٧ ٨ ٩). Both types of numerals need to be normalized.

When dealing with digital text, most of these orthographic features are readily dealt with by using letter normalization and diacritic and kashida removal. Some of these processes often lead to increased ambiguity. When dealing with printed text that requires Optical Character Recognition (OCR) prior to retrieval, orthographic features typically adversely impact OCR quality. For example, dots and diacritics are often confused with dust and speckle on the page. Connected characters that change shape depending on the position in the sentence require more complex OCR models. Further, some characters do not connect to the characters that follow them, even if they are within the same word, complicating the determination of word boundaries. Some example characters are ا (A - alef), د (d - dal), and و (w - wa), which can be used together in the word وَالده (wAldh - meaning “his father”).

Dealing with retrieval of handwritten text is an area where the literature is scant and is not addressed in this survey. However, a few quick notes can highlight some of the difficulties of dealing with handwritten text. Figure 2.2 shows an example of a handwritten Arabic book. As the example shows, the book uses multiple fonts, words are often overlaid, some letters are extended to use up empty spaces, side notes are misaligned with the main text, the text is diacritized in some parts and not others, elaborate separators are used, and many letters have alternate starting, ending, or middle forms. It is noteworthy, that the page presented in Figure 2.2 was written by a professional calligrapher and hence is far more readable than regular handwritten text. Other manuscripts, as in Figure 2.3 may contain colored pages, archaic fonts that may not have dots, misaligned text, decorations, and drawings.

## 2.2 Arabic Morphology

Arabic words are divided into three main types: nouns, verbs, and particles (1). Particles are connection words such as prepositions and pronouns. Arabic nouns, which include adjectives and adverbs, and verbs are derived from a closed set of around 10,000 roots, which are

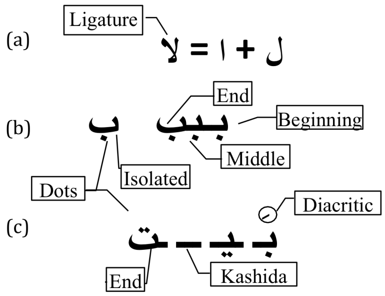
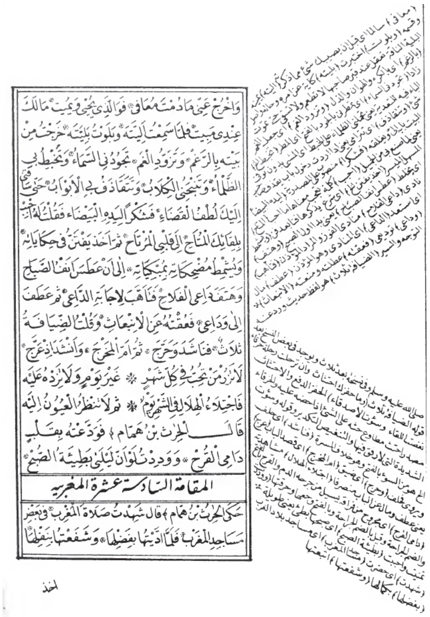


**Figure 2.1:** (a) Example of a ligature, (b) the different shapes of the letter “ba” and (c) example of a diacritic, kashida, and three letters which are distinguishable from each other only by dots

linguistic units of meaning composed of three, four, or five letters (44). Table 2.1 shows some of the words that can be generated from the root كتب (ktb). Arabic nouns and verbs are derived from roots by applying templates to the roots to generate stems. Applying templates often involves introducing infixes or deleting or replacing letters from the root. Table 2.2 shows some templates for triliteral roots.

Further, stems may accept multiple prefixes and/or suffixes to form words. Prefixes include coordinating conjunctions, determiners, and prepositions, and suffixes include attached pronouns and gender and number markers. Table 2.3 shows some of the possible prefixes and suffixes and their meanings. A word can have multiple prefixes and multiple suffixes. Further, plurals can be constructed using the addition of number markers as suffixes and this is often done using preset morphological transformations, producing so-called broken plurals. Some examples of singular to broken plural are: كتَاب (ktAb) → كتب (ktb); درة (drp) → درر (drr) or درّ (dr); and جَامع (jAmE) → جـوَامع (jwAmE). The number of possible Arabic words is estimated to be $6 \times 10^{10}$ words (9), with an estimated 1 million possible stems, and less than 10,000 roots. Lemmas are typically the units of meaning in Arabic, where a lemma is the canonical form of the word (e.g. “play” is the lemma of “plays” and “playing”). However, since finding the lemmas of words may be difficult, stems may serve as approximations for



واخرج عني ما دمت معافى * فوالذي يحيي ويميت ما لك
عندي مبيت * فلما سمعت اليته * وبلوت بليته * خرجت من
بيته بالرغم * وتزود الغم * تجود بي السماء * وتخبط بي
الظلماء * وتنبحني الكلاب * وتتقاذف بي الابواب * حتى ساقني
اليك لطف القضاء * فشكر اليد البيضاء * فقلت له انست
بلقائك المتاح * القلبي المرتاح * ثم اخذ يفتن في حكاياته
ويشمط مضحكاته بمبكياته * الى ان عطس انف الصباح
وهتف داعي الفلاح * فتاهب لاجابة الداعي * ثم عطف
الى وداعي * فعقته عن الانبعاث * وقلت الضيافة
ثلاث * فانشد وخرج * ثم ام المخرج * وانشد اذ عرج
لا تزر من تحب في كل شهر * غير يوم ولا تزده عليه
فاجتلاء الهلال في الشهر يوم * ثم لا تنظر العيون اليه
قال الحرث بن همام * فودعته بقلب
دامي القرح * ووددت لو ان ليلي بطيئة الصبح

المقامة السادسة عشرة المغربية

حكى الحرث بن همام * قال شهدت صلاة المغرب * في بعض
مساجد المغرب * فلما اديتها بفضلها * وشفعتها بنفلها

اخذ

**Figure 2.2:** Example of a handwritten book with notes on margins

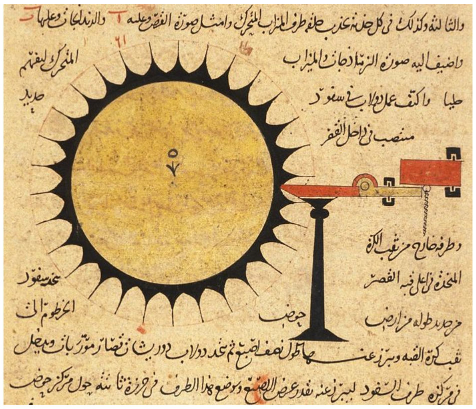


**Figure 2.3:** Example of a historical manuscript

| | | | | | |
|---|---|---|---|---|---|
| كتب (ktb) | He wrote | يـــكـــتـــب (yktb) | He is writing | أَكـــتـــب (>ktb) | I write |
| كَاتـــــب (kAtb) | Writer (masculine) | كـــتَـــاب (ktAb) | Book | كـــتَـــابـــه (ktAbh) | His book |
| وكتَابه (wktAbh) | And his book | كـــتَـــابهـــم (ktAbhm) | Their book (masculine) | كتب (ktb) | Books |

**Table 2.1:** Some of the words that can be derived from the root form كتب (ktb)

| | | | | | |
|---|---|---|---|---|---|
| فعل<br>CCC | كتب<br>(ktb - wrote) | فعَال<br>CCAC | كتَاب<br>(ktAb - book) | فَاعل<br>CACC | كَاتب<br>(kAtb - writer) |
| مفعول<br>mCCwC | مكتوب<br>(mktwb - something written) | فعَاعيل<br>CCACyC | كتَاتيب<br>(ktAtyb - Quran schools) | فعول<br>CCwC | كتوب<br>(ktwb - skilled writer) |

**Table 2.2:** Some templates to generate stems from roots with an examples from the root (كتب (ktb)). "C" stands for the letters that a part of the root.

| Examples of prefixes | | | | | |
|---|---|---|---|---|---|
| و (w) | And | ف (f) | Then | ال (Al) | The |
| ك (k) | Like | ل (l) | To | وَال (wAl) | and the |
| Examples of suffixes | | | | | |
| ه (h) | His | هم (hm) | Their | هَا (hA) | Her |
| ك (k) | your (singular) | كم (km) | your (plural) | ي (y) | My |

**Table 2.3:** Some examples of prefixes and suffixes and their meanings.

units of meaning. Hence, stems are very important. Arabic words may have multiple valid analyses, only one of which is typically correct in context. For example the word وليد (wlyd) could be the proper name "Waleed" or may mean "and to the hand of" و + ل + يد (w+l+yd).

## 2.3 Arabic Dialects

With spread of online social interaction in the Arab world, dialects started finding their way to online social interaction. There are 6 dominant dialects with many more variations of them and dozens more less spoken dialects. There are several factors that make dialects different. Different dialects may make different lexical choices to express concepts, though in many cases the lexical choices have proper Arabic roots. For example, the concept corresponding to "I want" is expressed as عاوز (EAwz) in Egyptian, أَبغي (>bgy) in Gulf, أبي (>by) in Iraqi, and بدي (bdy) in Levantine. The most popular MSA form is أريد (>ryd). In certain situations, some words are used to mean different or completely opposite things in different dialects. For example, the word تشكيل (t$kyl) means "applying diacritics" in most Arab countries, while it means "flirting" in Tunisia.

The pronunciations of different letters are often different from one dialect to another. One of the letters with most variations in pronunciation is the letter "qa" (ق (q)). In MSA, it is a voiceless uvular stop. However, it is pronounced as a glottal stop in Egyptian and Levantine, as a /g/ in Gulf, and as a voiced uvular stop in Sudanese (78). Another is the letter "jeem" (ج (j)) where it is pronounced as a soft "g" as in "gavel" in Egyptian and some Yemini dialects and as "j" like in "John" in most other dialects. Differing pronunciations of letters reflects on the way people spell dialectal text. For example, the word صدق (sdq – meaning "truth" and pronounced as "siddq") is often written in tweets from the Gulf as صج (sj – pronounced as "sijj").

Though the area of the Arab World is more than 35% larger than the area of the United States, large population centers are physically separated by seas (e.g. Red Sea) or large deserts (e.g. Sahara Desert). Such physical separation contributed to the independent development of Arabic dialects in different regions of the Arab world. Further, different parts of the Arab World have had strong interactions with other societies that do not speak Arabic. Some of the interactions have been the result of geographic juxtaposition or military conflict. For example, Lebanon was occupied by France, Egypt was occupied by the United Kingdom, Iraq has a large Kurdish minority and boarders Iran, and Algeria has a large Berber population. Such interactions have caused an influx of foreign words into the dialects of different countries. For example, in some of the Gulf countries you may hear the phrase شب اللّيت ($b Allyt - meaning "turn on the light"). In this particular example, شب ($b) is Arabic and اللّيت (Allyt) is a deformed form of the word "light". In more extreme situations as in countries like Algeria, people mix Arabic, French, and Berber together, while using grammatical rules of one language, typically Arabic, to construct sentences.

All these factors have complicated interactions between people from different parts of the Arab world. Social media platforms have caused large portions of populations to express themselves in writing. Though MSA was the traditional de facto modus operandi for writing Arabic, people became more inclined to use their dialectal Arabic in their written daily interactions on social media sites. Some notable trends started appearing in text used on social platforms, such as:

1. The writing of Arabic words using phonetically equivalent Latin characters, which is discussed in more detail in Section 2.4.

2. The mixed language text where many of the social platform users may speak multiple languages, mainly Arabic and English or Arabic and French. Consider the following tweet:

“أَهم حَاجة (>hm HAjp): do everything with pride”
where the Arabic is mixed with English to say “the most important thing: do everything with pride.”

3. The use of dialectal words that may have different morphological forms than MSA. For example, Egyptian Arabic uses a negation construct similar to the “ne pas” negation construction in French. Consider the Egyptian Arabic word ملعَبتش (mlEbt$ - meaning “I did not play”). It is composed of three parts “m+lEbt+$”. Such constructs are not handled by existing MSA morphological analyzers.
4. The use of phonetic transcription of words to match how words are pronounced in dialects. For example, the word صدق (Sdq - meaning “truth” or “honestly”) is often written as صج (Sj) to match the Gulf dialect.
5. Creative spellings, spelling mistakes, and word elongations are ubiquitous in social text.
6. The use of new words that do not exist in the language such as words expressing laughter (e.g. لول (lwl) corresponding to Laugh Out Loud (LOL)) and using Arabic words to mean new things (e.g. طحن (THn) meaning “grinding” in MSA, but intended to mean “very”).

All these factors may affect retrieval, particularly in the context of Arabic social media.

## 2.4 Arabizi

Arabic is sometimes written using Latin characters in transliterated form, which is often referred to as Arabizi, Arabish, Franco-Arab, and other names. Arabizi uses numerals to represent Arabic letters for which there is no phonetic equivalent in English or to account for the fact that Arabic has more letters than English. For example, “2” and “3” represent the letters أ (>) (that sounds like “a” as in apple) and ع (E) (that is a guttural “aa”) respectively. Arabizi is particularly popular in Arabic social media. Arabizi has grown out of a need to write Arabic on systems that do not support Arabic script natively. For example, Internet Explorer 5.0, which was released in March 1999, was the first version of the browser to support Arabic display natively.[1] Windows Mobile and Android did not support Arabic except through third party support until versions 6.5x and 3.x respectively. Despite the increasing Arabic support in many platforms, Arabizi continues to be popular due to the familiarity of users with it and the higher proficiency of users to use an English keyboard compared to an Arabic keyboard. Arabizi is used to present both MSA as well as different Arabic dialects, which lack spelling conventions and differ morphologically and phonetically from MSA. Additionally, due to the fact that many of the Arabic speakers are bilingual (with their second language being typically either English or French), another commonly observed phenomenon is the presence of English (or French) and Arabizi mixed together within sentences, where users code switch between both languages. Detecting and converting Arabizi to Arabic script would help: 1) ease the reading of the text, where Arabizi is difficult to read; 2) allow for the processing of Arabizi (post conversion) using existing NLP tools; and 3) normalize Arabic and Arabizi into a unified form for text processing and search. Detecting and converting Arabizi are complicated by the following challenges:

1. Due to the lack of spelling conventions for Arabizi and Arabic dialectal text, which Arabizi often encodes, building a comprehensive dictionary of Arabizi words is prohibitive. Consider the following examples:

    (a) The MSA word تحرير (tHryr - meaning “liberty”) has

[1] `http://en.wikipedia.org/wiki/Internet_Explorer`

    the following popular Arabizi spellings: ta7rir, t7rir, tahrir, ta7reer, tahreer, etc.

    (b) The dialectal equivalents to the MSA لَا يلعب (lA ylEb - meaning “he does not play”) could be مَابيلعبش (mAbylEb$), مَابلعبش (mAblEb$), ميلعبش (mylEb$), مَايلعبش (mAylEb$) ... etc. The resultant Arabizi could be: mayel3absh, mabyelaabsh, mabyel3absh, ... etc.

2. Some Arabizi and English words share a common spelling, making solely relying on an English dictionary insufficient to identify English words. Consider the following examples (ambiguous words are bolded):

    (a) Ana 3awez aroo7 **men** America leh Canada (meaning “I want to go from America to Canada”). The word “men” meaning “from” is also an English word, but with a different meaning.

    (b) I called **Mohamed** last night. “Mohamed” in this context is an English word, though it is a transliterated Arabic name.

## 2.5 Arabic Speech

Performing Automatic Speech Recognition (ASR) is commonly used as a component in retrieving audio documents. Morphology and the use of dialects significantly complicate Arabic ASR. As mentioned earlier, Arabic morphology is able to generate a very large number of word surface forms. This may adversely affect the coverage of underlying language models and underlying ASR dictionary with many unseen surface forms.

Another problem is that diacritics (mostly short vowels) are mostly not written by users, while they change the pronunciation of letters (23). For example, the letter ب (b) with different diacritics can be pronounced as:

1. بَ (ba) as in the English *ba*ck

2. بِ (bi) as in the English *bee*

3. بُ (bu) as in the English *boo*k

The letter can also be silent or stressed in combinations with other base diacritics. Also, when the base diacritics are doubled (using so-called "tanween"), an "n" sound is added after the letter. So بٍ (bin) is pronounced as *ben*. Therefore, the ASR may have to rely on an undiacritized language model and dictionary or both need to be diacritized, which is highly laborious. Other combinations of phonetic and orthographic features add more complexity. Two notable examples are:

1. the letter ة (p), which appears only at the end of words, is pronounced as "ha" if it appears in the last word in the sentence and as "ta" otherwise.

2. the letters ي (y) and ى (Y) are frequently (erroneously) interchanged by writers or both are consistently written as the latter. The first letter is a long vowel that is phonetically similar to that in "b*ee*" and the second, which only appears as the letter in a word, sounds like the long vowel ا (A) that is similar to that in "c*a*n". If a suffix is attached to a word containing ى (Y), ى (Y) is transformed to a ي (y) after attachment.

The problem with dialects is more profound, because:

1. Different dialects make different lexical choices leading to a large number of unique words.

2. Dialects do not have standard spellings. This aspect negatively impacts the creation of language models that may have multiple surface forms for the same word. Further, building suitable dialectal language models can be challenging.

3. Dialects are divergent in their pronunciation. As mentioned earlier, different dialects may pronounce some letters differently. For example the letter ق (q) can be pronounced as "ja", "a", or "q".

This has an effect on ASR's language and acoustic models.

4. MSA and dialects are often mixed during speech, and speakers may occasionally use foreign words.

## 2.6 Arabic on the Web

The size of Arabic presence on the web is not known. However, as of Google and Bing, they are estimated to index roughly 3 billion and 210 million Arabic pages respectively. To estimate the relative number of indexed Arabic pages, two Arabic stopwords (في (fy - meaning "in") and من (mn - meaning "from")) were used as queries to search in the two search engines. There are fundamental differences between the English and Arabic webs. Some of these differences are:

1. Unlike the English web, the Arabic web is sparsely connected. This makes features such as PageRank less useful.

2. The relative size of Arabic forum content is disproportionately larger compared to the English forum content.[2] English forum content is often considered of lower quality. However, such content is often of high quality in Arabic. Considering that most Arab companies and institutions lack web presence, many entities resort to forums to post information.

3. The size of the Arabic web is significantly smaller than the English web, and the size of available Arabic resources such as Wikipedia is dwarfed by those available for English. To contrast

[2] Based on communication with technical staff working on Arabic search at Google and Microsoft

both, Arabic Wikipedia has slightly more than 252 thousand articles compared to more than 4.4 million English articles.[3]

4. Much Arabic content is nestled inside large portals such as YouTube, WordPress, and Facebook. Short of indexing everything on the web in general and on such sites in specific, indexing the Arabic web specifically may be challenging.

There are other issues that are not necessarily unique to Arabic but would need handling nonetheless such as spam detection, query spell checking, web page segmentation, query reformulation, and results presentation.

[3] As of December 26, 2013

# 3

# Arabic Preprocessing and Indexing

This section presents core Arabic text preprocessing steps that have been shown to improve Arabic IR effectiveness. The preprocessing steps include: detecting Arabic texts in different encodings and converting texts to a common encoding; handling some of the orthographic features; performing some form of morphological analysis or stemming; identifying stopwords; and handling lexical and spelling variations. This section also introduces the best index terms for Arabic. Since many of experiments in the surveyed papers were conducted using varying experimental setups, it is hard to compare reported results in absolute terms.

## 3.1 Handling Encodings and Transliteration Schemes

There were multiple fragmented efforts to properly encode Arabic text in computing machinery. This fragmentation led to multiple divergent encodings for Arabic. Some of the encodings that encode Arabic letters are: ASMO-708, Windows CP1256, IBM420, ISO-8859-6, GB18030, and UTF-8. Though UTF8 has become the most dominant Arabic encoding on the web - where Arabic characters typically require two bytes and English characters require only one byte each, many documents continue to exist in single byte encodings - mainly Windows CP1256 and to a lesser extent ISO-8859-6.

