Term Selection for Searching Printed Arabic

Kareem Darwish
Electrical and Computer Engineering Dept.
University of Maryland, College Park
College Park, MD 20742
kareem@glue.umd.edu

Douglas W. Oard
College of Information Studies and UMIACS
University of Maryland, College Park
College Park, MD 20742
oard@glue.umd.edu

ABSTRACT

Since many Arabic documents are available only in print, automating retrieval from collections of scanned Arabic document images using Optical Character Recognition (OCR) is an interesting problem. Arabic combines rich morphology with a writing system that presents unique challenges to OCR systems. These factors must be considered when selecting terms for automatic indexing. In this paper, alternative choices of indexing terms are explored using both an existing electronic text collection and a newly developed collection built from images of actual printed Arabic documents. Character n-grams or lightly stemmed words were found to typically yield near-optimal retrieval effectiveness, and combining both types of terms resulted in robust performance across a broad range of conditions.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: # Information Search and Retrieval – s*earch process, selection process.*

General Terms: Algorithms, Design, Experimentation.

Keywords: OCR, Arabic, Information Retrieval, Term Selection.

1. INTRODUCTION

Since the advent of the printing press in the fifteenth century, the amount of printed text has grown to an overwhelming scale. Of course, a great deal of text is now generated in character-coded electronic formats (HTML, word processor files, etc.). But printed text remains important, in part because large collections of legacy documents exist only in printed form, and in part because printed text remains a ubiquitous distribution channel that can effectively deliver information without the technical infrastructure that is needed to deliver character-coded text. These factors are particularly important for Arabic, which is widely used in several countries where the installed computer infrastructure is quite limited. Furthermore, before the 1998 introduction of support for character-coded Arabic in the standard versions of Microsoft's Windows operating system, Arabic text on the World Wide Web was most often rendered as document images. Character-coded Arabic is now becoming more common on the Web, but it is still not unusual to find Arabic documents that are available only as document images.

Printed documents can be browsed relatively easily in limited quantities, but effective access to the contents of large collections of printed documents requires some form of automation. Although we focus on the effect of term selection on supporting search, term selection has important consequences for other applications (e.g., text classification and information visualization) that also exploit a "bag-of-terms" document representation. Three basic approaches to searching printed documents are possible:

- Hire catalogers to examine each document and assign metadata that describes the content of the document. This is the approach used in library catalogs.
- Hire typists to read each document and rekey the material in a way that duplicates (or perhaps summarizes) the content of the document. This is the approach used by the U.S. Foreign Broadcast Information Service.
- Scan the document to produce a document image, then perform Optical Character Recognition (OCR) to obtain a representation (possibly containing errors) that approximates the content of the document.

We have chosen to focus on the third approach because it clearly offers the greatest potential for affordably scaling up to process very large collections. Widely available sheet feed scanners can scan thousands of pages per hour, and specialized hardware can be used to acquire document images in other situations (e.g., document cameras can be used with bound volumes). Arabic OCR poses a number of challenges, however. In Arabic text, letters are connected, letters change shape depending on their position in a word, special forms for letter combinations and word elongations are often used, many letters include dots, diacritic marks may optionally be present, and both dots and diacritics might be easily confused with dust and/or speckle that was introduced during the scanning process. Arabic morphology is rich and complex, and many researchers have found that indexing terms obtained using morphological analysis can yield better retrieval effectiveness than indexing the surface form of each word (c.f., [3]). Character recognition errors might therefore adversely impact retrieval effectiveness in two ways: (1) by altering terms in a way that would preclude accurate morphological analysis, and (2) by altering meaning-bearing terms in ways that would prevent accurate matching between query terms and the terms found in a document. Our goal in this paper is therefore to reexamine the question of term selection for Arabic information retrieval in the context of OCR-degraded text.

SIGIR'02, August 11-15, 2002, Tampere, Finland.

Although the Text Retrieval Conference (TREC) has recently produced a large test collection for experiments with retrieval from character-coded Arabic text, we are not aware of any similar resource for printed Arabic documents. We therefore developed a small test collection of Arabic document images for use in our experiments.

The remainder of the paper is organized as follows. In the next section, we describe previous research on the use of OCR for information retrieval and on retrieval of character-coded Arabic text. Section 3 then develops the experimental framework for comparing the effect of alternative indexing terms on retrieval effectiveness. Our Arabic document image test collection is introduced in Section 4, along with results of experiments to characterize the suitability of that collection for the evaluation of information retrieval effectiveness. In section 5, we explore the effect of OCR-degraded Arabic text on retrieval effectiveness. Finally, we conclude with some more general observations on the selection of index terms for Arabic information retrieval and suggest some future research directions.

2. BACKGROUND

The goal of OCR is to transform a document image into character-coded text. The usual process is to automatically segment the document image into character images in the proper reading order using image analysis heuristics, apply an automatic classifier to determine the character codes that are most likely to correspond to each character image [20], 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) [6], the resolution at which the 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 printed document [5]. Arabic OCR presents several challenges, including:

- Connected characters, which change shape depending on their position in the word, make the isolation of individual character images challenging.
- Word elongations (kashida) and special forms for certain letter combinations (ligatures such as lam-alef (لا)) are often used in typed text [23], expanding the number of possibilities that the classifier must consider.
- 15 of the 28 Arabic letters include dots as an integral part of the character, and authors sometimes choose to additionally place diacritic marks on some letters. Dots and diacritic marks can easily be confused with speckle or dust, making detection of the correct character challenging.
- Due to the morphological complexity of Arabic, the number of legal words has been estimated to be 60 billion [1]. This limits the value of sequential context somewhat, since it would be impractical to store a complete vocabulary of that size.

