Fusion of Multiple Corrupted Transmissions and its effect on Information Retrieval

Walid Magdy[1], Kareem Darwish[2], and Mohsen Rashwan[1]
[1]Electronics and Communications Department, Faculty of Engineering, Cairo University, Egypt
[2]Faculty of Computers and Information, Cairo University, Egypt
wmagdy@eg.ibm.com,kareem@darwish.org,mrashwan@rdi-eg.com

Abstract – **Much previous work has focused on correction of OCR degraded text with little work addressing the possibility of fusing the generated text from different OCR systems, which are assumed to produce different types of errors. This paper explores text fusion, which involves the use of language modeling to determine which OCR system (if any) properly recognized individual words. The technique was applied on Arabic text that was synthetically degraded using different models of OCR degradation. The different degraded versions were consequently fused leading to a new version with significantly fewer errors than any of the original versions. Also, the effect of fusion on retrieval was examined.**

Index Terms – **Text Fusion, OCR, Language Modeling, and Information Retrieval**

I. INTRODUCTION

Since the advent of the printing press in the fifteenth century, the amount of printed text has grown overwhelmingly. Although a great deal of text is now generated in electronic character-coded formats (HTML, word processor files ... etc.), many documents available only in print remain important. This is due in part to the existence of large collections of legacy documents available only in print, and in part because printed text remains an important distribution channel that can effectively deliver information without the technical infrastructure that is required to deliver character-coded text. These factors are particularly important for Arabic, which is widely used in places where the installed computer infrastructure is often quite limited. Printed documents can be browsed and indexed for retrieval relatively easily in limited quantities, but effective access to the contents of large collections requires some form of automation.

One such form of automation is to scan the documents (to produce document images) and subsequently perform OCR on the document images to convert them into text. Typically, the OCR process introduces errors in the text representation of the document images. The introduced errors are more pronounced in Arabic OCR due to some of the orthographic and morphological features of Arabic and negatively impact retrieval effectiveness.

Previous work on OCRed text focused on two main aspects. The first involves improving Information Retrieval (IR) effectiveness on the degraded text using query garbling in conjunction with structured or balanced queries [1, 2, 11]. The second focuses on correcting OCR errors to improve IR effectiveness [3, 4, 5, 6].

Previously mentioned OCR correction work depends on the presence of only one source of degraded text. In this paper, a new technique is introduced that assumes the presence of more that one version of the degraded text, each with different types of errors. These different versions of degraded text are to be fused using a language model to try to determine which version (if any) has an uncorrupted version for each word in the text. The introduced technique in this paper is applied on synthetically degraded Arabic text using different OCR degradation models. The technique could be extended to cases where different types of degradation such as spelling errors, automatic speech recognition (ASR) errors … etc. are present. The effect of fusion on IR effectiveness is also examined.

This paper is organized as follows: Section 2 provides a detailed definition for text fusion; Section 3 describes the data set; Section 4 describes the experimental setup and reports the results; and Section 5 concludes the paper.

II. TEXT FUSION

Text fusion can be defined as follows: given a clean text set $S_0 = \{s_{01} \dots s_{0j} \dots s_{0m}\}$ and n degraded versions $S_i = \{s_{i1} \dots s_{ij} \dots s_{im}\}$, where $1 \leq i \leq n$ and s_{ij} is the degraded version of s_{0j}, S_i can be represented as $S_0 + \varepsilon_i$, where ε_i is the set of edit operations necessary to transform from the clean version to the degraded version and ε_i could result from to the data entry process (OCR, ASR, typing … etc). As illustrated in Fig. 1, the goal is to obtain a new version $S_0' = S_0 + \varepsilon_0'$, where S_0' is obtained by picking the closest s_{ij} to s_{0j} leading to $\varepsilon_0' <$ minimum(ε_j). In this work, a trigram language model is used to attempt to pick the closest s_{ij} to s_{0j} by finding S_0' that maximizes the language model probability.

