Novel Arabic OCR Degraded Text Retrieval Model

Mostafa Ezzat a, Tarek Elghazaly b, Mervat Gheith c

Computer Sciences Department, Institute of Statistical Studies and Research, Cairo University, Egypt
Mostafa.ezzat@cu.edu.eg, Tarek.ElGhazaly@cu.edu.eg, MGheith@issr.cu.edu.eg

Abstract

This paper provides a novel model enhances the Arabic OCR degraded text retrieval effectiveness. The model simulates the Arabic OCR recognition mistakes happens while the recognition process based on word based approach. Then using the expected OCR errors the model expands the user search query. The resulting expanded search query produced higher precision and recall in searching Arabic OCR-Degraded text rather than the original query. The new model showed a significant increase in the retrieval effectiveness over the previous models. The retrieval effectiveness of the new model is %97, while the best effectiveness published for word based approach was %84 and the best effectiveness for character based approach was %56. In addition, the new model overcomes several limitations of the current two existing models.

Keywords: Arabic OCR Degraded Text Retrieval, Arabic OCR-Degraded Text, Orthographic Query Expansion, Synthesize OCR-Degraded Text.

1. Introduction

The amount of printed material increased tremendously in the fifteenth century. Many documents continue to be available only in print, although the number of documents available as character-coded text is now increasing as a result of electronic publishing, especially for Arabic. There were unusual technological hurdles until recently, including limited computer infrastructure and lack of Arabic support in Web browsers and popular operating systems. Further, different proprietary formats for Arabic text encoding were done in many. All these factors mean that users often access only the printed documents or images (scanned images of the printed documents). In fact, finding character-coded Arabic text instead of Arabic document images on the Web is still uncommon. This makes searching documents by user query very complicated.

Without automated process, the user would be only able to manually find the desired documents or he can consult a person who is familiar with the searched documents. Since such search can easily be done by searching character coded documents, and we can automate the process by generating the character-coded representations of the documents. In principal, we can generate the character-coded representation of the documents by rekeying the documents’ text or creating metadata about the documents such as titles, summaries, or keywords. But these approaches would be labor intensive and impractical for large numbers of documents. Another automated way to produce a character-coded representation is to scan the documents and then use Optical Character Recognition (OCR), which is an automated process that converts document images into character-coded text. However, the results often contain errors. Further, Arabic properties like orthography, which is how words are written and complex morphology, which is how words are constructed, adversely affect the accuracy of OCR. On the other hand, the OCR process is inexpensive and thus well suited for large document collections.

1.1. Orthographic Properties of Arabic

Arabic is a right-to-left connected language that uses 28 letters, and its shapes change depending on their positions in words. Fifteen letters contain dots to differentiate them from other letters. And depending on the discretion of the document producer, these letters may or may not have diacritics (short vowels). Ligatures, which are special forms
for some character sequences, and kashidas, which are symbols that extend the length of words, are often employed in printed text. Having a connected script that uses ligatures and kashidas complicates isolation of individual letters in OCR, and dots and diacritics make OCR sensitive to document noise and speckle. Figure 1 shows some examples of letters with different dots, diacritic, and letters in different positions, kashida, and ligature [3]. Most Arabic OCR systems segment characters [4, 5, 6, 7], while a few opted to recognize words without segmenting characters [8, 9], and another system developed by BBN avoids character segmentation by dividing lines into slender vertical frames (and frames into cells) and uses an HMM recognizer to recognize character sequences [10].

![Diagram](image)

Fig. 1: (a) A ligature, (b) Shapes of the letter “ba” and (c) a diacritic and kashida[23]

This paper presents OCR-Degradation synthesizing model that is trained on different shapes of the words generated from the OCR recognition process and the model is independent of the technique used in the OCR process itself. Then using the training information, we reformate the user search query by inserting all the possible shapes of each query word to generate the search query that will return all the documents related to the original search query. And in the following section we will introduce the previous work done in this direction and also we will compare two different models that use a similar approach to enhance the degraded text information retrieval.