There are a number of commercial Arabic OCR systems, with Sakhr's Automatic Reader and Shonut's Omni Page being perhaps the most widely used [15].

Retrieval of OCR degraded text documents has been reported for many languages, including English, French, Spanish, and Chinese [12, 22, 24], but we are not aware of prior work on Arabic. Three methods have been used to produce test collections for OCR-degraded text:

- Systematically altering character-coded text using a character-level confusion model that is trained on aligned pairs of character-coded and OCR-degraded texts. Large test collections can be efficiently produced using this technique by starting with an existing test collection for which topics and relevance judgments are already available. However, the degree of insight that can be obtained depends on the fidelity of the character confusion model, which might model some aspects of the process (e.g., character replacement) better than others (e.g., the effect of document skew during scanning). Harding, et al. used OCR errors that were simulated in this way to examine the effect of indexing character *n*-grams on retrieval from four English document collections (with 423 to 12,380 documents), finding that *n*-grams outperformed words [12].
- Typesetting character-coded text to produce a document image, optionally degrading the image to simulate speckle, page skew, bleed-through, varying illumination, and other factors [6, 14], and then performing OCR. Although the operations on large document images adds some time to the process, large test collections can still be constructed relatively efficiently because it is possible to start with a collection for which topics and relevance judgments already exist. Baird used this technique to show that that retrieval effectiveness falls dramatically with increases in the character recognition error rate [5].
- Scanning a collection of printed documents, performing OCR, and then manually creating appropriate topics and relevance judgments. The size of a test collection created in this way will be limited by the resources available for the relevance judgment process. However, this technique can accurately model many aspects that may be present in real applications (e.g., unfamiliar fonts, damaged pages, and handwritten annotations). Taghva, et al. experimented on a 204-document English document image collection using this technique. The average length of the documents was 38 pages. He observed no significant effect of degradation on retrieval [21]. Tseng and Oard experimented with different combinations of n-grams on a Chinese collection of 8,438 document images. The documents images were scanned from printed material. They observed that combinations of character 1-grams and character 2-grams performed best [24].

Arabic words are derived from a closed set of approximately 10,000 roots by attaching prefixes, suffixes and infixes. Often, vowel replacement and letter omission are required to construct words. Roots are mostly 3 letters, often 4 letters, and rarely 5 letters. Stems are derived from roots by inserting infixes only [9]. Several types of index terms have been studied, including word surface forms, clusters of words [25], and results of morphological processing, such as stems and morphological roots [3, 4, 13], and character *n*-grams of various lengths [10, 16]. The effects of normalizing alternative characters, removal of diacritics and stop-word removal have also been explored [8, 10, 26]. Because the earliest of these studies used very small test collections and more recent results are based on a single large test collection (from the 2001 Text Retrieval Conference), relatively few general conclusions can be drawn regarding the optimal choice of indexing terms for character-coded Arabic text. The preponderance of the evidence does, however, suggest that some form of morphological analysis and/or the use of character *n*-grams substantially outperforms use of word surface forms, and that some form of character normalization is helpful. In the next section, we explore these questions in greater detail.

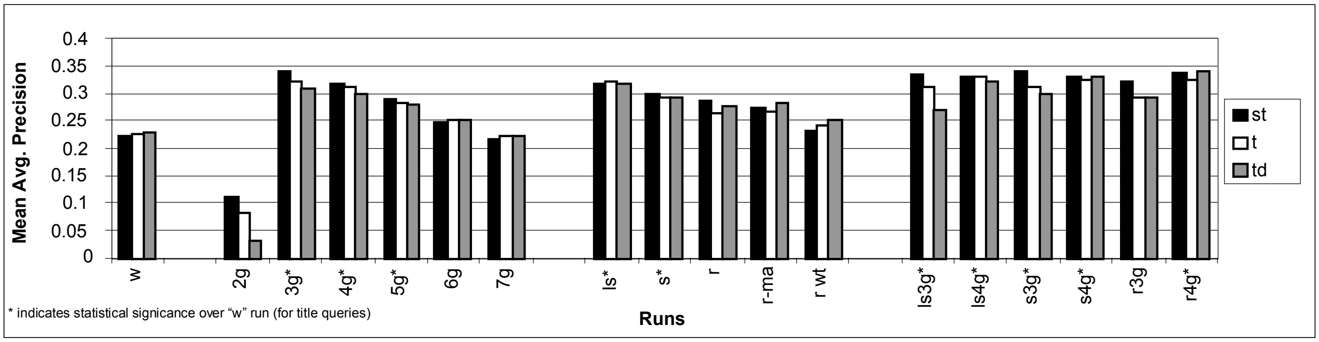


Figure 1: Runs on TREC collection with "short title" (st), "title" (t), and "title + description" (td) queries