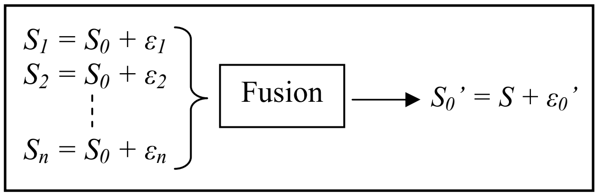
$S_1 = S_0 + \varepsilon_1$
$S_2 = S_0 + \varepsilon_2$
$S_n = S_0 + \varepsilon_n$
Fusion
$S_0' = S + \varepsilon_0'$

Fig. 1 Block diagram modelling Text Fusion process

III. DATA SET

Due to the unavailability of more than one Arabic OCR system to the authors, only Sakhr's Automatic Reader (version 4.0), an Arabic OCR system, was used to recognize a few scanned pages, which were scanned at 300x300 dpi, from the 14th century religious book (Zad Al-Me'ad). Eight pages, containing 4,236 words with Character Error Rate (CER) of 13.9% and Word Error Rate (WER) of 36.8%, were selected at random from the

OCR'ed text and manually corrected. The degraded and clean versions were used to build an error model that was subsequently used to train a garbler that attempts to introduce errors similar to those of the OCR system. OCR degradation was modeled using a noisy channel model in which the observed characters result from the application of some distortion function on the real characters [6, 8]. The model used here accounts for three character edit operations: insertion, deletion, and substitution. Formally, given a clean word $\#C_1..C_i..C_n\#$ and the resulting word after OCR degradation $\#D_1..D_j..D_m\#$, where D_j resulted from C_i, ε representing the null character, *L* representing the position of the letter in the word (beginning, middle, end, or isolated – Arabic characters change shape depending on their positions in words), and # marking word boundaries, the probability estimates for the three edit operations for the models, are:

$$P_{substitution}\ (C_i \rightarrow D_j) = \frac{count(C_i \rightarrow D_j \mid L_{C_i})}{count(C_i \mid L_{C_i})}$$

$$P_{deletion}\ (C_i \rightarrow \varepsilon) = \frac{count(C_i \rightarrow \varepsilon \mid L_{C_i})}{count(C_i \mid L_{C_i})}$$

$$P_{insertion}\ (\varepsilon \rightarrow D_j) = \frac{count(\varepsilon \rightarrow D_j)}{count(C)}$$

The resulting character-level alignments were used to create a garbler that reads in a clean word $\#C_1..C_i..C_n\#$ and synthesizes OCR degradation to produce $\#D'_1..D'_j..D'_m\#$. For a given character C_i, the garbler chooses a single edit operation to perform by sampling the estimated probability distribution over the possible edit operations. If an insertion operation is chosen, the model picks a character to be inserted prior to C_i by sampling the estimated probability distribution for possible insertions. Insertions before the # (end-of-word) marker are also allowed. If a substitution operation is chosen, the substituted character is selected by sampling the probability distribution of possible substitutions. If a deletion operation is chosen, the selected character is simply deleted.

To obtain different levels of degradation, the character error rate (CER) was tuned with tuning variable *k*.

$$P_{new}\ (C_i \rightarrow D_j) = \begin{cases} k \bullet P_{original}\ (C_i \rightarrow D_j) & , C_i \neq D_j \\ P_{original}\ (C_i \rightarrow D_j) + \\ (1-k) \bullet (1 - P_{original}\ (C_i \rightarrow D_j)), & C_i = D_j \end{cases}$$

Where $P_{original}$ and $\mathrm{CER}_{\mathrm{original}}$ are the calculated edit operation probability and original CER respectively, *k* is the tuning factor, and P_{new} is the new edit operation probability. $C_i = \varepsilon$ and $D_j = \varepsilon$ for insertion and deletion respectively. The new CER CER = $k \bullet \mathrm{CER}_{\mathrm{original}}$.

