Building a Heterogeneous Information Retrieval Test Collection of Arabic Document Images

Kareem Darwish, Walid Magdy, Ossama Emam

IBM Technology Development Center
P.O. Box 166 El-Ahram, Giza, Egypt
{darwishk,wmagdy,emam}@eg.ibm.com

Abdelrahim Abdelsapor

Institute of Statistical Research
Cairo University,
5 Ahmed Zuwail St., Giza, Egypt
ar_elmadany@hotmail.com

Noha Adly, Magdi Nagi

Bibliotheca Alexandrina
P.O. Box 138, Chatby,
Alexandria, Egypt
{magdy.nagi,noha.adly}@bibalex.org

Abstract

This paper describes the development of an Arabic document image collection containing 34,651 documents from 1,378 different books and 25 topics with their relevance judgments. The books from which the collection is obtained are a part of a larger collection 75,000 books being scanned for archival and retrieval at the Bibliotheca Alexandrina (BA). The documents in the collection vary widely in topics, fonts, and degradation levels. Initial baseline experiments were performed to examine the effectiveness of different index terms, with and without blind relevance feedback, on Arabic OCR degraded text.

1. Introduction

Since the advent of the printing press in 15th century the number of printed documents has overwhelmingly grown. Only recently has electronic text become ubiquitous. Electronic text is usually easy to search and retrieve which led to the development of many text search engines. Nonetheless, there remains a huge volume of legacy documents which are available in print only. One way to search and retrieve printed documents is by digitizing them and performing Optical Character Recognition (OCR) to transform the digitized printed documents (a.k.a. document images) into electronic text. Although the OCR process is not perfect and produces many errors, especially for orthographically and morphologically complex languages such as Arabic, it produces a text representation of the document images that can be searched.

As part of an international effort called the Million Book Project, the Bibliotheca Alexandrina (BA) was tasked with scanning and OCR'ing 75,000 Arabic books to make them accessible and searchable. A subset of the available document images was extracted to build a test collection to aid in the development of retrieval techniques that are suited for OCR degraded text retrieval. The test collection is composed of 34,651 document images from 1,378 different books. An effort was made to insure a diversity of degradation levels, genres, and fonts. This test collection is the largest such Arabic collection of document images with associated OCR output, topics, and relevance judgments.

This paper describes the test collection that was developed and reports on baseline runs exploring the retrieval effectiveness of indexing using words, light stems, character 3-grams, and character 4-grams with and without blind relevance feedback. The paper is organized as follows: Section 2 provides a background on previous work in OCR degraded text retrieval and test collection construction; Section 3 describes the collection and provides the experimental setup; Section 4 reports the results and discusses them; and section 5 concludes the paper.

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 (Singhal et al., 1996), and then to 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) (Baird, 2000), 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 (Baird, 1993). 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 (Trenkle et al., 2001), 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 (Ahmed, 2000). 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 (Kanungo et al., 1999). Retrieval of OCR degraded text documents has been reported for many languages, including English,

French, Spanish, Chinese, and Arabic (Darwish & Oard, 2002; Harding et al., 1997; Taghva et al., 1995; Tseng & Oard, 2001).

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 (Harding et al., 1997).
- 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 (Baird, 2000; Kanungo, 1996), 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 (Baird, 1993).
- 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 with 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 (Taghva et al., 1994). 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. Further, they reported that blind relevance feedback did not improve retrieval effectiveness (Tseng & Oard, 2001). Darwish and Oard experimented with a variety of Arabic index terms on an Arabic collection 2,730 document images. The documents were scanned from a single book. They reported that 3-grams and 4-grams are the best index terms for OCR degraded Arabic text (Darwish & Oard, 2002).