3. INDEX TERMS FOR CHARACTER-CODED ARABIC

In this section we experimentally explore the effectiveness of information retrieval using different Arabic index terms on a large collection. We explored three categories of Arabic index terms: character *n*-grams, terms obtained through morphological analysis, and combinations of both. In the first category, we looked at within-word character *n*-grams, with values of *n* ranging between 2 and 7 characters. In the second category, we examined word surface forms, light stemming (in which common prefixes and suffixes were removed), aggressive stemming (in which all recognizable prefixes and suffixes were removed), and morphological roots (in which any infixes were also removed). To see the difference between the four approaches, consider the word "وكريمتهم" (wkrymthm). The derived terms "كريمت" (krymt), "كريم" (krym), and "كرم" (krm) are the lightly stemmed version of the word, the stem, and the root respectively.[1] Before forming the character *n*-grams and performing morphological analysis, all words were stripped of diacritics and variant forms of the letter alef (bare alef, alef-hamza, alef-madd, waw-hamza, and ya-hamza – ا، أ، آ، إ، ئ، ؤ) were all normalized to bare alef. Similarly, alef-maqsoura and ya (ى، ي) were normalized to ya**.**

We used the Linguistic Data Consortium (LDC) LDC2001T55 collection, which was used in the TREC 2001 cross-language track. For brevity, we refer to this as the TREC collection. The collection contains 383,872 articles from the Agence France Press (AFP) Arabic newswire. Twenty-five topics were developed cooperatively by the LDC and the National Institute of Standards and Technology (NIST), and relevance judgments were developed at LDC by manually judging a pool of documents obtained from combining the top 70 documents from 20 runs submitted by 10 teams to TREC's Cross Language track in 2001. The number of known relevant documents ranges from 6 to 556, with an average of 165 relevant documents per topic. This is larger than is typical for a TREC collection, and there is some indication that there may still be a substantial number of undiscovered relevant documents [11]. Nevertheless, this is presently the best available large Arabic information retrieval test collection.

The TREC topic descriptions each 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 [11]. We constructed three types of queries from the TREC topics:

a. the title and description fields (*td*). This is intended to model the sort of statement that a searcher might initially make when asking an intermediary such as a librarian for help with a search.
b. the title field only (*t*). The title field in recent TREC collections is typically designed as a model for Web queries, which typically contain only 2 or 3 words. However, the average length if the *t* queries is about 6 words.
c. a short version of the title field *(st)* in which words that were deemed by the first author not to be typical terms in a brief Web query were deleted. The average length of the *st* queries is 3.5 words.

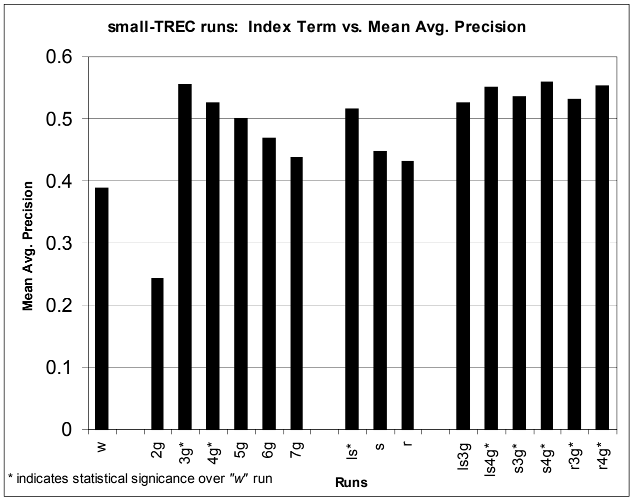
We performed experiments for each query length with the following index terms:

- *w*: words.
- *2g, 3g, ... 7g*: character n-grams (2-7 gram).
- *ls*: lightly stemmed words, obtained by using pattern matching to remove common prefixes and suffixes.
- *s*: aggressively stemmed words, found using the Sebawai morphological analyzer.
- three ways of obtaining roots:
  - *r*: the top root found by the Sebawai morphological analyzer, which produces a ranked list of possible roots [9].
  - *r-ma*: the highest ranked root found by Sebawai that was also produced by ALPNET [7], if ALPNET produced an analysis; otherwise the top root found by Sebawai. For words that it can analyze, ALPNET produces an unranked set of possible roots that almost always contains the correct one, but it fails to produce an analysis more often than Sebawai.
  - *r-wt*: Sebawai's top 2 roots, weighted by their likelihood ratio.
- combinations of:
  - *ls3g*: lightly stemmed words and 3-grams.
  - *ls4g*: lightly stemmed words and 4-grams.
  - *s3g*: aggressively stemmed words and 3-grams.
  - *s4g*: aggressively stemmed words and 4-grams.
  - *r3g*: roots and 3-grams.
  - *r4g*: roots and 4-grams.

[1] The transliteration scheme used in the paper is define in [9]

Table 1: Comparing TREC runs using the *t*-test's p-value. A "bold" p-value indicates statistical significance.