Beside the printed version, available was another version of the book in a clean (error free) electronic form. The electronic version consists of 2,730 separate documents. Associated with the documents are a set of 25 topics and relevance judgments, which were built by exhaustive judgment of the documents (which will be useful for IR tests). The number of relevant documents per topic ranges between 3 and 72 and averages around 20. The average query length is 5.4 words [7].

Degradation model was applied on the clean electronic version with different values of *k* (*k* = 1, 0.5, 0.66, 1.25, 2). For each value of k, two degraded versions were produced to check the reliability of the degradation model and the randomness of the generated errors. Results of the generated versions are listed in Table 1, which lists the CER, WER, and Out Of Vocabulary (OOV) words (not in the language model training set) for the original OCR text and the synthetically degraded versions. The synthetically degraded version where *k* = 1 has nearly identical CER, WER, and OOV to the original OCR text. For the rest of this paper, garbled versions will be referred to with the model number shown in Table 1 (Model-1 … Model-5). Between any two garbled versions (either using the same model or different models), there are *common word errors* (CWE) where both models misrecognized a given word, which means that the maximum text improvement with fusion will be limited by the CWE. For example, the two versions of Model-1 have a CWE of 17%, which means that the minimum WER after fusion of these two versions would be 17%.

Table 1 Produced versions of text set after applying error model with different CER

Data set	CER	WER	OOV
Original	13.9%	36.8%	20.9%
Model-1	13.9%	36.3%	21.1%
	13.9%	36.4%	21.1%
Model-2	7.0%	20.3%	11.9%
	7.0%	20.4%	11.9%
Model-3	9.3%	26.1%	15.2%
	9.3%	25.9%	15.2%
Model-4	17.4%	43.2%	25.0%
	17.4%	43.3%	24.9%
Model-5	27.9%	59.2%	33.8%
	27.9%	59.2%	33.7%

For all the generated versions, IR tests were performed with mean average precision as the figure of merit to check the effect of garbling on the retrieval effectiveness. Figure 1 shows the mean average precision of the garbled versions compared to the clean version.

The index term used for indexing and searching the collection was 4-grams. According to Darwish, character 4-grams are the best index term for Arabic OCR text [7]. Figure 2 shows that the retrieval effectiveness decreases as the WER increases in the collection set. For all degraded versions, IR effectiveness was observed to be statistically different from the clean version. A paired two-tailed t-test with p-value < 0.05 was used to indicate statistical significance.

IV. EXPERIMENTAL SETUP AND RESULTS

Text fusion is tested for the fusion of two and three different degraded versions with different errors. No tests were performed on fusing more than three different versions, as it is unlikely to obtain more than the three independent sources of the same text set.

A trigram language model was trained on a web-mined collection of religious books belonging to the teacher of the author of the documents at hand, to insure content similarity, using the SRILM toolkit [9].

The retrieval experiments were performed on the clean and fused versions of the text. The collections were indexed and searched using character 4-grams [7]. For all experiments, Indri search engine toolkit [10] was used with default parameters with no blind relevance feedback. Again, the figure of merit for evaluating retrieval results was mean average precision (MAP), with statistical significance testing done using a paired 2-tailed t-test.