2. Comprehensive Study for the Previous Work

Building a model that simulates the OCR-Degraded text covers an important part of information retrieval when the retrieved text is returned from the OCR process. The OCR process in this case is affected by many factors like the Paper quality which is affected by the time, human errors done during the scanning process, the accuracy of the OCR system used and the font which the document was initially printed with.

All these factors and others affects the information retrieval process done on the text resulted from the OCR process, so modeling the OCR-Degraded text is a major part in any research that deals with information retrieval. In this section we will present two models were developed by ElGhazaly and Darwish as parts of their research in the information retrieval in the same language or in different languages. To the best of our knowledge, they are the only models built specially or Arabic OCR-Degraded Text retrieval.

2.1. ElGhazaly OCR-Degraded Synthesizing Model

The model is a Word Based model that is trained and tested on complete words. ElGhazaly generated documents of single word per line, and then print, scan and OCR these documents using Sakhr OCR, then he manually align the OCR recognition results with the clean text document. During this manual alignment he checked if there is a deformation of the OCR-Degraded word, then he stored the deformed OCR-Degraded Word and its original word as a training pair in the OCR Errors Database [2, 22].

Then ElGhazaly used this OCR Error model to expand the query used in the information retrieval. This Orthographic Query Expansion approach attempts to find different misrecognized versions of a query word in the collection being searched. But ElGhazaly didn’t proof the retrieval effectiveness of his model.

2.2. Darwish OCR-Degraded Synthesizing Model

Darwish model is based on the character level not the word level as ElGhazally. The main idea of the model was based on the fact that the context affects the character shape in a word, because Arabic letters are connected and change shape depending on their position in the word and some special shapes are formed when special characters are sequenced in the same word, so his assumption was that the position of a letter being recognized and the letters surrounding it (“context”) would be important in developing a good model.

The model was simply based on aligning the OCR recognition results from the print(300X300), 200x200, and 200x100 versions of the Zad collection (the training and testing data set as we will mention later in this section) with the associated clean text version of the same collection. The alignment was done using SCLITE which is an application from the National Institute of Standards and Technology (NIST), which employs a dynamic programming string alignment algorithm that attempts to minimize the edit distance between two strings. Basically, the algorithm uses identical matches to anchor alignment, and then uses word position with respect to those anchors to estimate an optimal alignment on the remainder of the words. Two factors affected his alignment process. The printed and clean text versions in the Zad collection were obtained from different sources that exhibited minor differences (mostly substitution or deletion of particles such as in, from, or, and then), and secondly some areas in the scanned images of the printed page exhibited image distortions that resulted in relatively long runs of OCR errors.
Based on the alignment done using the SCLITE and the other algorithm Darwish implemented, he built a “garbler” tool which reads in the clean word C1,Ci,Cn and synthesize OCR degradation to produce a garbled word D’1,D’i,D’n. This “garbler” chooses and perform a random edit operation (insertion, substitution or deletion) based on the probability distribution for the possible edit operations. The garbler was the tool that simply takes any word and generates the different error shaped that would be produced if this word was printed and OCR-ed.

2.3. The Model Training

The training data is one of the most important factors that affects the success or failure of any experiment, if we designed the correct model but used bad data to train the model, the result will be the failure of the model, and in the other hand if we choose good training data for the model, this will raise the success rate of the model. In this section we will compare between the training data used by both ElGhazaly and Darwish with training their models.

2.3.1. ElGhazaly Training Data

ElGhazaly took 200 documents from the corpus documents he constructed, and then he formatted them to keep only one word per line. Then he used the “Adobe Acrobat printer driver to convert the plain text documents into PDF format. Then based on his tests on the different OCR systems- he decided to use SakhArabic OCR as the recognition system for the rest of the experiment. He then applied the OCR process on the 200 documents, with average 30 pages; the total number of pages was about 30,000 pages with single word per line.