3g	4g	5g	w	ls	s	r	ls3g	ls4g	s3g	s4g	r3g	r4g	
	.47	**.05**	**.00**	.87	.24	**.04**	.55	.58	.46	.84	.27	.87	**3g**
		.00	**.00**	.67	.56	.16	.98	.08	.94	**.02**	.61	.12	**4g**
			.02	**.06**	.78	.53	.30	**.00**	.36	**.00**	.78	**.00**	**5g**
				.00	**.05**	.19	**.01**	**.00**	**.01**	**.00**	.06	**.00**	**w**
					.18	**.02**	.71	.47	.65	.73	.39	.79	**ls**
						.02	.28	.22	.38	.28	.97	.29	**s**
							.02	.05	**.05**	.06	.25	.05	**r**
								.42	.81	.58	.30	.60	**ls3g**
									.38	.47	.24	.56	**ls4g**
										.51	.14	.52	**s3g**
											.32	.89	**s4g**
												.32	**r3g**
													r4g

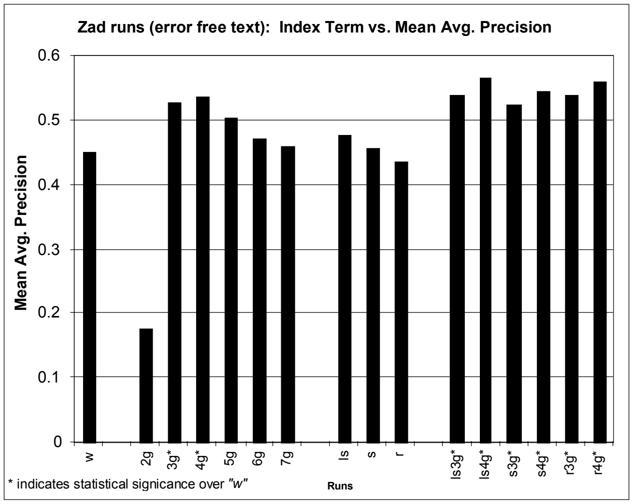


* indicates statistical signicance over "w" run

Figure 2: Runs on small-TREC collection

Table 2: Comparing small-TREC runs using the *t*-test's p-value.

3g	4g	5g	w	ls	s	r	ls3g	ls4g	s3g	s4g	r3g	r4g	
	.36	.09	**.00**	.29	**.04**	**.02**	.26	.93	.57	.91	.51	.98	**3g**
		.04	**.01**	.74	.09	.06	.98	.13	.72	.02	.87	.07	**4g**
			.02	.56	.27	.19	.39	**.00**	.27	**.00**	.43	***.01***	**5g**
				.00	.24	.39	**.00**	**.00**	**.00**	**.00**	**.01**	**.00**	**w**
					.10	.08	.73	.20	.55	.12	.71	.18	**ls**
						.27	**.05**	**.01**	**.02**	**.01**	**.03**	**.00**	**s**
							.03	**.01**	**.01**	**.01**	**.01**	**.00**	**r**
								.36	.43	.22	.81	.27	**ls3g**
									.49	.98	.41	.48	**ls4g**
										.40	.77	.46	**s3g**
											.39	.49	**s4g**
												.44	**r3g**
													r4g



* indicates statistical signicance over "w"

Figure 3: Runs on Zad collection

Table 3: Comparing Zad runs using the *t*-test's p-value.

3g	4g	5g	w	ls	s	r	ls3g	ls4g	s3g	s4g	r3g	r4g	
	.57	.34	**.03**	.09	**.04**	.07	.12	**.03**	.93	.15	.48	.06	**3g**
		.04	**.02**	.08	.08	.09	.85	**.03**	.69	.63	.93	.35	**4g**
			.08	.44	.29	.24	.16	**.01**	.45	.07	.23	**.03**	**5g**
				.52	.87	.76	**.02**	**.01**	**.04**	**.01**	**.02**	**.01**	**w**
					.58	.39	**.02**	**.01**	.11	**.02**	**.04**	**.01**	**ls**
						.47	**.02**	**.02**	**.02**	**.01**	**.01**	**.01**	**s**
							.04	**.03**	**.05**	**.03**	**.02**	**.01**	**r**
								.11	.35	.58	.95	.17	**ls3g**
									.23	.45	.32	.77	**ls4g**
										.21	.20	.06	**s3g**
											.71	.28	**s4g**
												.18	**r3g**
													r4g

We used a vector space information retrieval system with Okapi BM-25 term weights [18][2]. We chose to compare the effect of alternate indexing terms using mean uninterpolated average precision as our measure of retrieval effectiveness. We tested our results for statistical significance using a paired two-tailed *t*-test, claiming significance for *p* values at or below 0.05 (indicating—if the assumptions underlying the test were satisfied—that there would be no more than a 5% chance that the observed average precision values, paired by topic, were drawn from the same distribution). Sebawai, the light stemmer, and the IR system are freely distributable.

Figure 1 summarizes the results of these runs, and Table 1 presents the statistical significance test results for the *t* queries. These results indicate that if we were to pick a single index term, character 3-grams or 4-grams, or lightly or aggressively stemmed words would be good candidates. With few exceptions, all four

[2] IR engine, Sebawai and the Light Stemmer are available at www.glue.umd.edu/~kareem/research

types of terms statistically significantly outperformed words, roots, and character 5-grams, 6-grams and 7-grams on all three sets of queries. Any difference in effectiveness between the four types of terms could not be discerned using this test collection.

The average length of a stem (using aggressive stemming) is about 3.6 characters (the average length of an Arabic stem was derived by averaging the length of 250,000 stems produced by Sebawai). Therefore, it should not be surprising that 3 or 4 turn out to be the optimal *n*-gram length. Combining *n*-grams and with the results of morphological analysis in a single index was generally not harmful when compared to the use of either alone, although of course such a strategy necessarily results in a larger index. Interestingly, we found no evidence that longer queries do any better than shorter ones on this collection—in fact the trend is just the opposite (although none of the differences are statistically significant). For this reason, we focus on a single query type (title queries) in the remainder of our experiments.

