TEL AVIV UNIVERSITY אוניברסיטת תל-אביב

The Raymond and Beverly Sackler Faculty of Exact Sciences

The Blavatnik School of Computer Science

Improving Image Documents Retrievability through Image and OCR-Text Correction Assisted by Language Models

A thesis
submitted in partial fulfilment
of the requirements for the Degree of
Master of Science

by

Ido Kissos

under the supervision of

Prof. Nachum Dershowitz

Tel Aviv University

November 2016

Abstract

This thesis explores the use of learned classifiers for improving retrieval of image documents by increasing OCR accuracy and generating word-level correction-candidates.

The method suggests the simultaneous application of several image and text correction models, followed by a performance evaluation that enables the selection of the most efficient image-processing model for each image document and the most likely corrections for each word. The selection is assessed by two trained classifiers based on the output features of every model, mostly based on a language model. The highest-ranked image model is selected and applied to the image, while the top correction-candidates for every word are appended to the OCR word output and indexed at the same position. An additional and optional classifier aims to decide if an OCR word should be replaced by its correction-candidate, enabling the method to comply with non-retrieval purposes.

The presented method relies on a ground-truth corpus, comprising image documents and their transcription, in addition to an in-domain corpus to build the language model. It is designed to be applicable to any language that follows simple segmentation rules, in other words no compound words, and performs best on morphology-rich languages.

Experiments with an Arabic newspaper corpus show that this approach significantly improves OCR accuracy on the dataset by reducing its word error rate by 50%.

Table of Contents

Acknowledgments **iii**

1: Introduction **1**

2: Background **3**

2.1 The OCR Process . . . 3

2.1.1 Image Preparation . . . 3

2.1.1.1 Image Scaling Methods . . . 4

2.1.1.2 Binarization Methods and Thresholds . . . 4

2.1.1.3 Image Denoisement Methods . . . 5

2.1.2 Page layout Analysis . . . 5

2.1.3 Lexical Correction . . . 5

2.2 OCR Errors . . . 6

2.3 Data Resources . . . 8

3: Image Enhancement Methodology **10**

3.1 Image Enhancement Candidate Generation . . . 11

3.2 Candidate Selection . . . 12

3.2.1 Feature Extraction . . . 12

3.2.2 Ranking . . . 13

4: Text Correction Methodology **14**

4.1 OCR Text Tokenization . . . 14

4.2 Correction-Candidate Generation . . . 14

4.2.1 Modeling OCR errors . . . 15
4.2.2 Generating Candidates . . . 15
4.3 Candidate Ranking . . . 15
4.3.1 Feature Extraction . . . 16
4.3.2 Ranking . . . 17
4.4 Correction Decision Making . . . 17
4.4.1 Feature Extraction . . . 18
4.4.2 Decision Making . . . 18

5: Testing the Model 19

5.1 Image Enhancement . . . 19
5.1.1 Selecting the Algorithms and Thresholds . . . 19
5.1.2 Calculating the Language Model Score . . . 20
5.2 Text Correction . . . 21
5.2.1 Retrieving the Correct Word . . . 21
5.2.2 Candidate Ranking . . . 22
5.2.3 Correction Decision Making . . . 23
5.3 Overall Results . . . 23

6: Conclusions 26

References 28

Appendix A: Examples from the Dataset 30

Appendix B: Confusion Matrix Excerpt 36

List of Figures 51

List of Tables 52

Acknowledgments

I would like to thank the Blavatnik and Deutsch foundations for directly supporting this work, as well as the ISF and GIF foundations.

1 Introduction

Massive digitization of textual resources, such as books, newspaper articles and cultural archives has been underway for some decades, making these resources publicly available for research and cultural purposes. Institutions are converting document images into machine readable text via Optical Character Recognition (OCR), enabling a realistic way of exploring vast document corpora with automated tools, such as indexing for textual search and machine translation.

For reasons of low quality printing and scanning or physical deterioration, many of these corpora are of poor quality, making the OCR task notoriously difficult. Consequently, it is impossible to directly employ the obtained results for subsequent tasks without costly manual editing. Although contemporary OCR engines claim higher than 97% word accuracy for Arabic, for instance, the same datasets with low-resolution images or infrequent character classes can drop to lower than 80%.

In this thesis, we propose an OCR correction technique that consists of an image pre-processing and text post-correction pipeline based on a composite machine-learning classification. The technique wraps the core OCR engine and is in practice agnostic to it.

The image correction applies a small set of image enhancement algorithms on copies of the image document, which serve as input for the OCR engine. This set includes image scaling, binarization methods and parameter thresholds, and were chosen by their experimental accuracy gain. The potential gain for every algorithm is evaluated as the sum of the positive accuracy improvements, relying on a learned classifier for the selection of improved OCR text over a baseline OCR text, that is the output of the image with the OCR's engine's default pre-processing. Such a classifier was trained with a ground-truth set, relying on recognition confidence statistics and language model features to output an accuracy prediction.

The text post-correction applies a lexical spellchecker and potentially corrects single-error misspellings and a certain class of double-error misspellings, which are the major source of inaccurate recognition in most OCR use-cases. The novelty of this method is its ability to take into consideration several valuable word features, each giving additional information for a possible spelling correction. It is built out of two consecutive stages:

1. word expansion based on a confusion matrix, and

2. word selection by a regression model based on word features.

The confusion matrix and regression model are built from a transcribed set of images, while the word features rely on a language model built from a large textual dataset. The first stage generates correction candidates, ensuring high recall for a given word, while the second assures word-level precision by selecting the most probable word for a given position. Relying on features extracted from pre-existing knowledge, such as unigram and bigram document frequencies extracted from electronic dictionaries, as well as OCR metrics, such as recognition confidence and confusion matrix, we accomplished a significant improvement of text accuracy.

The correction methods, the image enhancement and text correction, implement an equivalent methodology in the sense that both begin by promoting recall, namely generating many correction candidates that some may improve the baseline result, and afterwards use prior knowledge and context to gain precision, namely select the best candidate. Initially, the research focused only on the text correction method, and it is its success that pushed the idea of implementing a similar methodology to the image processing, which by end turned out to have a larger accuracy gain.

This research is part of the digitization project of the "Arabic Press Archive" of the Moshe Dayan Center at Tel Aviv University, hence the method evaluation and adaptation is in the Arabic language. There are a number of open-source and commercial OCR systems trained for this language; we used NovoDynamics NovoVerus commercial version 4, evaluated as one of the leading OCR engines for Arabic scripts.

For evaluation purposes we use the Word Error Rate (WER) measure, or by its complement to 1 that we named OCR accuracy, which is adapted for subsequent applications of the OCR output such as information retrieval. Our correction method performs effectively, reducing faulty words by a rate of 50% on our dataset, which is a 8% absolute accuracy improvement. The overall results showed negligible false-positive errors, namely the method rarely rejects correct OCR words in favor of an erroneous correction, which is a major concern in spellcheckers. An analysis of classifier performance shows that bigram features have the highest impact on its accuracy, suggesting that the method is mainly context reliant.

This dissertation is organized as follows: Section 2 provides background information on the OCR process and OCR error correction. Section 3 presents the image enhancement methodology and Section 4 the text correction methodology. Section 5 reports and discusses experimental results; Section 6 concludes the paper and suggests some possible future directions. For readability reasons, examples of the created artifacts, such as confusion matrix and feature tables are presented in English instead of Arabic.

2 Background

2.1 The OCR Process

The goal of OCR is to extract the text, character by character, from a document image. The process proceeds in consecutive stages:

1. prepare the image document for character recognition, including, for example, document deskew, graphics and noise removal, etc.;

2. automatically segment a document image into images of individual characters in the proper reading order using image analysis heuristics known as page layout analysis; then

3. apply an automatic classifier to determine the characters that most likely correspond to each character image; and

4. exploit the surrounding context to select the most likely character in each position, based on a shallow language model or prior lexical knowledge of the document's language and subject domain.

Our paper focuses on stages 1, 2, and 4, which are done independently of stage 3. We elaborate on them in the remainder of this section.

2.1.1 Image Preparation

Preceding text extraction, image processing aims to transform the document image into its most suitable form for the subsequent task. A recognition system performs various image processing steps to produce a "cleaned" image, which is then examined to determine the gross orientation and the dominant alphabet on the page. The following sub-sections describe the main processing steps applied to an image prior to the character recognition stage.

2.1.1.1 Image Scaling Methods

Scaling refers to the resizing of a digital image, which can have a major impact on the character level characteristics that depend on their pixel form, thus affecting the text extraction recognition. Image size can be changed in several ways; these are the main methods adapted to textual documents.

1. Nearest-neighbor interpolation – One of the simpler ways of increasing the size is nearest-neighbor interpolation, replacing every pixel with a number of pixels of the same color: The resulting image is larger than the original, and preserves all the original detail, but has (possibly undesirable) jaggedness. Diagonal lines, for example, will show the "stairway" shape characteristic of nearest-neighbor interpolation.

2. Bilinear and bicubic interpolation – Bilinear algorithm works by interpolating 4 pixel (2x2) color values, introducing a continuous transition into the output even where the original material has discrete transitions. Although this is desirable for continuous-tone images, this algorithm reduces contrast (sharp edges) in a way that may be undesirable for text. Bicubic interpolation, based on 16 adjacent pixels (4x4), yields substantially better results on text, with only a small increase in computational complexity.

3. Box sampling – One weakness of bilinear, bicubic and related algorithms is that they sample a specific number of pixels. When downscaling below a certain threshold, such as more than twice for all bi-sampling algorithms, the algorithms will sample non-adjecent pixels, which results in losing data, and also causes unsmooth results. The trivial solution to this issue is box sampling, which is to consider the target pixel as a box on the original image, and sample all pixels inside the box. This ensures that all input pixels contribute to the output. The major weakness of this algorithm is that it is hard to optimize.

2.1.1.2 Binarization Methods and Thresholds

Image binarization converts an image of up to 256 gray levels to a black and white image. This is a necessary pre-processing step, as most OCR engines work on bi-level images. The use of a bi-level information decreases the computational load and enables the utilization of the simplified analysis methods. The simplest way to use image binarization is to choose a threshold value, and classify all pixels with values above this threshold as white, and all other pixels as black. The problem then is how to select the correct threshold. In cases of low resolution scanning, poor quality of the source document or complexity in the document structure (such as graphics is mixed with text), finding one threshold compatible to the entire image is impossible, therefore, adaptive image binarization is needed where an optimal threshold is chosen for each image area, as suggested by Sauvola's method [12]. When character recognition is performed, the melted sets of pixel clusters, which represent characters, are easily misinterpreted if binarization labeling has not successfully separated the clusters.

2.1.1.3 Image Denoisement Methods

The type of noise and artifacts seen in the documents is varied but tends to be characterized by splotches, specks, and streaks, generally resulting from a low-quality source document or scanning. The line height is used to derive the minimum number of pixels that a black blob must have to be considered text and not noise. A black blob is generally defined as a group of eight-connected black pixels. Blobs with smaller than a minimum area will be classified as noise and removed. In poor quality images, the benefits of removing more noise compensates for the loss of some letters [2]. In Arabic this tradeoff is aggravated with the use of diacritic dots to differentiate characters that can easily be taken by mistake as noise.

2.1.2 Page layout Analysis

The cleaned image page is analyzed, and the text zones are located and segmented into lines of text, separated from non-textual zones such as graphics. The algorithm used to perform this analysis detects the type of page layout, and uses the decomposition algorithm best suited to that layout. The layout analysis mode can be configured to simple mode, where the algorithm expects the page to be a single block of text having one or two columns, complex mode, where the algorithm expects a page with complex, multi-column layout, or no decomposition mode, where the system creates a single zone consisting of the entire page. There are two main approaches to document layout analysis.

1. Bottom-up approaches which iteratively parse a document based on the raw pixel data. These approaches typically first parse a document into connected regions of black and white, then these regions are grouped into words, then into text lines, and finally into text blocks [10]. Their main advantage is that they require no assumptions on the overall structure of the document.

2. Top-down approaches which attempt to iteratively cut up a document into columns and blocks based on white space and geometric information.They parse the global structure of a document directly, thus eliminating the need to iteratively cluster together the possibly hundreds or even thousands of characters/symbols which appear on a document. They tend to be faster, but in order for them to operate robustly they typically require a number of assumptions to be made about on the layout of the document[6].

2.1.3 Lexical Correction

There has been much research aimed at the automated correction of recognition errors for degraded collections. An early, useful survey is [5]; relevant methods for Arabic OCR are summarized in [8] and in collection [9].

In this work, we use language models on the character and word levels, plus lexicons. We do not apply morphological or syntactical analyses, nor passage-level or topic-based methods. Three language resources play a rôle:

- Dictionary lookup compares OCR-output with the words in a lexicon. When there is a mismatch, one looks for alternatives within a small edit (Levenshtein) distance, under the assumption that OCR errors are often due to character insertions, deletions, and/or substitutions. For this purpose, one commonly uses a noisy-channel model, a probabilistic confusion matrix for character substitutions, and term frequency lists [3], as we do here. One must, however, take into consideration unseen ("out of vocabulary") words, especially for morphologically-rich languages, like Greek, and even more so for *abjads*, like Arabic, in which vowels are not represented. The correct reading might not appear in the lexicon (even if it is not a named entity), while many mistaken readings will appear, because a large fraction of letter combinations form valid words. Morphological techniques could help here, of course. Dictionary lookup and shallow morphology are used in [8].

- We use the term k-mer[1] for the possible contiguous k-character substrings of words. By collecting statistics on the relative frequency of different k-mers for a particular language, one can often recognize unlikely readings. This technique was employed by BBN's OCR system for Arabic [7], as well as in [8].

- A language model, based on n-gram frequencies derived from a large corpus, is frequently used to estimate the likelihood of a reading in context [14].

2.2 OCR Errors

OCR accuracy is negatively influenced by poor image quality (e.g., scanning resolution, noise) and any mismatch between the instances on which the character image classifier was trained and the rendering of the characters in the printed document (e.g., font, size, spacing). Depending on the language and the image quality of the analyzed collection, there will be a different error distribution generated by an OCR process. These errors can be categorized according to the following types, listed in the order they occur during the OCR process:

- Word detection – failing to detect text in the image, commonly caused by poor image quality or text mixed with graphics.

- Word segmentation – failing to bound an individual word correctly, due to wrong interword space detection, generally due to different text alignments and spacing.

[1] Also known as "character n-gram". The term "k-mer", borrowed from bioinformatics, allows us to use "n-gram", unambiguously, for sequences of *words*.

- Character segmentation – failing to bound single characters in a segmented word. This is frequent for cursive or connected alphabets, such as printed Arabic or handwritten Latin-alphabet languages. It may also occur due to an analog process (e.g., printing and scanning speckles) that might disconnect connected components.

- Character recognition – failing to identify the correct character for a bounded character image.

An error can be classified into more that one category, resulting in error types that are partially overlapping. This implies that errors cannot always be mapped to a single OCR sub-process that should be improved, and such can generally be attributed to low image document quality or small random image effects that are hard to handle. This categorization, despite the fact it is not strict, facilitates the discussion on OCR errors and their possible correction.