Some transliteration schemes use Latin characters to represent Arabic characters such as Buckwalter transliteration which has a one-to-one mapping to Arabic characters (76). Arabizi is yet another way of representing Arabic that poses greater challenges, where Arabic words are written using numbers and Latin characters (e.g. 3' represents غ (g)). The problem of Arabizi is that it is free-form with *m*-to-*n* mappings between the Latin characters and Arabic characters. For example, though diacritics are rarely written in Arabic, they are often spelled out as vowels in Arabizi. For example, the word صبَاح (SbAH - meaning "morning"), could be written in Arabizi as "sabah" or "saba7".

To add to the problem, Arabic text can be nestled within other languages. These languages may be languages that use Latin characters, which would complicate the detection of Arabizi, or others that use an extended set of Arabic letters such as Urdu, and Kurdish, which share letters and often words with Arabic. Further, in social media, authors often use letters in the extended Arabic set to adorn words with letters that look similar to Arabic letters. These phenomena require the following:

- Determining if a sequence of bytes within a text document represents Arabic text or not. If so, Arabic text should be converted to a unified encoding (most likely UTF-8). These two steps already exist in most modern web browsers.

- Ascertaining if a word written in Latin characters is Arabic or not. If so, the Latin characters representing Arabizi text should be converted to Arabic characters. The detection of Arabizi is a straightforward language identification task. The detection is a bit more complicated if Arabizi is mixed with another language such as English or French. In such cases, the detection of Arabizi needs to be done at word level. Darwish published recent work on word-level Arabizi detection (54). In his work, he used a Conditional Random Fields (CRF) sequence labeler that was trained on a variety features such as English and Arabizi character n-gram language model scores, word n-gram language model scores, and the existance of letters and numerals in words. For the conversion from Arabizi to Arabic, there are several commercial tools that treat the task as a transliteration task. These systems include Yamli,[1] Maren[2] from Microsoft, and Ta3reeb[3] from Google. All three products are intended for online text entry and hence use limited language modeling (mostly unigram language models). Darwish (54) adapted transliteration work with language modeling to perform offline conversion of Arabizi to Arabic.

## 3.2 Handling Orthography

Prior to retrieval, the following features need to be handled: diacritics, kashidas, ligatures, and common spelling mistakes.

### Handling Diacritics:

Diacritics help disambiguate the meaning of words. For example, the two words عَلَم (Ealam - meaning “flag”) and عِلم (Eilm - meaning “knowledge”) share the same letters علم (Elm) but differ in diacritics. One possible solution is to perform diacritic recovery. However, this approach has many problems, namely: the accuracy of state-of-the-art Arabic diacrtizers on open domain text is typically below 90% for full diacritization including case endings (75); diacritization is computationally expensive, often causing indexing of large amounts of text to be prohibitive; diacritization of previously unseen words is generally intractable; and word sense disambiguation, which is akin

[1] http://www.yamli.com/ar

[2] http://www.getmaren.com/

[3] http://www.google.com/ta3reeb/

to diacritization, has been shown not to benefit retrieval (146). The more widely adopted approach is to remove all diacritics. Though this increases ambiguity, retrieval is generally tolerant of ambiguity (146). Further, this approach is computationally very efficient.

### Handling Kashidas and Ligatures:

Since kashidas are mere word elongation characters, they are typically removed. As for ligatures that are encoded as single characters in the code-page, they are transformed with the constituent letters. For example, the ligature لاَ (lA) is transformed to ل + ا (l+A). A ligature is transformed to its constituent letters only if the ligature is encoded as one character in the codepage. Normalization Form Compatibility Composition (NFKC), a unicode normalization form, can properly normalize most Arabic ligatures.[4]

### Common Spelling Mistakes and Variations:

When dealing with formal text such as newspaper articles, letter normalization is recommended to handle common spelling mistakes and spelling variations. Letter normalization pertains to four letters and their varying forms as follows:

- Different forms of the letter ا (A - alef), namely ا (A), آ (|), أَ (>), and إِ (<):
  - They are often erroneously confused by many who write Arabic.
  - Morphologically inflecting a word, e.g. changing the mood of a verb, may cause a change in the form of ا (A). For example, the verb ادرس (Adrs) in imperative mood (mean-

[4] http://unicode.org/reports/tr15/

ing “study” – in a command form) would turn to أَدرس (>drs) in first-person present tense (meaning “I study”).

- The letters ي (y) (ya) and ى (Y) (alef maqsoura):
  - They are often erroneously confused.
  - In some writing styles, a trailing ي (y) is always written as ى (Y).
  - A ى (Y), which only appears at the end of words, is most likely transformed into a ي (y) when a suffix is attached. For example, the word علَى (ElY - meaning “on”) is transformed to عليه (Elyh - meaning “on him”) with the attachment of the suffix ه (h).
  - In less common cases, ى (Y) is turned into the letter ا (A) with attachment of a suffix. For example, the word يرَى (yrY - meaning “he sees”) turns into يرَاه (yrAh - meaning “he sees him”).

- The letters ة (p - ta marbouta) and ه (h - ha):
  - They are often erroneously confused.
  - The letter ة (p), which appears strictly at the end of words, turns into the letter ت (t) with the attachment of suffixes. For example, the word لعبة (lEbp - meaning “toy”) turns into لعبته (lEbth - meaning “his toy”).

- Concerning the different forms of ء (’ - hamza), namely ؤ (&), ء (’), and ئ (}):
  - A standalone hamza ء (’) can change into a hamza on و (w) or hamza on ي (y) with the attachment of a suffix. For example, the word سَماء (smA’ - meaning “sky”) is turned

into either سَمَاؤُه (smA&h) or سَمَائِه (smA}h) both meaning “his sky” depending on the role of the word in the sentence (e.g. subject, object, etc.).

For the case of the varying forms of alef, ya and alef maqsoura, and ha and ta marbouta, it is possible to build a system that would correct these common mistakes with about 99% accuracy (63; 125). However, normalization is preferred for reasons that are similar to those associated with diacritic removal (as opposed to recovery). The most commonly used scheme for letter normalization is:

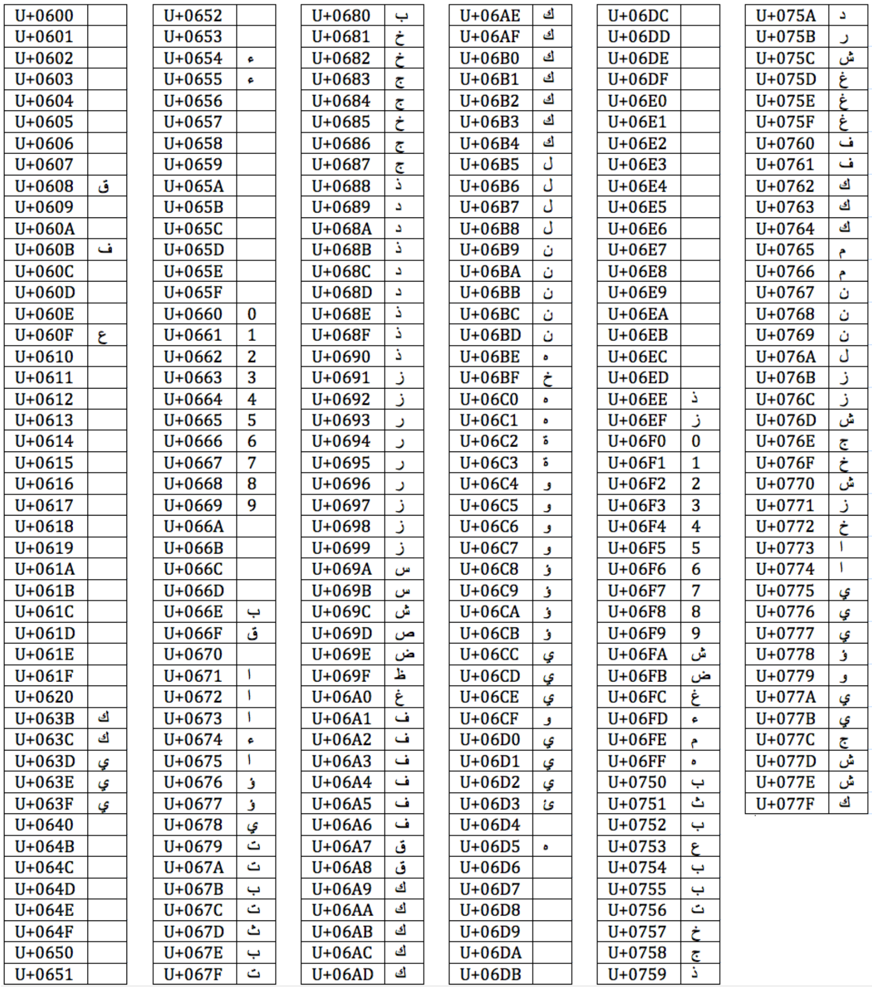
- ي (y) and ى (Y) are mapped to ي (y)
- ه (h) and ة (p) are mapped to ه (h)
- ا (A), أَ (|), آ (>), and إ (<) are mapped to ا (A)
- ؤ ( & ), ء ( ’ ), and ئ ( }) are mapped to ء ( ’ ) (47)

For informal text such as tweets and Facebook status messages, users may use an extended character set primarily to decorate the text. These additional characters are typically borrowed from Arabic-like alphabets such as Farsi and Urdu. For example, the word كبير (kbyr - meaning “large” or “big”) may appear in a tweet as گپير. Further, extra diacritics are available in the extended set. Figure 3.1 presents a full listing of the unicode entries of the extended Arabic characters and their mappings.

In addition, Arabic and Arabic-Indic numerals need to be normalized.

### Handling Word Elongation

In text from social media, authors routinely elongate words by repeating some of the characters in the word to express emotions or importance. For example, you may find words such as “cooooooool” and “loooool” in English tweets. In Arabic microblogs, there are two prevalent phenomena, namely:



| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| U+0600 | | U+0652 | | U+0680 | ب | U+06AE | ك | U+06DC | | U+075A | د |
| U+0601 | | U+0653 | | U+0681 | خ | U+06AF | ك | U+06DD | | U+075B | ر |
| U+0602 | | U+0654 | ء | U+0682 | خ | U+06B0 | ك | U+06DE | | U+075C | ش |
| U+0603 | | U+0655 | ء | U+0683 | ج | U+06B1 | ك | U+06DF | | U+075D | غ |
| U+0604 | | U+0656 | | U+0684 | ج | U+06B2 | ك | U+06E0 | | U+075E | غ |
| U+0605 | | U+0657 | | U+0685 | خ | U+06B3 | ك | U+06E1 | | U+075F | غ |
| U+0606 | | U+0658 | | U+0686 | ج | U+06B4 | ك | U+06E2 | | U+0760 | ف |
| U+0607 | | U+0659 | | U+0687 | ج | U+06B5 | ل | U+06E3 | | U+0761 | ف |
| U+0608 | ق | U+065A | | U+0688 | ذ | U+06B6 | ل | U+06E4 | | U+0762 | ك |
| U+0609 | | U+065B | | U+0689 | د | U+06B7 | ل | U+06E5 | | U+0763 | ك |
| U+060A | | U+065C | | U+068A | د | U+06B8 | ل | U+06E6 | | U+0764 | ك |
| U+060B | ف | U+065D | | U+068B | ذ | U+06B9 | ن | U+06E7 | | U+0765 | م |
| U+060C | | U+065E | | U+068C | د | U+06BA | ن | U+06E8 | | U+0766 | م |
| U+060D | | U+065F | | U+068D | د | U+06BB | ن | U+06E9 | | U+0767 | ن |
| U+060E | | U+0660 | 0 | U+068E | ذ | U+06BC | ن | U+06EA | | U+0768 | ن |
| U+060F | ع | U+0661 | 1 | U+068F | ذ | U+06BD | ن | U+06EB | | U+0769 | ن |
| U+0610 | | U+0662 | 2 | U+0690 | ذ | U+06BE | ه | U+06EC | | U+076A | ل |
| U+0611 | | U+0663 | 3 | U+0691 | ز | U+06BF | خ | U+06ED | | U+076B | ز |
| U+0612 | | U+0664 | 4 | U+0692 | ز | U+06C0 | ه | U+06EE | ذ | U+076C | ز |
| U+0613 | | U+0665 | 5 | U+0693 | ر | U+06C1 | ه | U+06EF | ز | U+076D | ش |
| U+0614 | | U+0666 | 6 | U+0694 | ر | U+06C2 | ة | U+06F0 | 0 | U+076E | ج |
| U+0615 | | U+0667 | 7 | U+0695 | ر | U+06C3 | ة | U+06F1 | 1 | U+076F | خ |
| U+0616 | | U+0668 | 8 | U+0696 | ر | U+06C4 | و | U+06F2 | 2 | U+0770 | ش |
| U+0617 | | U+0669 | 9 | U+0697 | ز | U+06C5 | و | U+06F3 | 3 | U+0771 | ز |
| U+0618 | | U+066A | | U+0698 | ز | U+06C6 | و | U+06F4 | 4 | U+0772 | خ |
| U+0619 | | U+066B | | U+0699 | ز | U+06C7 | و | U+06F5 | 5 | U+0773 | ا |
| U+061A | | U+066C | | U+069A | س | U+06C8 | ؤ | U+06F6 | 6 | U+0774 | ا |
| U+061B | | U+066D | | U+069B | س | U+06C9 | ؤ | U+06F7 | 7 | U+0775 | ي |
| U+061C | | U+066E | ب | U+069C | ش | U+06CA | ؤ | U+06F8 | 8 | U+0776 | ي |
| U+061D | | U+066F | ق | U+069D | ص | U+06CB | ؤ | U+06F9 | 9 | U+0777 | ي |
| U+061E | | U+0670 | | U+069E | ض | U+06CC | ي | U+06FA | ش | U+0778 | ؤ |
| U+061F | | U+0671 | ا | U+069F | ظ | U+06CD | ي | U+06FB | ض | U+0779 | و |
| U+0620 | | U+0672 | ا | U+06A0 | غ | U+06CE | ي | U+06FC | غ | U+077A | ي |
| U+063B | ك | U+0673 | ا | U+06A1 | ف | U+06CF | و | U+06FD | ء | U+077B | ي |
| U+063C | ك | U+0674 | ء | U+06A2 | ف | U+06D0 | ي | U+06FE | م | U+077C | ج |
| U+063D | ي | U+0675 | ا | U+06A3 | ف | U+06D1 | ي | U+06FF | ه | U+077D | ش |
| U+063E | ي | U+0676 | ؤ | U+06A4 | ف | U+06D2 | ي | U+0750 | ب | U+077E | ش |
| U+063F | ي | U+0677 | ؤ | U+06A5 | ف | U+06D3 | ئ | U+0751 | ث | U+077F | ك |
| U+0640 | | U+0678 | ي | U+06A6 | ف | U+06D4 | | U+0752 | ب | | |
| U+064B | | U+0679 | ت | U+06A7 | ق | U+06D5 | ه | U+0753 | ع | | |
| U+064C | | U+067A | ت | U+06A8 | ق | U+06D6 | | U+0754 | ب | | |
| U+064D | | U+067B | ب | U+06A9 | ك | U+06D7 | | U+0755 | ب | | |
| U+064E | | U+067C | ت | U+06AA | ك | U+06D8 | | U+0756 | ت | | |
| U+064F | | U+067D | ث | U+06AB | ك | U+06D9 | | U+0757 | خ | | |
| U+0650 | | U+067E | ب | U+06AC | ك | U+06DA | | U+0758 | ج | | |
| U+0651 | | U+067F | ت | U+06AD | ك | U+06DB | | U+0759 | ذ | | |

**Figure 3.1:** The unicode entries of the extended Arabic characters and their Arabic equivalents. If Arabic equivalent is blank, then characters should be removed

- Some letters or pairs of letters (e.g. لَالَالَالَا (lAlAlAlA - meaning “no”)) are often repeated multiple times.

- Some repeated letters in valid words are routinely omitted. Example: سعودين (sEwdyn) is shortened from سعوديّين (sEwdyyn - meaning “Saudis”).

Darwish et al. (56) proposed a method for compressing words with repeated characters and then recovering the most likely form based on unigram language models that are trained on a clean news corpus and on a tweet corpus.

## 3.3 Handling Morphology

Due to the morphological complexity of the Arabic language, some morphological processing would help recover the units of meaning or their proxies, such as stems (or perhaps roots). Most early Arabic morphological analyzers generally used finite state transducers (17; 16; 95). Their use is problematic for two reasons. First, they were designed to produce as many analyses as possible without indicating which analysis is most likely. This property of the analyzers complicates retrieval, because it introduces ambiguity in the indexing phase as well as the search phase. Second, the use of finite state transducers inherently limits coverage, which is the number of words that the analyzer can analyze, to the cases programmed into the transducers. Other similar approaches attempt to find all possible prefix and suffix combinations in a word and then try to match the remaining stem to a list of possible stems (94; 110). This approach has the same shortcomings as the finite transducer approach. Another approach to morphology is so-called light stemming. In this approach, leading and trailing letters in a word are removed if they match entries in lists of common prefixes and suffixes respectively. The advantage of this approach is that it requires no morphological processing and is hence very efficient. However, incorrect prefixes and suffixes are routinely removed. This approach was used to develop Arabic stemmers by Aljlayl et al. (11), Darwish and Oard (51), and Larkey et al. (100). Two commonly used light stemmers are:

- Al-Stem (51), which is fairly aggressive and removes the following prefixes and suffixes:
  - Prefixes: وَال (wAl), فَال (fAl), بَال (bAl), بت (bt), يت (yt), لت (lt), مت (mt), وت (wt), ست (st), نت (nt), بم (bm), لم (lm), وم (wm), كم (km), فم (fm), ال (Al), لّ (ll), وي (wy), لي (ly), سي (sy), في (fy), وَا (wA), فَا (fA), لَا (lA), و (w), and بَا (bA)
  - Suffixes: ات (At), وَا (wA), ون (wn), وه (wh), ان (An), تي (ty), ته (th), تم (tm), كم (km), هم (hm), هن (hn), هَا (hA), ية (yp), تك (tk), نَا (nA), ين (yn), يه (yh), ة (p), ه (h), ي (y), ا (A).
- Umass light10 stemmer (100), which removes the following prefixes and suffixes:
  - Prefixes: ال (Al), وَال (wAl), بَال (bAl), كَال (kAl), فَال (fAl), and و (w)
  - Suffixes: هَا (hA), ان (An), ات (At), ون (wn), ين (yn), يه (yh), ية (yp), ة (p), ه (h), and ي (y)

More recent analyzers can statistically perform deep word stemming. For example, Darwish attempted to solve this problem by developing a statistical morphological analyzer for Arabic called Sebawai that attempts to rank possible analyses to pick the most likely one (43). Lee et al. (104) developed IBM-LM, which adopted a trigram language model (LM) trained on a portion of the manually segmented LDC Arabic Treebank (109) in developing an Arabic morphology system, which attempts to improve the coverage and linguistic correctness over existing statistical analyzers such as Sebawai (43). IBM-LM's analyzer combined a trigram LM (to analyze a word within its context in the sentence) with a prefix-suffix filter (to eliminate illegal prefix suffix combinations, hence improving correctness) and unsupervised stem acquisition (to improve coverage). Lee et al. report a 2.9% error rate in analysis compared to 7.3% error reported by Darwish for Sebawai (104; 43). Diab (58) used an SVM classifier to ascertain the optimal segmentation for a word in context. The classifier was trained on the Arabic Penn Treebank data. Essentially, Diab treated the problem as a sequence labeling problem and reported a stemming error rate of about 1%. Although consistency is more important for IR applications than linguistic correctness, perhaps improved correctness would naturally yield greater consistency. Another analyzer that could potentially be useful for retrieval is MADA (72). MADA is a morphological tagger for MSA that has is used widely for processing Arabic in the context of machine translation (142).