Neither of the alternatives we tried for identifying roots outperformed Sebawai's top candidate, although neither difference is statistically significant for any query length. The fact that filtering Sebawai's candidates using ALPNET doesn't appear to help seems reasonable, since a prior study found that for words ALPNET could analyze, Sebewai's top candidate was right in 96% of the cases. The fact that weighting Sebawai's top two candidates did not help suggests that uncorrected likelihood values computed by Sebawai are not useful as probability estimates. For this reason, we present results only for Sebawai's first ranked roots (*r*) in the remainder of this paper. In some early work with small test collections [3, 13], roots had appeared to be a better choice than stems, but our experiments found just the opposite. One possible explanation for this is that earlier test collections contained at most a few hundred documents, and scaling up the size of the collection by several orders of magnitude might reward the choice of less ambiguous terms. An alternative explanation is that our morphological analysis might not be sufficiently accurate. In the earlier work, stems and roots had been obtained manually. For words that ALPNET fails to analyze (many of which are named entities), Sebawai often produces an incorrect analysis, ranking the correct root first in only 20% of the cases [9].

4. CREATING AN ARABIC DOCUMENT IMAGE COLLECTION

In this section we explore the development of an information retrieval test collection that contains Arabic document images. Our goal is to construct a collection with the following characteristics:

1. It should be built from actual printed sources (the third approach identified in Section 2), so that we can be confident that the effects of printing and scanning on the Arabic character recognition process have been accurately modeled;
2. It should include include accurate character-coded text (which we call "clean" text) for each document, so that we can check what we have learned about term selection using the TREC collection; and
3. It should either be available without charge or licensable at a reasonable price, so that others can build on our work.

We built our collection from *Zad Al-Me'ad*, a printed book that is free of copyright restrictions and for which an electronic copy could be obtained without charge from Al-Areeb Electronic Publishers [3]. 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. The first author of this paper, a native speaker of Arabic, 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. We refer to this collection as the Zad Collection.

We indexed and searched the clean (accurate character-coded) text in the Zad collection using the same range of character *n*-grams and terms obtained through morphological analysis as the TREC collection. Although the collection showed trends similar to those observed using title queries on the TREC collection, none of the differences were significantly significant. The failure to achieve statictical significance might be due to some obvious characteristic of the collection such as its size, or it might merely result from the fact that the two collections are drawn from different genres; perhaps affecting topic distinguishability (topics in religious texts might be diverse, but may frequently overlap because religious texts often have central underlying themes), or the way in which words are used (e.g., metaphorical usage may be more common in religious texts). To test the effect of collection size, we built a small collection (which we call small-TREC) by sampling documents from the full TREC collection in a manner that reflected the distribution of relevant documents in the Zad collection. We first randomly selected documents from the TREC collection to produce a 1% sample (a little over 2,900 documents). We then added relevant documents to approximate the distribution of relevant documents per topic in the Zad collection. The resulting small-TREC collection had 55,000 unique words, 4,200 unique roots, and an average document length of 186 words (compared to 62,000 unique words, 4,100 unique roots, and an average of 207 words per document for the Zad collection). We are not aware of any formal comparability measures for information retrieval test collections, but at least with respect to the parameters that we measured these two collections seem roughly comparable. Figure 2 and Table 2 show the results for the small-TREC collection using title queries. As with the full TREC collection, 3-grams, lightly stemmed words, and aggressively stemmed words typically outperformed words, roots, and 5-grams, with statistical significance in many cases. From this we conclude that our failure to obtain statistical significance on the Zad collection cannot be solely the result of collection size.

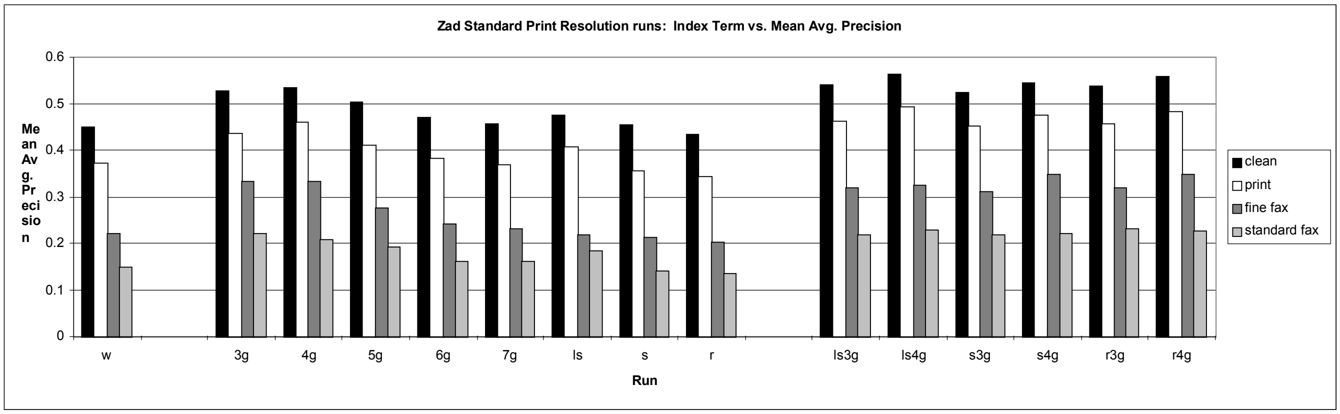


Figure 4: Runs on Zad at different degradation levels

Table 4: Comparing Zad--print resolution runs using the *t*-test's p-value.