To develop relevance judgments, there are several methods reported in the literature. Some of the methods reported are:

- Exhaustive search: due to the required amount of manual processing, relevance judgments developed using this method was restricted to small collections and was reported not be feasible for larger collections (Jones & Van Rijsbergen, 1975).
- Pooling: pooling involves the participation of a "significantly" diverse set systems in which the same topics are provided to all the systems and the top *n* retrieved results from each system are pooled and judged. This method is used by different evaluations such as the ones at TREC (Oard & Gey, 2002).
- Interactive Search and Judge (ISJ): ISJ technique, which was developed by Cormack et al., allows a judge to search the collection with different reformulations of topic expressions (Cormack et al., 1998). The judge continues to search until he/she is confident that all or most relevant documents are found.
- Iterative Search and Judge: in this technique, the judge is not required to manually reformulate topic expressions and the formulation is done automatically using relevance feedback. This method, which was developed and verified by Sanderson and Joho (2004), entails performing an initial search and then manually examining the top 100 retrieved documents. All the documents that are deemed relevant are used to reformulate the original queries. This process is repeated 5 times for each topic.

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 (Darwish, 2002).

Several types of index terms have been studied, including word surface forms, clusters of words (Larkey et al., 2002), resultants of morphological processing, such as stems and morphological roots (Al-Kharashi & Evens, 1994; Aljlayl et al., 2001; Hmeidi et al., 1997), and character n-grams of various lengths (Darwish et al., 2001; Mayfield et al., 2001). The effects of normalizing alternative characters, removal of diacritics and stop-words have also been explored (Chen & Gey, 2001; Darwish et al., 2001; Xu et al., 2001). The preponderance of the evidence suggests that some form of morphological analysis and/or the use of character n-grams substantially outperform use of word surface forms, and that some form of character normalization is helpful.

3. The Test Collection

This section describes the attributes of the test collection, the method of creating the topics and relevance judgments, and preliminary retrieval experiments on the collection.

3.1. The Collection

The collection was built by randomly picking approximately 25 pages from 1,378 Arabic books from the BA forming a set of 34,651 printed documents. The books cover a variety of topics including historical,

philosophical, cultural, and political and the printing dates of the books range from the early 1920's to the present. The documents were converted to document images by scanning them in black and white at 300x300 dpi using the Minolta PS 3000 book scanner. The scanning was done as a part of the Million Book Project in which the BA is responsible for scanning 75,000 Arabic documents. The document images were subsequently OCR'ed using Sakhr's Automatic reader (version 6). The OCR text had character error rates ranging between 1% and 21% for different books. The character error rate was estimated by manually examining a random page from each book. The fonts used in the books were divided into 12 different font classes, which correspond to the most popular fonts used in print, and a 13th class containing rare fonts. The variations in degradation levels, fonts, genres, the selective use of diacritics, and existence of non-textural graphics in many pages make the retrieval of the OCR'ed text more challenging. Figure 1 shows a sample image and Figure 2 shows the corresponding OCR output.

Figure 1: Sample page

من ناحيتى المراقبة والنيران على السهل الساحلى.
وطرق الاقتراب التى تسلكها أى قوات عربية من ناحية الشرتى تنح ر دى طرق خمسة أهمها الثلاثة التالية:
ا- الطريق ا لأول وهو ا لأقصر من بغداد- هـ 2233 " المفرتى.، أو ا لانحراف إلى
الرطبة قبل 3أ3 دمشق الأردن،
محاور ا لارابى من العرا إلى سوريا وا لأردد
2- الطريق الاثانى من بغداد " أبو كمال- بالميرا- دمشق- ا لأردد.
3- الطريق الثالسث وهو الأطول من بغ داد- الموصل- دير الزور- حملرو- دمشق- ا لأردن " 68

Figure 2: Sample OCR output

3.2. The Topics and Relevance Judgments

A set of 25 topics were developed for the collection along with relevance judgments, which map between the topics and the documents that relevant to them. Each of the topics includes a title field, which is similar to web queries, and a description field, which is a natural language statement that a user might give to a librarian. The topics were intentionally diverse historical, religious, sociological, legal, educational, and political issues. Figure 3 shows a sample topic.