2.3 Data Resources

The correction methodology is language, domain, scan quality and OCR engine agnostic; nevertheless the model itself is built upon a data corpus that resembles the test data.

Our experiment focused on OCR of printed Arabic newspaper articles of the Moshe Dayan Center at Tel Aviv University. The archive documents are fairly variable, containing text in different fonts and page layouts, such as tables and graphics. We made use of the following training resources[2]:

1. Two hundred fifty OCR document images and their ground truth transcription – each document is a newspaper article scanned at 300 dpi in grayscale. The newspaper article are relatively low-quality by origin, which emphasizes the OCR correction ability to deal with low quality images and enrich the word errors the noisy channel training. The set was manually transcribed at a document level and OCR processed by Novodynamics Verus version 4.0. Fifty documents were left aside during the training process for later testing and evaluation. The training set contains about 83,000 words with a WER of 17%. Figure 2.2 shows a sample image of a newspaper article. The articles in the set are all taken from "Al-Hayat" daily newspaper, printed in January 1994.

2. Gigaword Arabic[11] – This vast corpus contains about 3 million transcribed articles from various modern Arabic newspapers. A modeling process produced a unigram and a bigram frequency lists, which were later used as features for correction candidate ranking and classification. The size of the corpus and its thematic similarity to the dataset ensure relatively accurate features. The nature of language itself implies that the lexicon is not perfect, but we can take comfort that most imperfections lie in the low frequency tail, and do not have much impact on correction performance. An example transcription, alongside its originating article, can be seen in Figure 2.2.

[2] The dataset is publicly available at `https://github.com/idoki/ocr_correction`.

عنصرية معكوسة في أستراليا

■ كانبيرا - رويتر - قالت لجنة حقوق الانسان في استراليا امس ان استرالياً ابيض حصل على تعويض مقداره ١٢ الف دولار استرالي (٨٢٠٠ دولار اميركي) عن الالم والاهانة اللذين لحقا به من جراء التمييز العنصري الذي مارسه ضده احد السكان الاصليين.

وقالت متحدثة باسم اللجنة ان جورج بيل (٤٧ عاماً) وُصف بانه «عنصري ضيق الافق» ووجهت اليه شتائم وتلقى تهديدات بالايذاء البدني من زميل له من السكان الاصليين، وهو عضو في لجنة شكلت عام ١٩٩٠ للنظر في القضايا الخاصة بهؤلاء السكان.

Figure 2.1: “Al-Hayat” newspaper article, 06.01.1994

```
<DOC article="0033" id="HYT_ARB_199401060033" newspaper="HYT_ARB_19940106">
        <headline>عنصرية معكوسة في أستراليا</headline>
        <text>
            <p>قالت لجنة حقوق الانسان في استراليا أمس ان استراليا أبيض حصل
            على تعويض مقداره 12 ألف دولار استرالي ( 8200 دولار اميركي ) عن
            الألم والاهانة اللذين لحقا به من جراء التمييز العنصري الذي
            مارسه ضده أحد السكان الاصليين .</p>
            <p>وقالت متحدثة باسم اللجنة ان جورج بيل ( 47 عاما ) وصف بأنه
            عنصري ضيق الافق " ووجهت اليه شتائم وتلقى تهديدات بالايذاء "
            البدني من زميل له من السكان الاصليين , وهو عضو في لجنة شكلت
            عام 1990 للنظر في القضايا الخاصة بهؤلاء السكان .</p>
        </text>
    </DOC>
```

Figure 2.2: “Al-Hayat” newspaper transcription, 06.01.1994

3 Image Enhancement Methodology

The ability to consistently select the best image processing for every image document leans on the capability to reliably predict its performance, namely, its OCR text accuracy. To facilitate this task, this prediction can be based on the extracted text of each image and not on the image itself, suggesting that an a posteriori approach could outperform an a priori one.

The enhancement method requires one to move from a single-pass OCR engine, in which every document is processed once—and for which OCR engines are optimized, to multi-pass OCR. The latter enables an accuracy-performance trade-off, promoting better OCR results at the compromise of CPU resources, which is often a reasonable call for digitization projects. Having several output texts for a single image document, we can rank them and choose the most accurate, according to our prediction, for a specific image. The multi-pass architecture is built as a pipeline where each module applies a family of dependent algorithms, for example binarization methods and thresholds, and the sequential modules are independent one of the other. After every module there is an evaluation sequence which extracts the document image text and predicts its accuracy, then feeds its processed image to the next module, as shown in Figure 3.1. This implementation avoids the application of an unfeasible number of image processing sets, which is the sum of all possible algorithm combinations. Assuming independence between the modules, their application order has only a small significance.

Every module comprises three stages:

1. Enhancement candidate generation – Every algorithm in the set renders a processed image that serves as an input to the OCR engine.

2. Feature extraction – For each candidate, language model features and confidence statistics are extracted from its OCR text output.

3. Candidate ranking – This stage ranks the candidates according to their correctness probability, and selects the highest ranked candidate as the image to subsequent module, or the text to the post-correction task in case it is the last module.

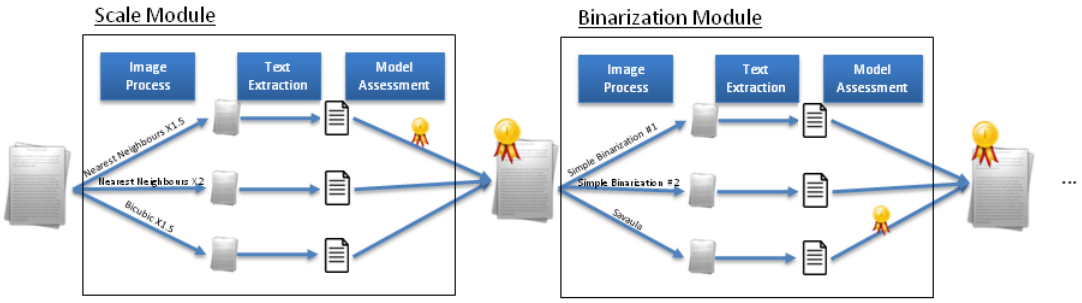

Figure 3.1: Image Enhancement Implementation

3.1 Image Enhancement Candidate Generation

The candidates are generated based on a set of image processing algorithms and thresholds we found had positive effect on the corpus' OCR accuracy. Finding this set required an evaluation of the potential gain of every algorithm. We benchmarked the performances of a large set of algorithms that are commonly used for OCR pre-processing, tuned their parameters and chose several configurations for each algorithm type i that produced the highest gains, which we denote by set X_i. For example, in Figure 3.1, X_i for the scale algorithm type is represented as follows:

$$X_{i=scale} = \{(NearestNeighbors, 1.5), (NearestNeighbors, 2), (Bicubic, 1.5)\}$$

In order to comply with the metric of improving the average accuracy, finding X required to solve an optimization problem that can be formulated as follows:

$$\begin{aligned} \underset{X_i}{\text{maximize}} \quad & \sum_{doc \in training\text{-}set} accuracy(doc|X_i) \\ \text{subject to} \quad & X_i \subseteq \{algorithm_{ij}\}\ \forall i \in Algorithm\ Type,\ \forall j \in Algorithm\ Configurations \\ & |X_i| \leq 3 \\ & Algorithm\ Type = \{\text{Scale, Binarization, Denoisment, Layout Analysis}\} \end{aligned}$$

Notice that maximizing on each of the members of the set separately would produce a different result, meaning that we do not necessarily choose the single most efficient algorithm, but a set that has a high inner variance gain. The limitation to a set of order 3 is conditioned for calculation purposes and empirical gain bounds. The approximation for $X_i$is obtained by trial and error and stopped when reaching negligible accuracy improvements.

Each family of dependent methods or thresholds is implemented in a separate module, resulting in a total of 4 modules. The following algorithm types and thresholds were applied:

1. Bicubic and K-Nearest-Neighbors methods, scaling from 1 to 3 with 0.25 stepsize.

2. Sauvola and threshold-based binarization algorithm, with thresholds varying from 100 to 250 with a step size of 25.

3. Image denoisement methods included 3 different filters: Mild, Median and None, enabled in NovoVerus.

4. Layout analysis modes included a Simple, None and Complex modes, enabled in NovoVerus.

Based on the above close-to-optimal algorithm set, the candidate generation module applies this set to the input image and extracts its text for the evaluation phase.

3.2 Candidate Selection

This stage evaluates the textual output from each of the image candidates at the final stage of every module. The evaluation is based on a learned linear regression that ranks the candidates according to their expected accuracy. This score does not necessarily have to be a normalized accuracy, but as a language model score, assessing which textual output is more probable to occur. As for a typical machine learning algorithm, we extract features and train a classifier upon them.

3.2.1 Feature Extraction

The language features are based on bigram occurences and frequencies, while confidence metrics are given at the character level by the OCR engine as an internal metric. This raw feature is aggregated to a document level statistic, which predicts the best the OCR accuracy. We compared 3 statistics for this task:

1. Character average – $\sigma_{doc} = \sum_{char \in doc} \frac{conf_{char}}{\|doc\|_c}$, a straightforward approach to estimate the document confidence-level.

2. Word average – $\sigma_{doc} = \sum_{word \in doc} \frac{\min_{char \in word} conf_{char}}{\|doc\|_w}$, matching more closely the accuracy measure, which is word-based and not character-based.

3. 0.8 character percentile – This approach ignores bias caused by outlying words with very low confidence, which would be the result of wrong region identification. For example text erroneously extracted from graphics, local deterioration or scanning problem.

3.2.2 Ranking

The ranker's rôle is to select the most accurate image out of all the module's images, that is, an inner order for a small image set. Even though a comparative metric between the different image instances should suffice for this task, we achieve this by the more general metric of the OCR accuracy prediction. We train the OCR accuracy predictor on the labeled set for which we know their real accuracy, and try different feature representation and models to achieve good results. Every image is scored independently from all other images, including its enhancement candidates. The loss-function the regression model is assessed upon is not the one that minimizes the mean-square error compared to the real accuracy, but the accuracy loss of faulty ranking, that is the 0-1 loss its weighted for as formulated in the following equation:

$$
\begin{aligned}
\underset{\hat{f}}{\text{minimize}} \quad & L(f,\hat{f}) = f - \hat{f} \\
\text{where} \quad & f_{doc} = \max_{\textit{image-candidate}} \textit{accuracy}(doc|\textit{image-candidate}) \\
& \hat{f}_{doc} = \textit{accuracy}(doc)
\end{aligned}
$$

4 Text Correction Methodology

The OCR error model is vital in suggesting and evaluating candidates. At the heart of the error model is a candidate generation for correction, based on a confusion matrix giving conditional probabilities of character edits. The possible error corrections include the error types listed in Section 2.2, except word detection problems, as the algorithm has no input image to correct. The latter has to be addressed with image pre-processing or detection robustness.

We focus the discussion on the word level at a certain position in the text, which is obtained by a standard tokenization of the OCR output text.

The error correction methodology comprises three stages:

1. Correction candidate generation – The original word is expanded by a confusion matrix and a dictionary lookup, forming all together a correction-candidates vector.

2. Feature extraction – Features are extracted for each word in the vector.

3. Word classification – A two-stage classification process, where the first stage ranks the correction candidates according to their correctness probability at this position, while the second selects the most probable between the original word and the highest-ranked candidate.

4.1 OCR Text Tokenization

In order to structure the OCR text to enable correction at a word level, the text is tokenized by a standard space delimiter tokenizer. This phase also parses the word recognition confidence produced by the OCR engine, forming a first level feature extraction.

4.2 Correction-Candidate Generation

This module is designed to generate correction candidates for a tokenized word in accordance with an observed OCR error model.

4.2.1 Modeling OCR errors

In order to model errors, we build a noisy channel to learn how OCR corrupts single characters or character segments. We align the ground truth image document to its respective OCR text at word-level, and for each pair of words we calculate their Levenshtein distance. This alignment includes segmentation errors in case the noisy channel detects a single word that was split into two. For each erroneous word we issue a segment correction instance, as can be seen in Appendix B. For example, given the aligned pairs (tlne, the), (amual, annual), the correction instances issued are: ln $\longrightarrow$ h, m $\longrightarrow$ nn. Subsequently, these instances are aggregated to a weighted confusion matrix, as shown in Appendix B.

The error model we implement supports the correction of erroneous character segmentation and recognition, as well as word segmentation. The former is handled by supporting primitive 1-Levenshtein distance[1] [13] plus 2:2 alignments, the latter by whitespace edition (deletion and insertion). Another way of formulating the error class is all 1-Levenshtein distance, plus 2-Levenshtein distance restricted to consecutive character edition. The limitation to primitive 1-Levenshtein distance plus 2:2 alignments corrections was based upon their relative ease of generalization, implied by the error class high proportion and recurrence in the erroneous word set, as well as its ease of implementation.

Specific care is given to segmentation errors, also known as spacing errors, when a whitespace character is omitted between two words or erroneously inserted between two characters in a single word. This error class is harder to generalize as segmentation errors are much more affected by text alignment and fonts than by preceding and following characters. Despite that, the correction segment for erroneous whitespace omission is comprised of the preceding and following characters, and its correction segment would insert the whitespace between both. For example the aligned pair (thegreat, the great) will issue a correction instance of: eg $\longrightarrow$ e g.

4.2.2 Generating Candidates

We use the confusion matrix to expand a tokenized word into its possible corrections, forming all together the correction-candidates vector. The candidate generation is rule-base, where every character segment in a word is looked up in the confusion matrix and replaced by a possible segment correction. An example of the generation process can be seen in Table 4.1, while the confusion matrix XML representation can be seen in Figure 4.1.

4.3 Candidate Ranking

The ranker's rôle is to produce an ordered word vector of correction candidates, calculating a score for each correction candidate, which correlates with how probable a correction is at a specific

[1] Based on modified Levenshtein distance where further primitive edit operations (character merge and split) are used, also known as 2:1 and 1:2 alignments.

Table 4.1: Example of the correction candidates generation

Correction candidates	tlna	grael	wollof	Chima
	the	great	walof	Cbina
	tlme	greet	wall of	Chna
OCR text	tlne	graat	wallof	China

```
<wrongSegment segment="a">
    <correction popularity="20">e</correction>
    <correction popularity="3">o</correction>
</wrongSegment>
<wrongSegment segment="h">
    <correction popularity="12">b</correction>
    <correction popularity="5">li</correction>
</wrongSegment>
```

Figure 4.1: Example XML representation of a confusion matrix

position. Every candidate is scored independently from all others in the word vector; then this candidate is compared to all the other correction-candidates. This stage does not take into account the original OCR output, as it has different features and will be considered in a secondary stage.

As a preliminary stage, the input vector is cleaned from all its non-dictionary words. As the dictionary is based on a large corpus, this procedure has only a negligible deleterious effect, while throwing away a considerable amount of irrelevant candidates hence facilitating the scoring task.

In a secondary stage the word score is calculated by a trained regression model using the word's features as input.