Follow on work by Darwish et al. (47) attempted to address shortcomings of existing stemmers that merely remove prefixes and suffixes. These shortcomings have to do with: a) words that are typically borrowed from other languages that do not have standard stem forms; and b) broken plurals. They used a generative character model to produce related stems and broken plurals, leading to significant improvements in retrieval effectiveness. Other fine tuning of stemming is likely required because it may change the intent of the query. For example, often masculine adjectives and nouns can be inflected into their feminine counterparts by adding the suffix ة (p). Consider the words سعيد (sEyd) and سعيدة (sEydp), which are the masculine and feminine versions of the adjective "happy". However, adding the same suffix can change the meaning of the word completely as in مكتب (mktb) and مكتبة (mktbp) meaning “office” and “library” respectively. Perhaps using corpus statistics and query-based stemming can be helpful to address such problems (135).

Though there has been some work on morphological analysis of Arabic dialects (80), particularly Egyptian and Levantine, the effect of such analysis is still unknown on retrieval. Most of the dialectal morphological phenomena primarily affect verbs with little effect on nouns, which are typically more important for retrieval. This continues to be an open area of research.

## 3.4 Handling Stopwords

Stopwords (or function words) perform different functions in sentences but are typically not useful for retrieval. They include prepositions, pronouns, and common nouns. Lists of Arabic particles are included with Sebawai (43) and Solr[5] or available online from a variety of sites.[6] One of the problems of Arabic stopwords is that they accept the attachment of prefixes and suffixes. For example, coordinating conjunctions and pronouns can be attached to the preposition من (mn - meaning “from”) leading to surface forms such as ومنه (wmnh - meaning “and from him”). Thus, identifying stopwords may require affix removal (stemming) first. Removing stopwords has been shown to be effective in retrieving news documents (36; 178), however it led to decreased effectiveness in retrieving Arabic microblogs (56). The decreased effectiveness was probably due to the effect of stopword removal on document length normalization, where microblogs are typically short and removing a few words from them has a major impact on their length.

## 3.5 Handling Spelling and Lexical Choice Variations

Due to regional and dialectal variations in pronunciation, authors of informal text (e.g. tweets) may choose to spell words in a way that

[5] http://lucene.apache.org/solr/

[6] e.g. http://www.ranks.nl/stopwords/arabic.html http://sourceforge.net/projects/arabicstopwords/

matches their pronunciation. As in the example given earlier, the word صدق (Sdq - meaning "truth" or "honestly") is often written as صج (Sj) to match the Gulf dialect. In many of these cases, the words have a proper MSA spelling. This problem is not yet addressed in the literature. It is likely that a spelling-correction-like approach would be required to overcome this problem.

Another related problem has to do with the varying lexical choices in different dialects. For example, the word of car in Egyptian and Tunisian is عربية (Erbyp) and كرهبة (krhbp) respectively while it is سيَارة (syArp) in most other dialects. There is recently published work from Columbia University in which they developed a 50 thousand concept dictionary with the equivalents in the different dialects which they use in turn to expand queries (136). Another potential approach is to translate dialectal text to MSA (142; 144; 152). Shatnawi et al. (152) reported some improvement using this approach.

## 3.6 Best Index Terms

### Using Morphology

Due to the morphological complexity of the Arabic language, much research has focused on the effect of morphology on Arabic IR. The goal of morphology in IR is to conflate words of similar or related meanings. Several early studies suggested that indexing Arabic text using roots significantly increases retrieval effectiveness over the use of words or stems (7; 13; 88). However, all these studies used small test collections of only hundreds of documents and the morphology in many of the studies was done manually.

A study by Aljlayl et al. (11) on a large Arabic collection of 383,872 documents suggested that lightly stemmed words, where only common prefixes and suffixes are stripped from them, were perhaps better index term for Arabic. Similar studies by Darwish and Oard (50) and Larkey et al. (100) also suggested that light stemming is indeed superior to morphological analysis in the context of IR. Darwish compared light stemming to using Sebawai (43) and Larkey et al. (100) compared to using the Buckwalter morphological analyzer (31). The reported shortcomings of morphology might be attributed to issues of coverage and correctness. Concerning coverage, analyzers typically fail to analyze Arabized or transliterated words, which may have prefixes and suffixes attached to them and are typically valuable in IR. As for correctness, the presence (or absence) of a prefix or suffix may significantly alter the analysis of a word. For example, the word الكسير (Alksyr) is unambiguously analyzed to the root كسر (ksr) and stem كسير (ksyr). However, removing the prefix ال (Al) introduces an additional analysis, namely to the root سير (syr) and the stem سير (syr). Perhaps such ambiguity can be reduced by using the context in which the word is mentioned. For example, for the word كسير (ksyr) in the sentence سَار كسير(sAr ksyr - meaning "he walked like"), the letter ك (k) is likely to be a prefix. The problem of coverage is practically eliminated by light stemming. However, light stemming yields greater consistency without regard to correctness.

However, a later study by Darwish et al. (49) suggested that using IBM-LM (104) statistically significantly improved retrieval effectiveness over using light stemming and other morphological analyzers. This is most likely due to the broad coverage of IBM-LM and the ability to rank the most likely analysis. Other work by Darwish and Ali (47) suggests that using AMIRA (58) and generating "similar" stems and broken plurals further improves retrieval effectiveness beyond other approaches due to the lower stemming error rate and broader coverage.

### Using Character N-grams

The use of character trigrams and 4-grams has been shown to be very effective in Arabic IR (50; 121). The estimated average length of an Arabic stem is about 3.6 characters. Darwish and Oard (50) showed that character n-grams in fact outperform the use of light stemming. Character n-grams are perhaps effective because:

- They consistently match stems of words
- They are not bound by a preset vocabulary like morphological analysis
- N-grams that include prefixes and suffixes appear more often than n-grams that include stems, and hence the use of inverse document frequency (IDF) would automatically demote the weight of n-grams that have prefixes and suffixes and promote the weight of n-grams that include stems.

The use of character n-grams should be done in conjunction with kashida and diacritic removal and performing letter normalization. The major downside of character n-grams is the increased processing of text and increased storage space requirements. For example, a 6 letter word is replaced with 4 tokens when character trigrams are used.

## 3.7 Retrieval Models

Though different retrieval models and ranking formulae have been used for Arabic IR, a thorough comparison of their effect on Arabic retrieval is not available. Some of the models that were used for Arabic IR include: statistical language modeling (124) and probabilistic models as implemented in InQuery (51; 99) or using Okapi BM25 similarity (149; 56). There is no indication that any of these models is inherently better suited for Arabic IR or that any special term weighting is necessary.

# 4

# Arabic IR in Shared-Task Evaluations

This section introduces different aspects of evaluation campaigns including their purpose and how evaluation sets are constructed. It also explores the presence of the Arabic language in different IR evaluation campaigns, such TREC, TDT, BOLT, and CLEF, and presents the different IR tasks and the associated test collections. The tasks in evaluation campaigns cover ad hoc retrieval, filtering, cross-language retrieval, topic detection and tracking, and question answering.

## 4.1 Evaluation Campaigns

### Evolution of Evaluation Campaigns in IR

Research in IR before the 1990's was relatively limited and immature compared to the period after that. This stemmed from the fact that only limited resources and data collections were available for the experimentation and evaluation of new methods and algorithms in IR. In the early 1990's, Donna Harman led the first IR evaluation campaign, which was called the Text REtrieval Conference (TREC) (140). Since then, IR research has developed significantly and rapidly. After TREC, several other IR evaluation campaigns, such as CLEF, INEX, NTCIR, and FIRE, were established to assist the improvement of IR systems for different languages and IR domains. These evaluation campaigns helped in the development of IR research because they worked on filling many of the gaps that IR researchers were facing. These evaluation campaigns and forums led to:

1. Creating standardized IR test collections that model different search tasks, while sometimes utilizing participants in the campaign to create ground truth judgments.

2. Providing standard methods and collections for evaluation, where different systems can be compared and evaluated fairly and effectively.

3. Offering resources and test data collections that are expensive to build (beyond the capability of any individual researcher or research group) for a small fee.

4. Publishing different algorithms by participants to help in the further development of better systems.

5. Organizing regular meetings for researchers who participate (or who are even interested) in the evaluation to meet and present their ideas and share their thoughts for different IR problems and tasks.

Such evaluation forums and campaigns were clearly impactful on the quality of IR research.

### The Design of an IR Test Collection

The design of the laboratory IR tests is usually based on the system-oriented Cranfield evaluation paradigm (40). The paradigm involves conducting IR evaluation in a controlled test environment that includes documents, queries, and relevance judgments. A system that ranks “relevant” documents higher than non-relevant ones is more desirable. This paradigm enables the replication of experiments in an easy manner, which allows for the rapid testing of different methods for IR with a fixed test set. Many current IR applications can adapt the paradigm of using a fixed test set for evaluating various IR approaches. However, some IR applications cannot be evaluated through this laboratory experimental setup due to their dependence on users for example (such as interactive IR) or to the dynamic nature of content (such as real-time web search). The design of laboratory IR test consists of three main parts:

1. Collection of Documents. Documents can be of varying types such as text (e.g., news articles, scientific publications, or web pages) or multimedia (e.g., images, videos, or maps). The collection size should preferably match the real world IR application that needs testing. For example, news test collections typically contain at least a few hundred thousand documents. Web search collections would have hundreds of millions or billions of documents.

2. Topics. The topics should represent typical user information needs and are often expressed in the form of queries. For assessment of relevance, some description of what constitutes relevance is often also available.

3. Relevance Assessments. These are the links between the topics and relevant documents. Relevance assessment provides the relevance information that is required for subsequent experimentation.

These three parts make up an IR test collection, which can be used to test several systems for their performance in retrieving the relevant documents in response to the set of topics.

### Building Test Collections

The cost of producing relevance judgments for a large collection is very high and dominates the cost of developing test collections (156). The document collection can be constructed from documents that are representative of those for the IR application. Sometimes, out-of-copyright collections or freely available documents are used. Examples of freely available and out-of-copyright resources are Wikipedia and old books in the public domain respectively. Alternatively, the acquisition of documents may require agreements with copyright holders. For example, many collections are constructed from recent news articles. Topics are usually constructed by typical users or domain experts to match real-life information needs.

The process of creating relevance assessment is usually laborious and costly. There are three main methods of developing relevance judgments:

1. The first is pooling, which involves manually assessing the relevance of the union of the top N documents from multiple retrieval systems for every topic (174). For example, developing the relevance judgments for the 2002 Text REtrieval Conference (TREC) cross-language track involved assessing up to 4,100 documents for each of the 50 topics (129).

2. The second method is a manual user-guided search in which a relevance assessor manually searches and assesses documents for a topic until the assessor thinks that most relevant documents have been found (176).

3. The third is exhaustively examining the documents for relevant ones (170).

There are methods that are reported in the literature to lower the cost of building standardized test collections. For example, Soboroff and Robertson (157) suggested a systematic approach involving: a) pooling the results of a limited number of systems; b) judging the top N documents; and c) using the relevant documents to expand the queries and to search again. This method can be repeated multiple times. A variation of their method involved using only one retrieval system (147). Carterette and Allen (34) proposed a new algorithm that is based on paired comparison of systems that was able to perform high rank correlation with a very small set of judgments. In essence, documents that contribute the most to measured difference between ranked lists are given priority in judging.

Many of the aforementioned methods often miss relevant documents, and relevance assessments are necessarily subjective. However, studies have suggested that relevance judgments can be reliably used to correctly differentiate between retrieval systems provided that a sufficient number of queries is used (30; 145; 174). Voorhees estimated the number of sufficient queries to be about 25 (174). Sanderson and Joho (147) concluded that using less than 25 topics is insufficient to evaluate the relative effectiveness of IR systems even when using statistical significance tests to compare the ranked-lists they produce.

In an effort to overcome the problem of incomplete judgements, there has been a push to introduce new effectiveness measures, such as *bpref* (29), that are tolerant of inherent incompleteness. Yilmaz et al. (179) introduced a method based on stratified sampling where documents in a pool are split into strata and sampled in such a way that optimizes judgment effort and better estimates effectiveness measures. They introduced so-called extended inferred average precision and inferred DCG. Similar stratified sampling has been used to estimate recall over document collections (130).

Test collections need to match the task at hand as much as possible in several aspects. Some of these aspects include:

1. Collection size: Collection size affects retrieval behavior, requiring a change in retrieval strategies. For example, performing deep morphology on Arabic to produce roots was shown to be the most effective strategy for searching small Arabic collections of several hundred documents (13). Later work on a large collection with hundreds of thousands of documents showed that using light stemming performed best (11). There are indications that even using light stemming may in fact hurt retrieval effectiveness at web-scale where the document collection is in the range of hundreds of millions of documents. In web search, users generally inspect the first several documents in a ranked list, which makes the task highly precision oriented.

2. Collection genre: Different genres exhibit different attributes that affect retrieval. For example, in searching news articles, document length normalization is typically important. However, when searching tweets, document length normalization is less important and often harmful (56).

3. Collection modality: Different collection modalities include text, images, videos, audio, document images, etc. An example of the effect of modality on retrieval has to do with document length normalization. In the cosine similarity equation:

$$similarity_{cosine} = cos(\theta) = \frac{\sum_{i=i}^{n} A_i B_i}{\sqrt{\sum_{i=i}^{n} A_i^2}\sqrt{\sum_{i=i}^{n} B_i^2}}$$

   document length takes into account document frequency. Hence, documents with rare words (with low DF) would have larger normalization factors. When searching OCRed text, misrecognized words may have low DF causing an artificial inflation of document length. Singhal et al. (154) solved the document length estimation problem using document byte length normalization. The Okapi BM25 similarly formula uses simple word count to perform normalization (138).

4. Document and collection structure: The structure of a document collection can complicate or ease retrieval. For example, news articles are typically disjoint with very few structural elements (headline, summary, article text, picture captions) that can be used to affect document ranking. Web documents on the other hand exhibit many structural features that can be used to enhance ranking. The most notable of these features is the existence of interlinks between pages (134).

5. Query formulation: It is important to construct queries that closely mimic queries that users actually issue. For example, web users often issue queries that contain spelling mistakes (8). Thus, constructing queries without spelling mistakes would hide real life phenomena. Commercial web search engines such as Google (www.google.com) and Bing (www.bing.com) are typically evaluated using actual queries from query logs (106). Observing spelling mistakes in queries led all major commercial web search engines to include query spelling correction in the form of automatic correction or with suggested correction as in “did you mean:”.

For more on IR evaluation, Sanderson (148) presents a thorough review of the subject.

## 4.2 Arabic in IR Evaluation Campaigns and Shared-Tasks

Evaluation tracks or shared-tasks in campaigns that pertain to Arabic are few compared to other languages, such as European and Indian languages. This is due to the absence of IR evaluation campaigns that focus on Arabic specifically. Other languages have dedicated campaigns, such as TREC for different IR tasks that are mostly in English, CLEF that focuses on the European languages, and FIRE that focuses on the Indian languages. Nonetheless, some of the tasks in these campaigns have included Arabic as one of the investigated languages, such as the TREC CLIR tasks in 2001/2002 and the CLEF INFILE tasks in 2008/2009.

Aside from the Arabic tasks in these campaigns, there were other initiatives from Arab institutes and researchers to create Arabic IR test collections that can be used for evaluating novel Arabic IR techniques in various IR domains. In this section, the different IR tasks that were investigated for Arabic are described.

There has been some initiatives and studies researching approaches for Arabic and to improve retrieval effectiveness for various Arabic IR tasks. Some of these initiatives were part of international evaluation campaign such as TREC and CLEF. Others were conducted by research groups to tackle unstudied domains in Arabic IR.

### TREC 2001/2002 Cross-Language IR Track

Most recent studies on Arabic retrieval have been based on this collection (68; 129). For brevity, the collection is henceforth referred to as the TREC collection. The collection contains 383,872 articles from the Agence France Press (AFP) Arabic newswire. Twenty five topics were developed for the 2001 evaluation and an additional fifty topics were developed for 2002. Relevance judgments were developed at the LDC by manually judging a pool of documents obtained from combining the top 100 documents from all the runs submitted by the participating teams in the TREC cross-language track. The 2001 topics and their relevance judgments are suspect due to the large number of relevant documents being contributed to the pool by only 1 of the participating teams, and the large drop in mean average precision on the rest of the runs when that run is removed (68). For the 2002 topics, the number of known relevant documents ranged from 10 to 523, with an average of 118 relevant documents per topic (129). This is presently the best available large Arabic IR test collection for the news domain. The TREC topics include a title field that briefly names the topic, a description field that usually consists of a single sentence description, and a narrative field that is intended to contain any information that would be needed by a human judge to accurately assess the relevance of a document (68).

### TDT Task in TREC

**TDT Collection:** The Topic Detection and Tracking (TDT) evaluation was designed to evaluate several IR tasks. TDT later became part of the DARPA Translingual Information Detection, Extraction, and Summarization (TIDES) program. The TDT document collections were constructed from radio, television, broadcast news, and newswire articles. When using audio documents, two transcripts of the audio were provided based on Automatic Speech Recognition (ASR) and manual transcription. A topic in TDT is a collection of events, where an event has an associated place, time, prerequisites, and consequences. TDT included five tasks: story link detection, clustering collection by topics (topic detection), topic tracking (information filtering), new events detection, and story segmentation. Topics for the topic-tracking task were defined by a small number of relevant documents, ranging from one to four. The task was to apply information filtering to find relevant stories to those topics in an incoming stream of documents (102). The collection contained documents in English, Chinese, and Arabic. Machine-translations into English were provided for all non-English stories to facilitate the multilingual setup for the task. The collections are as follows:

- TDT3: The collection contained 15,928 Arabic documents which are a subset of the TREC 2001/2002 Arabic collection with an associated 26 topics, that were translated from English and had matching documents in the collection.[1]

- TDT4: The collection contained 42,713 Arabic documents that included both newswire and broadcast news stories from AFP, Al-Hayat, An-Nahar, Voice of America, and Nile TV. These documents include those for TDT3. For the topics, 20 Arabic topics were added to the TDT3 topics.

- TDT5: The collection is available from LDC[2] and contains 72,905 Arabic documents from AFP, An-Nahar, Ummah Press, and Xinhua News Agency. Associated with the collection are 104 Arabic topics, of which 62 are monolingual and 42 are multilingual (available in English or both English and Chinese).[3]

### TRECvid

The TREC Video Retrieval Evaluation (TRECVid) is a shared-task dedicated to video retrieval. Some of the subtasks associated with TRECVid include interactive and non-interactive retrieval, either manual or automatic, shot and story boundary detection, and video semantic information detection (97; 155). For the 2005 and 2006 editions of the evaluation, Arabic and Chinese videos were introduced. The Arabic videos totaled 82.6 hours from the Lebanese Broadcasting Corporation and Alhurra TV. All the videos were automatically transcribed using ASR and were translated using automatic machine translation into English. The information needs of the retrieval task included finding specific people, locations, events, activities, items, or a combination of these. An example topic is: “Find shots of one or more people reading a newspaper”. An information need was expressed in text with pos-

[1] http://www.itl.nist.gov/iad/mig/tests/tdt/2000/

[2] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2005T16

[3] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T19

sible exemplars in the form images, video shots, or audio clips. The TRECVid data is available from LDC.