<num> Number: 24
<title>
علاقة جمال الدين الأفغاني بالدولة العثمانية
<desc> Description:
تطور العلاقة بين الدولة العثمانية وجمال الدين الأفغاني ورأي كلا الطرفين في الآخر

Figure 3: Sample topic about the relationship of Jamal-u-Deen Al-Afghani and the Ottoman Empire

To perform the judgments a special interactive retrieval web-based application was developed. The application allows users to search the collection using multiple query formulations while tracking document judgments that were made for each topic, selectively performing automatic query expansion, and viewing of the document images with queries terms highlighted. Figure 4 shows a screenshot of the system. The systems uses Lemur Language Modeling Toolkit (version 3.1), a vector space model IR engine, on the backend with OKAPI BM25 weighting scheme with Lemur's default parameters, and character 4-grams as index terms.

The number of relevant documents for the topics ranged between 4 and 95, with each topic having 37 relevant documents on average. The collection of documents with the associated topics and relevance judgments is larger than any previously reported on test collection of Arabic document images.

3.3. Preliminary Experiment Design

The collection was indexed 4 times using 4 different index terms namely words, light stems, character 3-grams, and character 3-grams. For all index terms, in both the documents and the queries, all forms of *alef* (*hamza*, *alef, alef maad*, *alef with hamza on top*, *hamza on w*a, *alef with hamza on the bottom*, and *hamza on ya*) were normalized to *alef* and *ya* and *alef maqsoura* were normalized to *ya*. For light stemming, Al-Stem was used without modification. Al-Stem removes a common list of prefixes and suffixes. For character 3-grams and 4-grams, a non-word overlapping sliding window would select 3 or 4 character respectively as index terms (example: jasmine => jasm, asmi, sami, amin, mine). Light stems and character and 3 and 4 grams were reported in earlier studies to be the best index terms for Arabic text with the character 3 and 4 grams being the best for OCR degraded Arabic text (Darwish & Oard, 2002).

For each of the index terms, two runs were performed. The first involved retrieving without blind relevance feedback, while the second involved using blind relevance feedback.

Figure 4: Web-based system for searching, displaying, and judging document images

Blind relevance feedback involves performing a search of the collection, then augmenting the user query with the *m* most valuable terms from the top *n* retrieved documents, and lastly searching using the new expanded query. In doing so, the top *n* retrieved documents are assumed to be relevant and the value of the terms is determined using the retrieval weighting scheme. For the experiments in this paper, the top 20 terms from the top 5 retrieved documents were used to augment the user's initial query.

For the experiments, the Indri vector space model information retrieval engine, which is a derivative of the Lemur Language Modeling Toolkit project, was used with the default parameters for the Okapi BM-25 weighting. For each topic, the top 1,000 document were retrieved and the retrieved effectiveness was evaluated using mean average precision. Also, the top 10 retrieved documents from each experiment were evaluated using precision 10. To determine if the difference between results was statistically significant, a paired two-tailed *t*-test was used with *p*-values less than or equal to 0.05 to claim significance.

4. Results

The results offer baseline retrieval effectiveness scores for the different index with and without blind relevance feedback on the test collection. Figure 5 and Figure 6 report on the results using mean average precision and precision at 10 respectively.

The following tables provide *t*-test *p*-values of comparing the different index terms with and without blind relevance feedback.

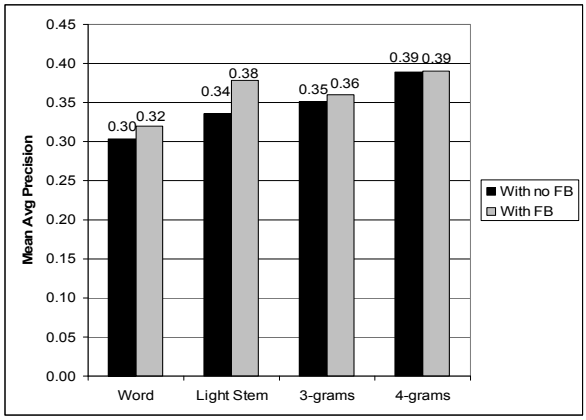

Figure 5: Retrieval Effectiveness for different index terms with and without blind relevance feedback using mean average precision