The result was having the original text for the documents and the result text from the OCR (the OCR-Degraded Text). Then he aligned manually the degraded documents returned from the OCR with the original clean text documents to have both Original & Degraded Text with the same line number.

2.3.2. Darwish Training Data

Darwish used Arabic book “Zad Al-Me’ad” which is available as a printed book and also available as electronic version for free. The book consists of 2,730 separate documents that address a variety of topics such as mannerisms, history, jurisprudence and medicine. Darwish scanned the printed version of Zad Al-Me’ad at 300x300 dpi (dots per inch) resolution, and then he manually zoned the images into multipage file to correspond exactly to the 2,730 documents in the character-coded clean copy of the collection. Then Darwish used Sakh’s Automatic Reader version 4.0 OCR engine to convert the images into plain text. Darwish measured the accuracy for the OCR-degraded text and it was computed with reference to the clean text using software from the University of Nevada at Las Vegas [11], obtaining 18.7%, for the images with 300 X 300 resolutions.

2.4. The Test Data

This section illustrates the test data collection that has been developed by both researchers in order to illustrating the success of their models.

2.4.1. ElGhazaly Test Data

ElGhazaly generated two data test sets selected from the training and testing pool, the first data set included 50 long documents consists of 26,579 words. The second test group contains 100 long documents consists of 51,765 words, and as ElGhazaly mentioned in his dissertation that there is no intersection between the training data sets and the test data sets collection. The following table contains the statistics presented by ElGhazaly on the testing data set.

2.4.2. Darwish Test Data

Darwish main goal for the OCR degradation model testing was to verify that the modeled OCR degradation and real OCR degradation have similar effects on information retrieval operation. Darwish decided to validate the effect of synthesized OCR degradation on retrieval effectiveness for the TREC collection [12] at the print (300X300) and 200x200 resolutions. TREC collection is the LDC LDC 2001 T55 collection, which was used in the Text RETrieval Conference (TREC) 2002 cross-language track. And for brevity Darwish called it TREC collection. The TREC collection contains 383,872 articles from the Agency France Press (AFP) Arabic newswire.

Darwish indexed the synthetically degraded TREC collections using words, lightly stemmed words, character n-grams- where n ranged between 3 and 5-, and combinations of n-grams and lightly stemmed words [23].

2.4.3. The Accuracy of the Models

It was important for us to see the accuracy of both models, because this will show us the direction of our research. Because each model has a different direction, one of them is based on word based approach and the other one is based on character based approach.

After training the model on 53,787 words, ElGhazaly model produced accuracy of 84.74% on test set contains 51,658 words [21]. Elghazaly introduced his own accuracy measure, which is the number of accurate replacements (with respect to the training set size) divided by the total number of OCR-Degraded words. In other words, if the mistaken OCR-degraded word is available in the training set with the correct original word, then this will be considered as accurate replacement. Otherwise, it will be considered as not accurate one. But ElGhazaly didn’t actually indexed and search his test data set calculate the precision, recall and mean average precision of his model [21].

On the other hand Darwish model produced accuracy for 3- gram or 4 gram character indexing was 87% [1, 24], but here we must illustrate the difference in both accuracy measurements, because Darwish was measuring according to the character level which means, For example, consider a page of 20 lines, each line has on average 10 words, and
each word has on average 5 characters. This means the page has 1000 characters in 200 words. If we consider the OCR output of this page to have only 20 character errors each in a separate word, this means character accuracy of 98% where it means word accuracy of 90%. Darwish also illustrated that the best mean average precision of his model which was “0.56” [1, 25].

3. The Proposed OCR-Degraded Arabic Text Retrieval Model

Our Proposed OCR-Degraded Arabic Text Retrieval model is based on two steps, the first step is synthesize the OCR-Degraded text, and then expands the user search query using the expected OCR error generated from the Arabic OCR error simulation. The first step in the model, which is, modeling the OCR degradation is an important point when handling electronic library information retrieval, unfortunately, despite the claims of commercial vendors; OCR error rates are far from perfect, particularly for challenging languages like Arabic [13]. The result is that the low accuracy of the Arabic OCR text affects directly the accuracy of the online retrieval process regarding these documents.