3g	4g	5g	w	ls	s	r	ls3g	ls4g	s3g	s4g	r3g	r4g	
	.29	.30	**.05**	.32	**.03**	**.05**	.14	**.01**	.51	.07	.32	**.02**	**3g**
		.01	**.00**	.12	**.02**	**.04**	.85	**.01**	.78	.42	.93	.19	**4g**
			.06	.95	.24	.23	.06	**.00**	.22	**.02**	.15	**.01**	**5g**
				.34	.71	.54	**.02**	**.00**	**.01**	**.01**	**.01**	**.00**	**w**
					.11	.13	**.01**	**.00**	.12	**.02**	.08	**.01**	**ls**
						.58	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**s**
							.01	**.01**	**.01**	**.01**	**.01**	**.00**	**r**
								.04	.44	.40	.71	.17	**ls3g**
									.26	.53	.36	.79	**ls4g**
										.18	.57	.06	**s3g**
											.41	.33	**s4g**
												.15	**r3g**
													r4g

Table 5: Comparing Zad—fine fax resolution--runs using the *t*-test's p-value.

3g	4g	5g	w	ls	s	r	ls3g	ls4g	s3g	s4g	r3g	r4g	
	.93	**.03**	**.00**	**.00**	**.00**	**.00**	.43	.58	.18	.22	.48	.38	**3g**
		.01	**.00**	**.00**	**.00**	**.00**	.57	.55	.42	.23	.57	.27	**4g**
			.00	.07	.08	**.04**	.15	**.05**	.20	**.01**	.09	**.00**	**5g**
				.94	.81	.54	**.01**	**.00**	**.01**	**.00**	**.00**	**.00**	**w**
					.81	.57	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**ls**
						.45	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**s**
							.00	**.00**	**.00**	**.00**	**.00**	**.00**	**r**
								.74	.71	.15	.95	.24	**ls3g**
									.57	.50	.89	.49	**ls4g**
										.05	.62	.10	**s3g**
											.15	.93	**s4g**
												.08	**r3g**
													r4g

One of the problems with the Zad collection is the existence of two topics with one relevant document and one topic with none[3]. Topics with no relevant documents certainly occur in real applications, but they cannot differentiate between systems. More importantly, uninterpolated average precision is subject to severe quantization effects when only a single relevant document is known (in this case, the value is always 1/rank, which can only take values of 1, 0.5, 0.33, 0.25, … etc.). We therefore decided to create new topics to replace these three problematic topics. The final set of 25 queries has an average length of 5.4 words and an average of 20 relevant documents per topic (with a minimum of 3 and a maximum of 72). Figure 3 and Table 3 show that after this change, statistical significance tests become informative. We now see combinations of *n*-grams and terms obtained though morphological analysis emerging as good choices, with the combination of 4-grams and lightly stemmed words producing results that are statistically significantly better than any single type of indexing term. From this we conclude that the modified Zad collection (which we henceforth refer to simply as the Zad collection) is a useful tool for exploring the effects of term selection on retrieval from scanned Arabic text.

5. RESULTS FOR RETRIEVAL OF OCR DEGRADED TEXT

In this section, we examine the effect of OCR on retrieval effectiveness. We scanned the 2,000 pages of the printed version of *Zad Al-Me'ad* using a Xerox Document Centre 460 Digital Copier, which is a high-volume copier and sheet feed scanner, in about two hours at 300x300 dpi (dots per inch) resolution. This corresponds to the resolution used commonly in older inkjet and laser printers, so we refer to this as "print" resolution. The images were then manually zoned into images of a single document (which might span multiple pages or parts of pages) to correspond exactly to the 2,730 documents in the character-coded clean copy of the collection. Two additional document image collections were then created using Corel PhotoPaint version 8.0 by resampling the zoned document images at both 203x196 and 203x98 dpi, which correspond to the fine and standard resolutions respectively of fax machines in the widely used "Group 3" standard. We then used Sakhr's Automatic Reader version 4.0 OCR engine [19] to convert the images into plain text. We computed the character error rate (insertions + deletions + substitutions) for the OCR degraded text with reference to the clean text using software from the University of Nevada at Las Vegas [17], obtaining 18.7%, 36.1%, and 42.4% for the print, fine fax, and standard fax resolutions respectively. Sakhr OCR engine is tuned for high resolution document images (corresponding to

[3] small-TREC had one topic with one relevant document and one topic with none.

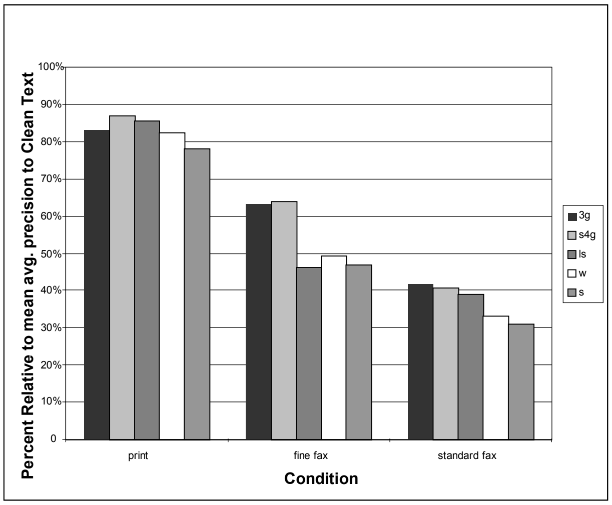


Figure 5: Condition vs. percent relative uniterpolated mean average precision to clean text

our print resolution), so the relatively high character error rate of the fine and standard fax resolution is understandable. Better results might be obtained if Sakhr's OCR engine was retrained on samples of the fine and standard fax resolutions prior to the recognition of the full collection (Sakhr's OCR software allows for manual retraining by the user).

Figure 4 compares the results of all the runs for all four version of the Zad collection (clean, print, fine fax, and standard fax versions), and Tables 4, 5, and 6 show the results of statistical significance testing for print, fine fax, and standard fax resolutions respectively. Higher character error rates clearly produced a substantial (and statistically significant) adverse effect on the retrieval effectiveness. Nonetheless, our system demonstrated the ability to reliably find some relevant documents with even the lowest resolution. This is consistent with results that have been obtained in other languages (c.f., [5]). Index terms that performed well for clean text continued to do so over the range of character error rates that we tested, with character 3-grams and 4-grams and all combinations of character n-grams with terms obtained using morphological analysis producing good results.

Table 6: Comparing Zad—standard fax resolution--runs using the *t*-test's p-value.

3g	4g	5g	w	ls	s	r	ls3g	ls4g	s3g	s4g	r3g	r4g	
	.54	.21	.03	.33	.03	.03	.89	.70	.95	.91	.30	.63	**3g**
		.14	.00	.36	.06	.06	.65	.19	.60	.27	.27	.22	**4g**
			.01	.67	.11	.10	.24	.09	.19	.04	.08	.02	**5g**
				.44	.80	.66	.03	.01	.02	.00	.01	.00	**w**
					.33	.26	.21	.11	.33	.20	.24	.20	**ls**
						.43	.01	.03	.01	.01	.01	.01	**s**
							.01	.03	.02	.01	.02	.01	**r**
								.59	.92	.75	.44	.52	**ls3g**
									.40	.77	.68	.76	**ls4g**
										.85	.07	.44	**s3g**
											.51	.56	**s4g**
												.67	**r3g**
													r4g

Figure 5 shows some of the same results grouped by resolution, illustrating that *n*-grams produce a substantial (and statistically significant) advantage over any term obtained using morphological analysis for the fine fax resolution. With the increase of error level, retrieval effectiveness of indexing using words, stems, and roots deteriorated faster than indexing using 3-grams, 4-grams, or combinations of n-grams and terms obtained using morphological analysis. This effect may result from weaknesses in the Sebawai morphological analyzer, which was trained using clean text, resulting in failure to analyze words that have one or more misrecognized characters. Character *n*-grams are relatively robust as long as the character error rate is low enough to yield many correct sequences of contiguous, but ultimately n-gram performance decays as well. Similar results have been observed in other languages [12, 24]. Consistent with previous research [5], Figure 5 suggests that retrieval effectiveness deteriorates faster for higher error rates. Surprisingly, however, the three character error rate values that we obtained did not allow us to distinguish between 3-grams and 4-grams.

6. CONCLUSIONS AND FUTURE WORK

We have described the development of a test collection that can be used to evaluate alternative techniques for searching scanned Arabic text and a set of experiments that were designed to identify the effect of alternative indexing terms on retrieval effectiveness. From our experiments, we can conclude that character *n*-grams (specifically 3-grams and 4-grams) are well suited to Arabic document image retrieval applications, and that combining *n*-grams with terms obtained through some form of morphological analysis (particularly lightly stemmed words) can produce a robust system that is effective over a range of genres, collection sizes, and character error rates.

Searching printed Arabic is an important problem, in part because a great deal of Arabic text that is presently available only in printed form. We expect that our new Zad test collection will provide a useful basis for further work including:

- Improved morphological analysis. Because of the way it was trained, Sebawai does well on words that ALPNET can analyze, but poorly on most other words. Many of these problematic words are named entities, so we expect that substantial improvements could be made if named entities could be reliably detected and routed to an analysis system that is tuned to work well on such terms.
- Improved retrieval algorithms. Singhal has shown that for English byte length normalization is more robust to character recognition errors than the cosine normalization usually used in vector space retrieval systems [20], and Tseng and Oard have seen similar results for Chinese [24].
- Larger test collections. The development of good error models for Arabic OCR would make it possible to generate a large test collection directly from the TREC collection or from another collection of similar size. When this becomes possible, the insights that we have gained using the Zad collection can serve as a useful point of departure for further exploration of alternate indexing terms.
- Automatic layout analysis. In our experiments, individual documents within *Zad Al-Me'ad* were segmented manually, and no special processing was required to determine the appropriate reading order. Automatic layout analysis will,

however, be needed in many practical applications (e.g., searching printed newspapers).

- Image enhancement for low-resolution applications. Faxes, video captions, and scene text in video have significantly lower resolution than ordinary scanned documents, and video applications often also include unusual background characteristics. Image processing techniques such as mathematical morphology and multi-frame integration can be helpful in such cases.

Arabic is the eighth most widely spoken language in the world, and improved access to Arabic text could have profound implications for cross-cultural communication and economic development. At present, the dissemination of Arabic text is dominated by printed rather than character-coded electronic documents, which makes retrieval from document images a particularly salient task for Arabic. Both the morphology and the orthography of Arabic make this particularly challenging, but we believe that the results reported in this paper form a solid basis for future work on this important problem.

ACKNOWLDGEMENTS

The authors are grateful to Dr. Phillip Resnik, Mona Diab, Nizar Habash, and David Doermann for his valuable input. This work was supported in part by DARPA cooperative agreement N660010028910.

BIBLIOGRAPHY

1 Ahmed, Mohamed Attia. A Large-Scale Computational Processor of the Arabic Morphology, and Applications. Master's Thesis, Faculty of Engineering, Cairo University, Cairo, Egypt, 2000.

2 Al-Areeb Electronic Publishers, LLC. 16013 Malcolm Dr., Laurel, MD 20707, USA

3 Al-Kharashi, Ibrahim and Martha Evens. Comparing Words, Stems, and Roots as Index Terms in an Arabic Information Retrieval System. JASIS. 45 (8): 548-560, 1994.

4 Aljlayl, M., S. Beitzel, E. Jensen, A. Chowdhury, D. Holmes, M. Lee, D. Grossman, and O. Frieder. IIT at TREC-10. TREC-2001, 2001.

5 Baird, Henry. Document Image Defects Models and their Uses. Proceedings of the Second International Conference on Document Analysis and Recognition (ICDAR), 62-67, 1993.

6 Baird, Henry. State of the Art of Document Image Degradation Modeling. Proceedings of the 4th IAPR Workshop on Document Analysis Systems (DAS 2000), 2000.

7 Beesley, Kenneth. Arabic Finite-State Morphological Analysis and Generation. COLING-96, 1996.

8 Chen, Aitao and F. Gey. Translation Term Weighting and Combining Translation Resources in Cross-Language Retrieval TREC-2001, 2001.

9 Darwish, Kareem. Building a Shallow Morphological Analyzer in One Day. To appear in ACL 2002 Workshop on Computational Approaches to Semitic Languages, July 11, 2002.

10 Darwish, Kareem, D. Doermann, R. Jones, D. Oard, and M. Rautiainen. TREC-10 Experiments at Maryland: CLIR and Video. TREC-2001, 2001.

11 Gey, Fredric and D. Oard. The TREC-2001 Cross Language Retrieval Track: Searching Arabic using English, French, and Arabic Queries. TREC-2001, 2001.

12 Harding, S., W. Croft, and C. Weir. Probabilistic Retrieval of OCR Degraded Text Using N-Grams. European Conference on Digital Libraries, 1997

13 Hmeidi, Ismail, Ghassan Kanaan, and Martha Evens. Design and Implementation of Automatic Indexing for Information Retrieval with Arabic Documents. JASIS. 48 (10): 867-881, 1997.

14 Kanungo, Tapas. Document Degradation Models and Methodology for Degradation Model Validation. Ph.D. Thesis, Electrical Engineering Department, University of Washington, 1996.

15 Kanungo, Tapas, Gregory Marton, and Osama Bulbul. OmniPage vs. Sakhr: Paired Model Evaluation of Two Arabic OCR Products. Proceedings of SPIE Conference on Document Recognition and Retrieval (VI), Vol. 3651, San Jose, California, Jan. 27-28, 1999.

16 Mayfield, James, P. McNamee, C. Costello, C. Piatko, and A. Banerjee. JHU/APL at TREC 2001: Experiments in Filtering and in Arabic, Video, and Web Retrieval. TREC-2001, 2001.

17 Rice, S., Frank R. Jenkins, and Thomas A. Nartker. The fifth annual test of OCR accuracy. Technical Report 96-01Information Science Research Institute, University of Nevada, Las Vegas, April 1996.

18 Robertson, S. and K. S. Jones. Simple proven approaches to text retrieval. Tech. Rep. TR356, Cambridge University Computer Laboratory, 1997

19 Sakhr Technologies, Cairo, Egypt www.sakhr.com

20 Singhal, Amit, Gerard Salton, and Chris Buckley. Length Normalization in Degraded Text Collections. Proceedings of 5th Annual Symposium on Document Analysis and Information Retrieval, 149-162, April 15-17, 1996.

21 Taghva, Kazem, Julie Borasack, Allen Condit, and Jeff Gilbreth. Results and Implications of the Noisy Data Projects. Technical Report 94-01, Information Science Research Institute, University of Nevada, Las Vegas, 1994.

22 Taghva, Kazem, Julie Borasack, Allen Condit, and Padma Inaparthy. Querying Short OCR'd Documents. Technical Report 94-10, Information Science Research Institute 1995.

23 Trenkle, John, Andrew Gillies, Erik Erlandson, Steve Schlosser, and Stan Cavin. Advances in Arabic Text Recognition. Proceeding of Symposium on Document Image Understanding Technology, Columbia, Maryland, April 23-25, 2001.

24 Tseng, Yuen-Hsien and Douglas Oard. Document Image Retrieval Techniques for Chinese. Proceeding of Symposium on Document Image Understanding Technology, Columbia, Maryland, April 23-25, 2001.

25 Xu, Jinxi, A. Fraser, and R. Weischedel. TREC 2001 Cross-Lingual Retrieval at BBN. TREC-2001, 2001.