Table 1: *p*-values of comparing different index terms without blind relevance feedback using mean average precision (shaded *p*-values indicate statistical significance)

Stem	3-gram	4-gram	
0.23	0.13	0.00	**Word**
	0.54	0.03	**Light Stem**
		0.02	**3-gram**

Table 2: *p*-values of comparing different index terms with blind relevance feedback using mean average precision (shaded *p*-values indicate statistical significance)

Stem	3-gram	4-gram	
0.08	0.28	0.04	**Word**
	0.56	0.58	**Light Stem**
		0.20	**3-gram**

Table 3: *p*-values of comparing effect of blind relevance feedback on different index terms using mean average precision (shaded *p*-values indicate statistical significance)

Word	Light Stem	3-gram	4-gram
0.34	0.02	0.71	0.94

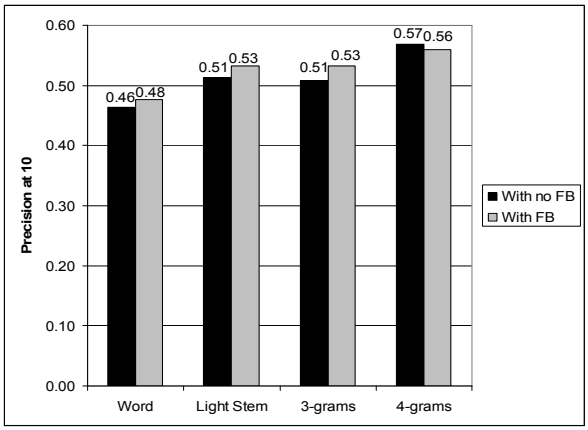

Figure 6: Retrieval Effectiveness for different index terms with and without blind relevance feedback using precision @ 10

Table 4: *p*-values of comparing different index terms without blind relevance feedback using precision @ 10 (shaded *p*-values indicate statistical significance)

Stem	3-gram	4-gram	
0.29	0.30	0.01	**Word**
	0.93	0.18	**Light Stem**
		0.05	**3-gram**

Table 5: *p*-values of comparing different index terms with blind relevance feedback using precision @ 10 (shaded *p*-values indicate statistical significance)

Stem	3-gram	4-gram	
0.16	0.30	0.03	**Word**
	1.00	0.40	**Light Stem**
		0.51	**3-gram**

Table 6: *p*-values of comparing effect of blind relevance feedback on different index terms using precision @ 10 (shaded *p*-values indicate statistical significance)

Word	Light Stem	3-gram	4-gram
0.73	0.38	0.38	0.81

The results show that character 4-grams are the best index terms, statistically significantly outperforming all other index terms when no blind relevance feedback is used regardless of the effectiveness metric used (except for light stems when measuring effectiveness using precision at 10). The results also show that blind relevance feedback had a small and statistically insignificant effect on retrieval effectiveness. This result is inline with previous findings reported in the literature for Arabic and Chinese (Darwish & Emam, 2005; Tseng & Oard, 2001).

5. Conclusion

This paper presented the development of an Arabic OCR degraded collection and provided baseline results on the effect of using different index terms with and without using blind relevance feedback. The results show that character 4-grams provided the best results and that blind relevance feedback had a minor effect on retrieval effectiveness.

This collection is the largest known test collection of document images with associated OCR text, topics, and relevance judgments. The collection is a valuable resource for exploring different methods for improving the retrieval of document images and OCR degraded documents.

6. References

Ahmed, M. A large-scale computational processor of the Arabic morphology, and Applications. *Master's Thesis, Faculty of Engineering, Cairo University*, Cairo, Egypt, 2000.

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

Al-Kharashi, I. and M. Evens. Comparing words, stems, and roots as index terms in an Arabic information retrieval system. *JASIS.* 45 (8): 548-560, 1994.