### INFILE Track in CLEF 2008/2009

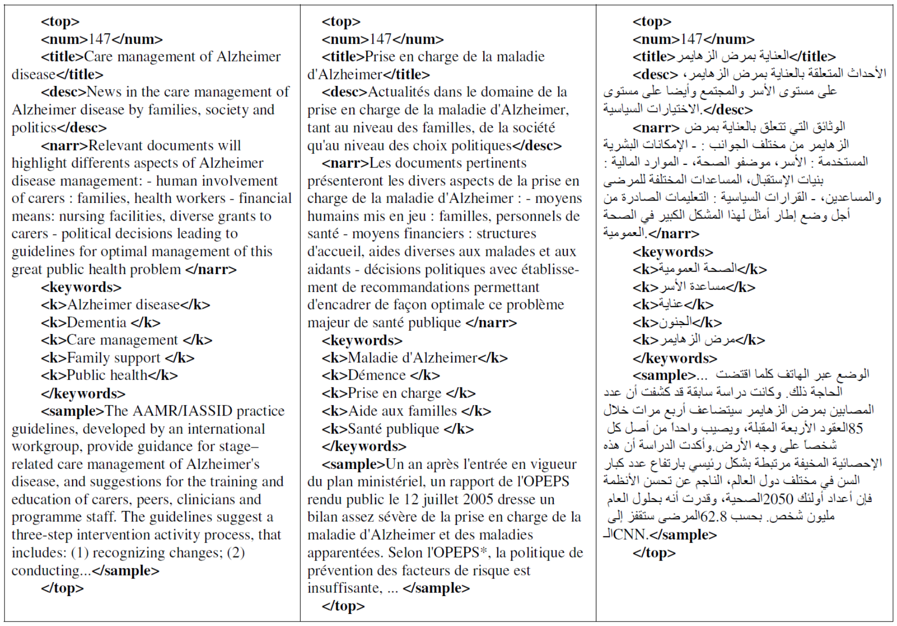
The INFILE (INformation FILtering Evaluation) track was part of the CLEF campaign for the year: 2008 and 2009 (20; 21). The main objective of that track was to evaluate the cross-language adaptive filtering systems by measuring the ability of automated systems to filter out irrelevant documents from an incoming stream of documents with respect to a given topic. The task was designed for applying information filtering in three languages: English, French, and Arabic.

The document collection contained 300 thousand articles from the AFP newswire that were published in the period between 2004 and 2006. This is different than the TREC 2001/2002 collection mentioned earlier. The document collection is composed of 100 thousand articles from each of the three languages. Two sets of topics totaling 50 topics related to the articles in that period were prepared. One set relating to news and events had 30 topics; and the other set relating to scientific and technological subjects had 20 topics. Figure 4.1 shows an example topic in the three languages. The track organizers prepared relevance judgments by searching the collection with different combinations of the topic fields using four different search engines. This generated a pool of retrieval results from 28 different runs, which were manually assessed for relevance.

Unfortunately, since no participation in the Arabic task was received in either of the two years in which the track ran (20; 21). Nonetheless, the task data collection is available for usage in Arabic IR experimentation.

### QA4MRE Track in CLEF 2012/2013

Arabic was included as one of the seven languages to be evaluated in the Question Answering for Machine Reading Evaluation (QA4MRE) task starting from 2012 (22). The task involves answering a multiple-choice question (MCQ) on documents concerned with a set of four specific topics, namely, climate change, music and society, Alzheimer's



| | | |
|---|---|---|
| <top><br><num>147</num><br><title>Care management of Alzheimer disease</title><br><desc>News in the care management of Alzheimer disease by families, society and politics</desc><br><narr>Relevant documents will highlight differents aspects of Alzheimer disease management: - human involvement of carers : families, health workers - financial means: nursing facilities, diverse grants to carers - political decisions leading to guidelines for optimal management of this great public health problem </narr><br><keywords><br><k>Alzheimer disease</k><br><k>Dementia </k><br><k>Care management </k><br><k>Family support </k><br><k>Public health</k><br></keywords><br><sample>The AAMR/IASSID practice guidelines, developed by an international workgroup, provide guidance for stage–related care management of Alzheimer's disease, and suggestions for the training and education of carers, peers, clinicians and programme staff. The guidelines suggest a three-step intervention activity process, that includes: (1) recognizing changes; (2) conducting...</sample><br></top> | <top><br><num>147</num><br><title>Prise en charge de la maladie d'Alzheimer</title><br><desc>Actualités dans le domaine de la prise en charge de la maladie d'Alzheimer, tant au niveau des familles, de la société qu'au niveau des choix politiques</desc><br><narr>Les documents pertinents présenteront les divers aspects de la prise en charge de la maladie d'Alzheimer : - moyens humains mis en jeu : familles, personnels de santé - moyens financiers : structures d'accueil, aides diverses aux malades et aux aidants - décisions politiques avec établissement de recommandations permettant d'encadrer de façon optimale ce problème majeur de santé publique </narr><br><keywords><br><k>Maladie d'Alzheimer</k><br><k>Démence </k><br><k>Prise en charge </k><br><k>Aide aux familles </k><br><k>Santé publique </k><br></keywords><br><sample>Un an après l'entrée en vigueur du plan ministériel, un rapport de l'OPEPS rendu public le 12 juillet 2005 dresse un bilan assez sévère de la prise en charge de la maladie d'Alzheimer et des maladies apparentées. Selon l'OPEPS*, la politique de prévention des facteurs de risque est insuffisante, ... </sample><br></top> | <top><br><num>147</num><br><title>العناية بمرض الزهايمر</title><br><desc> الأحداث المتعلقة بالعناية بمرض الزهايمر، على مستوى الأسر والمجتمع وأيضا على مستوى الاختيارات السياسية.</desc><br><narr> الوثائق التي تتعلق بالعناية بمرض الزهايمر من مختلف الجوانب : - الإمكانات البشرية المستخدمة : الأسر، موضفو الصحة، - الموارد المالية : بنيات الإستقبال، المساعدات المختلفة للمرضى والمساعدين، - القرارات السياسية : التعليمات الصادرة من أجل وضع إطار أمثل لهذا المشكل الكبير في الصحة العمومية.</narr><br><keywords><br><k>الصحة العمومية</k><br><k>مساعدة الأسر</k><br><k>عناية</k><br><k>الجنون</k><br><k>مرض الزهايمر</k><br></keywords><br><sample>... الوضع عبر الهاتف كلما اقتضت الحاجة ذلك. وكانت دراسة سابقة قد كشفت أن عدد المصابين بمرض الزهايمر سيتضاعف أربع مرات خلال العقود الأربعة المقبلة، ويصيب واحداً من أصل كل 85 شخصاً على وجه الأرض.وأكدت الدراسة أن هذه الإحصائية المخيفة مرتبطة بشكل رئيسي بارتفاع عدد كبار السن في مختلف دول العالم، الناجم عن تحسن الأنظمة الصحية، وقدرت أنه بحلول العام 2050 فإن أعداد أولئك المرضى ستقفز إلى 62.8 مليون شخص. بحسب الـCNN.</sample><br></top> |

**Figure 4.1:** An example topic from INFILE track 2008/2008 in three language: English, French, and Arabic (21)

disease, and AIDS. The user is provided with a question and a set of five answers, with only one of them being correct. The objective is to apply IR/NLP techniques to parse the question and retrieve the correct answer from the document set.

The Arabic document collection in the track included a set of 19,300, 120,600, 10,200, 15,700 documents related to the four topics Alzheimer, AIDS, climate change, and music and society respectively. A set of 40 questions was prepared for each topic, leading to a test set of 160 questions. Some of these questions required using NLP methods for getting the answers. These questions required inference resolution for some named entities or required inference from multiple sentences and paragraphs to get the answer. In 2012, two groups submitted runs for the Arabic task. The average scores of the Arabic runs were the lowest among all runs in the seven languages that were evaluated in the track (22).

## GALE

The Global Autonomous Language Exploitation (GALE) Program is a DARPA program that aims to transcribe, translate, analyze, and distill textual and speech data in Arabic and Chinese for decision makers (132). Much of the focus of the program related to improving machine translation and coupling it with automatic transcription (for speech translation). The program produced significant training corpora for translation and transcription that cover both MSA and dialectal data and produced new tools such as MADA for dialects (80). So-called distillation was applied to transcribed and translated material. Distillation involved identifying relevant information, removing redundancy while maintaing proper citation, and providing a "functional" presentation of such information. Some of the operational engines involving speech search and question answering are presented in Section 5.5 and Section 5.6 respectively. Olive et al. (132) offer a thorough description of the work done the GALE program.

## Arabic IR in BOLT

In 2011, DARPA launched the Broad Operational Language Translation (BOLT) program in an attempt to create new techniques for automated translation and linguistic analysis that can be applied to informal text and speech that are common online and in-person communication[4].

There is particular focus on enabling access of non-English resources through:

- Allowing English-speakers to understand foreign-language sources of all genres, including chat, messaging and informal conversation.
- Providing English-speakers the ability to quickly identify targeted information in foreign-language sources using natural-language queries.

[4] `http://www.darpa.mil/Our_Work/I2O/Programs/Broad_Operational_Language_Translation_%28BOLT%29.aspx`

- Enabling multi-turn communication in text and speech with non-English speakers.

BOLT has gone through two phases thus far. Phases 1, held in 2012, targeted performing complex question answering tasks on discussion forum threads in English, Arabic, and Chinese. The task entailed identifying so-called "nuggets", which are text snippets, that support different facets (aspects) of a topic. The evaluations involved identifying responses in threads in a specific language or in different languages. The Arabic collection in phase 1 consisted of 43,000 Arabic forum threads with more than 154 topics (both test and dry run topics) of which 26 were intended to retrieve Arabic documents.[5] Arabic forum threads were mostly dialectal Egyptian dialect.

Phase 2, held in 2013, introduced an additional task, beside identifying nuggets, in which participants were asked to cluster retrieved results topically, temporally, or geographically.[6] The topics in phase 2 (150 topics of which 22 were intended to retrieve Arabic documents) specified the desired language of the returned documents. Since BOLT is not a public evaluation, information about the program and the results of the participants is relatively scant. Future phases of BOLT are expected to include evaluation on Arabic SMS data some of which has Arabizi also.

[5] http://www.nist.gov/itl/iad/mig/upload/bolt-ir-guidelines-v5-0-April-15-2012.pdf

[6] www.nist.gov/itl/iad/mig/upload/BOLT_P2_IR-guidelines-v1-3.pdf

# 5

# Domain-specific IR

In the following, we explore IR applications for different applications and domains. These include cross-language IR, document image retrieval, social search, general web search, questions answering, image retrieval, and speech search. The section addresses some of the challenges associated with different applications and some of the solutions that are reported in the literature.

## 5.1 Arabic-English CLIR

Cross-Language IR (CLIR) is the process of finding documents in one language based on queries in a different language (128). One of the central issues in CLIR pertains to the translation of the query to the language of the documents or the translation of the documents to the language of the query. Due to the fact that queries are typically short and documents are typically large, query translation is the more commonly used approach in experimental settings. However, there are indications that doing both query and document translation yields better results (122). Query translation has been explored extensively in the context of CLIR, where a query is supplied in a source language to retrieve results in a target language. Two of the most popular query translation approaches are Dictionary Based Translation (DBT) methods (105) and Machine Translation (MT) (177). DBT methods usually involve replacing each of the source language words with equivalent target language word(s). Since a source language word may have multiple translations, optionally the most popular translation or $n$ best translations are used. Since web search engines typically use an AND operator by default, using multiple translations may cause translated queries to return no results. Another alternative is to use a synonym operator, which has been shown to be effective in CLIR (105; 137). A synonym operator can be approximated in web search by using an OR operator between different translations. Some online search engines have synonym operators, such as the "word" operator in Bing. However, the use of a weighted synonym operator, where each translation is assigned a confidence score, is not supported in popular web search engines, though it has been shown to be effective in cross-language search (175), and it is implemented in some open source search toolkits such as the "#wsyn" operator in Indri (162). MT has been widely used for query translation (172; 177). Wu et al. (177) claim that MT outperforms DBT. Their claim is sensible in the context of web search for two reasons: a) MT attempts to optimally reorder words after translation and web search engines are typically sensitive to word order; and b) MT produces a single translation without any synonym or OR operators, for which the rankers of web search engines are not tuned. There are a few publicly available translation services such those from Google[1] and Microsoft.[2] There is also a fair amount of parallel text that is available from LDC to train a machine translation system (74; 142). More recent work from the GALE program involved using MT to enable search in multilingual broadcast news feeds (132). More discussion of this is in Section 5.5 about Arabic Speech Search.

Another approach of interest here is the so-called cross-lingual query suggestion (65; 67). This approach involves finding related translations for a source language query in a large web query log in the

[1] `http://translate.google.com/`

[2] `http://www.bing.com/translator/`

target language. Gao et al. (65) proposed a cross-language query suggestion framework that uses a discriminative model trained on cross-language query similarity, cross-language word co-occurrence in snippets of search results, co-clicks from both queries to the same URL, and monolingual query suggestion. Recent work by Hefny et al. (87) extends the work of Gao et al. (65) by using alternative features, such as phonetic similarity, relative query lengths, and cross-language coverage.

The second issue involves merging multilingual results, where results from multiple languages that the user may speak are combined into a single ranked list. For results merging in general, there are several simple techniques in the literature such as score-based merging, round-robin merging, and normalized score based merging (107). Score-based merging assumes that scores in different ranked lists are comparable, which cannot be guaranteed. This can be addressed to some degree by normalizing scores from different ranked lists. Round robin merging assumes that different ranked lists have a comparable number of relevant documents (107). Si and Callan (153) used a logistic regression based model to combine results from different multilingual ranked lists in the context of ad hoc search. Tsai et al. (169) also addressed the problem in the context of ad hoc search. They used document, query, and translation based word-level features to train an FRank-based ranker whose score was then linearly combined with the BM25 score of each document. Rankers of web search engines typically consider many more features such as link-graph, document structure, and query log-based features. Gao et al. (66) used a Boltzman machine to learn a merging model that takes cross-language relations between retrieved documents into account. They tested their approach on ad hoc as well as web search. In the context of web search, they evaluated their approach using queries that have equivalents in the query log of the target language, which means that these queries would likely benefit from cross-language results merging. Hefny et al. (87) employed a supervised learning rank model, namely SVMRank, to merge multiple ranked lists into a single list. SVMRank was trained using the relevance judgments for query-result pairs which were used to extract pairwise order constraints.

The third central issue is ascertaining when cross-language web search would yield good results. The literature is relatively sparse on this issue. Kishida (96) examined "ease of search" and translation quality as means for cross-language query performance prediction. Lee et al. (103) examined the issue of when to translate query words or not. Hefny et al. (87) proposed the use of query logs, through so-called query picking, to determine if cross-language search would be effective or not. Essentially, given a source language query, if an equivalent query is found in a large query log, then the query would likely produce good results.

Most of the work on Arabic-English CLIR was conducted as part of the TREC 2001 and 2002 CLIR track (68; 129). The track experiments were conducted on the aforementioned collection 383,872 news articles from the AFP Arabic newswire with 75 topics and relevance judgments, and it focused exclusively on searching an Arabic corpus using English queries. Studies on this collection focused on a variety of issues such as:

- Handling named entities. Named entities are often central to queries and it is estimated that more than two thirds of queries contain named entities (71). Despite their importance and prevalence, named entities are less likely to be covered by dictionaries and translation resources than regular words. One way to improve the coverage is to transliterate named entities, specially those with no translations. Doing so has been shown to improve CLIR effectiveness (2; 101). Another approach involves performing transliteration mining between the source query and the top retrieved results in the target language to attempt to identify transliterations of untranslated words, which are typically named entities (172). Translation resources can be enriched using automatically mined named entity transliterations from parallel text (62) or comparable text (61; 173).

- Stemming. Due to the morphological and orthographic complexity of Arabic, stemming can greatly improve the coverage of translation resources. Stemming has been shown to positively impact

Arabic-English CLIR (12; 36) and other translation related tasks such as machine translation (74; 126). If machine translation is employed in CLIR, improving machine translation quality using stemming would likely lead to improved CLIR effectiveness.

- Combining different translation resources. Combining multiple translation/transliteration resources improves translation coverage and improves CLIR effectiveness (51; 87; 101).

- Combining multiple translations in DBT. Since words are often ambiguous, with multiple valid translations, combining different translation may improve retrieval effectiveness. The translations can be combined using a balanced translation approach. Using multiple translations can be done by replacing each query word using the top $n$ translations or using structured query methods, which have been shown to outperform balanced translation (44; 175).

- Interactive retrieval. In interactive retrieval, a user can provide feedback to the retrieval system. For example, a user can guide query translation or may indicate the relevance of some of the results (86).

- Blind relevance feedback. Blind relevance feedback has been shown to improve retrieval effectiveness when applied before and after translation, with pre-translation expansion being more effective, particularly when translation resources are limited (123).

Hefny et al. in (87) focused on searching the English web using Arabic queries. Their work addressed other problems related to combing multilingual results and cross-language query performance prediction.

## 5.2 Arabic Document Image Retrieval

There has been major efforts aimed at digitizing large volumes of historical manuscripts, either in print or handwritten. One such effort for Arabic manuscripts was carried out at the Bibliotheca Alexandrina in the context of the Million Book Project (60). The most notable method for searching digitized manuscripts involves recognizing the text in them. For handwritten manuscripts, there has been recent efforts for recognizing handwriting such as Darpa's MADCAT program (161). Saleem et al. (143) reported recognition error rates of approximately 30%, while Cao et al. (33) reported error rates in excess of 45%. Some work has been done on retrieving offline Arabic handwritten document including historical manuscripts (35; 37). In online handwriting recognition, the system has access to the strokes of the user, while in offline recognition, the system does not and only the final written forms are available. Retrieval of such documents focused mainly on identifying lines or regions containing specific words (35; 37).

Much more work was done on retrieving OCRed printed documents. Although OCR is fast, OCR output typically contains errors. The errors are even more pronounced in OCRed Arabic text due to Arabic's orthographic and morphological properties. The introduced errors adversely affect linguistic processing and retrieval of OCRed documents, which leads to degradation in the retrieval performance. Some techniques have been shown to improve Arabic document image retrieval. In this section we describe these techniques.

### Arabic OCR

The goal of OCR is to transform a document image into character-coded text. The usual process is to automatically segment a document image into character images in the proper reading order using image analysis heuristics, apply an automatic classifier to determine the character codes that most likely correspond to each character image, and then exploit sequential context (e.g., preceding and following characters and a list of possible words) to select the most likely character in each position. The character error rate can be influenced by reproduction quality (e.g., original documents are typically better than photocopies), the resolution at which a document was scanned, and any mismatch between the instances on which the character image classifier was trained and the rendering of the characters in the document image. Arabic

OCR presents several challenges, including:

- Arabic's cursive script in which most characters are connected and their shapes vary with position in the word. Further, multiple connected characters may resemble other single characters or combinations of characters. For example, the letter ش ($) may resemble نت (nt).

- The optional use of word elongations and ligatures, which are special forms of certain letter sequences.

- The presence of dots in 15 of the 28 characters to distinguish between different letters, and the optional use of diacritic which can be confused with dirt, dust, and speckle (50). The orthographic features of Arabic lead to some characters being more prone to OCR errors than others.

- The morphological complexity of Arabic, which results in an estimated 60 billion possible surface forms, complicates dictionary-based error correction.

There are a number of commercial Arabic OCR systems including:

- Sakhr Automatic Reader (92; 93).