There are many challenges for the OCR to give accurate outputs. Also, there are many models for correcting the OCR outputs. Many models for simulating the OCR errors have been analyzed. They are mainly depending on 1-gram and sometimes n-gram character replacement algorithm. However, in Arabic, as character shape defers up to its position in the words (begin, middle, end, isolated). So, it is too difficult to include all these variables (7-gram character for example) plus the character position in one model. However, even if the 7-gram is reached, the 8-grams will not be covered and so on.

The next section (3.1) describes the OCR-Degradation synthesizing model, which is a word based model trained and tested on complete words. This model supports the maximum n-grams in the training set with respect to the words’ positions. Then in section 3.2 we will show how we used this model to enhance Arabic OCR Degraded Text retrieval.

3.1. The OCR-Degradation Synthesizing Model

As we mentioned, the proposed OCR-Degradation synthesizing model is a word based model that is trained and tested on complete words. This model supports the maximum n-grams in the training set with respect to the characters’ positions. The model main idea is to align the OCR degraded words and the clean text words, the alignment operation is done based on the edit distance which is the insertions, deletions, and substitutions operations required to correct the degraded text. Calculating the edit distance between two Arabic words requires caution, because Arabic words may be formatted with prefix letters which sometimes forms half the word length, which means, if we calculated the edit distance between two words contains the same prefixes but the remaining of the two words is completely different, our model will consider them as the same word but some degradation happened while the OCR. To solve this problem our model ignores the Arabic prefixes letters while computing the edit distance between any words.

The model starts with checking each word in each document of the OCR-Degraded documents’ set against the corresponding original word, if the edit distance between the two words less than or equal a selected value , the model stores the deformed shape linked with the original word as a training pair in the OCR errors database. This enables the model to be trained on both high and low resolution, which means specifying the deformation level the user of the model wants to cover in his training data set. Figure 2 illustrates the OCR-degradation synthesizing model.

The model starts by fetching the first word from the clean the OCR-Degrade text files, and then the model calculates the edit distance between both words after removing the prefixes. Here we have three cases we shall handle, the first case, the edit distance value is less than or equal the accepted value, the second case, the value is larger than the accepted value but less than or equal to one, and the third and final case, the value is larger than one. The first case means that the edit distance is within the accepted value specified by the model user in the application interface; so the model adds both words to the training database. The second case means that either the degraded word is part of the original word, but the OCR application split it while the recognition process, which happens sometimes for long words, or the recognition process was bad for this word and the system recognized only few characters of the word. This point formed a limitation in the word based models and the character based models we checked, they didn’t consider if the word was recognized as two and sometimes three parts [1, 2, 22, 25].

We decided to cover this limitation in our model. This means, if the OCR system recognized one word into two or three words, our system will consider these parts as one word and store them in the training database as one shape corresponding to the original shape. Finally, the third case , if the edit distance between both words is larger than one, which means that both words and completely different, this means that the model has lost the correct position of anchors corresponding to each file, in this case our model realign the anchors until both files anchors are pointing on the same word.

And to implement the proposed model we built a software application The “Aligner”, which performs all the actions illustrated in the model and finally generates the training database. the “Aligner” takes both Original and OCR-ed text files and tokenize them to words, the tool align the words based on the edit distance between them, the finally the tool updates our training database that stores the original word shape and the corresponding shapes of the words appeared in the OCR version. Figure 3 displays the aligner application interface.
During the alignment process the "Aligner" updates the training database with degraded shapes with its corresponding original shapes; the database structure was designed to keep only one entry of the original word and any number, but unique, of the deformed shapes of this word. The database also stored the context of the deformed shapes, like the preceding word and the succeeding word. This information will be used in our future work which is checking how the OCR-Degraded word context affects the retrieval effectiveness.