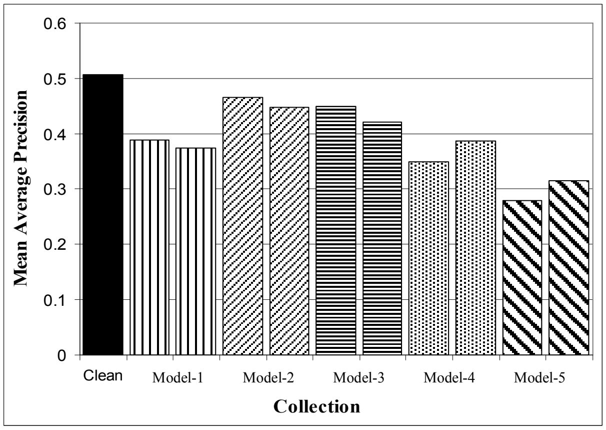

Fig. 2 Results in MAP of searching different versions of the collection

Table 2 shows the fusion results of pairs of fused versions coming from different models. The original WER in each model is mentioned under the model name. The resulting WER after fusion and the CWE rate are listed in the top and bottom parts of the cell respectively. Results show that the language model usually selects the proper word between the two candidate words (when at least one of them is the proper word). In many of resulting fusions, the WER was more than halved. Another observation is that fusion is always useful even when fusing a lightly degraded set and a highly degraded set, but the improvement in text quality decreases as the degradation of the fused text sets increases. Fusion can be used in tandem with automatic error correction using a character error model and language modeling, which typically remove more than 50% of the errors [6], to further eliminate more errors.

Table 3 shows the fusion results when fusing 3 degraded versions, which yields better results compared to fusing 2 degraded versions.

Figure 3 shows the information retrieval results on all the previously mentioned fused versions, where M$_{ij}$ returns to fusion output from Model-i and Model-j respectively. For all output sets, the MAP of the fused set is better than each individual set, but not all the output sets from fusion were statistically significant better than the individual sets. Considering that character based error correction has been reported to have little effect on retrieval effectiveness [8], achieving retrieval effectiveness that is statistically indistinguishable from the retrieval effectiveness when searching the clean text version in a few cases is very promising. Further, combining fusion with character level and language model based correction may have significant impact on retrieval effectiveness. Perhaps, OOV words, which contribute more than half the remaining errors after fusion, can be the primary targets of such correction.

Table 2 Results of fusion of each two different text set versions. Each cell is formed of upper and lower values, upper value is the WER after fusion of the two models opposite to the cell, lower value represents the common error between these two versions of text and which is the limit if WER

Model-1	17.4%				
36.3%	17.0%				
Model-2	10.0%	6.0%			
20.3%	9.6%	5.7%			
Model-3	12.7%	7.4%	9.3%		
26.1%	12.2%	7.0%	8.9%		
Model-4	20.4%	11.7%	14.9%	23.9%	
43.2%	19.9%	11.2%	14.4%	23.4%	
Model-5	26.8%	15.4%	19.5%	31.6%	42.2%
59.2%	26.3%	14.9%	19.0%	31.0%	41.6%
	Model-1	Model-2	Model-3	Model-4	Model-5
	36.4%	20.4%	25.9%	43.3%	59.2%

Table 3 Results of fusion of some triples of models, Model-1-1-2 returns to fusion of two different versions of model-1 and one version of model-2

	Common Error	WER
Model-1-1-2	5.3%	5.9%
Model-1-3-4	7.8%	8.4%
Model-1-2-5	7.8%	8.5%
Model-3-4-5	11.7%	12.3%