4.3.1 Feature Extraction

The following features are extracted at word-level:

- Confusion weight – The weight attribute of the corruption-correction pair in the confusion matrix, which is the number of occurrences of this pair calculated by the noisy channel on the training set. This feature reflects the OCR engine's specific error model, very much affected by the characters' graphical resemblence.

- Unigram frequency – The unigram document frequency, providing a thematic domain and language feature independent of adjacent words or document context.

- Backward/Forward bigram frequency – The maximal document frequency of the bigram formed by a correction candidate and any candidate at the preceding/following position. This feature is valuable as it contains an intersection between language model and domain context, but is non-existent for many of the bigrams and is redundant if one of the unigrams does not exist. Although the bigrams should have been calculated in comparison to all the

Table 4.2: Example of a training vector for the OCR word “graat”

Candidates	Confusion weight	Unigram frequency	Backward bigram frequency	Forward bigram frequency	Output
graat → great	41 (a→e)	17,222 (great)	1,238 (the great)	73 (great wall)	**1**
graat → greet	5 (aa→ee)	3,124 (greet)	27 (the greet)	0 (greet wall)	0

correction candidates, it was taken only on the OCR output due to calculation complexity and the relative rarity of sequential word errors. Furthermore, we set a cutoff frequency to overcome performances issues in the extraction stage.

No subsequent normalization procedure had to be made in order to linearize the feature effect for later linear regression modeling. In other words, the confusion weight behaves linearly, as well as the term frequency features that proportionally promote frequent corrections relative to their appearance in a similar corpus. Table 4.2 demonstrates the candidates feature extraction result.

4.3.2 Ranking

The ranker is trained from the OCR erroneous word set. Note that this set comprises solely words with extended single-error misspellings, words that the candidate generator supposedly generates. We used the training words to generate their correction-candidates vector and with their extracted features, with the single correct candidate marked with a positive output, as can be seen in Table 4.2.

The appending of these vectors creates a large training set used to create a regression model that attributes a continuous score to every correction candidate. This model is used to rank the correction-candidate vector and to sort it in descending order.

The choice of a ranker over a classifier was made to permit further applications of the ranked vector, such as outputting several words for information retrieval purposes or using them in a secondary correction process as we did. The scores could also be used to evaluate the process itself or to expose the correction confidence to the user.

4.4 Correction Decision Making

The correction decision maker is a classifier that decides whether a replacement should be made of the OCR word with its highest ranked correction-candidate. Such a replacement is made in case the candidate is more likely to be the correct word at this position, as represented in Table 4.3. We will refer to the OCR word and its highest-ranked correction-candidate as an “correction pair”.

Table 4.3: A schematic example of a correction decision training observation

Correction pair	Inverse proportions				OCR confidence	Confusion weight	Decision
(OCR word, top candidate)	unigram	backward bigram	forward bigram	term frequency			
(graat,great)	100	10,000	20,000	500	0.4	15	**1**

4.4.1 Feature Extraction

The decision is calculated by a trained regression model using the correction pair features as input:

- Confidence – An OCR output metric at a character level, which is generalized to a word level by taking the minimal confidence of the characters forming the word.
- Term frequency – The term frequency in the document, calculated by its frequency in the OCR text. This gives document level contextual information, as words forming a document tend to repeat themselves. A common problem of this feature is its bias to consistent OCR mistakes, thus it must be dealt with precaution.
- Proportional dictionary features – The same feature as used above. The proportion metric was included in order to adapt the features to comparative features that have a linear sense. A simple smoothing method was used to handle null-occurrences.

4.4.2 Decision Making

The correction decision is made by a model trained on the total transcribed corpus of correction pairs. Pairs with erroneous OCR word and correct candidate were marked with a positive output, as shown in Table 4.3, indicating that these cases are suitable for replacement.

5 Testing the Model

The model was tested on 50 articles containing a total of 22,000 words. The evaluation of the method was done by a ceiling analysis to understand the performance of every phase independently, as well as a conclusive evaluation for the entire process.

5.1 Image Enhancement

In order to reach the maximal gain of this stage, we need first to see what algorithm settings can bring an accuracy improvement for some of the documents, and, second, to check if we can predict which of the document processing methods is better, namely predicting its accuracy.

5.1.1 Selecting the Algorithms and Thresholds

We ran each of the algorithms proposed in the previous chapter iteratively with different thresholds, saving their accuracy at the document level. An exemplar output of such iterations are presented in Figure 5.1 and Figure 5.2, in which solely positive gains should be considered; to be noted, these improvements have a large variance, induced by few documents with a considerable improvements, while the rest of the corpus' accuracy remains unaffected or decreases. A first threshold screening selected the thresholds with significant potential gain, that is, summing up its positive gains. In order to find the most efficient algorithm set, we solve the optimization problem for each module, that is algorithm family, by calculating the gains of the combinatoric possible sets of power 3.

The approximated improvement of the scale module, which is the solution to the problem maximization presented in Chapter 3, can be seen in Figure 5.3 in a per article view. Out of the 211 training articles 87 had some improvement, leading to an average of absolute 5.6% accuracy improvement on the dataset, which is significantly high. Most of this gain is related to articles with an exceptional improvement, and not an average improvement, and as such this average improvement accuracy should be taken with care as it is very sensitive to the data. A conclusive performance table can be seen in Table 5.1. Note that every module runs independently from the other, implying that the overall gain is not the summation of the different modules. An impirical analysis is shown in Section 5.3.

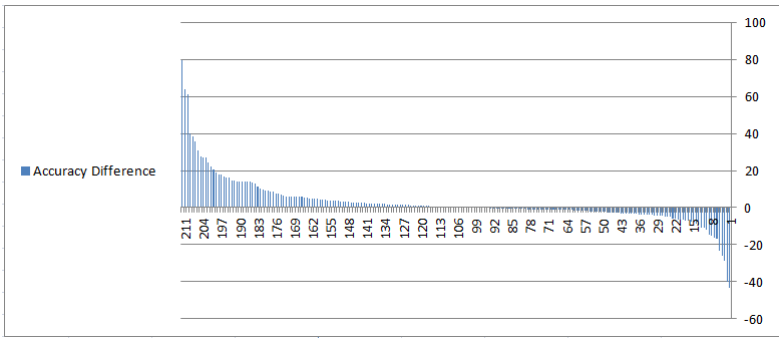

Figure 5.1: Accuracy difference per document. Algorithm = Bi-cubical Scale, Threshold = 1.5. As a single algorithm it was found to have the largest average positive gain of an absolute 4.8% accuracy gain, namely almost 30% WER decrease.

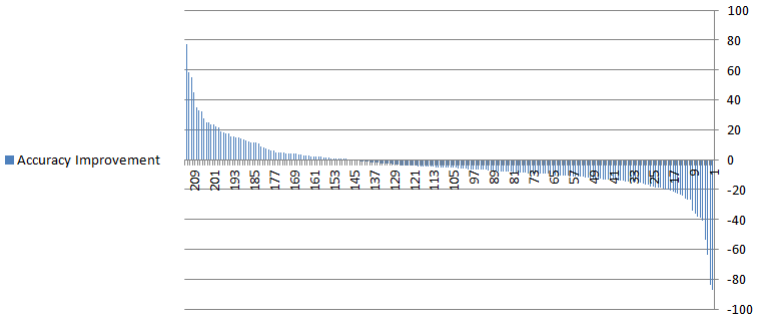

Figure 5.2: Accuracy difference per document. Algorithm = Bi-cubical Scale, Threshold = 2.25. Average positive gain is 4.0%.

As the two last modules did not have much of accuracy gain, we decided to omit them, reducing the processing time and code complexity.

5.1.2 Calculating the Language Model Score

The text confidence information is given by the OCR engine at a character level, so a first stage assessed the accuracy prediction of different statistics to find which of these represents best a document confidence. We used a set of likely statistics, as presented in Section 3.2.1. The results were similar one to another, implying that there is not much information gain playing with this statistic, hence we selected the word-average statistic for simplicity reasons, as it will also be used later on during the text post-correction process. We calculated a simple linear regression to predict the document accuracy, and discontinued the work on the additional bigram feature, as it had a negligible improvement potential. The loss functions demonstrate the model's efficiency in Table 5.2.

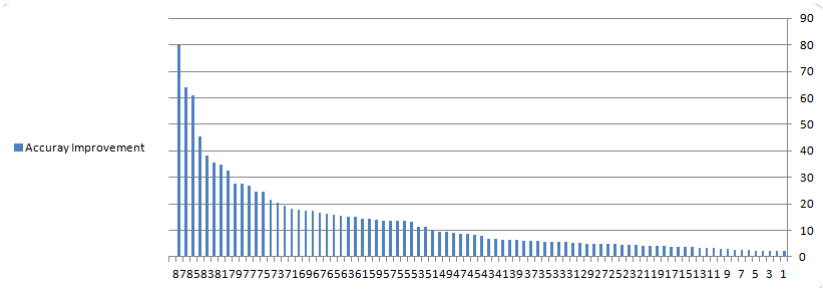

Figure 5.3: Improvement gain per document. Algorithm = Bi-cubical Scale, Threshold = argmax(Accuracy(threshold)). Average positive gain is 5.6%.

5.2 Text Correction

This stage assumes the OCR text of the optimal image as input, not implying any obligation for the first stage to occur for the lexical correction, but rather to present the results in a standardized way.

The method is a pipe of 3 subsequent modules that can be seen as a funnel that narrows down from the correction-candidate generation through the ranking of these words to the classifier that decides of the replacement of the original word with the highest ranked correction candidate. Therefore, these stages are evaluated independently, allowing a ceiling analysis for ongoing improvements.

5.2.1 Retrieving the Correct Word

An analysis of the error type distribution on the test set demonstrates that 60% of the erroneous words had been retrieved in their correct spelling in the correction-candidate generation process. The non-corrected words either did not belong to primitive 1-Levenshtein misspellings, or their correction instance did not occur in the training set. This fair result suggests that the OCR errors belong to a wider error set than the one trained on and can be attributed to random text variability, such as noise or deterioration, or to the existance of low graphical resemblence between large sets of characters.

The confusion matrix that models the OCR errors, which is a representation of the OCR engine's error model, tells us a lot about our corpus and OCR engine. As can be seen in Appendix B, the matrix is sparse. The confusion weights distribution is very characteristic: every character has a small subset of characters they've ever been confused with (explaining 0-sparsity), out of which a smaller subset, up to 4 characters, has relatively high weights and represent the graphic-similar characters. The rest of the subset is larger, and is comprised of characters that have been confused only a single time (1-sparsity), a type of fault that is hard to train for as they are replacements with weak patterns and as such will have a weak signal in the ranking process. Faulty word

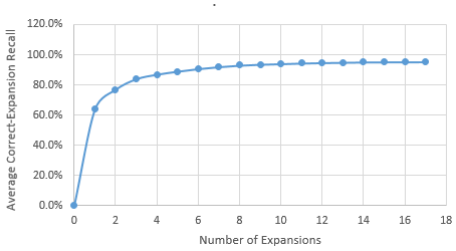

Figure 5.4: Average recall on erroneous words as function of correction-candidates

segmentation is especially sparse, reasoning to think that a straightforward approach that splits or concatenates words and looks them up in the dictionary would improve the process. A further analysis of this matrix shows that more than 85% of the errors occur more than once, half of the errors belong to the class of character substitution, 30% character deletion or insertion, 10% 2:1 and 1:2 alignments and 10% spacing errors.

This result sets an upper bound to the correction efficiency, that would be reached only if the subsequent correction tasks, namely ranking and correction decision, are fully efficient. Improving it could be acquired by enlarging the training set or by generating candidates by an additional fuzzy logic against a dictionary or a trigram dictionary to retrieve possible word completions.

5.2.2 Candidate Ranking

The preliminary dictionary lookup qualification stage leaves on average 30 candidates to rank for a typical word as above, as the non-dictionary words are left out.

We tried various ranking techniques on our training set and validated the results using a validation set. A logistic regression model outperformed other models, yielding the results shown in Figure 5.4.

Calculated for words that have a valid candidate, the best model is able to find the proper correction within the top 5 proposed candidates for 90% of the words, and within the highest ranked candidate for 64% of the words. This result is lower than expectations and it seems the model does not exploit the the existence of the word's bigrams as feature. Rarely an erroneous candidates have a left or right bigram in the dictionary, while correct candidates frequently do, as suggested in 5.2.3. Improving this result may be achieved by a better model, such as a non-linear one, or by expanding the training set in order to enhance the confusion weight feature. Another way to overcome this caveat is taking into account more than the top candidate and cancelling the next phase. The text output would contain multiple words on the same position, complying with the goal of improving retrievabiliy on image documents.

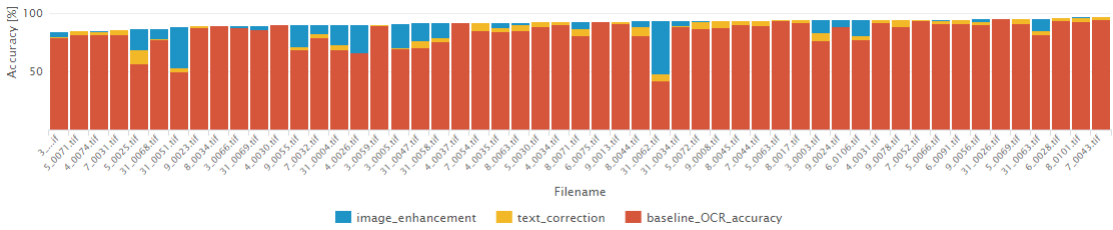

Figure 5.5: Accuracy over the test set

5.2.3 Correction Decision Making

Table 5.3 reports the decision model performance over all words in text. The critical factor in this stage is the false positive rate, namely rejecting a correct OCR word in favor of its correction-candidate, as most of OCR words are correct and such rejections would significantly harm the reliability of the method. Therefore, the trained model gives preference to false positive rate diminution over false negative diminution. The main reason for this excellent result, implying a very efficient classification model, is the bigram proportion feature. In case left or right bigram exists, as occurs in vast majority of cases on correct words thanks to the large corpus on which is based our language model, the respective feature has a high impact on the classifier and would generally lead to a righteous correction.

5.3 Overall Results

An overall representation of the results over the test set is shown in Figure 5.5. The baseline OCR text WER on the test collection is 16.5% on average; applying image enhancement reduces it to 10.4%, while applying on top of that the text correction results to a 7.3% WER, reducing the overall error by over than 50%. This is a considerable improvement given that improvement is harder as WER gets lower. This relative accuracy improvement suggests that this method belongs to the state-of-the-art algorithms for OCR correction. A further examination of the uncorrected errors demonstrates that most originate from deteriorated zones or significant inaccuracies in OCR recognition. The rigorous implementation of the image enhancement and lexical correction methods, shown in [5] and other works presented in the background, and especially their combination by machine-learning techniques, bring most of the additive improvement gains suggested in these.

Image enhancement improves almost twice as much compared to lexical correction. That can be explained by the fact that improving input is generally better than correcting the output, as information is added to the process and exploited in the subsequent OCR tasks. The overall results of the image enhancement demonstrates that the algorithms family are rather dependent by the fact that the overall accuracy gain is not very close to the addition of its two modules, 5.7% vs 7.3%.