- Novodynamics NovoVerus[3]

- ABBYY FineReader[4]

- IRIS ReadIRIS[5]

[3] `http://www.novodynamics.com/`
[4] `http://ocrsdk.com`
[5] `http://www.irislink.com`

- Tesseract OCR[6]

Most Arabic OCR systems segment characters (69; 83; 84; 93), while a few opted to recognize words without segmenting characters (14; 108). A system developed by BBN avoids character segmentation by dividing lines into slender vertical frames (and frames into cells) and uses an Hidden Markov Model (HMM) recognizer to recognize character sequences (108).

### OCR Degraded Text Retrieval

Much work has been reported on different approaches for the retrieval of OCR degraded text documents for many languages, including English (82; 91; 164; 165); Chinese (170); and Arabic (50). For Arabic, most work has been done on the ZAD collection (50). The ZAD collection was built from Zad Al-Mead, a medieval book that is free of copyright restrictions and for which a free electronic copy is available. The book, written in the 14th century by a Muslim theologian, consists of 2,730 separate documents that address a variety of topics such as mannerisms, history, jurisprudence and medicine. A native speaker of Arabic (50) developed 25 topics and exhaustively searched the collection for relevant documents. The number of relevant documents per topic ranges from zero (for one topic) to 72, averaging 18. The average query length is 5.5 words. This collection is composed to document images (from a printed copy of the book) along with Optical Character Recognition (OCR) output and a typeset version of the documents.

Generally, the approaches used for degraded text retrieval are considered language independent, where the same approach can be used across different languages but with special configuration for each. It has been reported that OCRed text with recognition errors of more than 5% Character Error Rate (CER) leads to a statistically significant drop in retrieval effectiveness (when compared to clean text i.e. text with no errors) (59; 85). This finding motivated the introduction of the confusion track in TREC-5 (91). In this track, a set of approximately 50,000 English documents from the Federal Register were degraded by ap-

[6] `https://code.google.com/p/tesseract-ocr/`

plying random edit operations to random characters in the documents to simulate OCR degraded text. Experiments showed that retrieval effectiveness is adversely affected by the increase in degradation and decrease in redundancy of search terms in the documents (59; 91).

One of the introduced solutions for overcoming errors in OCR text was the use of character n-grams for indexing and searching the document collections. This approach proved to be successful in improving the retrieval effectiveness for different languages, such as English (82), Chinese (170), and Arabic (43). Simply, the character n-gram representation for a term involves splitting the word into n-characters; for example, the 3-gram representation of the term "retrieval" is {#re, ret, etr, tri, rie, iev, eva, val, al#}. This method gives a higher chance for a term with misrecognized characters to match the correctly spelled term provided in the user's query.

Another common approach for improving OCR degraded text retrieval is to apply text correction in an attempt to correct errors in OCRed text. Reducing the number of errors in the text may lead to improved IR effectiveness. There are two main approaches to error correction, namely, word level and passage level. Some of the kinds of word-level post-processing include the use of dictionary lookup (28), language modeling (89) frequency analysis, and morphological analysis (131). Passage-level post-processing techniques include the use of word n-grams (113), word collocations (89), grammar (10), conceptual closeness (89), passage-level word clustering (166) (which requires handling of affixes for Arabic (57)), and linguistic and visual context (89).

Dictionary lookup is used to compare recognized words with words in a lexicon (28; 39; 89; 90). Finding the closest matches to every OCRed word in the dictionary is attempted, and the matches are then ordered using a character-level error model in conjunction with either a unigram probability of the matches in the text (90) or an n-gram language model (113; 167).

In the next part, the work reported on Arabic document image retrieval is described in detail including the approaches, data collections used, and reported results.

### Approaches for Arabic Document Image Retrieval

There are several approaches that have been shown to improve retrieval of Arabic document images. Most of these approaches are geared towards overcoming errors that are introduced due to OCR. Some of these approaches are:

**Searching using character n-grams:** Using character trigrams and 4-grams have been shown to significantly improve retrieval effectiveness over using words or stems (50). Character n-grams are relatively robust as long as the character error rate is low enough to yield many correct sequences of contiguous characters. Another advantage of character n-grams is that they overcome the need for morphological analysis, which is expected to be adversely affected by errors in the text.

**Query garbling:** In query garbling, degraded forms of query words are generated to match the degradation in the documents. Garbling can be performed at either word or stem level (52) or at character n-gram level (44). In this approach, an error model is used to generate garbled versions of query words (53). Query garbling is akin to translating queries to the document space, and much of the work on CLIR would apply. The translated (or garbled) version of the query words, or their character n-grams, can be used in multiple ways. In one approach, multiple garbled versions replace (or augment) the original version (52). This would be akin to using balanced query translation. In another approach, the possible garbled forms of the word, or constituent n-grams, are combined using structured queries where they are treated as synonyms of equal weight or as weighted synonyms (44). In the absence of a character degradation model, another method involves identifying OCRed words in the document set that share common character n-grams with the query word and using all of them as synonyms (82).

**OCR Error Correction:** This method attempts to correct misrecognized OCRed words. The correction is often done using a noisy channel model to learn how OCR corrupts single characters or character segments, producing a character level confusion model, and language model to determine the prior probability of the candidate corrections. The alignment can be performed using different methods such as weighted edit distance (113) or automatic alignment (as in machine translation alignment). For language modeling, a unigram language model, a dictionary (90), or an n-gram language model (113; 167) is used. Correction was shown to improve retrieval effectiveness (114; 118) and combining correction with query garbling yields even further improvement (53).

In the absence of training data, a simple edit distance based model can be used to as a stand-in for the confusion model (115).

**Fusing the output of multiple OCR systems:** Another approach suggested by Magdy et al. (117) involves the use of multi-OCR output fusion. In this approach multiple OCR systems, which typically have different classification engines with different training data, are used to recognize the same text. The output of the different OCR systems is then fused by picking the most likely recognized sequence of tokens using language modeling (117). This approach was shown to yield improved error correction and consequently better retrieval effectiveness.

**OCR-less retrieval:** Aside from the methods that rely on OCR, an OCR-less document image retrieval avoids using OCR (119). The basic idea is that similar connected characters in document images are clustered and represented with IDs. The IDs are indexed using an IR engine. Then given a query word, the connected characters in the word are rendered as an image. The resultant image is used to find the most similar cluster(s) to it. Then the cluster ID(s) are used in place of the connected character segments in the query. Structured queries can then be used to combine multiple cluster IDs. Though this method was shown to achieve only 61% relative effectiveness compared to OCR retrieval (119), it does not require OCR, which is a major advantage.

## 5.3 Arabic Web Search

| Feature | Google | Bing |
|---|---|---|
| Diacritic removal | $\surd$ | $\surd$ |
| Kashida removal | $\surd$ | $\surd$ |
| Letter normalization | Partial | $\surd$ |
| Light stemming | X | X |
| stopword removal | X | X |

**Table 5.1:** Arabic processing in Google and Bing

Aside from the work of Hefny et al. (87), to best of our knowledge, there is no publicly published work on Arabic web search. Stopwords can be used as search queries to ascertain the relative number of indexed Arabic pages in a web search engine. Based on this, Google and Bing are believed to index roughly 3 billion and 210 million Arabic pages respectively. There are several challenges that need to be addressed for Arabic search, namely:

1. Efficiency: Due to the large scale of the web, all processing must be very efficient. Hence, performing complex processing such as morphological analysis becomes prohibitive.

2. Language handling: Due to efficiency related constraints, minimalist language processing is done. Table 5.1 summarizes current support in Google and Bing. This information was obtained by searching using different queries that were designed to test the specific features in the engines. Both search engines emulate stemming using query alteration where some sort of expanding query words using their synonyms or morphological variants is used.[7]

3. Ranking: There are fundamental differences between the English and Arabic webs that can affect static ranking features that relate

[7] This is based on communication with people working on Arabic web search at Google and Bing

to the interconnection between pages such as PageRank. As mentioned earlier, the Arabic web is sparsely connected, with much forum content, and is much smaller than the English web.

4. Market size: Due to the significantly smaller Arabic market as manifested by the number of users and commercial viability, particularly in terms of online advertisement revenue, Arabic index coverage and search facilities typically garner less priority compared to large markets, such as the US and UK markets, from major search engines.

Hefny et al. (87) reported on Arabic-English cross-language web retrieval results using a set of 200 cross-language queries that were run against Bing. However, their collection is not public.

## 5.4 Arabic Social Search

Social media has been instrumental in facilitating the launch of the so-called "Arab Spring". Since then, the penetration of social media has been steadily increasing. The number of Facebook users in the Arab countries is estimated to be 42.4 million, representing 14.8% of the population. This number has increased by 10% between September 2011 and February 2012.[8] Based on interaction with people at Twitter, the estimated number of Arabic microblogs on Twitter is in excess of 15 million per day. Arabic social media exhibits the dialectal phenomena described in Section 2.3. Microblog retrieval has attracted some interest in recent years. TREC introduced a Microblog track focused on English microblog retrieval (133). The track provided a collection of about 14 million tweets with a set of 50 topics and their relevance judgments. A recent preliminary paper by Darwish et al. (56) attempted to address some of the issues associated with Arabic microblog retrieval, but many more issues are yet to be tackled. In their work they devised an expanded letter normalization scheme, modified tokenization to handle microblog specific tokens (e.g. smilies, user mentions, and hashtags), corrected word elongations and contractions, and introduced a new

[8] www.internetworldstats.com

stopword list. Their work does not handle dialect specific stemming though. Pasha et al. (136) constructed a thesaurus of equivalent words across dialects to expand queries to enable improve retrieval of dialectal text. This could potentially be helpful in Arabic microblog search where dialects are commonly used.

Darwish et al. (56) also performed a study on Arabic microblog retrieval using a large collection of Arabic microblogs containing 112 million tweets. The tweets were scraped from Twitter between Nov. 20, 2011 and Jan. 9, 2012 using the query "lang:ar". Associated with the collection are 35 topics, with each topic having a title query and a relevant exemplar. Binary relevance judgments were constructed by manually judging all top 30 results from several runs. Duplicate or near duplicate results were eliminated, and judgments were propagated from judged tweets to all duplicate or near duplicate tweets. Roughly 566 judgements were made per topic per query on average, with an average of 267 relevant tweets per topic. They are making the collection publicly available for research purposes in the form of tweets IDs, queries, and relevance judgements.

Magdy et al. (112) presented a retrieval system for aggregating microblogs that match a user's query to construct a multi-faceted presentation that includes tag-clouds of top terms in them, query terms time-series, most popular microblogs, most shared videos and images, and jokes. An instance of the system was used to build a filtering system that tracks user topics in microblog streams (111).

## 5.5 Arabic Speech Search

Speech search involves two main tasks: searching in audio content; and searching using spoken queries. Both tasks typically involve the use of ASR. In this section, we briefly describe some of the issues associated with Arabic ASR and then we describe some of the efforts associated with searching ASR output.

Performing ASR involves using three key elements, namely: an acoustic model, a dictionary, and a language model. As mentioned in Section 2.5, morphology, orthography, optional short vowels, and dialects can complicate ASR. Most ASR work has focused on handling MSA speech, typically in the form of broadcast news and conversational programs. Broadcast news is typically strictly MSA, while conversational programs may include some dialectal speech. Much work on Arabic ASR was performed under the GALE program and is documented in Section 3.6 of (132).

To handle diacritics (or vocalization), there are two common approaches. The first involves using diacritics in building the acoustic models and then identifying the correct diacritized forms of words (using an automatic diacritizer or a morphological analyzer). Most GALE participants used this approach (132). Variations of this included performing back-off to undiacritized models or using joint models (98; 120; 158). There is some indication that undiacritized acoustic models performed better than diacritized ones (25; 120), but a combined model would perform better (120). Pronunciation dictionaries, which contain phonetic transcriptions of words, are typically manually crafted for languages such as English. Manually crafting such dictionaries for morphologically rich languages such as Arabic is prohibitive due to the large number of possible surface forms. However, due to the regularity of Arabic phonetics, each undiacritized word form can be diacritized automatically and then a phonetic transcription of the different diacritized forms can be generated (24).

To handle dialects, the use of dialectal models, either specifically trained or adapted, is required. Further, dialects have been shown to be sufficiently different, and using training data for one dialect (or MSA) to recognize another dialect typically leads to high recognition word error rate (24; 159). Due to the difference between dialects, the automatic identification of a speaker's dialect to invoke the most appropriate ASR models would lead to lower recognition error rates (24; 25). For dialectal training data, Biadsy et al. (25) collected nearly 240 hours of training data for each of five different dialects (Egyptian, Lebanese, Jordanian, Saudi, and Emirates). In (24), Biadsy used available training data from LDC and Appen that cover Levantine (which covers both Lebanese and Jordanian), Egyptian, Gulf (which covers Saudi and Emirates), and Iraqi. He also used an automated method to identify Levantine speech within the GALE data. For language modeling, he trained a language identifier to extract dialectal text from a large text corpus (24), while in later work by Biadsy et al. (25), on building a system for searching using audio queries, they used a large Google query log from different countries to build their language models.

As previously mentioned, speech search focuses on searching audio content or using audio queries. The first formal evaluation involving retrieval of Arabic audio content was in the Topic Detection and Tracking (TDT4) evaluation where some of the documents included automatically and manually transcribed broadcast news stories. In later work in the context of the GALE program, IBM collaborated with several universities to develop the Rosetta distillation multilingual system which integrates ASR, search, machine translation, summarization, information extraction, and question answering. The system was used to process and index video feeds in multiple languages (Arabic, Chinese, English, Spanish) (132). The system would allow an analyst to issue keyword based queries against the acquired feeds. Another system was developed by BBN and dubbed the Broadcast Monitoring System (132). Their system would also continuously monitor a video channel, converts its audio to text, and then index it for later retrieval. Other commercial systems are available from different vendors such as SAIC.[9]

As for spoken queries, Google has been working on expanding on their Voice Search[10] system for Arabic (25).

## 5.6 Question Answering

There has been limited work on Arabic question answering. In the context of the GALE program, the aforementioned Rosetta system allowed the use of so-called template mode, in which a user would populate arguments in a template to perform queries (or question answering) along 15 different facets. An interactive search (or information gathering) evaluation was performed, and it was found that the template-

[9] https://www.saic.com/linguistics/media.html

[10] http://www.google.com/mobile/voice-search/

based system was more effective than the standard search mode. As for the QA4MRE track in CLEF-2012, it included Arabic as one of the languages for the track and there were two participating systems for Arabic. The task involved answering multiple choice questions by identifying supporting evidence for the correct answer in a document collection. For Arabic, there were two participating system. The first system was IDRAAQ (6), which had a question classifier and a passage retrieval module. For passage retrieval they employed query expansion, synonym expansion (using WordNet), and word n-gram similarity. The second system, by Trigui et al. (168), used a question classifier, a passage retrieval module to identify passages that may contain the correct answer, and then an alignment module that attempted to align the passages to the answer choices. It is notable that both systems achieved the lowest scores among all the participating systems across all languages in QA4MRE.

There was work on Arabic question answering prior to QA4MRE. QARAB used a question analyzer that identified question types using the question word (who, when, etc.), a retrieval module that searches using word roots for documents that may contain the answer, and a POS tagger along with a named entity recognizer to help identify correct answers in the documents (81). Benajiba and Rosso (18) introduced the ArabicQA system which includes a passage retrieval system, a named entity recognizer, and an answer extractor that identifies the type of answer from the question word and proper target from the named entity recognizer. In prior work by Abouenour et al., they focused on expansion using WordNet (4) and passage retrieval (5).

As mentioned in Section 4.2, the QA evaluation part of the BOLT program focused on identifying so-called “nuggets”, which are text snippets, that support different topic facets in forum threads in different languages that include Arabic.

There has been a fair amount of work on these named entity recognition (19; 54; 141) and named entity linking (116; 150) all of which can be important for question answering.

## 5.7 Image Retrieval

A popular approach for image retrieval involves retrieval on textual information that is associated with images such as captions, user tags, web page context, or metadata files (127; 171).

However, Arabic content is limited on the web and most images on the web do not have Arabic textual information associated with them. Luckily, the contents of the images are generally language independent. There is some indication that using cross-language search to retrieve images may yield much better search results. To illustrate this, Figures 5.1 and 5.2 show the the results of search results for “الاَبَامَا ” (AlAbAmA - Alabama in Arabic) and “Alabama” on Google image search. Clearly search in English yields much better results. There was an initiative in CLEF 2004 cross-language image retrieval task, where English topics were provided with manual translations in 12 languages, including Arabic. Though no runs were submitted for the Arabic topics (41), Clough et al. (42) submitted Arabic runs to the iCLEF 2006 (70). In their work, they provided an interface to search Flickr images with tags in multiple languages using Arabic queries (42). However, the work on images retrieval using Arabic queries continues to be scant. Many research questions linger including:

- What are the effective text-based methods for image retrieval of Arabic-tagged images? And how do they compare to methods used in other languages?

- What are the effective CLIR methods for retrieving Arabic-tagged images? How can monolingual and cross-language results be merged?

- What are the best morphological processing on image-captions or on textual contexts of images that can lead to high retrieval effectiveness for images with Arabic context?

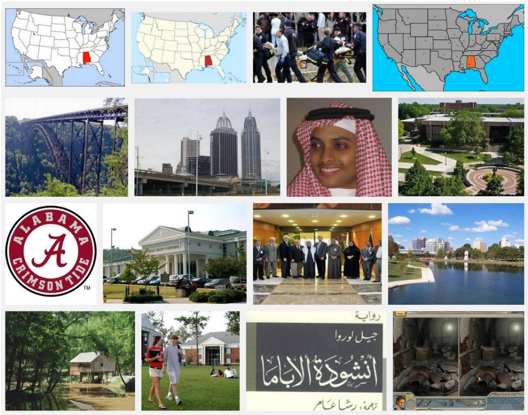


**Figure 5.1:** Searching for images of الأَبَامَا (AlAbAmA - Alabama in Arabic) on Google

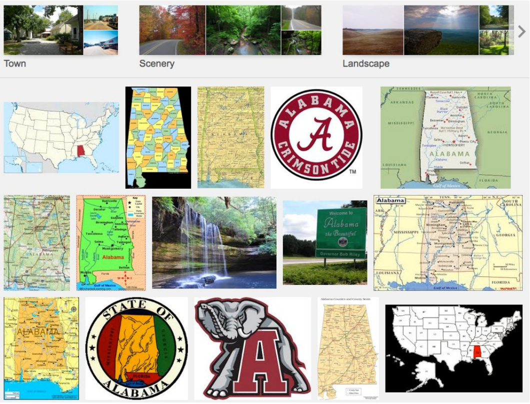


**Figure 5.2:** Searching for images of Alabama (in English) on Google

# 6

# Open Research Areas in Arabic IR

Though much research has been conducted thus far, Arabic IR research continues to lag behind research conducted for several other languages. Further, some of the peculiarities of Arabic would impact IR for different domains, many of which are unexplored or not thoroughly investigated.

In this section, we highlight some of the open areas for Arabic IR. The nature of each task is mentioned and the challenges are listed. Potential research investigations for these tasks are discussed and possible solutions are explored.

## 6.1 Ad Hoc IR

Although the ad hoc retrieval task is the most studied Arabic IR task, the amount of research reported for this task is considerably limited compared to what is done in other less inflected languages such as English. As was shown earlier in Section 4, there is only one standard large collection for this task, namely the TREC 2001/2002 CLIR collection, and it covers a single genre (news) from a single news source (AFP). Another publicly available collection is the ZAD collection, which contains religious documents.