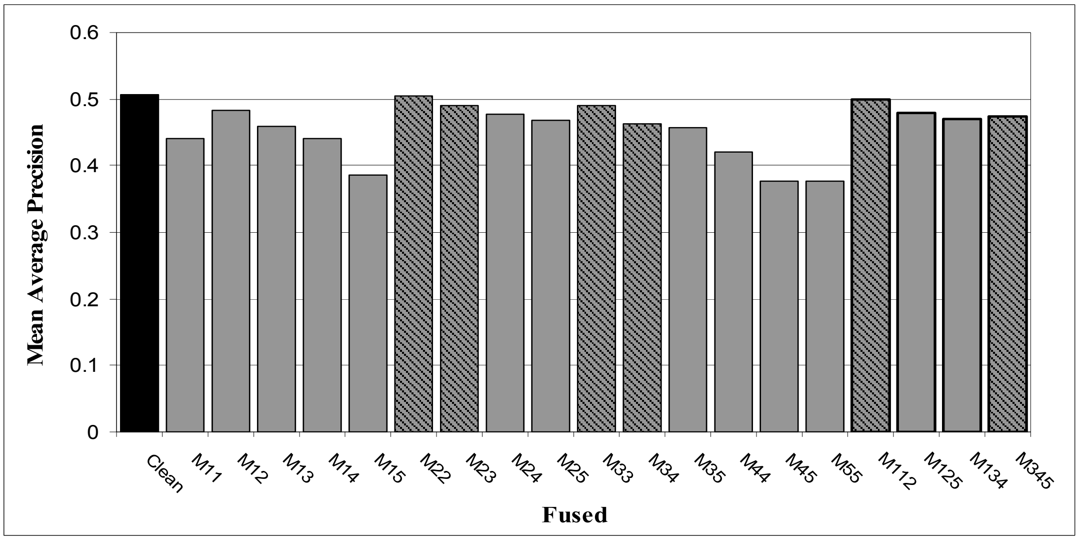

Fig. 3 Results in MAP of searching different fused models, hashed bars referes to statistical significant retrieval results better than the original degraded versions

V. CONCLUSION AND FUTURE WORK

The paper examined the effect of text fusion on the reduction of WER and the improvement in information retrieval effectiveness. The results clearly show that the proposed technique selects the proper word from among different candidates. The results suggest that text fusion is never harmful, and its effect on WER reduction and IR effectiveness improvement depends upon the quality of the fused versions of text and the CWE between them. Results also show that using more than two different text versions in the fusion process further reduces the WER of the resultant version.

One of the shortcomings of the paper is that the fusion technique was tested on synthetically degraded text, and testing on real degraded data from either OCR, speech recognition, or other sources may be warranted. Nonetheless, the results are promising. Another problem that would surface when using real data is the problem of aligning the words in the texts from different sources. The problem becomes more acute when degradation levels increase. Both problems need to be addressed in future work. Lastly, applying fusion in conjunction with character level and language modeling based correction can further eliminate much of the errors in the text.

REFERENCES

[1] Oard, D. W. and F. Ertunc. Translation-Based Indexing for Cross-Language Retrieval. In ECIR 2002 (2002).

[2] Darwish, K. and D. Oard. Probabilistic Structured Query Methods. In SIGIR-2003 (2003).

[3] Taghva, K., J. Borsack, and A. Condit. An Expert System for Automatically Correcting OCR Output. In SPIE - Document Recognition (1994).

[4] Tseng, Y. and D. Oard. Document Image Retrieval Techniques for Chinese. In Symposium on Document Image Understanding Technology, Columbia, MD (2001).

[5] Lu, Z., I. Bazzi, A. Kornai, J. Makhoul, P. Natarajan, and R. Schwartz. A Robust, Language-Independent OCR System. In the 27th AIPR Workshop: Advances in Computer Assisted Recognition, SPIE (1999).

[6] Magdy, W. and K. Darwish. Arabic OCR Error Correction Using Character Segment Correction, Language Modeling, and Shallow Morphology. In EMNLP, pp. 408 – 414, (2006).

[7] Darwish, K. and D. Oard. Term Selection for Searching Printed Arabic. In SIGIR-2002 (2002).

[8] Magdy, W. and K. Darwish. Arabic Word-Based Correction for Retrieval of Arabic OCR Degraded Documents. In SPIRE (2006)

[9] Stolcke, A. SRILM - An Extensible Language Modeling Toolkit. Proceedings of the Workshop on Statistical Machine Translation, pp. 72—77 (2002)

[10] Abdul-Jaleel, N., J. Allan, B. Croft, F. Diaz, L. Larkey, X. Li, D. Metzler, T. Strohman, H. Turtle, and C. Wade. UMass at TREC 2004: Notebook. In TREC 2004, pp. 657, (2004).

[11] Darwish, K. and D. Oard. Improving OCR-Degraded Text Retrieval Using Query Translation. In Symposium on Document Image Understanding Technology, pp. 181-187, (2003).