3.2. Orthographic Query Expansion

After synthesizing the OCR-Degraded text, here comes the second step of the orthographic query expansion model, which is expanding the user search using the expected OCR error simulation model. To expand the user query we built a query generation tool, this tool takes the original user query and generates the deformed OCR-Degraded query. The following figure (Figure 4) illustrates the Query Generation user interface.

---

Fig. 2: The OCR Degradation Synthesizing Model

Fig. 3: The "Aligner" Application interface
3.3. Training the OCR Degradation Synthesizing Model

The model was trained on the electronic version of the Arabic three volume book “Abgad El Eloum” or “Alphabet of science” and for short naming we will call it “ABGAD”. For the experiment we printed the first two volumes of the book, which contained 205,761 words, then we scanned and OCR-ed the first two volumes using the best Arabic OCR application, Sakhr Automatic reader version 10. Sakhr OCR produces 99.8% accuracy for high-quality documents 96% accuracy for low-quality documents [14]. In this stage we have the original electronic version of the book, and the corresponding OCR-ed version of the same electronic text.

Then using the “Aligner” we processed both the clean text and the OCR generated text and generates the training database, the training database generated constructed from 50 documents (205,696 words), consists of 388,40 unique words, 182,647 words read correctly and 230,49 word read wrong.

3.4. Testing the Orthographic Query Expansion Model

This section describes the group of tests done to verify that modeled OCR degradation increases the accuracy of the degraded text information retrieval. Generally; evaluating retrieval effectiveness requires the availability of a test document collection with an associated set of topics and relevance judgments. Relevance judgments are the mappings between topics and documents in the collection that are relevant to them. The cost of producing relevance judgments for a large collection is very high and dominates the cost of developing test collections [15]. There are three ways to produce relevance judgment, the first one is “pooling”, which is assessing manually the relevance of the union of the top n documents from multiple retrieval systems for every topic [16]. The second way is manual user-guided search, where a relevance judge manually searches and assesses documents for a topic until the judge is convinced that all relevant documents are found [17]. The third way is exhaustively searching the documents for relevant documents [18].

These three ways often miss some relevant documents, and assessment of relevance is necessarily subjective, but studies shown that relevance judgments can be reliably used to correctly differentiate between retrieval systems provided that a sufficient number of queries are used [4, 19, 16]. Voorhees estimated the number of sufficient queries to be about 25 [16]. But here in our test we extended the search query to be 35 queries to collect more data about the retrieval accuracy.

3.4.1. Orthographic Query Expansion Model Test Set Statistics.

We based our test on two test sets have been selected from the Training and test pool. The first one includes 20 documents (66,985 words), which is the third volume of “ABGAD”. And to memorize the training set was the first two volumes only of the book. And the second data test set contains 300 long documents containing 621,763 words. There is no intersection between the Training and Test sets. The second test data is “ZAD” data collection, which is a 14th century religious book called Zad Al-Me’ad, which is free of copyright restrictions and for which an accurately character-coded electronic version exists [20]. We here must illustrate that this is the same test set used by Darwish in his model [1].

3.5. Testing the Orthographic Query Expansion Model Accuracy.

This section describes 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 the proposed model on retrieval effectiveness. IR evaluation measures are concerned with precision and recall given:

<table>
<thead>
<tr>
<th>Not Relevant</th>
<th>Relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Retrieved</td>
<td>C</td>
</tr>
<tr>
<td>Retrieved</td>
<td>A</td>
</tr>
</tbody>
</table>

Precision = \( \frac{A}{A + C} \)

Recall = \( \frac{A}{A + D} \)

Precision measures the fraction of retrieved documents that are relevant. Recall measures the fraction of all relevant documents that are actually retrieved. IR-effectiveness measures use precision and recall in different ways. For example, precision at n measures the precision after a fixed number of documents have been retrieved. Another is precision at specific recall levels, which is the precision after a fraction of relevant documents are retrieved.