Many other domains still require investigation. Some of these domains include:

- **Religious texts.** Religious texts cover a broad spectrum of content and present interesting kinds of issues. Some of the content types include:
  *Prophetic traditions:* A prophetic tradition has two parts, namely the text of the tradition and the chain of narration, which indicates how and who transmitted the tradition. Though text of the tradition is unstructured, some structure can be inferred from the chains of narration by properly identifying the narrators in the chain. Further, the chain of narration can be linked to extensive biographies of the individuals, which can be used to ascertain their truthfulness and soundness of memory to help determine the authenticity of the tradition. Recent work by Shatnawi et al. (151) constructed a set of 17,000 Prophetic traditions for automatic verification of authenticity.
  *Commentary:* This includes commentaries and explanations of either verses, prophetic traditions, and proverbs. The interesting part of the text is that there is typically a short piece of text with an ensuing explanation of variable length. Some of the search scenarios may include: finding an explanations of a piece of text; recommending a verse or a tradition that matches a particular meaning of interest; or linking between multiple items that address a particular topic.
  *Jurisprudence:* Such texts present rulings and often discussions of particular issues of jurisprudence. They typically involve a statement of issue, a ruling or rulings, an explanation, and supporting evidences.
  The previous mentioned Zad collection is composed of religious texts, but it only covers one book and it is very small (2,730 documents).

- **Classical texts:** Such texts cover more than 14 centuries of Arabic literature and may include literary works such as: short stories, prose, poetry, biographies, history, and social sciences. Due to their age, they may differ from modern texts in style, organization, and lexical choices.

- **Wikipedia:** Wikipedia pages contain semi-structured content that includes titles, synopses, so-called info-boxes, categories, cross-language links, internal and external links, references, and pictures. Arabic Wikipedia currently contains 252k pages.[1] Creating an IR test collection based on Arabic Wikipedia pages can be fairly straightforward (55). There was some work on cross document entity matching in the context of the Automatic Content Extraction (ACE) evaluation (160). Performing such matching, aided by Wikipedia links, can enrich Wikipedia-based search tasks.

- **Online forums:** As mentioned earlier, Arabic forum content is disproportionately larger compared to the forum content in other languages and often contains high quality content. Handling forum content presents challenges related to content. As mentioned earlier, the BOLT IR tasks for 2012 and 2013 involved identifying nuggets in forums.

- **News search:** Though the largest available standard test collection is based on news articles, the collection is constructed from a single news source. Creating a collection that covers different news sources from different countries is required. There are some Arabic news aggregator on the web that can be useful for collecting news articles for this task. Some of the aggregators are Google News[2], Moheet[3], Johaina[4], and Maghress[5].

To properly develop algorithms and methods to retrieve such content, standard test sets and clear usage scenarios are required. Such test sets are currently not available.

[1] As of December 2013

[2] `https://news.google.com/`

[3] `http://moheet.com/`

[4] `http://johaina.sakhr.com/`

[5] `http://www.maghress.com/`

## 6.2 Question Answering

There are two related tasks that have not been addressed in the literature. Both these tasks involve finding existing answers to a user's question.

The first involves finding answers in online forums or specialized community question answering forums such as Google Ejabat (meaning "answers" in Arabic), which is akin to Yahoo! Answers. The task has been explored for Yahoo! Answers with work on identifying existing answers (163) and finding the most authoritative contributors (27). We are not aware of any such work on Arabic question answering communities. Identifying questions and their answers in generic online forums is a broader problem and perhaps more challenging.

The second involves finding answers to religious questions particularly as they pertain to Islamic jurisprudence. Such answers are generally referred to as "Fatwas". Due to the growth of the Internet in the last decade, many websites have been established as a hub for Muslims to ask their questions and get answers regarding their religion from experts. Some of these sites such as IslamWeb[6] contain tens of thousands of fatwas. Though most are in Arabic, there are many fatwa repositories in other languages, such as English, French, German, Dutch, Italian, Hindi, and many others. Figure 6.1 shows two sample fatwas. Like community question answering sites, user questions have most likely been asked previously by others. However, questions are typically situational where a person describes a problem, an interaction, or an experience. Thus, abstracting the underlying question, and hence similar questions, is difficult. The situation is more challenging when searching for similar answered questions in different languages. This specific QA task for Arabic is of large interest by many Muslims around the world.[7]

In both tasks, these problems can be listed as research issues as follows:

1. Detecting and abstracting a question.

[6] http://www.islamweb.net

[7] Muslims population is over 1.6 billion over the world: http://www.pewforum.org/The-Future-of-the-Global-Muslim-Population.aspx

2. Determining if a question is new or has been answered before.
3. Ranking possible questions and answers.
4. Classifying answers on different aspects such as authoritativeness, supporting evidence, ... etc.
5. Summarizing answers to highlight answer
6. Identifying complementary questions and answers
7. Searching across languages and identifying cross-language and perhaps cross cultural issues.

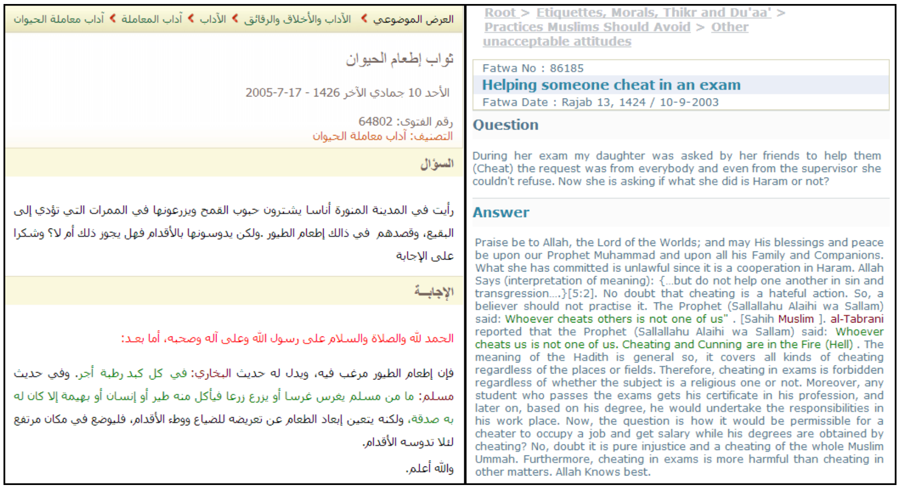
Obtaining a data collection for evaluating fatwa QA task in Arabic and other languages could be done by scraping fatwa repositories such as IslamWeb, which provides answers to hundreds of questions per day and contains more than 100 thousand Arabic, 20 thousand English, 4 thousand French, and 1 thousand German questions.[8] This is a very valuable dataset that could be utilized and used for creating test sets for different tasks that can cover the aforementioned research questions. Some effort would be required to prepare a test set for the evaluation.

## 6.3 Social Search

Despite the work done on Arabic social search that was shown in Section 5, there are many open problems that need to be addressed to improve the effectiveness of Arabic social search.

An essential processing step for effective Arabic social search is a dialectal Arabic stemmer. Currently, most available Arabic stemmers are designed for MSA. An effective stemmer for dialectal Arabic is currently missing. There are challenges for stemming dialectal Arabic due to the varying linguistic features of different dialects and the lack of spelling standards of the dialects. There has been some recent effort pertaining to establishing standardized spelling of dialects (79) and for developing NLP tools including morphological analyzers and parsers for dialects (38).

[8] `http://www.islamweb.net`



العرض الموضوعي ‹ الآداب والأخلاق والرقائق ‹ الآداب ‹ آداب المعاملة ‹ آداب معاملة الحيوان

ثواب إطعام الحيوان

الأحد 10 جمادي الآخر 1426 - 2005-7-17

رقم الفتوى: 64802
التصنيف: آداب معاملة الحيوان

السؤال

رأيت في المدينة المنورة أناسا يشترون حبوب القمح ويزرعونها في الممرات التي تؤدي إلى البقيع، وقصدهم  في ذالك إطعام الطيور .ولكن يدوسونها بالأقدام فهل يجوز ذلك أم لا؟ وشكرا على الإجابة

الإجابـــة

الحمد لله والصلاة والسلام على رسول الله وعلى آله وصحبه، أما بعـد:

فإن إطعام الطيور مرغب فيه، ويدل له حديث البخاري: في كل كبد رطبة أجر. وفي حديث مسلم: ما من مسلم يغرس غرسا أو يزرع زرعا فيأكل منه طير أو إنسان أو بهيمة إلا كان له به صدقة، ولكنه يتعين إبعاد الطعام عن تعريضه للضياع ووطء الأقدام، فليوضع في مكان مرتفع لئلا تدوسه الأقدام.

والله أعلم.

Root > Etiquettes, Morals, Thikr and Du'aa' > Practices Muslims Should Avoid > Other unacceptable attitudes

Fatwa No : 86185

Helping someone cheat in an exam

Fatwa Date : Rajab 13, 1424 / 10-9-2003

Question

During her exam my daughter was asked by her friends to help them (Cheat) the request was from everybody and even from the supervisor she couldn't refuse. Now she is asking if what she did is Haram or not?

Answer

Praise be to Allah, the Lord of the Worlds; and may His blessings and peace be upon our Prophet Muhammad and upon all his Family and Companions. What she has committed is unlawful since it is a cooperation in Haram. Allah Says (interpretation of meaning): {...but do not help one another in sin and transgression....}[5:2]. No doubt that cheating is a hateful action. So, a believer should not practise it. The Prophet (Sallallahu Alaihi wa Sallam) said: Whoever cheats others is not one of us" . [Sahih Muslim ]. al-Tabrani reported that the Prophet (Sallallahu Alaihi wa Sallam) said: Whoever cheats us is not one of us. Cheating and Cunning are in the Fire (Hell) . The meaning of the Hadith is general so, it covers all kinds of cheating regardless of the places or fields. Therefore, cheating in exams is forbidden regardless of whether the subject is a religious one or not. Moreover, any student who passes the exams gets his certificate in his profession, and later on, based on his degree, he would undertake the responsibilities in his work place. Now, the question is how it would be permissible for a cheater to occupy a job and get salary while his degrees are obtained by cheating? No, doubt it is pure injustice and a cheating of the whole Muslim Ummah. Furthermore, cheating in exams is more harmful than cheating in other matters. Allah Knows best.

**Figure 6.1:** Examples of Fatwa QA in Arabic and English.

In addition, cross-language social search for Arabic and different languages can be of interest in many applications, especially given that a large percentage of current Arabic Internet users are at least bilingual. Methods for overcoming the language barrier between social Arabic text (such as microblogs and status updates on social networking sites) and other languages can thus be of interest.

## 6.4 Web Search

Arabic web search has mostly been confined to the commercial domain. Currently, the two biggest providers of Arabic web search are Google and Bing. There were several attempts to develop specialized Arabic search engines such as Araby, Ayna, and onkosh, but they have mostly gone out of business. There remain many research issues that need to be addressed for Arabic web search including:

- *Crawling:* Google and Bing engines may specify a quota for a particular language, but would typically use language independent algorithms to crawl pages in all languages. Given the difference in topology between the Arabic and English webs, proper crawling of Arabic pages needs to identify, prioritize, and scrape the pages of interest to users. A crawler may include subcomponents that are tuned for Arabic to perform: page prioritization, page cleaning, and meta feature extraction.

- *Indexing:* Some language-dependent aspects of indexing include: tokenization, stemming, page segmentation, and page filtering (e.g. adult and spam filtering).
- *Search interface:* Some language-dependent parts of the search interface include: spell checking, query suggestions, query expansion, results interleaving from different document types, and layout design.

To date, the largest publicly available large crawl of the Arabic web is a collection of 29.2 million Arabic pages that are part of the ClueWeb09 collection (32). However, the collection does not have any associated Arabic topics or relevance judgements on Arabic documents.

# 7

# Conclusions

In this survey, we reviewed Arabic IR including the nature of the Arabic language, the techniques used for pre-processing the language, the latest research in Arabic IR in different domains, and the open areas in Arabic IR. Arabic language is ranked as the seventh largest language on the Internet. However, it has been the fastest growing language in the last decade in terms of users. Given the current growth rate of Internet penetration among the Arabic speaking population, Arabic users should have the fourth largest user population on the Internet by 2020. This gives a special importance to the language and emphasizes the need for effective IR approaches for enabling effective search of Arabic documents.

In section 2, entitled "Arabic Features Affecting Retrieval", some of the peculiarities of Arabic were described, including morphology, orthography, phonetics, formal and informal language usage, encodings, and writing schemes. It was explained how the morphology of Arabic is highly inflected, where the same word in Arabic may differ in spelling according to the position in the sentence. The possible Arabic words are estimated to be 60 billion words that are derived from approximately 10,000 roots. Furthermore, the language processing becomes even more challenging when considering the language used in social networking and microblogging sites, where dialects are heavily used. Arabic dialects differ from Modern Standard Arabic that is typically used in formal communications such as news articles. Further, dialects differ from one region to the next. Moreover, Arabizi is sometimes used in Arabic social media, where Arabic words are typed using Latin characters. This adds more challenges for searching and retrieving documents written in this manner.

Section 3, entitled “Arabic Preprocessing and Indexing”, presented the core Arabic preprocessing steps. Pre-processing goes beyond case-folding, stemming, and stopword removal, which are often applied for English documents. Arabic preprocessing includes handling different encodings, orthography, morphology, lexical and spelling variations, and stopwords. For example, diacritics and khashidas are removed; normalization is applied to conflate different sets of letters together, such as different forms of the letter “Alef”; statistical stemming is preferred; and stopwords lists vary from MSA to dialectal Arabic. The section also explored the effective index terms for Arabic, which are stems and characters n-grams.

Section 4, entitled “Arabic IR in Shared-Tasks”, introduced evaluation tasks that contain Arabic in different IR evaluation campaigns such as TREC, TDT, and CLEF. These evaluation campaigns covered a variety of Arabic IR applications including ad hoc retrieval, filtering, cross-language retrieval, topic detection and tracking, and question answering. Nonetheless, the number of tasks, and participation in these evaluation tasks, are low compared to other languages including European, Asian, and Indian languages. This suggests that there is a need for promoting Arabic IR including community building and creating Arabic IR datasets and search tasks in the evaluation campaigns.

Additional work in different domain-specific IR applications for Arabic were reported in Section 5, entitled “Domain-specific IR”. This included cross-language IR, document image retrieval, image search, speech search, social search, and web search. Some of these IR tasks were either not sufficiently covered in evaluation campaigns or not addressed at all. For example, document image retrieval for Arabic poses interesting problems. The number of Arabic documents that do not exist in electronic form is large, and the OCR process for Arabic is less accurate compared to other languages. This creates a challenge for retrieving these kinds of documents. Different approaches for document image retrieval in Arabic were described while showing the advantages and disadvantages of each. Also, Arabic social search is not yet covered by public evaluation campaigns.

Finally, Section 6, entitled “Open Research Areas in Arabic IR”, showcased some of the open areas of research in Arabic IR that have received limited attention. These areas potentially need to be explored to create effective retrieval systems. The proposed Arabic IR tasks were described and illustrated with some examples. Also, the potential test collections for such tasks were suggested. The Arabic IR field can gain considerably if these tasks and others receive additional attention from researchers.

A list of some of the available Arabic resources is showcased in Appendix A. Some of these resources can be used in Arabic IR research and Arabic NLP in general. The listed resources include Arabic test collections, stemmers, stopword lists, WordNets and other resources.

In conclusion, ad hoc Arabic search in the news domain is fairly well explored. However, substantial work is required: in different domains such as social search and web search; for different genres such as religious texts and web forums; and for different Arabic dialects. There is a clear lack of standardized test sets and tools for most of these tasks and genres. Significant effort is required to build such collections and tools. There is also a lack of dedicated evaluation campaigns for Arabic IR.

# Appendices

# A

# Arabic IR Resources

This section focuses on listing and providing links to Arabic resources that can be useful for IR such as test collections, stemmers, index tools, and translation tools.

## A.1 Test Collections

**For ad hoc Arabic and cross-language news retrieval :**
The LDC Arabic Newswire Part 1 collection (LDC2001T55), which contains 383,872 from the AFP newswire:

1. The document collection:
   `http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2001T55`

2. The topics and relevance judgments:

   (a) TREC 2002 cross-language topics in Arabic:
   `http://trec.nist.gov/data/topics_noneng/CL.topics.arabic.trec11.txt`

   (b) TREC 2002 cross-language topics in English:

`http://trec.nist.gov/data/topics_noneng/CL.topics.english.trec11.txt`

- (c) TREC 2001 cross-language topics in Arabic:
  `http://trec.nist.gov/data/topics_noneng/arabic_topics.txt`
- (d) TREC 2001 cross-language topics in English:
  `http://trec.nist.gov/data/topics_noneng/english_topics.txt`
- (e) TREC 2001 cross-language topics in French:
  `http://trec.nist.gov/data/topics_noneng/french_topics.txt`

Relevance judgements:

- (a) TREC 2002 qrels:
  `http://trec.nist.gov/data/qrels_noneng/qrels.trec11.xlingual.txt`
- (b) TREC 2001 qrels:
  `http://trec.nist.gov/data/qrels_noneng/xlingual_t10qrels.txt`

Topic Detection and Tracking (TDT):

1. The document collection:
   - (a) TDT3: A subset of the The LDC Arabic Newswire Part 1 collection (LDC2001T55)
   - (b) TDT4 and annotations:
     `http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2005T16`
   - (c) TDT5:
     `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T19`

2. The topics and relevance judgments:

   (a) TDT5 topics and annotations: `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T19`

**Filtering:**

1. INFILE multilingual filtering: `http://www.trebleclef.eu/infile.php`

**Question answering:**

1. QA4MRE: `http://celct.fbk.eu/QA4MRE/index.php`

**OCR degraded test collection:**

1. The Zad collection (50)

**Social Search:**

1. Collection of 112 million tweets with 35 topics and qrels (56)

**Web Search:**

1. The ClueWeb09 web crawl collection contains 29.2 million Arabic webpages, but no associated Arabic topics or relevance judgments on Arabic documents (32).

**Video Search:** The TRECVid video data containing Arabic videos are available from LDC as follows:

1. **LDC2007V01** TRECVID 2005 Keyframes and Transcripts: `http://catalog.ldc.upenn.edu/LDC2007V01` **LDC2010V02** TRECVID 2006 Keyframes: `http://catalog.ldc.upenn.edu/LDC2010V02`

## A.2 Stemming

The following is a subset of the available Arabic stemmers.

1. Shereen Khoja stemmer:
   `http://sourceforge.net/projects/arabicstemmer/`

2. Light10 stemmer:
   implemented in Lemur:
   `http://sourceforge.net/p/lemur/wiki/Parser%20Applications/`
   and Solr:
   `http://wiki.apache.org/solr/LanguageAnalysis`.

3. AMIRA 2.0:
   `http://nlp.ldeo.columbia.edu/amira/`

4. QCRI's Arabic processing library that includes a tokenizer, word segmenter, POS tagger, and NER:
   `http://alt.qcri.org/tools/` (46)

5. Al-Stem (51)

6. Alexander Fraser (while at BBN):
   `http://tides.umiacs.umd.edu/software/stem_aggressive.tar`

7. Buckwalter analyzer:
   `http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2004L02`
   with Java reimplementation:
   `http://sourceforge.net/projects/aramorph/`

8. MADA+TOKAN:
   `http://www1.cs.columbia.edu/~rambow/software-downloads/MADA_Distribution.html` (77)

9. MADA for dialects (80)

10. Sebawai (43)

11. IBM word segmenter (104)

A more comprehensive survey of available Arabic morphological engines is in (15).

## A.3 Stopwords

Some of the stemmers mentioned before contain stopword lists (e.g. Sebawai) and some stemming integrations in search engines like Lemur and Solr contain Arabic stopword lists. Other sources of Arabic stopwords are:

1. Al-mostabaadat:
   `http://sourceforge.net/projects/arabicstopwords/`
2. Anton Balucha stopword list:
   `https://code.google.com/p/stop-words/`
3. Arabic stopword list from UniNE:
   `http://members.unine.ch/jacques.savoy/clef/index.html`
4. Wael Salloum stopword list:
   `https://www.academia.edu/2663620/A_Modern_Standard_Arabic_Closed-Class_Word_List`
5. Dialectal stopword list (56)

It is noteworthy that Arabic stopwords may accept prefixes and suffixes. Thus, the identification of stopwords may require stemming.