Baird, H. Document image defects models and their uses. *Proceedings of the Second International Conference on Document Analysis and Recognition (ICDAR)*, 62-67, 1993.

Baird, H. State of the art of document image degradation modelling. *Proceedings of the 4th IAPR Workshop on Document Analysis Systems (DAS 2000)*, 2000.

Chen, A. and F. Gey. Translation term weighting and combining translation resources in cross-language retrieval. *TREC-2001*, 2001.

Cormack, G., C. Palmer, and C. Clarke. Efficient Construction of Large Test Collections. *In the Proceedings of the 21st ACM SIGIR Conference*, 282-289, 1998.

Darwish, K. Building a shallow morphological analyzer in one day. *ACL 2002 Workshop on Computational Approaches to Semitic Languages*, July 11, 2002.

Darwish, K. and O. Emam. The Effect of Blind Relevance Feedback on a New Arabic OCR Degraded Text Collection. *In International Conference on Machine Intelligence: Special Session on Arabic Document Image Analysis*, 2005.

Darwish, K., D. Doermann, R. Jones, D. Oard, and M. Rautiainen. TREC-10 experiments at Maryland: CLIR and video. *TREC-2001*, 2001.

Darwish, K., D. Oard, Term selection for searching printed Arabic. *In the Proceedings of the 25th ACM SIGIR Conference*, page 261 - 268, 2002.

Darwish, K., Probabilistic Methods for Search OCR Degraded Arabic Text. *Ph.D. dissertation*, *University of Maryland, College Park*, 2003.

Harding, S., W. Croft, and C. Weir. Probabilistic retrieval of OCR degraded text using n-grams. *European Conference on Digital Libraries*, 1997

Hmeidi, I., G. Kanaan, and M. Evens. Design and implementation of automatic indexing for information retrieval with Arabic documents. *JASIS*. 48 (10): 867-881, 1997.

Jones, K. S. and C. J. Van Rijsbergen. Report on the need for and previous of 'ideal' test collection, *TR #365, University Computer Laboratory*, Cambridge, 1975.

Kanungo, T. Document degradation models and methodology for degradation model validation. *Ph.D. Thesis, Electrical Engineering Department, University of Washington*, 1996.

Kanungo, T., G. Marton, and O. 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.

Larkey, L., L. Ballesteros, and M. Connell. Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis. *In SIGIR, 2002*, Tampere, Finland. pp. 275-282

Mayfield, J., 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.

Oard, D. and F. Gey. The TREC-2002 Arabic/English CLIR Track. *TREC-2002*, 2002.

Sanderson, M. and H. Joho. Forming Test Collection with no System Pooling. *In the Proceedings of the 27th ACM SIGIR Conference*, page 33-40, 2004

Singhal, A., G. Salton, and C. Buckley. Length normalization in degraded text collections. *Proceedings of 5th Annual Symposium on Document Analysis and Information Retrieval*, 149-162, April 15-17, 1996.

Taghva, K., J. Borasack, A. Condit, and J. Gilbreth. Results and implications of the noisy data projects. *Technical Report 94-01, Information Science Research Institute, University of Nevada, Las Vegas*, 1994.

Taghva, K., J. Borasack, A. Condit, and P. Inaparthy. Querying short OCR'd documents. *Technical Report 94-10, Information Science Research Institute* 1995.

Trenkle, J., A. Gillies, E. Erlandson, S. Schlosser, and S. Cavin. Advances in Arabic text recognition. *Proceeding of Symposium on Document Image Understanding Technology*, Columbia, Maryland, April 23-25, 2001.

Tseng, Y. and D. Oard. Document image retrieval techniques for Chinese. *Proceeding of Symposium on Document Image Understanding Technology*, Columbia, Maryland, April 23-25, 2001.

Xu, J., A. Fraser, and R. Weischedel. TREC 2001 cross-lingual retrieval at BBN. *TREC-2001*, 2001.