The ceiling analysis for the lexical correction clearly designates the correction candidate generation as a weak link, due to the fact that it does not handle out of primitive 1-Levenshtein misspellings, as well as its relatively low generalization on specific error type, such as spacing errors, missing in total more than 35% of the true candidates in their generation process. Adding other correction methods to the current noisy channel one, training based as well as unsupervised methods, would greatly improve the overall process. The ranker could also be improved by working on its accuracy for the highest ranked candidates, as for 35% of erroneous words their correction is among the top 5 ranked candidates but does not make it to the top candidate, which is the only one to make it to the subsequent correction decision. As the current model gives a sufficient result for the top 5 ranking, one may suggest keeping all the top corrections for retrieval tasks. The correction decision maker is effective; with its large training set and indicative features one can expect similar results for different datasets.

Table 5.1: Algorithm Set Selection with Average Accuracy Gain

Module	Algorithm [:Threshold]	Avg. Accuracy Gain
Scale	Default:1, Bicubical:1.5, Bicubical:2.25	5.6%
Binarization	Default, Sauvola:170, Sauvola:230	1.7%
Denoisement	Default:none, Mild, Median	0.8%
Layout Analysis	Default:none, Simple, Complex	0.7%

Table 5.2: Performance of the image candidate classifier

	Selection of most performant image candidate
Using 0-1 Loss	13/211
Using Avg. Weighted Loss	0.4%

Table 5.3: Performance of the decision model for word correction

	OCR word is actually	
	correct	incorrect
Reject OCR word	2%	94%
Accept OCR word	98%	6%

6 Conclusions

This work examined the use of machine-learning techniques for improving OCR accuracy by using the combination of a number of features to enhance an image for OCR and to correct misspelled OCR words. The relative independence of the features, issuing from the language model, OCR model and document context, enables a reliable spelling model that can be trained for many languages and domains. The results of the experiment on Arabic OCR text show an improvement in accuracy for every additional feature, implying the superiority of our multi-feature approach over traditional single-feature approaches that most spelling correction techniques rely on. We can infer from the bigram feature significance that the contextual word completion is a reliable method for a machine as well for the human eye. Lastly, we would like to emphasize the similarity between image enhancement and text correction. Even though both are considered unrelated domains, viz. vision and language, this work and its results demonstrate the mutual significance of the two and mostly the ability to apply a similar correction methodology to both.

For future work, correction-candidate generation and ranking improvements need to be considered further—as implied by the ceiling analysis, while it can be inferred that the image enhancement reached its improvement potential. The caveat of the text correction is the correction-candidate generation based on the confusion matrix. The matrix sparsity in the word segmentation can be fairly easily be generalized to any word by splitting and concatenating every word or bigram, and adding them as correction candidates. A further improvement of candidate correction logic could be achieved by an unsupervised process, facilitating and generalizing the training task without having to rely on an annotated ground truth. Such candidate generation could be achieved by a dictionary-based expansion using a fuzzy unigram logic or a "gap 3- gram" to generate correction candidates based on conditional left and right neighbors [1]. Another and more recent method could be the usage of word embeddings, such as Word2Vec or GloVe trained on a large OCR corpus, to find word similarities between erroneous words and their correction. A less significant improvement could be achieved in candidate ranking by building more complex models than the linear one.

The strength of this method is its ability to "squeeze" the performance of any out-of-the-box OCR engine. Although new OCR methods based on deep learning techniques are emerging and start to commercialize, taking a step further the standard OCR engines' accuracy, a sample testing [4] showed that these state-of-the-art techniques only compare to the results presented in

this work, while a combination of this method to these recent techniques may bring even further improvement.

References

[1] John Evershed and Kent Fitch. Correcting noisy OCR: Context beats confusion. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, DATeCH '14, pages 45–51, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2588-2. doi: 10.1145/2595188.2595200. URL `http://doi.acm.org/10.1145/2595188.2595200`.

[2] Maya R. Gupta, Nathaniel P. Jacobson, and Eric K. Garcia. {OCR} binarization and image pre-processing for searching historical documents. *Pattern Recognition*, 40(2):389 – 397, 2007. ISSN 0031-3203. doi: http://dx.doi.org/10.1016/j.patcog.2006.04.043. URL `http://www.sciencedirect.com/science/article/pii/S0031320306002202`.

[3] Dan Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International, Englewood Cliffs, NJ, 2. ed., [pearson international edition] edition, 2009. ISBN 0-13-504196-1 ; 978-0-13-504196-3.

[4] Ido Kissos. Ocr quality report over jpress corpus for the national israeli library, Jul 2016.

[5] Karen Kukich. Techniques for automatically correcting words in text. *ACM Comput. Surv.*, 24(4):377–439, December 1992. ISSN 0360-0300. doi: 10.1145/146370.146380. URL `http://doi.acm.org/10.1145/146370.146380`.

[6] Seong-Whan Lee and Dae-Seok Ryu. Parameter-free geometric document layout analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1240–1256, Nov 2001. ISSN 0162-8828. doi: 10.1109/34.969115.

[7] Zhidong Lu, Issam Bazzi, András Kornai, John Makhoul, Premkumar Natarajan, and Richard Schwartz. A robust, language-independent OCR system. In Robert J. Mericsko, editor, *Proceedings of the 27th AIPR Workshop: Advances in Computer-Assisted Recognition*, volume 3584. SPIE, 1999.

[8] Walid Magdy and Kareem Darwish. Arabic OCR error correction using character segment correction, language modeling, and shallow morphology. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 408–414, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. ISBN 1-932432-73-6. URL `http://dl.acm.org/citation.cfm?id=1610075.1610132`.

[9] Volker Märgner and Haikal El Abed, editors. *Guide to OCR for Arabic Scripts*. Springer, London, 2012. ISBN 1447140710, 9781447140719.

[10] L. O'Gorman. The document spectrum for page layout analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1162–1173, Nov 1993. ISSN 0162-8828. doi: 10.1109/34.244677.

[11] Parker and Robert et al. Arabic gigaword fifth edition, philadelphia. web download. linguistic data consortium, 2011.

[12] J. Sauvola and M. Pietikinen. Adaptive document image binarization. *Pattern Recognition*, 33(2):225 – 236, 2000. ISSN 0031-3203. doi: http://dx.doi.org/10.1016/S0031-3203(99)00055-2. URL `http://www.sciencedirect.com/science/article/pii/S0031320399000552`.

[13] Klaus U. Schulz and Stoyan Mihov. Fast string correction with Levenshtein automata. *International Journal on Document Analysis and Recognition*, 5(1):67–85, 2002. ISSN 1433-2833. doi: 10.1007/s10032-002-0082-8. URL `http://dx.doi.org/10.1007/s10032-002-0082-8`.

[14] Mikael Tillenius. Efficient generation and ranking of spelling error corrections. Report TRITA-NA-E9621, NADA, 1996.

A Examples from the Dataset

فوز لظافر حامل اللقب

الهزيمة الأولى للبستان في الدوري العماني للكرة

□ مسقط – من سالم الحبسي:

■ تعرض فريق البستان لخسارته الأولى في الدوري العماني لكرة القدم عندما فاز عليه فريق الشرطة بهدف يتيم أحرزه المهاجم السنغالي حسام انجاي في الدقيقة الـ ١١ من المباراة التي أجريت أول من أمس في قمة الأسبوع التاسع.

وبقي البستان رغم خسارته متصدراً ترتيب الدوري للأسبوع الثاني على التوالي على حساب الشرطة الذي كان احتل القمة ثلاثة أسابيع، وتساوى الفريقان في الرصيد ولكل منهما ١١ نقطة، ويتقدم البستان بفارق الأهداف.

وكان هدف الشرطة جاء مبكراً، سعى البستان الى تعديل النتيجة، الا ان كل المحاولات باءت بالفشل، بما فيها ركلة الجزاء التي احتسبت للفريق في الدقيقة الـ ٢٠ وأهدرها الطيب عبدالنور.

وحقق ظفار حامل اللقب فوزاً مهماً على غريمه مرباط بهدف لشريف سنجور في مباراة أقيمت أمام مدرجات خالية تنفيذاً لعقوبة فرضها اتحاد اللعبة على ظفار بعد احداث الشغب التي أثارها جمهوره في بداية الموسم.

واستمر التساوي بين الخابورة وعمان بعد تعادلها السلبي وتقاسمها نقطتي المباراة وارتفع رصيد كل منهما الى ٨ نقاط.

وفجر السيب المتقوقع في قاع الدوري مفاجأة الأسبوع عندما هزم العروبة بطل الكأس (٢/١) محققاً فوزه الثاني هذا الموسم. وبه خرج من القاع ورفع رصيده الى ٨ نقاط، فيما توقف رصيد العروبة عند ٩ نقاط.

وتقدم الفائز بهدفين أحرزهما عبدالحميد عبدالله من ضربة رأسية ويونس الفهدي، اما هدف العروبة فأحرزه سعيد ناصر الفارسي في الدقيقة الأخيرة من المباراة.

وفاز النصر على سداب بهدف أحرزه مروان رجب في الدقيقة الـ ٨٣، فيما دخل صور بطل الكأس في الموسم الماضي في مأزق وتعرض لخسارة غير متوقعة امام الهلال بهدف أحرزه أحمد سعيد.

```
<?xml version='1.0' encoding='utf8'?>
<DOC article="0004" id="HYT_ARB_199401030004" newspaper="HYT_ARB_19940103">
        <headline>فوز لظافر حامل اللقب . الهزيمة الأولى للبستان في الدوري العماني للكرة
</headline>
        <dateline>سالم الحبسي</dateline>
        <text>
            <p>تعرض فريق البستان لخسارته الأولى في الدوري العماني لكرة القدم عندما فاز عليه
فريق الشرطة بهدف يتيم أحرزه المهاجم السنغالي حسام أنجاي في الدقيقة ال 11 من المباراة التي
أجريت أول من أمس في قمة الأسبوع التاسع .</p>
            <p>وبقي البستان رغم خسارته متصدرا ترتيب الدوري للأسبوع الثاني على التوالي على
, حساب الشرطة الذي كان احتل القمة ثلاثة أسابيع , وتساوى الفريقان في الرصيد ولكل منهما 11 نقطة
. ويتقدم البستان بفارق الأهداف .</p>
            <p>وكان هدف الشرطة جاء مبكرا , سعى البستان الى تعديل النتيجة , الا ان كل المحاولات
باءت بالفشل , بما فيها ركلة الجزاء التي احتسبت للفريق في الدقيقة ال 20 وأهدرها الطيب
. عبدالنور</p>
            <p>وحقق ظفار حامل اللقب فوزا مهما على غريمه مرباط بهدف لشريف سنجور في مباراة
أقيمت أمام مدرجات خالية تنفيذا لعقوبة فرضها اتحاد اللعبة على ظفار بعد احداث الشغب التي
أثارها جمهوره في بداية الموسم .</p>
            <p>واستمر التساوي بين الخابورة وعمان بعد تعادلها السلبي وتقاسمها نقطتي المباراة
. وارتفع رصيد كل منهما الى 8 نقاط .</p>
            <p>وفجر السيب المتقوقع في قاع الدوري مفاجأة الأسبوع عندما هزم العروبة بطل الكأس
محققا فوزه الثاني هذا الموسم . وبه خرج من القاع ورفع رصيده الى 8 نقاط , فيما توقف ( 1 / 2
. رصيد العروبة عند 9 نقاط .</p>
            <p>وتقدم الفائز بهدفين أحرزهما عبدالحميد عبدالله من ضربة رأسية ويونس الفهدي , اما
. هدف العروبة فأحرزه سعيد ناصر الفارسي في الدقيقة الأخيرة من المباراة .</p>
            <p>وفاز النصر على سداب بهدف أحرزه مروان رجب في الدقيقة ال 83 , فيما دخل صور بطل
. الكأس في الموسم الماضي في مأزق وتعرض لخسارة غير متوقعة امام الهلال بهدف أحرزه أحمد سعيد .</p>
        </text>
    </DOC>
```

Figure A.1: “Al-Hayat” newspaper article, 03.01.1994

داني تلهو مع الدلفين. (ريكس)

داني تحد زوج المستقبل وتتحدى حماتها

□ نيويورك – ريكس فيتشرز:

■ بدأ العد التنازلي لحفلة زفاف مغنية البوب والممثلة الأسترالية داني مينوغ وخطيبها جوليان ماكماهون المقررة في كانون الثاني (يناير) الجاري.

الا ان داني التي شقت طريقها في عالم الغناء مقتدية بشقيقتها كايلي، وجدت نفسها في ما يشبه حلقة من حلقات مسلسل «ديناستي» الاميركي حين شنت عليها والدة خطيبها السيدة الارستقراطية الليدي سونيا ماكماهون حرباً استخدمت فيها سلاح الصمت. ويبدو ان الليدي سونيا معترضة على زواج ابنها «العريق» من داني (٢٢ عاماً) ابنة الطبقة العاملة.

اندلعت الحرب يوم عيد ميلاد داني الثاني والعشرين حين تلقت عبر الفاكس نسخة عن موضوع نشرته مجلة عنوانه: «الليدي سونيا تقول: لا تتزوج داني الديسكو». وعلى الرغم من ان داني وخطيبها وصفا الموضوع بانه «جارح ومؤلم» فانهما لم ينكرا ما ورد فيه.

نجاحات

وفي ما عدا الحرب العائلية الدائرة، تسير حياة داني من نجاح الى آخر، فسجلت البوماً غنائياً ناجحاً، وأنهت دورها في فيلم «أسرار» الذي حاز اعجاب النقاد، وأطلقت مجموعة من الأزياء التي تحمل اسمها، والأهم من كل ذلك انها عثرت على رجل أحلامها.

وتحاول داني ان تبقي تفاصيل حفلة الزفاف سرية، الا ان البعض يؤكد انها ستقام في ثلاثة طوابق في فندق خمسة نجوم في ملبورن. كما اتفقت مع شركة خاصة في لندن تمدها بعدد من الحراس للتأكد من هويات المدعوين.