## A.4 Arabic WordNet

Arabic WordNet (AWN) (26; 139) can be used in different IR applications as presented earlier. AWN is not as well developed as those for other languages.

1. The Arabic WordNet Project
   `http://www.talp.upc.edu/index.php/technology/resources/multilingual-lexicons-and-machine-translation-resources/multilingual-lexicons/72-awn`

## A.5 Other Resources

CLIR and speech retrieval require translation resources and ASR. Much training data was provided as part of the GALE program. The data is as follows:

1. `http://projects.ldc.upenn.edu/gale/data/kickoff1-contents.html`
2. `http://projects.ldc.upenn.edu/gale/data/DataMatrix.html`
3. `http://projects.ldc.upenn.edu/gale/data/Catalog.html`

There is a translation probability table that was trained on United Nations parallel data by BBN (178).

## A.6 Buckwalter transliteration

| Buckwalter | Arabic | Buckwalter | Arabic | Buckwalter | Arabic |
|---|---|---|---|---|---|
| A | ا | b | ب | t | ت |
| v | ث | j | ج | H | ح |
| x | خ | d | د | * | ذ |
| r | ر | z | ز | s | س |
| $ | ش | S | ص | D | ض |
| T | ط | Z | ظ | E | ع |
| g | غ | f | ف | q | ق |
| k | ك | l | ل | m | م |
| n | ن | h | ه | w | و |
| y | ي | Y | ى | ’ | ء |
| \| | آ | > | أ | < | إ |
| & | ؤ | } | ىء | p | ة |

## A.7 List of Acronyms

| **Acronym** | **Meaning** |
|---|---|
| ACE | Automatic Content Extraction |
| AFP | Agence France Press |
| ASR | Automatic Speech Recognition |
| AWN | Arabic Word Net |
| BOLT | Broad Operational Language Translation |
| CER | Character Error Rate |
| CLEF | Cross-Language Evaluation Forum |
| CLIR | Cross-Language Information Retrieval |
| CRF | Conditional Random Fields |
| DARPA | Defense Advanced Research Projects Agency |
| DBT | Dictionary-based Translation |
| DF | Document Frequency |
| FIRE | Forum for Information Retrieval Evaluation |
| GALE | Global Autonomous Language Exploitation |
| HMM | Hidden Markov Model |
| IDF | Inverse Document Frequency |
| INEX | INitiative for the Evaluation of XML retrieval |
| INFILE | INformation FILtering Evaluation |
| IR | Information Retrieval |
| LDC | Linguistic Data Consortium |
| LM | Language Model |
| MCQ | Multiple Choice Question |
| MSA | Modern Standard Arabic |
| MT | Machine Translation |
| NFKC | Normalization Form Compatibility Composition |
| NLP | Natural Language Processing |
| NTCIR | NII Tesbeds and Community for Information access Research |
| OCR | Optical Character Recognition |
| POS | Part-Of-Speech |
| QA | Question Answering |
| QA4MRE | Question Answering for Machine Reading Evaluation |
| SVM | Support Vector Machines |

| | |
|---|---|
| TDT | Topic Detection and Tracking |
| TIDES | Translingual Information Detection, Extraction, and Summarization |
| TREC | Text REtrieval Conference |

# Bibliography

[1] Abdul-Monem Abdul-Al-Aal. 1987. An-nahw ashamil. Maktabat Annahda Al-Masriya, Cairo, Egypt, 1987.

[2] Nasreen AbdulJaleel, Leah S. Larkey. 2003. Statistical transliteration for English-Arabic cross-language information retrieval. In Proceedings of the 2003 Conference on Information and Knowledge Management (CIKM), New Orleans, Louisiana, USA.

[3] Abdelrahim Abdelsapor, Noha Adly, Kareem Darwish, Ossama Emam, Walid Magdy, and Magdy Nagi. 2006. Building a heterogeneous information retrieval collection of printed Arabic documents. In Proceedings of the 2006 Language Resources and Evaluation Conference.

[4] Lahsen Abouenour, Karim Bouzoubaa, and Paolo Rosso. 2008. Improving Q/A using Arabic Wordnet. In Proceedings The 2008 International Arab Conference on Information Technology (ACIT'2008), Tunisia.

[5] Lahsen Abouenour, Karim Bouzoubaa, and Paolo Rosso. 2009. Three-level approach for passage retrieval in Arabic question/answering systems. In Proceedings Of the 3rd International Conference on Arabic Language Processing CITALA2009, Rabat, Morocco.

[6] Lahsen Abouenour, Karim Bouzoubaa, and Paolo Rosso. 2012. IDRAAQ: new Arabic question answering system based on query expansion and passage retrieval." In Proceedings of the 2012 Conference and Labs of the Evaluation Forum (CLEF) (Online Working Notes/Labs/Workshop).

[7] Hani Abu-Salem, Mahmoud Al-Omari, and Martha Evens. 1999. Stemming methodologies over individual query words for Arabic information retrieval. Journal of the American Society for Information Science and Technology Vol. 50(6): p.524-529.

[8] Farooq Ahmad and Grzegorz Kondrak. 2005. Learning a spelling error model from search query logs. In Proceedings of the 2005 Human Language Technology (HLT).

[9] Mohamed Attia Ahmed. 2000. A large-scale computational processor of the Arabic morphology, and applications. Masters thesis in Faculty of Engineering, Cairo University, Cairo, Egypt.

[10] Eneko Agirre, Koldo Gojenola, Kepa Sarasola, and A. Voutilainen. 1998. Towards a single proposal in spelling correction. In the Proceedings of the 1998 International Conference on Computational Linguistics: The Association for Computational Linguistics (COLING ACL '98). San Francisco, CA, pp. 22-28.

[11] Mohammed Aljlayl, Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David O. Holmes, M. Lee, David A. Grossman, and Ophir Frieder. 2001. IIT at TREC-10. In 2001 Text REtrieval Conference (TREC). Gaithersburg, MD.

[12] Mohammed Aljlayl and Ophir Frieder. 2002. On Arabic search: improving the retrieval effectiveness via a light stemming approach. In Proceedings of 2002 Conference on Information and Knowledge Management (CIKM).

[13] Ibrahim Al-Kharashi and Martha Evens. 1994. Comparing words, stems, and roots as index terms in an Arabic information retrieval system. Journal of the American Society for Information Science and Technology, 45(8): pp. 548-560.

[14] May Allam. 1995. Segmentation versus segmentation-free for recognizing Arabic text. In IS&T/SPIE's Symposium on Electronic Imaging: Science & Technology, pp. 228-235.

[15] Imad Al-Sughaiyer and Ibrahim A. Al-Kharashi. 2004. Arabic morphological analysis techniques: A comprehensive survey. In Journal of the American Society for Information Science and Technology Archive, Vol. 55 Issue 3.

[16] Kenneth Beesley. 1996. Arabic finite-state morphological analysis and generation. In the Proceedings of the 1996 International Conference on Computational Linguistics (COLING).

[17] Kenneth Beesley, Tim Buckwalter, and Stuart Newton. 1989. Two-level finite-state analysis of Arabic morphology. In the Seminar on Bilingual Computing in Arabic and English, Cambridge, England.

[18] Yassine Benajiba and Paolo Rosso. 2007. Arabic question answering. Diploma of advanced studies. Technical University of Valencia, Spain.

[19] Yassine Benajiba, Mona Diab, and Paolo Rosso. 2008. Arabic named entity recognition using optimized feature sets. In the Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 284âĂŞ293.

[20] Romaric Besançon, Stéphane Chaudiron, Djamel Mostefa, Olivier Hamon, Ismaïl Timimi, and Khalid Choukri. 2009. Overview of CLEF 2008 INFILE Pilot Track. Evaluating Systems for Multilingual and Multimodal Information Access, in the 2009 Cross-Language Evaluation Forum (CLEF), pp. 939-946.

[21] Romaric Besançon, Stéphane Chaudiron, Djamel Mostefa, Olivier Hamon, Ismaïl Timimi, and Khalid Choukri. 2010. Information filtering evaluation: Overview of CLEF 2009 INFILE Track. In 2010 Conference on Multilingual and Multimodal Information Access Evaluation (CLEF-2010). Text Retrieval Experiments. Springer Berlin Heidelberg, 342-353.

[22] Pinaki Bhaskar, Partha Pakray, Somnath Banerjee, Samadrita Banerjee, Sivaji Bandyopadhyay, and Alexander F. Gelbukh. 2012. Overview of QA4MRE at CLEF 2012: Question answering for machine reading evaluation. In 2012 Conference and Labs of the Evaluation Forum (CLEF) (Online Working Notes/Labs/Workshop).

[23] Fadi Biadsy, Nizar Habash, and Julia Hirschberg. 2009. Improving the Arabic pronunciation dictionary for phone and word recognition with linguistically-based pronunciation rules. Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL (HLT NAACL '09), pp. 397-405.

[24] Fadi Biadsy. 2011. Automatic dialect and accent recognition and its application to speech recognition. Ph.D. Thesis, Columbia University.

[25] Fadi Biadsy, Pedro J. Moreno, and Martin Jansche. 2012. Google's cross dialect Arabic voice search. In the Proceedings of the 2012 International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[26] William Black, Sabri Elkateb, Horacio Rodriguez, Musa Alkhalifa, Piek Vossen, Adam Pease, and Christiane Fellbaum. 2006. Introducing the Arabic WordNet Project. In Proceedings of the third International WordNet Conference (GWC-06).

[27] Mohamed Bouguessa, Benoît Dumoulin, and Shengrui Wang. 2008. Identifying authoritative actors in question-answering forums: The case of Yahoo! answers. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). pp. 866-874.

[28] Eric Brill and Robert Moore. 2000. An improved error model for noisy channel spelling correction. In the Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL), Hong Kong, pp. 286-293.

[29] Chris Buckley and Ellen M. Voorhees. 2004. Retrieval evaluation with incomplete information. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval.

[30] Robert Burgin. 1992. Variations in relevance judgments and the evaluation of retrieval performance. Information Processing & Management, Vol. 28(5): pp. 619-627.

[31] Tim Buckwalter. 2002. Buckwalter Arabic morphological analyzer version 1.0. Technical Report LDC2002L49, Linguistic Data Consortium.

[32] Jamie Callan, Mark Hoy, Changkuk Yoo, and Le Zhao. 2009. ClueWeb09 data set. `http://lemurproject.org/clueweb09/`.

[33] Huaigu Cao, Rohit Prasad, and Prem Natarajan. 2011. Handwritten and typewritten text identification and recognition using hidden Markov models. In 2011 International Conference on Document Analysis and Recognition (ICDAR), pp. 744-748.

[34] Ben Carterette, James Allan, and Ramesh Sitaraman. 2006. Minimal test collections for retrieval evaluation. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 268-275.

[35] Jim Chan, Celal Ziftci, and David Forsyth. 2006. Searching off-line Arabic documents. In the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 2.

[36] Aitao Chen and Fredric Gey. 2002. Building an Arabic stemmer for information retrieval. In 2002 Text REtrieval Conference (TREC), Gaithersburg, MD.

[37] Jinying Chen, Rohit Prasad, Huaigu Cao, and Premkumar Natarajan. 2013. Detecting OOV names in Arabic handwritten data. In the 12th International Conference on Document Analysis and Recognition.

[38] David Chiang, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. 2006. Parsing Arabic dialects. In Proceedings of the European Chapter of ACL (EACL), Vol. 111, pp. 112.

[39] Kenneth Church and William Gale. 1991. Probability scoring for spelling correction. Statistics and Computing, 1, pp. 93-103.
Christopher Cieri, David Graff, Mark Liberman, Nii Martey and Stephanie Strassel. 2000. Large, multilingual, broadcast news corpora for cooperative research in topic detection and tracking: The TDT-2 and TDT-3 corpus efforts. In Proceedings of the 2000 Language Resources and Evaluation Conference.

[40] Cyril Cleverdon. 1997. The Cranfield tests on index language devices. In: Spärck-Jones, Karen; Willett, Peter (Eds.): Readings in Information Retrieval. pp. 47-59.

[41] Paul Clough, Henning Müller, and Mark Sanderson. 2005. The CLEF 2004 cross-language image retrieval track. Multilingual Information Access for Text, Speech and Images, Cross-Language Evaluation Forum, pp. 597-613.

[42] Paul Clough, Azzah Al-Maskari, and Kareem Darwish. 2007. Providing multilingual access to Flickr for Arabic users. Evaluation of Multilingual and Multi-modal Information Retrieval. Springer Berlin Heidelberg, 205-216.

[43] Kareem Darwish. 2002. Building a shallow morphological analyzer in one day. In Proceedings of the ACL-2002 Workshop on Computational Approaches to Semitic Languages.

[44] Kareem Darwish. 2003. Probabilistic methods for searching OCR-degraded Arabic text. Ph.D. Thesis, Electrical and Computer Engineering Department, University of Maryland, College Park.

[45] Kareem Darwish. 2013. Arabizi detection and conversion to Arabic. CoRR abs/1306.6755

[46] Kareem Darwish and Ahmed Abdelali. 2014. Using Stem-Templates to improve Arabic POS and Gender/Number Tagging. In International Conference on Language Resources and Evaluation (LREC-2014).

[47] Kareem Darwish and Ahmed Ali. 2012. Arabic retrieval revisited: Morphological hole filling. In the Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL).

[48] Kareem Darwish and Ossama Emam. 2005. The effect of blind relevance feedback on a new Arabic OCR degraded text collection. International Conference on Machine Intelligence: Special Session on Arabic Document Image Analysis.

[49] Kareem Darwish, Hany Hassan, and Ossama Emam. 2005. Examining the effect of improved context sensitive morphology on Arabic information retrieval. In Proceedings of the ACL-2005 Workshop on Computational Approaches to Semitic Languages, pp. 25-30.

[50] Kareem Darwish and Douglas W. Oard. 2002. Term selection for searching printed Arabic. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, pp. 261-268.

[51] Kareem Darwish and Douglas W. Oard. 2002. CLIR experiments at Maryland for TREC 2002: Evidence combination for Arabic-English retrieval. In 2002 Text REtrieval Conference (TREC), Gaithersburg, MD.

[52] Kareem Darwish and Douglas W. Oard. 2003. Balanced query methods for improving OCR-based retrieval. Proceedings 2003 Symposium on Document Image Understanding Technology.

[53] Kareem Darwish and Walid Magdy. 2007. Error correction vs. query garbling for Arabic OCR document retrieval. In ACM Transactions on Information Systems (TOIS), Vol. 26.

[54] Kareem Darwish. 2013. Named Entity Recognition using Cross-lingual Resources: Arabic as an Example. In the Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL), pp. 1558-1567.

[55] Ludovic Denoyer and Patrick Gallinari. 2006. The Wikipedia XML corpus. ACM Special Interest Group on Information Retrieval Forum, 40 (1).

[56] Kareem Darwish, Walid Magdy, and Ahmed Mourad. 2012. Language processing for Arabic microblog retrieval. In Proceedings of Conference on Information and Knowledge Management (CIKM).

[57] Anne De Roeck and Waleed El-Fares. 2000. A morphologically sensitive clustering algorithm for identifying Arabic roots. In the Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL), Hong Kong, pp. 199-206.

[58] Mona Diab. 2009. Second generation tools (AMIRA 2.0): Fast and robust tokenization, POS tagging, and base phrase chunking. 2nd International Conference on Arabic Language Resources and Tools.

[59] David Doermann. 1998. The indexing and retrieval of document images: A survey. Computer Vision and Image Understanding, 70(3): pp. 287-298.

[60] Youssef El-Dakar, Khalid El-Gazzar, Noha Adly, Magdy Nagi. 2005. The Million Book Project at Bibliotheca Alexandrina. Journal of Zhejiang University-Science A 6, no. 11 (2005): 1327-1340.

[61] Ali El Kahki, Kareem Darwish, Ahmed Saad El Din, and Mohamed Abd El-Wahab. 2012. Transliteration mining using large training and test sets. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT '12). pp. 243-252.

[62] Ali El Kahki, Kareem Darwish, Ahmed Saad El Din, Mohamed Abd El-Wahab, Ahmed Hefny, and Waleed Ammar. 2011. Improved transliteration mining using graph reinforcement. In the 2011 Empirical Methods in Natural Language Processing (EMNLP). pp. 1384-1393.

[63] Ahmed El-Kholy and Nizar Habash. 2010. Techniques for Arabic morphological detokenization and orthographic denormalization. In Proceedings of the 2000 Language Resources and Evaluation Conference.

[64] Alexander Fraser, Jinxi Xu, and Ralph M. Weischedel. 2002. TREC 2002 cross-lingual retrieval at BBN. In 2002 Text REtrieval Conference (TREC), Gaithersburg, MD.

[65] Wei Gao, Cheng Niu, Jian-Yun Nie, Ming Zhou, Jian Hu, Kam-Fai Wong, and Hsiao-Wuen Hon. 2007. Cross-lingual query suggestion using query logs of different languages. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 463-470.

[66] Wei Gao, Cheng Niu, Ming Zhou, and Kam-Fai Wong. 2009. Joint ranking for multilingual web search. In the 2009 European Conference on Information Retrieval (ECIR), pp. 114-125.

[67] Wei Gao, Cheng Niu, Jian-Yum Nie, Ming Zhou, Kam-Fai Wong, and Hsiao-Wuen Hon. 2010. Exploiting query logs for cross-lingual query suggestions. ACM Transactions on Information Systems (TOIS), Vol. 28(2), 6.

[68] Fredric Gey and Douglas W. Oard. 2001. The TREC-2001 Cross- Language Information Retrieval Track: Searching Arabic using English, French or Arabic queries. In 2001 Text REtrieval Conference (TREC), Gaithersburg, MD, pp. 16-23.

[69] Andrew Gillies, Erik Erlandson, John Trenkle, and Steve Schlosser. 1997. Arabic text recognition system. The Symposium on Document Image Understanding Technology.

[70] Julio Gonzalo, Jussi Karlgren, and Paul Clough. 2007. iCLEF 2006 overview: Searching the Flickr WWW photo-sharing repository. Evaluation of Multilingual and Multi-modal Information Retrieval, In 2007 Cross-Language Evaluation Forum (CLEF).

[71] Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 267-274.

[72] Nizar Habash and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In Proceedings of the Conference of American Association for Computational Linguistics.

[73] Nizar Habash and Owen Rambow. "MAGEAD: a morphological analyzer and generator for the Arabic dialects." Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics.

[74] Nizar Habash and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers (NAACL-Short '06), pp. 49-52.

[75] Nizar Habash and Owen Rambow. 2007. Arabic Diacritization through Full morphological tagging. In Proceedings of the 2007 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT '07), Companion Volume, pp. 53-56, Rochester, NY.

[76] Nizar Habash, Abdelhadi Soudi, and Timothy Buckwalter. 2007. On Arabic transliteration. Arabic Computational Morphology. Text, Speech and Language Technology, Vol. 38, 2007, pp. 15-22.

[77] Nizar Habash, Owen Rambow, and Ryan Roth. 2009. Mada+Tokan: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt.

[78] Nizar Habash. 2010. Introduction to Arabic language processing. Synthesis Lectures on Human Language Technologies 3(1), pp. 1-187.

[79] Nizar Habash, Mona T. Diab, and Owen Rambow. 2012. Conventional orthography for dialectal Arabic. In Proceedings of the 2012 Language Resources and Evaluation Conference.