For our experiment we indexed our clean and OCR-Degraded documents on the best Arabic search engine, IDRISI 6.0[14]. One of IDRISI features is creating separate search collection for different group of documents depending on the user requirements. So we created a separate collection for the clean text documents and another collection for the OCR-Degraded documents. The author of the paper, a native speaker of Arabic, developed 35 topics and exhaustively searched the collection for relevant documents.

Then using the “Query Generation” tool that takes the user clean text query and based on the training database generates the relevant OCR-Degraded text query, we passed both the original search query and the OCR.
degraded query to IDRISI to search the clean text collection and the OCR-Degraded text collection separately. Using this way we can compare the results of both collections and get the precision and recall to check our model accuracy.

After completing the experiment on the thirty five queries we analyzed the results and found that. For test data set 1, the number of relevant documents per topic ranged from one (for one topic) to eighteen, averaging 14. For test data set 2 the number of relevant documents per topic ranged from two (for one topic) to 224, averaging 121. The average query length used for the test data set 1 is 4.1 words and for test data set 2 is 5.2 words. The following figure displays the search precision and recall returned from the search engine relevant to the number of retrieved document for test data set 1.

![Fig. 5 Test Data Set 1 Precision and Recall relevant to no. of documents retrieved](image)

Unfortunately Darwish didn’t mention the precision and recall measurements for his model. He was most concerned with the mean average precision value of his information revival system, the mean average precision is the most commonly used evaluation metric which eases comparison between systems [16]. The best mean average precision resulted by Darwish was “0.56” [1] for 3-Gram and 4-gram steamed words, in the other hand the mean average precision resulted by the proposed model for test data set 1 is “0.96” and for test data set 2 is “0.97”, which means a significant improvement in the information retrieval effectiveness.

4. Conclusion

This paper advanced the state of the art in Arabic OCR-Degraded text retrieval in specific and IR in general. Perhaps the most far reaching contribution was the development of new OCR degraded synthesizing model that improves retrieval effectiveness. The proposed model was shown to be more effective than previously developed n-gram or word level models.

<table>
<thead>
<tr>
<th>Table 1. Comparison Between The Three Models</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Comparison Point</strong></td>
</tr>
<tr>
<td>The model considered word context error that may affect the recognition operation.</td>
</tr>
<tr>
<td>The model is word based and covers any word with any length once it was trained on it</td>
</tr>
<tr>
<td>The accuracy measurement used based on the word level not on character level</td>
</tr>
<tr>
<td>The model solved the word concatenation problem</td>
</tr>
<tr>
<td>The model tried to simulate different degradation levels</td>
</tr>
<tr>
<td>The model idea can potentially be useful in widespread applications</td>
</tr>
<tr>
<td>The model retrieval effectiveness is tested on indexed data set</td>
</tr>
<tr>
<td>The model considered enhancing the scanned images before OCR</td>
</tr>
<tr>
<td>The model accuracy is not affected by the training set size</td>
</tr>
</tbody>
</table>
The model was tested on two different data sets, on the first data set, which were 20 documents containing 66,985 words. And the second data set were 300 long documents containing 621,763 words. There were no intersection between the training data sets and test data sets. When the model was applied on the first data set, the mean average precision produced was “0.96”, and when the model applied on the second data set the mean average precision produced was “0.97”. In the other hand Darwish shown that the best mean average precision produced by his model for 3-gram indexed data set was “0.56” [1]. We here must illustrate that ElGhazaly didn’t measure his model performance against indexed data set [2] and he depended only on calculating the number of OCR-Degraded words in the training set as an accuracy measurement. Table 1 illustrates the main comparison points between the three models mentioned in this paper.

The main limitation of the proposed OCR-Degraded synthesizing model is its dependency on the training set size, as it’s a word based model. On the other hand the character based model proposed by Darwish doesn’t have this limitation.

References