تقول داني التي تعيش حالياً في نيويورك: «كل شيء في حياتي يبدو على ما يرام. أصحو من النوم في الصباح وابتسامة كبيرة مرسومة على وجهي. أمشي في شوارع نيويورك وأنا ابتسم للمارة».

```
<?xml version='1.0' encoding='utf8'?>
<DOC article="0028" id="HYT_ARB_199401030028" newspaper="HYT_ARB_19940103">
        <headline>داني تجد زوج المستقبل وتتحدى حماتها</headline>
        <dateline>ريكس فيتشرز</dateline>
        <text>
            <p>بدأ العد التنازلي لحفلة زفاف مغنية البوب والممثلة الاسترالية داني مينوغ
وخطيبها جوليان ماكماهون المقررة في كانون الثاني ( يناير ) الجاري .</p>
            <p>الا ان داني التي شقت طريقها في عالم الغناء مقتدية بشقيقتها كايلي , وجدت نفسها
في ما يشبه حلقة من حلقات مسلسل " ديناستي " الاميركي حين شنت عليها والدة خطيبها السيدة
الارستقراطية الليدي سونيا ماكماهون حربا استخدمت فيها سلاح الصمت . ويبدو ان الليدي سونيا معترضة
على زواج ابنها " العريق " من داني ( 22 عاما ) ابنة الطبقة العاملة .</p>
            <p>اندلعت الحرب يوم عيد ميلاد داني الثاني والعشرين حين تلقت عبر الفاكس نسخة عن
موضوع نشرته مجلة عنوانه : " الليدي سونيا تقول : لا تتزوج داني الديسكو " . وعلى الرغم من ان
داني وخطيبها وصفا الموضوع بأنه " جارح ومؤلم " فانهما لم ينكرا ما ورد فيه .</p>
            <p>نجاحات</p>
            <p>وفي ما عدا الحرب العائلية الدائرة , تسير حياة داني من نجاح الى آخر , فسجلت
ألبوما غنائيا ناجحا , وأنهت دورها في فيلم " أسرار " الذي حاز اعجاب النقاد , وأطلقت مجموعة من
الأزياء التي تحمل اسمها , والأهم من كل ذلك انها عثرت على رجل أحلامها .</p>
            <p>وتحاول داني ان تبقي تفاصيل حفلة الزفاف سرية , الا ان البعض يؤكد انها ستقام في
ثلاثة طوابق في فندق خمسة نجوم في ملبورن . كما اتفقت مع شركة خاصة في لندن تمدها بعدد من الحراس
للتأكد من هويات المدعوين .</p>
            <p>تقول داني التي تعيش حاليا في نيويورك : " كل شيء في حياتي يبدو على ما يرام
أصحو من النوم في الصباح وابتسامة كبيرة مرسومة على وجهي . أمشي في شوارع نيويورك وأنا ابتسم
للمارة " .</p>
        </text>
    </DOC>
```

Figure A.2: “Al-Hayat” newspaper article, 03.01.1994

١٠ أعوام لحل مشكلة ملايين المعاقين في فيتنام

■ هانوي – ا ف ب – اكد تقرير لوزارة العمل نشرته وكالة الاعلام الفيتنامية أول من أمس ان عدد المعاقين عقلياً وجسدياً في فيتنام يراوح بين خمسة ملايين وسبعة ملايين، منهم ٢,٢ مليون يحتاجون الى اطراف صناعية واعادة تاهيل.

واضاف التقرير ان حوالي ٢٠٠ الف شخص مبتوري الذرع او الارجل يحتاجون الى اطراف صناعية، وان هناك ٤٠ الف معاق على مقاعد متحركة. كما ان هناك ٦٠ الفاً الى ٨٠ الف معاق آخرين يحتاجون الى معالجة مرتبطة بتقويم الاعضاء.

وذكر التقرير ان مراكز اعادة التاهيل الفيتنامية اجرت ٢٣٤٤ جراحة بين ١٩٩٠ و١٩٩٣ وعالجت ٦٣ الف مريض ووفرت ٥٢ الف طرف صناعي وألف مقعد متحرك.

وأضاف ان ١٣٢٥٠ معاقاً تلقوا معاقاً تلقوا علاجاً في مراكز اعادة التاهيل الاربعة عشر في فيتنام ومنها ثلاثة مخصصة للاطفال المشلولين والمصابين بتشوه عضلي أو الذين يعانون تشوهات ناجمة عن سوء التغذية.

وتتلقى المراكز مساعدات مالية وفنية من منظمات اجنبية ولا سيما أميركية غير ان خدماتها لا تغطي سوى ٣٠ الى ٤٠ في المئة من الحاجة الى الاطراف الصناعية من جراء نقص الاعتمادات، كما اوضح التقرير الذي أشار الى ان فيتنام تحتاج الى عشرة اعوام لحل المشكلة.

وتفيد الاحصاءات الرسمية المتوافرة في وزارة العمل ان عدد معاقي حرب فيتنام يبلغ مليون شخص.

```
<?xml version='1.0' encoding='utf8'?>
<DOC article="0029" id="HYT_ARB_199401030029" newspaper="HYT_ARB_19940103">
        <headline>10 أعوام لحل مشكلة ملايين المعاقين في فيتنام</headline>
        <dateline>هانوي</dateline>
        <text>
            <p>أكد تقرير لوزارة العمل نشرته وكالة الاعلام الفيتنامية أول من أمس ان عدد المعاقين عقليا وجسديا في فيتنام يراوح بين خمسة ملايين وسبعة ملايين , منهم 2,2 مليون يحتاجون الى اطراف صناعية واعادة تأهيل .</p>
            <p>وأضاف التقرير ان حوالي 200 ألف شخص مبتوري الاذرع أو الارجل يحتاجون الى اطراف صناعية , وان هناك 40 ألف معاق على مقاعد متحركة . كما ان هناك 60 ألفا الى 80 ألف معاق آخرين يحتاجون الى معالجة مرتبطة بتقويم الاعضاء .</p>
            <p>وذكر التقرير ان مراكز اعادة التأهيل الفيتنامية اجرت 2344 جراحة بين 1990 و1993 وعالجت 63 ألف مريض ووفرت 52 ألف طرف صناعي وألف مقعد متحرك .</p>
            <p>وأضاف ان 13250 معاقا تلقوا معاقا تلقوا علاجا في مراكز اعادة التأهيل الاربعة عشر في فيتنام ومنها ثلاثة مخصصة للاطفال المشلولين والمصابين بتشوه عضلي أو الذين يعانون تشوهات ناجمة عن سوء التغذية .</p>
            <p>وتتلقى المراكز مساعدات مالية وفنية من منظمات اجنبية ولا سيما أميركية غير ان خدماتها لا تغطي سوى 30 الى 40 في المئة من الحاجة الى الاطراف الصناعية من جراء نقص الاعتمادات , كما اوضح التقرير الذي أشار الى ان فيتنام تحتاج الى عشرة اعوام لحل المشكلة .</p>
            <p>وتفيد الاحصاءات الرسمية المتوافرة في وزارة العمل ان عدد معاقي حرب فيتنام يبلغ مليون شخص .</p>
        </text>
    </DOC>
```

Figure A.3: “Al-Hayat” newspaper article, 03.01.1994

حطت الطائرة في لندن وعلى متنها مسافر جديد

■ لندن – أ ف ب – أعلنت شركة الخطوط الجوية التايلاندية ان طائرة بوينغ ٧٤٧ تابعة لها هبطت أول من أمس في مطار هيثرو في لندن، وعلى متنها شخص لم يكن اسمه مدرجاً في لوائح المسافرين لدى اقلاعها من مطار نيودلهي... بعدما ولدت إمرأة على متنها طفلاً.

فبعد ثلاث ساعات من اقلاع الطائرة، أي حوالي الساعة ٢,٣٠ بتوقيت غرينتش، بدأت المسافرة الهندية التي لم تكشف هويتها تشعر بالام المخاض، ومن حسن حظها ان طبيبين كانا في عداد المسافرين فسهلا عملية الولادة التي تمت في «عيادة» أعدت على عجل في إحدى مقصورات الدرجة الأولى.

وقد أخطر قبطان الطائرة التي كانت تقوم بالرحلة «تي جي ٩١٤» مكتب الخطوط الجوية التايلاندية في مطار كوبنهاغن الأقرب الى خط سير البوينغ ٧٤٧ بالولادة. ومن مطار كوبنهاغن أبرق مكتب الخطوط الجوية التايلاندية الى مطار هيثرو طالباً اتخاذ الاجراءات الضرورية لاستقبال «المسافر» الاضافي.

```
<?xml version='1.0' encoding='utf8'?>
<DOC article="0032" id="HYT_ARB_199401030032" newspaper="HYT_ARB_19940103">
        <headline>حطت الطائرة في لندن وعلى متنها مسافر جديد</headline>
        <dateline>لندن</dateline>
        <text>
            <p>أعلنت شركة الخطوط الجوية التايلاندية ان طائرة بوينغ 747 تابعة لها هبطت أول من
أمس في مطار هيثرو في لندن , وعلى متنها شخص لم يكن اسمه مدرجا في لوائح المسافرين لدى اقلاعها
. من مطار نيودلهي . . . بعدما ولدت إمرأة على متنها طفلا</p>
            <p>فبعد ثلاث ساعات من اقلاع الطائرة , أي حوالي الساعة 2.30 بتوقيت غرينتش , بدأت
المسافرة الهندية التي لم تكشف هويتها تشعر بآلام المخاض , ومن حسن حظها ان طبيبين كانا في عداد
المسافرين فسهلا عملية الولادة التي تمت في " عيادة " أعدت على عجل في إحدى مقصورات الدرجة الأولى
.
</p>
            <p>وقد أخطر قبطان الطائرة التي كانت تقوم بالرحلة " تي جي 914 " مكتب الخطوط
الجوية التايلاندية في مطار كوبنهاغن الأقرب الى خط سير البوينغ 747 بالولادة . ومن مطار كوبنهاغن
" أبرق مكتب الخطوط الجوية التايلاندية الى مطار هيثرو طالبا اتخاذ الاجراءات الضرورية لاستقبال
. المسافر " الاضافي</p>
            <p>اول الكلام :</p>
            <p>: " من همزية " أحمد شوقي</p>
            <p>الله فوق الخلق . . . فيها وحده -</p>
            <p>والناس تحت لوائها : أكفاء</p>
            <p>والدين يسر , والخلافة بيعة</p>
            <p>! والأمر شورى , والحقوق قضاء</p>
            <p>في زحام الشعارات , والأشكال السياسية التي تدعي " الديموقراطية " والحفاوة برأي
الشعب وسماع صوته . . . وهي في حقيقتها تعاني ايضا من : أشكال في الأزمات الاقتصادية , والواقع
. . . الأمني , والبناء التنموي , ومساحة الحرية التي لا بد أن يتمتع بها كل مواطن هناك</p>
            <p>في هذا الزحام الذي تبدو خسائر الشعوب فيه أكثر من مكاسبهم , ومن حصولهم على
الحرية . . . ينطلق هذا الانسان العربي من أعظم البقاع على الأرض , وبكل الإتساع في صحرائه : هادئا
. . . رزينا , مستعصما بتشريع دينه الاسلامي العظيم . . . متسلحا بما حصل عليه من علوم , وثقافات
نهل الكثير منها من جامعاته , ومن جامعات العالم المتقدمة بدراساتها وبأطروحاتها . . . فإذا هو
يطلع الى التاريخ من خلال تسجيله لحدث , قد يبدو غريبا على العالم الذي مارس كل أشكال
```

Figure A.4: “Al-Hayat” newspaper article, 03.01.1994

رابين يتهم عرفات بنقض «وثيقة القاهرة» والاردن يعطي المنظمة «فرصة أخيرة» للتنسيق

□ القدس المحتلة – من ربى الحصري:
□ عمان – من سلامه نعمات:
□ تونس –من رشيدخشانة:

■ تصاعدت الأزمة أمس بين اسرائيل ومنظمة التحرير الفلسطينية حول استئناف محادثات غزة – أريحا، فيما شنت اسرائيل هجوماً على رئيس المنظمة ياسر عرفات واتهمته بـ «التراجع» عن اتفاقات يبرمها مبعوثوه معها وأعلنت انها لن تستأنف المحادثات في طابا الا على اساس «الاتفاق الذي تم التوصل اليه في القاهرة».

كذلك صعّد الأردن لهجته ازاء منظمة التحرير، اذ أمهل الملك حسين الرئيس الفلسطيني «فرصة أخيرة» لاستئناف التنسيق بين الأردن والمنظمة بشأن الترتيبات الاقتصادية في الاراضي المحتلة ومسألة المعابر بين أريحا والأردن.

كما طالب العاهل الاردني عرفات، في تصريحاته التي بثها التلفزيون الاردني، بان «يتوقف نهائياً عن الحديث عن انشاء اتحاد كونفيديرالي مع الأردن ولا باي شكل من الأشكال ولا باية صيغة أخرى». واشار الى ضرورة «البناء من القاعدة الى القمة» في سياق صوغ العلاقات المستقبلية بين الأردن والكيان الفلسطيني في الأراضي المحتلة.

وعلم في تونس ان قيادة منظمة التحرير اجرت اخيراً اتصالات بالادارة الاميركية طلباً لتوسطها في المفاوضات الحالية، لكن واشنطن لم تتحمس للرغبة الفلسطينية. وقال العاهل الاردني في خطاب وجهه لكبار ضباط القوات المسلحة انه أوضح للزعيم

التتمة في الصفحة (٤)

فلسطينية تحتج لدى جنود اسرائيليين على اطلاق الرصاص على الاطفال في غزة امس.(ا ف ب)


```
<?xml version='1.0' encoding='utf8'?>
<DOC article="0035" id="HYT_ARB_199401030035" newspaper="HYT_ARB_19940103">
        <headline>" رابين يتهم عرفات بنقض " وثيقة القاهرة " والاردن يعطي المنظمة " فرصة أخيرة
للتنسيق</headline>
        <dateline>ربى الحصري - سلامه نعمات - رشيد خشانة</dateline>
        <text>
            <p>تصاعدت الأزمة أمس بين اسرائيل ومنظمة التحرير الفلسطينية حول استئناف محادثات
غزة - أريحا , فيما شنت اسرائيل هجوما على رئيس المنظمة ياسر عرفات واتهمته ب " التراجع " عن
اتفاقات يبرمها مبعوثوه معها وأعلنت انها لن تستأنف المحادثات في طابا الا على اساس " الاتفاق
الذي تم التوصل اليه في القاهرة " .</p>
            <p>كذلك صعد الأردن لهجته ازاء منظمة التحرير , اذ أمهل الملك حسين الرئيس الفلسطيني
فرصة أخيرة " لاستئناف التنسيق بين الأردن والمنظمة بشأن الترتيبات الاقتصادية في الاراضي المحتلة "
ومسألة المعابر بين أريحا والأردن .</p>
            <p>" كما طالب العاهل الاردني عرفات , في تصريحاته التي بثها التلفزيون الأردني , بأن
يتوقف نهائيا عن الحديث عن انشاء اتحاد كونفيديرالي مع الأردن ولا بأي شكل من الأشكال ولا بأية صيغة
أخرى " . وأشار الى ضرورة " البناء من القاعدة الى القمة " في سياق صوغ العلاقات المستقبلية بين
الاردن والكيان الفلسطيني في الأراضي المحتلة .</p>
            <p>وعلم في تونس ان قيادة منظمة التحرير اجرت اخيرا اتصالات بالادارة الاميركية طلبا
لتوسطها في المفاوضات الحالية , لكن واشنطن لم تتحمس للرغبة الفلسطينية . وقال العاهل الاردني في
خطاب وجهه لكبار ضباط القوات المسلحة انه أوضح للزعيم الفلسطيني خلال لقائه به يوم الخميس الماضي
ان " هذه هي آخر فرصة لمعالجة الاولويات المطلوبة وبالسرعة الممكنة " , محذرا من انه في حال عدم
التزام منظمة التحرير فإن " كل طرف سيتحمل المسؤولية على حدة " . وأكد العاهل الأردني بأنه " لا
يمكن بعد الآن القول ان هناك تنسيقا ما لم يكن هناك تنسيق أو ان هناك اتفاقا ما لم يكن هناك
اتفاق " .</p>
            <p>ومن المقرر ان يبدأ مسؤولون في منظمة التحرير برئاسة السيد فاروق القدومي , مدير
الدائرة السياسية في منظمة التحرير , محادثات مع المسؤولين الاردنيين غدا الثلثاء أو بعد غد
الاربعاء لتسوية الخلافات التي نتجت عن غياب التنسيق الثنائي ورفض السيد عرفات اقرار اتفاق
اقتصادي تم التوصل اليه مع الأردن .</p>
            <p>وفي اسرائيل صرح رئيس الوزراء اسحق رابين للصحافيين بعد اجتماع الحكومة
الاسرائيلية الاسبوعي , أمس , بأن اسرائيل بعثت برسالة الى تونس أمس " أكدنا فيها بوضوح ان
" محادثات طابا لن تستأنف الا على أساس وثيقة التفاهم التي تم التوصل اليها في مفاوضات القاهرة
```


Figure A.5: “Al-Hayat” newspaper article, 03.01.1994

B Confusion Matrix Excerpt

Description	CorrectNormalizedWord	OcrNormalizedWord
החלפת אות	الثاني	الثاتي
החלפת אות	هذا	كذا
החלפת אות	بعض	بنض
החלפת אות	اضطهد	0ضطهد
החלפת אות	فلسطين	للسطين
החלפת אות	يراد	براد
החלפת אות	للمسلمين	للفسلمين
החלפת אות	وربما	وريما
החלפת אות	هنا	حنا
הוספת אות	الاراضي	الاراضبي
הוספת אות	والعنف	والعنفب
הוספת אות	والعنف	والعنفب
הוספת אות	لا	لتا
הוספת אות	المعتقلين	المعتسقلين
הוספת אות	لتتوج	لتتورج
הוספת אות	فالاحري	فسالاحري
הוספת אות	اقتصاديه	ا0قتصاديه
הוספת אות	وحتي	وحتيو
הוספת אות	فقط	ففقط
הוספת אות	اجيالا	اجيالا0
הוספת אות	المحبه	المحشبه
הוספת אות	المنطقه	المنطقيه
הוספת אות	ليست	ليسمت
הוספת אות	الغربي	الغربتي
מחיקת אות	المحتله	الحتله
החלפת אות בשתי אותיות	ان	اب0
החלפת אות בשתי אותיות	مهدوره	ننهدوره
החלפת אות בשתי אותיות	السياق	السياو
החלפת שתי אותיות באות	العالم	العاط

Figure B.1: Noisy Channel output for an article

```
<?xml version="1.0" encoding="utf-8"?>
<config xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xsd="http://www.w3.
  <mistake type="letter subsititution">
  <mistake type="letter insertion">
  <mistake type="letter deletion">
  <mistake type="one letter to two">
  <mistake type="two letters to one">
</config>
```

Figure B.2: Confusion Matrix Error Types

```
<?xml version="1.0" encoding="utf-8"?>
<config xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xsd=
"http://www.w3.org/2001/XMLSchema">
  <mistake type="letter subsititution">
    <wrongSegment string="و">
      <correctSegment popularity="56">ر</correctSegment>
      <correctSegment popularity="8">ا</correctSegment>
      <correctSegment popularity="7">ب</correctSegment>
      <correctSegment popularity="6">ي</correctSegment>
      <correctSegment popularity="5">ز</correctSegment>
      <correctSegment popularity="4">ق</correctSegment>
      <correctSegment popularity="3">ه</correctSegment>
      <correctSegment popularity="3">ل</correctSegment>
      <correctSegment popularity="3">ف</correctSegment>
      <correctSegment popularity="2">ش</correctSegment>
      <correctSegment popularity="2">ك</correctSegment>
      <correctSegment popularity="1">ط</correctSegment>
      <correctSegment popularity="1">ء</correctSegment>
      <correctSegment popularity="1">س</correctSegment>
      <correctSegment popularity="1">ذ</correctSegment>
      <correctSegment popularity="1">ن</correctSegment>
      <correctSegment popularity="1">ج</correctSegment>
    </wrongSegment>
    <wrongSegment string="ك">
      <correctSegment popularity="6">ه</correctSegment>
      <correctSegment popularity="6">ف</correctSegment>
      <correctSegment popularity="4">ا</correctSegment>
      <correctSegment popularity="3">ل</correctSegment>
      <correctSegment popularity="2">ي</correctSegment>
      <correctSegment popularity="2">ق</correctSegment>
      <correctSegment popularity="2">ء</correctSegment>
      <correctSegment popularity="1">ع</correctSegment>
      <correctSegment popularity="1">ن</correctSegment>
      <correctSegment popularity="1">ج</correctSegment>
      <correctSegment popularity="1">ض</correctSegment>
      <correctSegment popularity="1">خ</correctSegment>
      <correctSegment popularity="1">غ</correctSegment>
      <correctSegment popularity="1">ب</correctSegment>
      <correctSegment popularity="1">م</correctSegment>
      <correctSegment popularity="1">ث</correctSegment>
      <correctSegment popularity="1">د</correctSegment>
    </wrongSegment>
    <wrongSegment string="ذ">
      <correctSegment popularity="36">د</correctSegment>
      <correctSegment popularity="14">ت</correctSegment>
      <correctSegment popularity="13">ي</correctSegment>
      <correctSegment popularity="2">ا</correctSegment>
      <correctSegment popularity="2">ن</correctSegment>
      <correctSegment popularity="2">م</correctSegment>
      <correctSegment popularity="2">ز</correctSegment>
      <correctSegment popularity="1">ه</correctSegment>
      <correctSegment popularity="1">ث</correctSegment>
      <correctSegment popularity="1">ل</correctSegment>
      <correctSegment popularity="1">ف</correctSegment>
      <correctSegment popularity="1">ب</correctSegment>
      <correctSegment popularity="1">ح</correctSegment>
    </wrongSegment>
```

```
<wrongSegment string="ي">
  <correctSegment popularity="49">ب</correctSegment>
  <correctSegment popularity="15">ن</correctSegment>
  <correctSegment popularity="10">ه</correctSegment>
  <correctSegment popularity="10">ت</correctSegment>
  <correctSegment popularity="9">ل</correctSegment>
  <correctSegment popularity="9">ح</correctSegment>
  <correctSegment popularity="7">ا</correctSegment>
  <correctSegment popularity="7">ر</correctSegment>
  <correctSegment popularity="6">م</correctSegment>
  <correctSegment popularity="6">و</correctSegment>
  <correctSegment popularity="4">د</correctSegment>
  <correctSegment popularity="3">ع</correctSegment>
  <correctSegment popularity="2">ذ</correctSegment>
  <correctSegment popularity="2">س</correctSegment>
  <correctSegment popularity="2">ث</correctSegment>
  <correctSegment popularity="2">ف</correctSegment>
  <correctSegment popularity="2">ك</correctSegment>
  <correctSegment popularity="1">ق</correctSegment>
  <correctSegment popularity="1">چ</correctSegment>
  <correctSegment popularity="1">ش</correctSegment>
</wrongSegment>
<wrongSegment string="ه">
  <correctSegment popularity="2">ي</correctSegment>
  <correctSegment popularity="1">د</correctSegment>
  <correctSegment popularity="1">ب</correctSegment>
</wrongSegment>
<wrongSegment string="ج">
  <correctSegment popularity="32">ح</correctSegment>
  <correctSegment popularity="19">ع</correctSegment>
  <correctSegment popularity="9">ب</correctSegment>
  <correctSegment popularity="6">ي</correctSegment>
  <correctSegment popularity="6">م</correctSegment>
  <correctSegment popularity="2">ن</correctSegment>
  <correctSegment popularity="1">د</correctSegment>
  <correctSegment popularity="1">ل</correctSegment>
  <correctSegment popularity="1">ط</correctSegment>
</wrongSegment>
<wrongSegment string="ف">
  <correctSegment popularity="27">م</correctSegment>
  <correctSegment popularity="9">ت</correctSegment>
  <correctSegment popularity="6">ي</correctSegment>
  <correctSegment popularity="6">و</correctSegment>
  <correctSegment popularity="5">ب</correctSegment>
  <correctSegment popularity="5">ا</correctSegment>
  <correctSegment popularity="3">ل</correctSegment>
  <correctSegment popularity="3">غ</correctSegment>
  <correctSegment popularity="3">ذ</correctSegment>
  <correctSegment popularity="3">ع</correctSegment>
  <correctSegment popularity="2">د</correctSegment>
  <correctSegment popularity="2">ز</correctSegment>
  <correctSegment popularity="2">ج</correctSegment>
  <correctSegment popularity="1">ط</correctSegment>
  <correctSegment popularity="1">ه</correctSegment>
  <correctSegment popularity="1">ح</correctSegment>
  <correctSegment popularity="1">ق</correctSegment>
</wrongSegment>
```

```
    <wrongSegment string="ز">
      <correctSegment popularity="55">ر</correctSegment>
      <correctSegment popularity="1">د</correctSegment>
      <correctSegment popularity="1">ل</correctSegment>
      <correctSegment popularity="1">ه</correctSegment>
      <correctSegment popularity="1">ف</correctSegment>
      <correctSegment popularity="1">و</correctSegment>
      <correctSegment popularity="1">ن</correctSegment>
      <correctSegment popularity="1">غ</correctSegment>
    </wrongSegment>
    <wrongSegment string="ت">
      <correctSegment popularity="40">ء</correctSegment>
      <correctSegment popularity="39">ث</correctSegment>
      <correctSegment popularity="32">ن</correctSegment>
      <correctSegment popularity="18">ي</correctSegment>
      <correctSegment popularity="17">ذ</correctSegment>
      <correctSegment popularity="11">ل</correctSegment>
      <correctSegment popularity="6">د</correctSegment>
      <correctSegment popularity="5">ع</correctSegment>
      <correctSegment popularity="4">ق</correctSegment>
      <correctSegment popularity="4">م</correctSegment>
      <correctSegment popularity="2">ب</correctSegment>
      <correctSegment popularity="2">غ</correctSegment>
      <correctSegment popularity="1">و</correctSegment>
      <correctSegment popularity="1">ف</correctSegment>
      <correctSegment popularity="1">س</correctSegment>
      <correctSegment popularity="1">ا</correctSegment>
      <correctSegment popularity="1">ك</correctSegment>
      <correctSegment popularity="1">ج</correctSegment>
      <correctSegment popularity="1">ص</correctSegment>
      <correctSegment popularity="1">ط</correctSegment>
    </wrongSegment>
    <wrongSegment string="ل">
      <correctSegment popularity="16">ا</correctSegment>
      <correctSegment popularity="10">ن</correctSegment>
      <correctSegment popularity="9">ب</correctSegment>
      <correctSegment popularity="6">ذ</correctSegment>
      <correctSegment popularity="6">ت</correctSegment>
      <correctSegment popularity="6">د</correctSegment>
      <correctSegment popularity="5">م</correctSegment>
      <correctSegment popularity="4">ك</correctSegment>
      <correctSegment popularity="4">ي</correctSegment>
      <correctSegment popularity="4">و</correctSegment>
      <correctSegment popularity="4">ف</correctSegment>
      <correctSegment popularity="3">ه</correctSegment>
      <correctSegment popularity="2">ر</correctSegment>
      <correctSegment popularity="2">ع</correctSegment>
      <correctSegment popularity="2">ق</correctSegment>
      <correctSegment popularity="1">غ</correctSegment>
      <correctSegment popularity="1">ط</correctSegment>
      <correctSegment popularity="1">ح</correctSegment>
      <correctSegment popularity="1">ء</correctSegment>
      <correctSegment popularity="1">س</correctSegment>
      <correctSegment popularity="1">ش</correctSegment>
    </wrongSegment>
    <wrongSegment string="د">
      <correctSegment popularity="4">ل</correctSegment>
```

```
      <correctSegment popularity="4">ه</correctSegment>
      <correctSegment popularity="4">م</correctSegment>
      <correctSegment popularity="3">ف</correctSegment>
      <correctSegment popularity="3">ن</correctSegment>
      <correctSegment popularity="3">و</correctSegment>
      <correctSegment popularity="3">ر</correctSegment>
      <correctSegment popularity="3">ذ</correctSegment>
      <correctSegment popularity="2">ا</correctSegment>
      <correctSegment popularity="2">ي</correctSegment>
      <correctSegment popularity="2">ض</correctSegment>
      <correctSegment popularity="1">ك</correctSegment>
      <correctSegment popularity="1">ث</correctSegment>
      <correctSegment popularity="1">چ</correctSegment>
      <correctSegment popularity="1">س</correctSegment>
      <correctSegment popularity="1">ش</correctSegment>
      <correctSegment popularity="1">ء</correctSegment>
    </wrongSegment>
    <wrongSegment string="غ">
      <correctSegment popularity="51">ع</correctSegment>
      <correctSegment popularity="4">ا</correctSegment>
      <correctSegment popularity="2">ن</correctSegment>
      <correctSegment popularity="1">ف</correctSegment>
      <correctSegment popularity="1">ت</correctSegment>
      <correctSegment popularity="1">ق</correctSegment>
      <correctSegment popularity="1">د</correctSegment>
    </wrongSegment>
    <wrongSegment string="خ">
      <correctSegment popularity="48">ح</correctSegment>
      <correctSegment popularity="5">ع</correctSegment>
      <correctSegment popularity="4">غ</correctSegment>
      <correctSegment popularity="4">ت</correctSegment>
      <correctSegment popularity="3">ن</correctSegment>
      <correctSegment popularity="2">چ</correctSegment>
      <correctSegment popularity="2">ك</correctSegment>
      <correctSegment popularity="1">ل</correctSegment>
      <correctSegment popularity="1">ق</correctSegment>
      <correctSegment popularity="1">ء</correctSegment>
      <correctSegment popularity="1">ث</correctSegment>
      <correctSegment popularity="1">د</correctSegment>
      <correctSegment popularity="1">ف</correctSegment>
      <correctSegment popularity="1">ش</correctSegment>
      <correctSegment popularity="1">ي</correctSegment>
    </wrongSegment>
    <wrongSegment string="ظ">
      <correctSegment popularity="9">ط</correctSegment>
      <correctSegment popularity="1">ء</correctSegment>
      <correctSegment popularity="1">م</correctSegment>
      <correctSegment popularity="1">ف</correctSegment>
    </wrongSegment>
    <wrongSegment string="ض">
      <correctSegment popularity="26">ص</correctSegment>
      <correctSegment popularity="16">م</correctSegment>
      <correctSegment popularity="2">د</correctSegment>
      <correctSegment popularity="1">ع</correctSegment>
    </wrongSegment>
    <wrongSegment string="س">
      <correctSegment popularity="9">م</correctSegment>
```

```
      <correctSegment popularity="4">ن</correctSegment>
      <correctSegment popularity="3">ع</correctSegment>
      <correctSegment popularity="3">ي</correctSegment>
      <correctSegment popularity="2">ب</correctSegment>
      <correctSegment popularity="2">ت</correctSegment>
      <correctSegment popularity="2">ر</correctSegment>
      <correctSegment popularity="2">ا</correctSegment>
      <correctSegment popularity="1">ث</correctSegment>
      <correctSegment popularity="1">و</correctSegment>
      <correctSegment popularity="1">ه</correctSegment>
      <correctSegment popularity="1">د</correctSegment>
      <correctSegment popularity="1">ء</correctSegment>
      <correctSegment popularity="1">ذ</correctSegment>
      <correctSegment popularity="1">ش</correctSegment>
      <correctSegment popularity="1">س</correctSegment>
      <correctSegment popularity="1">ن</correctSegment>
      <correctSegment popularity="1">ف</correctSegment>
      <correctSegment popularity="1">ل</correctSegment>
    </wrongSegment>
    <wrongSegment string="ا">
      <correctSegment popularity="42">ع</correctSegment>
      <correctSegment popularity="33">ي</correctSegment>
      <correctSegment popularity="28">ل</correctSegment>
      <correctSegment popularity="22">ه</correctSegment>
      <correctSegment popularity="16">و</correctSegment>
      <correctSegment popularity="13">ن</correctSegment>
      <correctSegment popularity="6">د</correctSegment>
      <correctSegment popularity="3">س</correctSegment>
      <correctSegment popularity="3">م</correctSegment>
      <correctSegment popularity="2">ر</correctSegment>
      <correctSegment popularity="2">ب</correctSegment>
      <correctSegment popularity="2">ك</correctSegment>
      <correctSegment popularity="2">ص</correctSegment>
      <correctSegment popularity="2">ف</correctSegment>
      <correctSegment popularity="2">چ</correctSegment>
      <correctSegment popularity="1">ط</correctSegment>
      <correctSegment popularity="1">غ</correctSegment>
      <correctSegment popularity="1">ق</correctSegment>
      <correctSegment popularity="1">ح</correctSegment>
    </wrongSegment>
    <wrongSegment string="م">
      <correctSegment popularity="15">ي</correctSegment>
      <correctSegment popularity="10">ه</correctSegment>
      <correctSegment popularity="10">ا</correctSegment>
      <correctSegment popularity="9">ن</correctSegment>
      <correctSegment popularity="8">ب</correctSegment>
      <correctSegment popularity="7">ع</correctSegment>
      <correctSegment popularity="4">ل</correctSegment>
      <correctSegment popularity="3">ف</correctSegment>
      <correctSegment popularity="2">و</correctSegment>
      <correctSegment popularity="2">ذ</correctSegment>
      <correctSegment popularity="2">س</correctSegment>
      <correctSegment popularity="1">ش</correctSegment>
      <correctSegment popularity="1">ت</correctSegment>
      <correctSegment popularity="1">ء</correctSegment>
      <correctSegment popularity="1">ح</correctSegment>
      <correctSegment popularity="1">ض</correctSegment>
```

```
      <correctSegment popularity="1">ز</correctSegment>
      <correctSegment popularity="1">خ</correctSegment>
      <correctSegment popularity="1">س</correctSegment>
      <correctSegment popularity="1">د</correctSegment>
    </wrongSegment>
    <wrongSegment string="ن">
      <correctSegment popularity="35">ت</correctSegment>
      <correctSegment popularity="25">د</correctSegment>
      <correctSegment popularity="23">ي</correctSegment>
      <correctSegment popularity="12">ب</correctSegment>
      <correctSegment popularity="9">ل</correctSegment>
      <correctSegment popularity="7">ه</correctSegment>
      <correctSegment popularity="7">ث</correctSegment>
      <correctSegment popularity="6">ع</correctSegment>
      <correctSegment popularity="5">م</correctSegment>
      <correctSegment popularity="5">ء</correctSegment>
      <correctSegment popularity="5">ذ</correctSegment>
      <correctSegment popularity="4">ر</correctSegment>
      <correctSegment popularity="4">س</correctSegment>
      <correctSegment popularity="3">ا</correctSegment>
      <correctSegment popularity="2">ف</correctSegment>
      <correctSegment popularity="2">ق</correctSegment>
      <correctSegment popularity="1">ز</correctSegment>
      <correctSegment popularity="1">چ</correctSegment>
      <correctSegment popularity="1">ح</correctSegment>
      <correctSegment popularity="1">خ</correctSegment>
    </wrongSegment>
    <wrongSegment string="ط">
      <correctSegment popularity="5">ت</correctSegment>
      <correctSegment popularity="5">ل</correctSegment>
      <correctSegment popularity="3">د</correctSegment>
      <correctSegment popularity="3">م</correctSegment>
      <correctSegment popularity="2">ن</correctSegment>
      <correctSegment popularity="2">ش</correctSegment>
      <correctSegment popularity="2">ب</correctSegment>
      <correctSegment popularity="2">ء</correctSegment>
      <correctSegment popularity="1">ك</correctSegment>
      <correctSegment popularity="1">ه</correctSegment>
      <correctSegment popularity="1">ق</correctSegment>
      <correctSegment popularity="1">ا</correctSegment>
    </wrongSegment>
    <wrongSegment string="س">
      <correctSegment popularity="17">م</correctSegment>
      <correctSegment popularity="3">و</correctSegment>
      <correctSegment popularity="2">ض</correctSegment>
      <correctSegment popularity="2">ر</correctSegment>
      <correctSegment popularity="2">ت</correctSegment>
      <correctSegment popularity="1">ب</correctSegment>
      <correctSegment popularity="1">ث</correctSegment>
      <correctSegment popularity="1">س</correctSegment>
      <correctSegment popularity="1">ي</correctSegment>
      <correctSegment popularity="1">ل</correctSegment>
      <correctSegment popularity="1">د</correctSegment>
    </wrongSegment>
    <wrongSegment string="٠">
      <correctSegment popularity="18">ا</correctSegment>
      <correctSegment popularity="7">و</correctSegment>
```

```
      <correctSegment popularity="3">ع</correctSegment>
      <correctSegment popularity="2">ي</correctSegment>
      <correctSegment popularity="2">ن</correctSegment>
      <correctSegment popularity="2">ل</correctSegment>
      <correctSegment popularity="1">ك</correctSegment>
      <correctSegment popularity="1">2</correctSegment>
      <correctSegment popularity="1">4</correctSegment>
    </wrongSegment>
    <wrongSegment string="ع">
      <correctSegment popularity="40">ا</correctSegment>
      <correctSegment popularity="7">م</correctSegment>
      <correctSegment popularity="5">غ</correctSegment>
      <correctSegment popularity="3">ب</correctSegment>
      <correctSegment popularity="3">ن</correctSegment>
      <correctSegment popularity="3">ي</correctSegment>
      <correctSegment popularity="2">ه</correctSegment>
      <correctSegment popularity="2">ء</correctSegment>
      <correctSegment popularity="2">د</correctSegment>
      <correctSegment popularity="1">س</correctSegment>
      <correctSegment popularity="1">ل</correctSegment>
      <correctSegment popularity="1">ح</correctSegment>
      <correctSegment popularity="1">ذ</correctSegment>
      <correctSegment popularity="1">ق</correctSegment>
    </wrongSegment>
    <wrongSegment string="ب">
      <correctSegment popularity="35">ي</correctSegment>
      <correctSegment popularity="8">م</correctSegment>
      <correctSegment popularity="6">ل</correctSegment>
      <correctSegment popularity="6">ا</correctSegment>
      <correctSegment popularity="4">ن</correctSegment>
      <correctSegment popularity="3">ك</correctSegment>
      <correctSegment popularity="3">ح</correctSegment>
      <correctSegment popularity="2">س</correctSegment>
      <correctSegment popularity="2">ت</correctSegment>
      <correctSegment popularity="2">چ</correctSegment>
      <correctSegment popularity="1">ث</correctSegment>
      <correctSegment popularity="1">ق</correctSegment>
      <correctSegment popularity="1">ر</correctSegment>
      <correctSegment popularity="1">ف</correctSegment>
      <correctSegment popularity="1">ه</correctSegment>
    </wrongSegment>
    <wrongSegment string="ر">
      <correctSegment popularity="5">ز</correctSegment>
      <correctSegment popularity="4">س</correctSegment>
      <correctSegment popularity="3">ب</correctSegment>
      <correctSegment popularity="3">و</correctSegment>
      <correctSegment popularity="2">ن</correctSegment>
      <correctSegment popularity="2">ل</correctSegment>
      <correctSegment popularity="2">د</correctSegment>
      <correctSegment popularity="1">چ</correctSegment>
      <correctSegment popularity="1">ا</correctSegment>
      <correctSegment popularity="1">ع</correctSegment>
      <correctSegment popularity="1">ه</correctSegment>
      <correctSegment popularity="1">غ</correctSegment>
    </wrongSegment>
    <wrongSegment string="ه">
      <correctSegment popularity="25">م</correctSegment>
```

```
      <correctSegment popularity="24">ي</correctSegment>
      <correctSegment popularity="14">ا</correctSegment>
      <correctSegment popularity="5">ن</correctSegment>
      <correctSegment popularity="3">ف</correctSegment>
      <correctSegment popularity="3">و</correctSegment>
      <correctSegment popularity="3">ر</correctSegment>
      <correctSegment popularity="3">ب</correctSegment>
      <correctSegment popularity="2">ع</correctSegment>
      <correctSegment popularity="2">د</correctSegment>
      <correctSegment popularity="2">ك</correctSegment>
      <correctSegment popularity="2">ت</correctSegment>
      <correctSegment popularity="1">ء</correctSegment>
      <correctSegment popularity="1">ض</correctSegment>
      <correctSegment popularity="1">ق</correctSegment>
      <correctSegment popularity="1">ذ</correctSegment>
      <correctSegment popularity="1">س</correctSegment>
      <correctSegment popularity="1">ل</correctSegment>
      <correctSegment popularity="1">ح</correctSegment>
    </wrongSegment>
    <wrongSegment string="ق">
      <correctSegment popularity="32">ف</correctSegment>
      <correctSegment popularity="10">ت</correctSegment>
      <correctSegment popularity="5">ي</correctSegment>
      <correctSegment popularity="4">م</correctSegment>
      <correctSegment popularity="3">ع</correctSegment>
      <correctSegment popularity="3">غ</correctSegment>
      <correctSegment popularity="3">ن</correctSegment>
      <correctSegment popularity="2">ا</correctSegment>
      <correctSegment popularity="2">ذ</correctSegment>
      <correctSegment popularity="2">ء</correctSegment>
      <correctSegment popularity="2">ث</correctSegment>
      <correctSegment popularity="2">خ</correctSegment>
      <correctSegment popularity="1">ص</correctSegment>
      <correctSegment popularity="1">ه</correctSegment>
      <correctSegment popularity="1">ك</correctSegment>
      <correctSegment popularity="1">ر</correctSegment>
    </wrongSegment>
    <wrongSegment string="ث">
      <correctSegment popularity="8">ت</correctSegment>
      <correctSegment popularity="4">ء</correctSegment>
      <correctSegment popularity="2">ب</correctSegment>
      <correctSegment popularity="1">ف</correctSegment>
      <correctSegment popularity="1">ن</correctSegment>
      <correctSegment popularity="1">ي</correctSegment>
    </wrongSegment>
    <wrongSegment string="ؤ">
      <correctSegment popularity="4">و</correctSegment>
      <correctSegment popularity="1">ز</correctSegment>
    </wrongSegment>
    <wrongSegment string="ش">
      <correctSegment popularity="11">ث</correctSegment>
      <correctSegment popularity="7">س</correctSegment>
      <correctSegment popularity="5">ي</correctSegment>
      <correctSegment popularity="4">م</correctSegment>
      <correctSegment popularity="4">ء</correctSegment>
      <correctSegment popularity="3">ك</correctSegment>
      <correctSegment popularity="2">ن</correctSegment>
```

```
      <correctSegment popularity="2">س</correctSegment>
      <correctSegment popularity="2">د</correctSegment>
      <correctSegment popularity="2">ع</correctSegment>
      <correctSegment popularity="1">ف</correctSegment>
      <correctSegment popularity="1">ت</correctSegment>
      <correctSegment popularity="1">غ</correctSegment>
      <correctSegment popularity="1">ل</correctSegment>
      <correctSegment popularity="1">ب</correctSegment>
    </wrongSegment>
    <wrongSegment string="ج">
      <correctSegment popularity="1">ج</correctSegment>
    </wrongSegment>
    <wrongSegment string="٢">
      <correctSegment popularity="1">ا</correctSegment>
      <correctSegment popularity="1">1</correctSegment>
    </wrongSegment>
    <wrongSegment string="ى">
      <correctSegment popularity="4">م</correctSegment>
      <correctSegment popularity="3">ر</correctSegment>
      <correctSegment popularity="2">ا</correctSegment>
      <correctSegment popularity="2">ه</correctSegment>
      <correctSegment popularity="1">ن</correctSegment>
      <correctSegment popularity="1">ف</correctSegment>
    </wrongSegment>
    <wrongSegment string="6">
      <correctSegment popularity="1">ا</correctSegment>
    </wrongSegment>
    <wrongSegment string="٢">
      <correctSegment popularity="1">ا</correctSegment>
    </wrongSegment>
    <wrongSegment string="ح">
      <correctSegment popularity="8">م</correctSegment>
      <correctSegment popularity="5">ج</correctSegment>
      <correctSegment popularity="5">ي</correctSegment>
      <correctSegment popularity="4">ع</correctSegment>
      <correctSegment popularity="3">ن</correctSegment>
      <correctSegment popularity="2">ب</correctSegment>
      <correctSegment popularity="1">ل</correctSegment>
      <correctSegment popularity="1">ه</correctSegment>
      <correctSegment popularity="1">ك</correctSegment>
      <correctSegment popularity="1">ش</correctSegment>
      <correctSegment popularity="1">ت</correctSegment>
      <correctSegment popularity="1">ق</correctSegment>
      <correctSegment popularity="1">س</correctSegment>
      <correctSegment popularity="1">ذ</correctSegment>
      <correctSegment popularity="1">ا</correctSegment>
    </wrongSegment>
    <wrongSegment string="ة">
      <correctSegment popularity="12">ا</correctSegment>
      <correctSegment popularity="4">ي</correctSegment>
      <correctSegment popularity="3">ن</correctSegment>
      <correctSegment popularity="1">ل</correctSegment>
      <correctSegment popularity="1">م</correctSegment>
      <correctSegment popularity="1">ء</correctSegment>
      <correctSegment popularity="1">ش</correctSegment>
      <correctSegment popularity="1">ق</correctSegment>
    </wrongSegment>
```

```
<wrongSegment string="څ">
  <correctSegment popularity="1">ه</correctSegment>
</wrongSegment>
<wrongSegment string="ء">
  <correctSegment popularity="1">ه</correctSegment>
  <correctSegment popularity="1">ا</correctSegment>
  <correctSegment popularity="1">ن</correctSegment>
  <correctSegment popularity="1">ع</correctSegment>
</wrongSegment>
<wrongSegment string="٥">
  <correctSegment popularity="1">4</correctSegment>
  <correctSegment popularity="1">ه</correctSegment>
</wrongSegment>
<wrongSegment string="أ">
  <correctSegment popularity="2">ل</correctSegment>
  <correctSegment popularity="1">ت</correctSegment>
  <correctSegment popularity="1">م</correctSegment>
  <correctSegment popularity="1">ب</correctSegment>
</wrongSegment>
<wrongSegment string="ئ">
  <correctSegment popularity="2">ن</correctSegment>
  <correctSegment popularity="1">م</correctSegment>
</wrongSegment>
<wrongSegment string="٤">
  <correctSegment popularity="1">1</correctSegment>
  <correctSegment popularity="1">م</correctSegment>
</wrongSegment>
<wrongSegment string="پ">
  <correctSegment popularity="3">ي</correctSegment>
</wrongSegment>
<wrongSegment string="ى">
  <correctSegment popularity="1">ي</correctSegment>
</wrongSegment>
<wrongSegment string="٩">
  <correctSegment popularity="1">8</correctSegment>
</wrongSegment>
<wrongSegment string="9">
  <correctSegment popularity="1">ه</correctSegment>
</wrongSegment>
<wrongSegment string="ک">
  <correctSegment popularity="1">ك</correctSegment>
  <correctSegment popularity="1">ء</correctSegment>
  <correctSegment popularity="1">ف</correctSegment>
</wrongSegment>
<wrongSegment string="إ">
  <correctSegment popularity="1">ع</correctSegment>
</wrongSegment>
<wrongSegment string="ر">
  <correctSegment popularity="1">ل</correctSegment>
</wrongSegment>
<wrongSegment string="4">
  <correctSegment popularity="2">ه</correctSegment>
  <correctSegment popularity="1">ا</correctSegment>
</wrongSegment>
<wrongSegment string="څ">
  <correctSegment popularity="1">م</correctSegment>
</wrongSegment>
```

```
    <wrongSegment string="۶">
      <correctSegment popularity="1">ن</correctSegment>
    </wrongSegment>
  </mistake>
  <mistake type="letter insertion">
    <wrongSegment string="جيد">
      <correctSegment popularity="2">جد</correctSegment>
    </wrongSegment>
    <wrongSegment string="قند">
      <correctSegment popularity="2">قد</correctSegment>
    </wrongSegment>
    <wrongSegment string=" ءا">
      <correctSegment popularity="1"> ا</correctSegment>
    </wrongSegment>
    <wrongSegment string="۰د ">
      <correctSegment popularity="10">د </correctSegment>
    </wrongSegment>
    <wrongSegment string="نـن ">
      <correctSegment popularity="3">ن </correctSegment>
    </wrongSegment>
    <wrongSegment string="خسر">
      <correctSegment popularity="19">خر</correctSegment>
    </wrongSegment>
    <wrongSegment string="منا">
      <correctSegment popularity="5">ما</correctSegment>
    </wrongSegment>
    <wrongSegment string="لـر ">
      <correctSegment popularity="2">ل </correctSegment>
    </wrongSegment>
    <wrongSegment string="او ">
      <correctSegment popularity="11">ا </correctSegment>
    </wrongSegment>
    <wrongSegment string="راي">
      <correctSegment popularity="5">ري</correctSegment>
    </wrongSegment>
    <wrongSegment string="كسا">
      <correctSegment popularity="2">كا</correctSegment>
    </wrongSegment>
    <wrongSegment string="صتي">
      <correctSegment popularity="1">صي</correctSegment>
    </wrongSegment>
    <wrongSegment string="نـن ">
      <correctSegment popularity="7">ى </correctSegment>
    </wrongSegment>
    <wrongSegment string="لـنس">
      <correctSegment popularity="5">لس</correctSegment>
    </wrongSegment>
    <wrongSegment string="فت ">
      <correctSegment popularity="11">ف </correctSegment>
    </wrongSegment>
    <wrongSegment string=" لـك">
      <correctSegment popularity="3"> ك</correctSegment>
    </wrongSegment>
    <wrongSegment string=" لأ">
      <correctSegment popularity="4"> أ</correctSegment>
    </wrongSegment>
    <wrongSegment string="يعه">
```

```
    <correctSegment popularity="1">يـه</correctSegment>
  </wrongSegment>
  <wrongSegment string=" ذك">
    <correctSegment popularity="1"> ك</correctSegment>
  </wrongSegment>
  <wrongSegment string=" ال">
    <correctSegment popularity="14"> ل</correctSegment>
  </wrongSegment>
  <wrongSegment string="نـذ ">
    <correctSegment popularity="16">ن </correctSegment>
  </wrongSegment>
  <wrongSegment string="ينة">
    <correctSegment popularity="2">يـة</correctSegment>
  </wrongSegment>
  <wrongSegment string="فـا ">
    <correctSegment popularity="16">ف </correctSegment>
  </wrongSegment>
  <wrongSegment string=" آت">
    <correctSegment popularity="1"> ت</correctSegment>
  </wrongSegment>
  <wrongSegment string="يسد">
    <correctSegment popularity="1">يـد</correctSegment>
  </wrongSegment>
  <wrongSegment string="بين">
    <correctSegment popularity="5">بـن</correctSegment>
  </wrongSegment>
  <wrongSegment string="هسا">
    <correctSegment popularity="4">هـا</correctSegment>
  </wrongSegment>
  <wrongSegment string=" ٠ب">
    <correctSegment popularity="11"> ب</correctSegment>
  </wrongSegment>
  <wrongSegment string="دوي">
    <correctSegment popularity="1">دي</correctSegment>
  </wrongSegment>
  <wrongSegment string="تـت ">
    <correctSegment popularity="56">ت </correctSegment>
  </wrongSegment>
  <wrongSegment string="ركي">
    <correctSegment popularity="2">ري</correctSegment>
  </wrongSegment>
  <wrongSegment string="تسي">
    <correctSegment popularity="6">تـي</correctSegment>
  </wrongSegment>
  <wrongSegment string="يب ">
    <correctSegment popularity="1">ي </correctSegment>
  </wrongSegment>
  <wrongSegment string=" خا">
    <correctSegment popularity="2"> ا</correctSegment>
  </wrongSegment>
  <wrongSegment string="سكا">
    <correctSegment popularity="1">سـا</correctSegment>
  </wrongSegment>
  <wrongSegment string="أ٠م">
    <correctSegment popularity="2">أم</correctSegment>
  </wrongSegment>
  <wrongSegment string="تـب ">
```

```
    <correctSegment popularity="4">ت </correctSegment>
  </wrongSegment>
  <wrongSegment string="سني">
    <correctSegment popularity="10">سي</correctSegment>
  </wrongSegment>
  <wrongSegment string="ينو">
    <correctSegment popularity="1">يو</correctSegment>
  </wrongSegment>
  <wrongSegment string="ديم">
    <correctSegment popularity="1">دم</correctSegment>
  </wrongSegment>
  <wrongSegment string="صنو">
    <correctSegment popularity="2">صو</correctSegment>
  </wrongSegment>
  <wrongSegment string="ي• ">
    <correctSegment popularity="18">ي </correctSegment>
  </wrongSegment>
  <wrongSegment string="تمو">
    <correctSegment popularity="2">تو</correctSegment>
  </wrongSegment>
  <wrongSegment string="ستا">
    <correctSegment popularity="1">سا</correctSegment>
  </wrongSegment>
  <wrongSegment string="فني">
    <correctSegment popularity="26">في</correctSegment>
  </wrongSegment>
  <wrongSegment string=" ه•">
    <correctSegment popularity="9"> ه</correctSegment>
  </wrongSegment>
  <wrongSegment string=" أه">
    <correctSegment popularity="1"> ه</correctSegment>
  </wrongSegment>
  <wrongSegment string="يا ">
    <correctSegment popularity="6">ي </correctSegment>
  </wrongSegment>
  <wrongSegment string="د•ي">
    <correctSegment popularity="1">دي</correctSegment>
  </wrongSegment>
  <wrongSegment string=" ت•">
    <correctSegment popularity="4"> ت</correctSegment>
  </wrongSegment>
  <wrongSegment string="لفا">
    <correctSegment popularity="1">لا</correctSegment>
  </wrongSegment>
  <wrongSegment string=" ضم">
    <correctSegment popularity="3"> م</correctSegment>
  </wrongSegment>
  <wrongSegment string="يي ">
    <correctSegment popularity="28">ي </correctSegment>
  </wrongSegment>
  <wrongSegment string="فن ">
    <correctSegment popularity="1">ف </correctSegment>
  </wrongSegment>
  <wrongSegment string="لى ">
    <correctSegment popularity="4">ل </correctSegment>
  </wrongSegment>
  <wrongSegment string="بسر">
```

```
      <correctSegment popularity="1">بـر</correctSegment>
    </wrongSegment>
    <wrongSegment string="مـنش">
      <correctSegment popularity="3">مـش</correctSegment>
    </wrongSegment>
    <wrongSegment string=" كـا">
      <correctSegment popularity="8"> ا</correctSegment>
    </wrongSegment>
    <wrongSegment string="شس ">
      <correctSegment popularity="25">ش </correctSegment>
    </wrongSegment>
    <wrongSegment string="مـت ">
      <correctSegment popularity="1">م </correctSegment>
    </wrongSegment>
    <wrongSegment string="٧٠٥">
      <correctSegment popularity="1">٧٥</correctSegment>
    </wrongSegment>
    <wrongSegment string=" وت">
      <correctSegment popularity="4"> ت</correctSegment>
    </wrongSegment>
    <wrongSegment string=" وا">
      <correctSegment popularity="28"> ا</correctSegment>
    </wrongSegment>
    <wrongSegment string="زه ">
      <correctSegment popularity="1">ز </correctSegment>
    </wrongSegment>
    <wrongSegment string="تـصي">
      <correctSegment popularity="1">تـي</correctSegment>
    </wrongSegment>
    <wrongSegment string="يـيك">
      <correctSegment popularity="2">يـك</correctSegment>
    </wrongSegment>
    <wrongSegment string=" وف">
      <correctSegment popularity="16"> ف</correctSegment>
    </wrongSegment>
    <wrongSegment string="سبـي">
      <correctSegment popularity="3">سي</correctSegment>
    </wrongSegment>
    <wrongSegment string="8٠ ">
      <correctSegment popularity="2">8 </correctSegment>
    </wrongSegment>
    <wrongSegment string="لـيق">
      <correctSegment popularity="2">لـق</correctSegment>
    </wrongSegment>
    <wrongSegment string="ذوك">
      <correctSegment popularity="1">ذك</correctSegment>
    </wrongSegment>
    <wrongSegment string="لـي ">
      <correctSegment popularity="12">ل </correctSegment>
    </wrongSegment>
    <wrongSegment string="بـه ">
      <correctSegment popularity="12">ى </correctSegment>
    </wrongSegment>
    <wrongSegment string="مـيل">
      <correctSegment popularity="1">مـل</correctSegment>
    </wrongSegment>
    <wrongSegment string="عسد">
```

List of Figures

2.1 "Al-Hayat" newspaper article, 06.01.1994 9

2.2 "Al-Hayat" newspaper transcription, 06.01.1994 9

3.1 Image Enhancement Implementation 11

4.1 Example XML representation of a confusion matrix 16

5.1 Accuracy difference per document. Algorithm = Bi-cubical Scale, Threshold = 1.5. As a single algorithm it was found to have the largest average positive gain of an absolute 4.8% accuracy gain, namely almost 30% WER decrease. 20

5.2 Accuracy difference per document. Algorithm = Bi-cubical Scale, Threshold = 2.25. Average positive gain is 4.0%. 20

5.3 Improvement gain per document. Algorithm = Bi-cubical Scale, Threshold = argmax(Accuracy(threshold)). Average positive gain is 5.6%. 21

5.4 Average recall on erroneous words as function of correction-candidates 22

5.5 Accuracy over the test set 23

A.1 "Al-Hayat" newspaper article, 03.01.1994 31

A.2 "Al-Hayat" newspaper article, 03.01.1994 32

A.3 "Al-Hayat" newspaper article, 03.01.1994 33

A.4 "Al-Hayat" newspaper article, 03.01.1994 34

A.5 "Al-Hayat" newspaper article, 03.01.1994 35

B.1 Noisy Channel output for an article 36

B.2 Confusion Matrix Error Types 36

List of Tables

4.1 Example of the correction candidates generation 16

4.2 Example of a training vector for the OCR word “graat” 17

4.3 A schematic example of a correction decision training observation 18

5.1 Algorithm Set Selection with Average Accuracy Gain 25

5.2 Performance of the image candidate classifier 25

5.3 Performance of the decision model for word correction 25

אוניברסיטת תל-אביב TEL AVIV UNIVERSITY

הפקולטה למדעים מדויקים ע"ש ריימונד ובברלי סאקלר

בית הספר בלווטניק למדעי המחשב

שיפור אחזוריות מסמכים סרוקים באמצעות תיקון תמונה ופלט OCRבעזרת מודל שפה

חיבור זה מוגש בתור עבודת מחקר לקראת תואר מוסמך בחקר ביצועים על-ידי

עידו קיסוס

העבודה נכתבה בבית הספר למדעי המחשב

בהנחיית פר' נחום דרשוביץ

נובמבר 2016

תשרי התשע"ז

תקציר

התיזה מתחקה אחר השימוש במסווגים שעברו למידה לצורך שיפור האחזוריות של מסמכים סרוקים ע"י שיפור דיוק ה OCR וחילול תיקונים אפשריים ברמת המילה.

השיטה מציעה יישום בו-זמני של מספר מודלי תיקון תמונה וטקסט, שלאחריהם שלב הערכה שמאפשר את בחירת המודל היעיל ביותר לתמונת הקלט ואת התיקונים הסבירים ביותר לכל מילה. בחירה זו מגובה בשני מסווגים שעברו למידה בהתבסס על מאפיינים של הפלט הטקסטואלי של ה OCR שעיקרם בנויים על מודל שפה. מודל תיקון התמונה הטוב ביותר מיושם על התמונה, בעוד שהתיקונים הסבירים ביותר לכל מילה מצורפים למילה מפלט ה OCR ומאונדקסים באותו מיקום. מסווג נוסף ואופציונאלי מכריע האם מילת OCR צריכה להיות מוחלפת ע"י מועמד התיקון הסביר ביותר, דבר המאפשר לשיטה להתאים למטרות שאינן אחזוריות.

השיטה נשענת על סט מסמכים מתויג, המורכב מתמונה ותעתיק שלה, בנוסף לקורפוס הלקוח מאותו עולם תוכן על-בסיסו נבנה מודל השפה. השיטה ניתנת ליישום לכל שפה שעוקבת אחר כללי סגמנטציית מילים פשוטים, כלומר בלי מילים מורכבות (compound) ומתאימה במיוחד לשפות עשירות מורפולוגית.

ניסויים על קורפוס עיתונות ערבית מראה שגישה זו משפרת באופן משמעותי את דיוק ה OCR כשהיא מפחיתה בקורפוס את מספר השגיאות ב 50%