[80] Nizar Habash, Ryan Roth, Owen Rambow, Ramy Eskander, and Nadi Tomeh. 2013. Morphological analysis and disambiguation for dialectal Arabic. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT '13), pp. 426-432.

[81] Bassam Hammo, Hani Abu-Salem, and Steven Lytinen. 2002. QARAB: a question answering system to support the Arabic language. In Proceedings of the ACL-2002 Workshop on Computational Approaches to Semitic Languages.

[82] Stephen M. Harding, W. Bruce Croft, and C. Weir. Probabilistic retrieval of OCR-degraded text using n-grams. European Conference on Digital Libraries.

[83] Khosrow M. Hassibi. 1994a. Machine printed Arabic OCR. The 22nd AIPR Workshop: Interdisciplinary Computer Vision, SPIE Proceedings.

[84] Khosrow M. Hassibi. 1994b. Machine printed Arabic OCR using neural networks. The 4th International Conference on Multi-lingual Computing, London.

[85] David Hawking. 1996. Document retrieval in OCR-scanned text. 6th Parallel Computing Workshop, Kawasaki, Japan.

[86] Daqing He, Douglas W. Oard, Jianqiang Wang, Jun Luo, Dina Demner-Fushman, Kareem Darwish, Philip Resnik, Sanjeev Khudanpur, Michael Nossal, Michael Subotin, and Anton Leuski. 2003. Making MIRACLEs: Interactive translingual search for Cebuano and Hindi. ACM Transactions on Asian Language Information Processing (TALIP) Vol. 2 Issue 3.

[87] Ahmed Hefny, Kareem Darwish, and Ali Alkahky. 2011. Is a query worth translating: Ask the users! In the 2011 European Conference on Information Retrieval (ECIR), pp. 238-250.

[88] Ismail Hmeidi, Ghassan Kanaan, and Martha Evens. 1997. Design and implementation of automatic indexing for information retrieval with Arabic documents. Journal of the American Society for Information Science and Technology Vol. 48(10): pp. 867-881.

[89] Tao Hong. 1995. Degraded text recognition using visual and linguistic context. Ph.D. thesis, Computer Science Department, SUNY Buffalo, Buffalo, NY.

[90] Dan Jurafsky and James Martin. 2000. Speech and language processing. Prentice Hall.

[91] Paul Kantor and Ellen Voorhees. 1996. Report on the TREC-5 Confusion Track. In 1996 Text REtrieval Conference (TREC), Gaithersburg, MD.

[92] Tapas Kanungo, Gregory Marton, and Osama Bulbul. 1999. OmniPage vs. Sakhr: Paired model evaluation of two Arabic OCR products. in SPIE Conference on Document Recognition and Retrieval (VI), San Jose, California.

[93] Tapas Kanungo, Osama Bulbul, Gregory Marton, and Doe-Wan Kim. 1997. Arabic OCR systems: State of the art. Symposium on Document Image Understanding Technology, Annapolis, MD.

[94] Shereen Khoja and Roger Garside. 2001. Automatic tagging of an Arabic corpus using APT. The Arabic Linguistic Symposium (ALS), University of Utah, Salt Lake City, Utah.

[95] George Kiraz. 1998. Arabic computation morphology in the West. in The 6th International Conference and Exhibition on Multi-lingual Computing.

[96] Kazuaki Kishida. 2008. Prediction of performance of cross-language information retrieval using automatic evaluation of translation. Library & Info. Science Research. Vol. 30 (2), pp. 138-144.

[97] Wessel Kraaij, Paul Over, and A. Smeaton. 2006. TRECVID 2006-an introduction. In 2006 TREC Video Retrieval Evaluation.

[98] Lori Lamel, Abdelkhalek Messaoudi, and Jean-Luc Gauvain. 2007. Improved acoustic modeling for transcribing Arabic broadcast data. Interspeech'07, pp. 2077-2080.

[99] Leah S. Larkey, James Allan, Margaret E. Connell, Alvaro Bolivar, and Courtney Wade. 2002. UMass at TREC 2002: Cross language and novelty tracks. In 2002 Text REtrieval Conference (TREC).

[100] Leah S. Larkey, Lisa Ballesteros, and Margaret E. Connell. 2002. Improving stemming for Arabic information retrieval: Light stemming and co-occurrence analysis. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 275-282. 77777

[101] Leah S. Larkey, Nasreen AbdulJaleel, and Margaret Connell. 2003. What's in a name?: Proper names in Arabic cross-language information retrieval. Technical report, CIIR Technical Report, IR-278.

[102] Leah S. Larkey, Fangfang Feng, Margaret Connell, and Victor Lavrenko. 2004. Language-specific models in multilingual topic tracking. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 402-409.

[103] Chia-Jung Lee, Chin-Hui Chen, Shao-Hang Kao, and Pu-Jen Cheng. 2010. To translate or not to translate? In Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

[104] Young-Suk Lee, Kishore Papineni, Salim Roukos, Ossama Emam, and Hany Hassan. 2003. Language model based Arabic word segmentation. In the Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, July 2003, Sapporo, Japan. pp. 399-406.

[105] Gina Anne Levow, Douglas W. Oard, and Philip Resnik. 2005. Dictionary-based techniques for cross-language information retrieval. Information Processing & Management Journal, Vol. 41 Issue 3.

[106] Dirk Lewandowski. 2012. Web search engine research. Series editor Amanda Spink. Vol. 4. Emerald Group Publishing.

[107] Wen-Cheng Lin and Hsin-Hsi Chen. 2003. Merging mechanisms in multilingual information retrieval. Lecture notes in computer science, pp. 175-186.

[108] Zhidong A. Lu, Issam Bazzi, Andras Kornai, John Makhoul, Premkumar S. Natarajan, and Richard Schwartz. 1999. A robust, language-independent OCR system. in The 27th AIPR Workshop: Advances in Computer Assisted Recognition, SPIE.

[109] Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The Penn Arabic treebank: Building a large-scale annotated Arabic corpus. In NEMLAR Conference on Arabic Language Resources and Tools, pp. 102-109.

[110] Mohamed Maamouri, Dave Graff, Basma Bouziri, Sondos Krouna, Ann Bies, and Seth Kulick. 2010. LDC Standard Arabic Morphological Analyzer (SAMA) version 3.1. Linguistics Data Consortium, Catalog No. LDC2010L01.

[111] Walid Magdy. 2013. TweetMogaz: a news portal of tweets. In Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1095-1096.

[112] Walid Magdy, Ahmed Ali, and Kareem Darwish. 2012. A summarization tool for time-sensitive social media. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM), pp. 2695-2697.

[113] Walid Magdy and Kareem Darwish. 2006. Arabic OCR error correction using character segment correction, language modeling, and shallow morphology. In the 2006 Empirical Methods in Natural Language Processing (EMNLP). Sydney, Australia, pp. 408-414.

[114] Walid Magdy and Kareem Darwish. 2006. Word-based correction for retrieval of Arabic OCR degraded documents. String Processing and Information Retrieval (SPIRE).

[115] Walid Magdy and Kareem Darwish. 2010. Omni font OCR error correction with effect on retrieval. International Conference on Intelligent Systems Design and Applications (ISDA), 2010, pp. 415-420.

[116] Walid Magdy, Kareem Darwish, Ossama Emam, and Hany Hassan. 2007. Arabic cross-document person name normalization. In Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, pp. 25-32.

[117] Walid Magdy, Kareem Darwish, and Mohsen Rashwan. 2007. Fusion of multiple corrupted transmissions and its effect on information retrieval. In the Conference of the Egyptian Society of Language Engineering (ESOLE) 2007.

[118] Walid Magdy and Kareem Darwish. 2008. Effect of OCR error correction on Arabic retrieval. Information Retrieval Journal. 11, 5, 405-425

[119] Walid Magdy, Kareem Darwish, and Motaz El-Saban. 2009. Efficient language-independent retrieval of printed documents without OCR. String Processing and Information Retrieval (SPIRE).

[120] Lidia Mangu, Hong-Kwang Kuo, Stephen Chu, Brian Kingsbury, George Saon, Hagen Soltau, and Fadi Biadsy. 2011. The IBM 2011 GALE Arabic speech transcription system. 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 272-277.

[121] James Mayfield, Paul McNamee, C. Costello, C. Piatko, and A. Banerjee. 2001. JHU/APL at TREC 2001: Experiments in filtering and in Arabic, video, andWeb retrieval. In 2001 Text REtrieval Conference (TREC), Gaithersburg, MD.

[122] J. Scott McCarley. 1999. Should we translate the documents or the queries in cross-language information retrieval? Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics.

[123] Paul McNamee and James Mayfield. 2002. Comparing cross-language query expansion techniques by degrading translation resources. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

[124] Paul McNamee, C. Piatko, James Mayfield. 2002. JHU/APL at TREC 2002: Experiments in filtering and Arabic retrieval. In 2002 Text REtrieval Conference (TREC).

[125] Mohammed Moussa, Mohamed Waleed Fakhr, and Kareem Darwish. 2012. Statistical denormalization for Arabic Text. In Empirical Methods in Natural Language Processing, pp. 228. 2012.

[126] ThuyLinh Nguyen, Stephan Vogel, and Noah A. Smith. 2010. Nonparametric word segmentation for machine translation. In Proceedings of the 2010 International Conference on Computational Linguistics (COLING), pp. 815-823.

[127] Stefanie Nowak, and Peter Dunker. 2010. Overview of the CLEF 2009 Large Scale Visual Concept Detection and Annotation Task. Multilingual Information Access Evaluation II. Multimedia Experiments Lecture Notes in Computer Science Vol. 6242, pp. 94-109.

[128] Douglas W. Oard, Bonnie Dorr. 1996. A survey of multilingual text retrieval. UMIACS, University of Maryland, College Park.

[129] Douglas W. Oard and Fredric Gey. 2002. The TREC 2002 Arabic/English CLIR Track. In 2002 Text REtrieval Conference (TREC), Gaithersburg, MD.

[130] Douglas W. Oard and William Webber. 2013. Information retrieval for e-discovery. Foundations and Trends in Information Retrieval, Vol. 7, No 1 (2013) 1-145.

[131] Kemal Oflazer. 1996. Error-tolerant finite state recognition with applications to morphological analysis and spelling correction. Computational Linguistics, 22(1), 73-89.

[132] Joseph Olive, Caitlin Christianson, and John McCary. 2011. Handbook of natural language processing and machine translation. Springer ISBN 978-1-4419-7712-0

[133] Iadh Ounis, Craig Macdonald, Jimmy Lin, and Ian Soboroff. 2011. Overview of the TREC-2011 Microblog Track. In 2011 Text REtrieval Conference (TREC).

[134] Lawrence Page. 1998. Method for node ranking in a linked database. US Patent No. 6285999

[135] Jiaul H. Paik, Dipasree Pal, and Swapan K. Parui. 2011. A novel corpus-based stemming algorithm using co-occurrence statistics. In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval.

[136] Arfath Pasha, Mohammad Al-Badrashiny, Mohamed Altantawy, Nizar Habash, Manoj Pooleery, Owen Rambow and Ryan M. Roth. 2013. DIRA: Dialectal Arabic Information Retrieval Assistant. The Companion Volume of the Proceedings of International Joint Conference on Natural Language Processing (IJCNLP) 2013: System Demonstrations, pp. 13-16.

[137] Ari Pirkola. 1998. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 55-63.

[138] Stephen Robertson and Karen Spärck Jones. 1996. Simple, proven approaches to text-retrieval. Technical Report 356, Computer Laboratory, University of Cambridge, Cambridge, England.

[139] Horacio Rodríguez, David Farwell, Javi Farreres, Manuel Bertran, Musa Alkhalifa, M. Antonia Martí, William Black, Sabri Elkateb, J. Kirk, Adam Pease, Piek Vossen, and Christiane Fellbaum. 2008. Arabic WordNet: Current state and future extensions. In The Fourth Global WordNet Conference, Szeged, Hungary.

[140] Gregory Tassey, Brent R. Rowe, Dallas W. Wood, Albert N. Link, and Diglio A. Simoni. 2010. Economic impact assessment of NIST's Text REtrieval Conference (TREC) Program. Report prepared for National Institute of Technology (NIST).

[141] Khaled Shaalan and Hafsa Raza. 2007. Person Name Entity Recognition for Arabic. In Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, pp. 17âĂŞ24, Prague, Czech Republic.

[142] Hassan Sajjad, Kareem Darwish, and Yonatan Belinkov. 2013. Translating dialectal Arabic to English. The 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria.

[143] Shirin Saleem, Huaigu Cao, Krishna Subramanian, Matin Kamali, Rohit Prasad, Prem Natarajan. 2009. Improvements in BBN's HMM-based offline Arabic handwriting recognition system. 10th International Conference on Document Analysis and Recognition.

[144] Wael Salloum and Nizar Habash. 2012. Elissa: A dialectal to standard arabic machine translation system. In Proceedings of the 2012 International Conference on Computational Linguistics (COLING). Mumbai, India.

[145] Gerard Salton and M. Lesk. 1969. Relevance assessments and retrieval system evaluation. Information Storage and Retrieval, 1969 (4), pp. 343-359.

[146] Mark Sanderson. 1994. Word sense disambiguation and information retrieval. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 142-151.

[147] Mark Sanderson and H. Joho. 2004. Forming test collections with no system pooling. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, South Yorkshire, UK.

[148] Mark Sanderson. 2010. Test collection based evaluation of information retrieval systems. Foundations and Trends in Information Retrieval, Vol. 4, No. 4, pp. 247-375.

[149] Jacques Savoy, and Yves Rasolofo. 2002. Report on the TREC 11 experiment: Arabic, named page and topic distillation searches. In 2002 Text REtrieval Conference (TREC).

[150] Asad Sayeed, Tamer Elsayed, Nikesh Garera, David Alexander, Tan Xu, Douglas W. Oard, David Yarowsky, and Christine Piatko. 2009. Arabic cross-document coreference detection. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pp. 357-360.

[151] Mohammed Q. Shatnawi, Qusai Q. Abuein, and Omar Darwish. 2011. Verification hadith correctness in islamic web pages using information retrieval techniques. Proceedings of International Conference on Information & Communication Systems, Irbid, Jordan.

[152] Mohammed Q. Shatnawi, Muneer Bani Yassein, and Reem Mahafza. 2012. A framework for retrieving Arabic documents based on queries written in Arabic slang language. Journal of Information Science, Vol. 38, no. 4: 350-365.

[153] Luo Si and Jamie Callan. 2006. CLEF 2005: Multilingual retrieval by combining multiple multilingual ranked lists. Accessing Multilingual Information Repositories Lecture Notes in Computer Science, Vol. 4022, pp. 121-130.

[154] Amit Singhal, Gerard Salton, and Chris Buckley. 1996. Length normalization in degraded text collections. The 5th Annual Symposium on Document Analysis and Information Retrieval.

[155] Alan F. Smeaton, Paul Over, and Wessel Kraaij. 2006. Evaluation campaigns and TRECVid. In Proceedings of the 8th ACM international workshop on Multimedia information retrieval, pp. 321-330.

[156] Ian Soboroff, Charles Nicholas, and Patrick Cahan. 2001. Ranking retrieval systems without relevance judgments. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

[157] Ian Soboroff and Stephen Robertson. 2003. Building a filtering test collection for TREC 2002. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 243-250.

[158] Hagen Soltau, George Saon, Brian Kingsbury, Jeff Kuo, Lidia Mangu, Daniel Povey, and Geoffrey Zweig. 2007. The IBM 2006 GALE Arabic ASR system. In Proceedings of the 2007 International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 4, pp. 349-352.

[159] Hagen Soltau, Lidia Mangu, and Fadi Biadsy. 2011. From Modern Standard Arabic to Levantine ASR: Leveraging GALE for dialects. In Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on, pp. 266-271.

[160] Stephanie Strassel, Mark A. Przybocki, Kay Peterson, Zhiyi Song, and Kazuaki Maeda. 2008. Linguistic resources and evaluation techniques for evaluation of cross-document automatic content extraction. In Proceedings of the 2008 Language Resources and Evaluation Conference.

[161] Stephanie Strassel. 2009. Linguistic resources for Arabic handwriting recognition. In Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt.

[162] Trevor Strohman, Donald Metzler, Howard Turtle, and W. Bruce Croft. 2005. Indri: A language model-based search engine for complex queries. In Proceedings of the International Conference on Intelligent Analysis, Vol. 2, No. 6, pp. 2-6.

[163] Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2008. Learning to rank answers on large online QA collections. In Proceedings of the 46th Annual Meeting for the Association for Computational Linguistics: Human Language Technologies (ACL HLT '08), pp. 719-727.

[164] Kazem Taghva, Julie Borasack, Allen Condit, and Jeff Gilbreth. 1994. Results and implications of the noisy data projects. 1994, Information Science Research Institute, University of Nevada, Las Vegas.

[165] Kazem Taghva, Julie Borasack, Allen Condit, and Padma Inaparthy. 1995. Querying short OCR'd documents. 1995, Information Science Research Institute, University of Nevada, Las Vegas.

[166] Kazem Taghva, Julie Borasack, and Allen Condit. 1994. An expert system for automatically correcting OCR output. In SPIE-Document Recognition.

[167] Mikael Tillenius. 1996. Efficient generation and ranking of spelling error corrections. NADA tech. report TRITA-NA-E9621.

[168] Omar Trigui, Lamia Hadrich Belguith, Paolo Rosso, Hichem Ben Amor, and Bilel Gafsaoui. 2012. Arabic QA4MRE at CLEF 2012: Arabic Question Answering for Machine Reading Evaluation. CLEF (Online Working Notes/Labs/Workshop).

[169] Ming-Feng Tsai, Yu-Ting Wang, and Hsin-Hsi Chen. 2008. A study of learning a merge model for multilingual information retrieval. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

[170] Yuen-Hsien Tseng and Douglas W. Oard. 2001. Document image retrieval techniques for Chinese. In Symposium on Document Image Understanding Technology (SDIUT). Columbia, MD, pp. 151-158.

[171] Theodora Tsikrika, and Jana Kludas. 2009. Overview of the WikipediaMM Task at ImageCLEF 2009. Multilingual Information Access Evaluation II. Multimedia Experiments, Lecture Notes in Computer Science Vol. 6242, pp 60-71.

[172] Raghavendra Udupa, K. Saravanan, A. Bakalov, and A. Bhole. 2009. "They are out there, if you know where to look": Mining transliterations of OOV query terms for cross-language information retrieval. In the 2009 European Conference on Information Retrieval (ECIR), LNCS 5478, pp. 437-448.

[173] Raghavendra Udupa, K. Saravanan, A. Kumaran, and Jagadeesh Jagarlamudi. 2009. Mint: A method for effective and scalable mining of named entity transliterations from large comparable corpora. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL), pp. 799-807.

[174] Ellen Voorhees. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia.

[175] Jianqiang Wang and Douglas W. Oard. 2006. Combining bidirectional translation and synonymy for cross-language information retrieval. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 202-209.

[176] Charles L. Wayne. 1998. Detection & tracking: A case study in corpus creation & evaluation methodologies. Language Resources and Evaluation Conference, Granada, Spain.

[177] Dan Wu, Daqing He, Heng Ji, and Ralph Grishman. 2008. A study of using an out-of-box commercial MT system for query translation in CLIR. Workshop on Improving non-English web searching, Proceedings of 2008 Conference on Information and Knowledge Management (CIKM).

[178] Jinxi Xu, Alexander Fraser, and Ralph Weischedel. 2002. Empirical studies in strategies for Arabic retrieval. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval.

[179] Emine Yilmaz, Evangelos Kanoulas, and Javed A. Aslam. 2008. A simple and efficient sampling method for estimating AP and NDCG. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval.