Enhancing Image-based Arabic Document Translation Using a Noisy Channel Correction Model

Yi Chang, Ying Zhang, Stephan Vogel, Jie Yang

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
USA
{changyi, joy+, vogel+, yang+}@cs.cmu.edu

Abstract

An image-based document translation system consists of several components, among which OCR (Optical Character Recognition) plays an important role. However, existing OCR software is not robust against environmental variations. Furthermore, OCR errors are often propagated into the translation component and cause, causing poor end-to-end performance. In this paper, we propose an image-based document translation using an error correction model to correct misrecognized words from OCR output. We train our correction model from synthetic data with different fonts and sizes to simulate real world situations. We further enhance our correction model with bigrams to improve our word segmentation error correction. Experimental results show substantial improvements in both word recognition accuracy and translation quality. For instance, in an experiment using Arabic Transparent Font, the BLEU score increases from 18.70 to 33.47 with the use of our noisy channel model.

Introduction

In an information society, we communicate with people and information systems through diverse media in increasingly varied environments. Advanced technologies have bridged many gaps for such communication. While Internet technology provides a shortcut overcoming distance barriers, machine translation (MT) technology helps us to overcome barriers to communicate with people who use different languages. Much information is in written form embedded in various environments, such as on a paper, on a wall, on a bulletin board, etc. As digital cameras become popular, an image-based MT system is able to capture information in a variety of environments and translate it from a source language into another target language for different applications. For example, we have developed a sign translation system to translate Chinese into English for tourist applications (Yang et al., 2001; Yang et al., 2002). In this research, we are developing an image-based system that translates Arabic document images into English. The system works as follows. After an image is captured from a digital camera, the system preprocesses the image to account for fonts, skew, rotation, illumination, shadows, glare, reflection, and other sources of variability. Subsequently, it automatically detects text regions in the image, performs recognition using off-the-shelf OCR (Optical Character Recognition) software on the text regions, and then translates the text strings into English using our state-of-the-art statistical MT system (Zhang and Vogel, 2007).

An image-based MT system consists of many components. Its performance relies on not only the machine translation (text-to-text) technology, but also other component technologies. A good image-based document translation system requires robust technologies for text detection, OCR, and language translation. Traditional pipeline approaches yield error propagation, and errors in any component of an image based document translation system can affect the end-to-end performance of the entire system. The errors can be propagated through the system and even amplified during the propagation process, e.g., a misrecognized word or prophase can result a translation error or even multiple translation errors. Due to the peculiarities of languages, an effective solution to error reduction in language translation cannot be expected to be a purely sequential connection of OCR and MT components. Like ASR (Automatic Speech Recognition) in a speech-to-speech translation system, OCR plays an important role in a document image translation system. However, when the quality of an image to be recognized does not match the system's training conditions, OCR software performs poorly, which directly affects translation performance. For example, in one of our experiments, the BLEU score of image text translation using gold standard character recognition is 43.12, while a 10.7% word recognition error rate severely drops the BLEU score to 28.56.

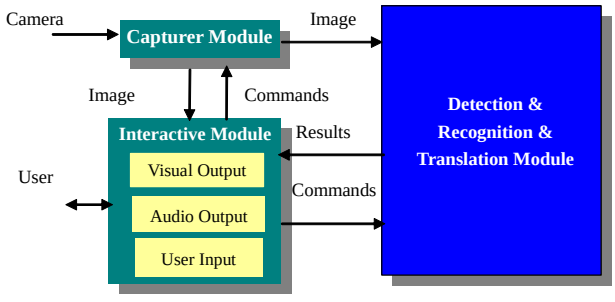

Figure 1: The architecture of the image-based document translation system

In this paper, we propose an enhancement of image-based document translation using an error correction model to correct misrecognized words. We deploy a noisy channel model that is trained from synthetic data with different fonts and sizes to simulate real world situations. We further enhance our correction model with bigrams to improve our word segmentation error correction. We perform experiments to demonstrate the proposed methods. The experimental results show a significant improvement in translation quality. For instance, in the experiment using Arabic Transparent Font, the BLEU

score increases from 18.70 to 33.47 with the use of our noisy channel model.

The remainder of this paper is organized as follows: in section 2, we introduce the architecture of our image-based document translation system. Then, in section 3 we describe the existing problems that will impair the accuracy of our image-based document translation system, and we explain our correction model in Section 4. Finally, we present experimental results in Section 5, which is followed by the related work and conclusions in the last two sections.

System Architecture

As shown in Figure 1, the image-based Arabic document translation system consists of three modules: an image capture module; a detection, recognition, and translation module; and an interactive module. The capture module handles image input, and it is hardware dependent. The input image is then fed into the detection, recognition, and translation module for processing. This module is a key part of the system. It first performs text detection and locates text regions in an image, then further processes these regions and feeds them into the OCR engine, which recognizes the contents of the areas in the source language. Then, the recognition results are sent to the translation system to obtain an interpretation in the target language. The interactive module provides an interface between a user and the system. As a user-centered system, a user-friendly interface is also important. It provides necessary information to the user through an appropriate modality. It also allows the user to interact with the system if needed.

Figure 2 illustrates a prototype of the image-based Arabic document translation system. The system can process input images from a file or a digital camera. The upper left window in the interface shows the input image and the bottom left window is the detection results. The recognized text is at the right upper window and the translation results are shown in the right bottom window.

Figure2: An illustrate of the prototype system

It is a challenging problem to automatically detect text from an image. To work around variations in an image, we have developed a robust text detection module that uses a hierarchical detection framework that embeds multi-resolution and multi-scale edge detection, adaptive searching, color analysis, and affine rectification algorithms (Chen et al., 2004). We combine multi-resolution and multi-scale edge detection techniques to effectively detect text in different sizes. We employ a Gaussian mixture model (GMM) to represent background and foreground, and perform color segmentation in selected color spaces. We use affine rectification to recover deformation of the text regions caused by an inappropriate camera view angle. After affine rectification for each text region in the image, we perform text detection again in rectified regions within the image to refine detection results.

In our system implementation, the recognition module incorporates an off-the-shelf OCR software package, Sakhr Automatic Reader version 8.0 (Platinum Edition), which is one of the most commonly used Arabic OCR products.

The PanDoRA system (Zhang and Vogel, 2007) is used for the translation module. The image-based document translation system is expected to be used on hand-held devices such as digital camera and PDAs. PanDoRA is a phrase-based statistical machine translation system with compact data structure which makes it possible to be run on hand-held devices. There are two decoding mode in PanDoRA: monotonic decoding and ITG-style reordering decoding. For our experiments, we use monotonic decoding as the testing sentences are usually very short.

Recognition Problems in Document Translation from Images

OCR is one of the most successful applications in the pattern recognition field. It is a common belief that OCR is a solved problem because so many papers and patents have claimed recognition rates as high as 99.99%. Although many commercial OCR systems work well on high quality scanned documents under controlled conditions, they fail in many tasks, such as video OCR, license plate OCR, and sign OCR. Current video OCR is limited to recognizing captions in video images for video indexing, or to identify license plates on vehicles for various applications. Even 99% accuracy would generate about 30 errors on a typical printed page of 3000 characters. Rice et al. classified errors produced by OCR systems and their causes (Rice et al., 1999). OCR errors have been organized into four major classes: Imaging Defects, Similar Symbols, Punctuation and Typography.

In our image-based Arabic document translation system, we have found four problems related to recognition accuracy that could severely impair the end-to-end system performance.

Different Fonts in Arabic

Arabic is one of the languages with the most complex print form. Arabic is written from right to left. Most of the word characters in a word are connected, and each character has 4 different forms: isolated form, beginning form, middle form and end form. All these variations are multiplied with different fonts. We choose the three most commonly used Arabic fonts in the Windows OS: Arabic Transparent, Simplified Arabic and Traditional Arabic. Figure 3, 4 and 5 illustrate examples of these fonts with the same sentence. Note that characters of Traditional Arabic Font are largely different from the characters of the other fonts. In addition, the word segmentation is quite vague in these fonts. Without knowledge of Arabic, it is quite difficult for us to precisely segment the word boundary from a single image. Correspondingly, the segmentation vagueness is also an obstacle for the OCR software.

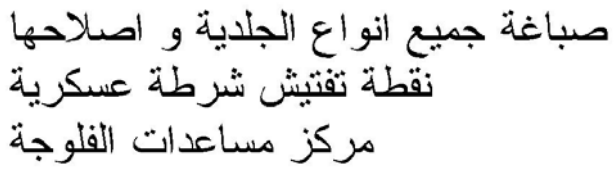

Figure 3: An example of Arabic Transparent Font

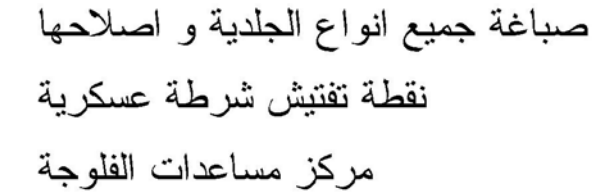

Figure 4: An example of Simple Arabic Font

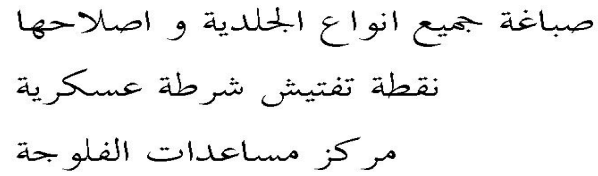

Figure 5: An example of Traditional Arabic Font

To the best of our knowledge, most commercial OCR products only work well within certain fonts. If text in a font which is far different from those fixed fonts needs to be recognized, the accuracy of recognition will severely drop. Nevertheless, unsupported fonts could be possible and common in our image-based document translation scenario due to the diversity of fonts in the real world.

The Limitation of OCR

Although most of commercial OCR companies claim that their products have almost perfect recognition accuracy near 99.9%, their accuracy rates are hardly to be achieved. According to (Kanungo et al., 1999), the absolute page accuracy rate of Sakhr Automatic Reader version 3.0 is 90.33% with reasonably high quality images, where the page accuracy rate is based on character level.

To verify the performance of Sakhr OCR, we also performed a set of experiments. We manually generated a set of high resolution images (600 dpi) with different fonts and different sizes. Without any noise on the images, we believe we can achieve the upper bound of Sakhr OCR in the ideal case. The evaluation results are summarized in Table 1 and Table 2.

Table 1 illustrates the character error rate with different character fonts and sizes, and the character errors are computed by edit distance. According to the table, we can observe that the accuracy of OCR is very sensitive to both font and size changes, and the recognition accuracy only could reach 99% with some specific parameters. For example, OCR performs the best in the Tradition Arabic Font on images with 108 pixels, and performs much worse with 36 pixels. But it performs differently for the other two fonts. It is difficult to find one single set of parameters to get the most satisfactory recognition accuracy with all fonts. Although the data set here is too small to draw a conclusion, it helped us to find suitable parameter settings for our following experiments. Table 2 shows the corresponding word error rate with the same parameters. We found that the word recognition error rate was about 4 times larger than the character recognition error rate.

	36 pixels	60 pixels	84 pixels	108 pixels
Arabic Transparent	11.6%	4.0%	3.1%	4.8%
Simplified Arabic	17.2%	4.5%	5.6%	5.1%
Traditional Arabic	28.2%	3.1%	0.3%	0.3%

Table 1: Character Recognition Error Rate

	36 pixels	60 pixels	84 pixels	108 pixels
Arabic Transparent	46.7%	11.3%	12.9%	22.6%
Simplified Arabic	53.2%	17.7%	19.4%	21.0%
Traditional Arabic	64.5%	9.7%	1.6%	1.6%

Table 2: Word Recognition Error Rate

The Limited Image Quality

Sakhr OCR software claims to be able to achieve the best performance with images above 300 dpi resolution. However, in our application, due to the size of the images captured by the camera, the small text regions in the image hardly satisfy the resolution requirement of 300 dpi. Image preprocessing also leads to further information loss. We deploy a set of algorithms, such as bilinear interpolation, image binarization and affination transformation, etc. to handle skew, rotation, illumination, shadows, glare, and reflection of images. The accumulated information loss in image preprocessing degrades the quality of the image that is input to the recognition module below the requirement of the OCR software. As a result, the accuracy of OCR might be worse than expected with the limited quality images in our image-based document translation system.

The OCR Error Propagated in Statistical Machine Translation

In this paper, we use the BLEU score to measure translation quality. BLEU averages the precision for unigrams, bigrams and up to 4-grams and applies a length penalty if the generated sentence is shorter than the best matching (in length) reference translation (Papineni et al., 2001). Due to the definition of the BLEU score, a recognition error not only punishes the unigram score, but also hurts the bigram, trigram and 4-gram scores, which have larger contributions to the final BLEU score. As a result, OCR errors are propagated in the current machine translation metrics. Our experiment shows that a 10.7%

error rate in word recognition severely drops the BLEU score from 43.12 to 28.56.

Although OCR errors might not be a problem for many other applications such as information retrieval (IR), it causes serious troubles for a MT task. Several studies (Taghva et al., 1994; Croft et al., 1993; Mittendorf et al., 1995) indicate that there is no statistical difference in retrieving original text and English OCR text, if the OCR performance is good enough. Applying probabilistic IR, we can retrieve the most relevant OCR-generated documents using approximate matching techniques even without correcting OCR errors. However, due to the four problems above, we can not ignore OCR errors in our document image translation scenario, and an effective method for OCR error correction is essential to achieve high quality MT for an image-based MT task.

OCR Correction with Noisy Channel Model

Noisy Channel Models in the Image-based Document Translation System

Noisy channel models are widely used in AI problems such as speech recognition and machine translation. It assumes that the source input sequence *I* is encoded by a noisy channel and we observe the output sequence *O*. The task is to estimate the source message by a decoder $\hat{I} = \arg\max P(I \mid O)$. In our application, there are two noisy channels: the translation channel and the image generation channel (Figure 6). The translation model encodes the source language (English) to the target language (Arabic), and the image generation channel encodes the Arabic texts into images. The decoding process is shown in Figure 7.

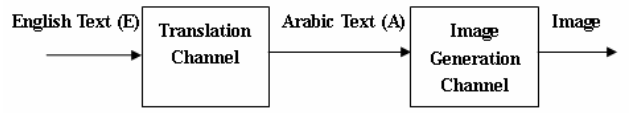

Figure 6: The two-noisy-channel-model system

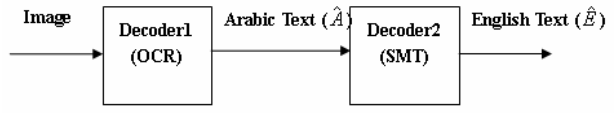

Figure 7: The decoding process of the noisy channel model in Figure 6

Due to the limitations of the OCR system, the output from Decoder 1 is not perfect and OCR errors will be propagated in Decoder 2. One possible solution is to consider all possible errors from OCR and update the statistical machine translation model by adding the OCR errors in the training dataset. However, this method requires converting all Arabic training data into images and then generating the recognized erroneous training data from the OCR engine, which is time consuming or even impossible when the training data is too large. Another possible solution is to take the Divide and Conquer strategy, adding an error correction module while keeping the existing SMT systems that were well trained and tuned.

In this paper, we propose an alternative approach by adding another noisy channel model for correction of the OCR output, as Figure 8 below. The OCR error is modeled in the text transform channel. And the decoding process is shown in Figure 9. With the new hierarchy we can assume that the decoded message of Decoder1 (OCR) is perfect ($\hat{A}' = A'$) and also keep the SMT model independent. What we are focusing on is the text correction process.

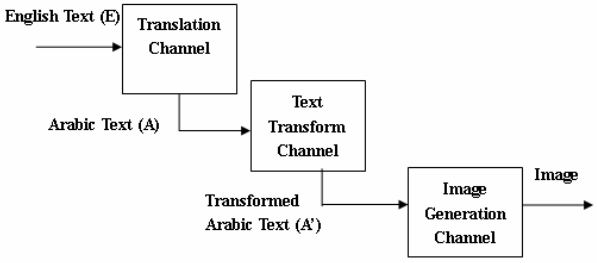

Figure 8: The three-noisy-channel-model system

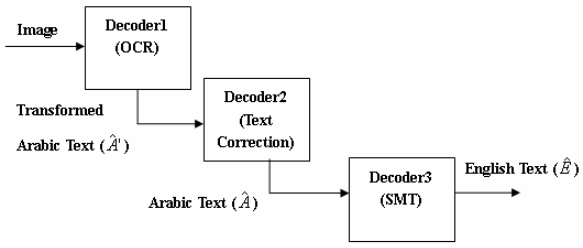

Figure 9: The decoding process of the noisy channel model in Figure 8

Text Correction

To decode the correct text sequence from the OCR output, the Bayesian rule is applied:

$$\begin{aligned}\hat{A} &= \arg\max_A P(A \mid \hat{A}') = \arg\max_A P(A \mid A') \\ &= \arg\max_A \frac{P(A) * P(A' \mid A)}{P(A')} \\ &= \arg\max_A P(A) * P(A' \mid A),\end{aligned} \quad (1)$$

Where *P(A)* is the Arabic language model, which can be trained from a large Arabic dataset. *P(A'|A)* is the transformation model. To simplify the problem we assume that words are independent of each other and the transformation model can be written as:

$$P(A' \mid A) = \prod_k P(a_k' \mid A) \quad , \quad (2)$$

where a_k' is the *k*th word in *A'*.

We future assume that the number of words in *A'* and *A* is the same and a_k' only depends on the *k*th word in *A* (a_k):

$$P(A' \mid A) \approx \prod_k P(a_k' \mid a_k) \quad . \quad (3)$$

To evaluate $P(a_k' \mid a_k)$, we synthesize different images from a given word a_k and use OCR to get the transformed texts. $P(a_k' \mid a_k)$ can be computed by maximum likelihood estimation:

$$P(a_k' \mid a_k) = \frac{count(a_k', a_k)}{count(a_k)} \quad . \qquad (4)$$

A small probability is applied for unseen word pairs. We use the well known Viterbi algorithm to decode the most probable correct word sequence given the output of OCR.

Experiments

Data

We evaluate our OCR correction module based on an Arabic-English corpus called the Basic Travel Expression Corpus (BTEC) (Eck & Hori, 2005). The training data contains 20,000 sentence pairs and 131,711 Arabic words. The total vocabulary size is 26,116. To generate the training data, we synthesize the training data vocabulary into multiple images based on different fonts. The development data contain 506 Arabic sentences and 2662 words, and the testing data contain 500 Arabic sentences and 2566 words. The height of text in each image is 108 pixels, and the resolution of the image is 96 dpi.

Experimental Results

Table 3 illustrates a comparison of word recognition error rate with different fonts. The first row of the table is measured with the output of OCR directly, and the second row is based on the result of our noisy channel correction model. The recognition of Simplified Arabic Font, which is the most accurate one, can make an improvement of 4.02% by correction, while the recognition of Arabic Transparent Font could improve 7.05% from our correction module.

	Arabic Transparent	Simplified Arabic	Traditional Arabic
OCR	17.81%	11.42%	22.21%
Correction	10.76%	7.40%	19.49%
Enhanced Correction	5.96%	5.96%	19.25%
Perfect Segmentation	1.71%	1.60%	1.29%

Table 3: A comparison of word recognition error rate

	Arabic Transparent	Simplified Arabic	Traditional Arabic
OCR	18.70	25.90	18.73
Correction	25.19	31.13	21.61
Enhanced Correction	33.47	34.01	21.80
Perfect Segmentation	42.10	42.25	42.65

Table 4: A comparison of BLEU scores

In the following step, we translate all OCR and correction results into English using the same statistical machine translation module. Translating the testing data directly, the BLEU score is 43.12, and we can regard it as the translation upper bound. Table 4 demonstrates the corresponding machine translation BLEU scores.

According to Table 3 and Table 4, a 4.02% improvement in word recognition with Simplified Arabic Font brings a 6.23 BLEU score improvement. The highest BLEU score after correction can reach to 31.13. Interestingly, the BLEU score is not consistent with the word recognition error. A cell with larger word recognition error rate might also have larger BLEU score in these tables. That is, we have to consider both BLEU score and word recognition error as the evaluation metric for our correction model in our image-based document translation scenario.

Error Analysis

Although our correction model makes reasonable improvement in the BLEU score, it is still much lower than the translation upper bound. After analyzing the errors, we find there are mainly two types of errors from the OCR output, character recognition errors and word segmentation errors.

Character recognition errors come from when a character in an image is misrecognized into one or more characters, or even skipped. Our noisy channel model aims to solve this type of error. Word segmentation errors occur when the OCR misrecognizes one single word in an image as multiple words or misrecognizes multiple adjacent words in an image as one single word.

أود أن أرى
أوفى أن ألى لى
أود أن ألى لى

Figure 10: An example of recognition & correction.

ما رأيك في تناول شراب
ما رأيك في تناول شر اب
ما رأيك في تناول شعر اب

Figure 11: Recognition & correction example II.

Figure 10 illustrates an example of an OCR error. The first line in the figure is the ground truth text, which means “I'd like to see it”. The second line in the figure is the result of the recognition, where the leftmost word on the first line was incorrectly segmented into two words, and the rightmost word was also misrecognized. The translation result is “<unk> I <unk> I”, where “<unk>” is the translation of an out-of-vocabulary word. The third line in the figure is the result of text correction. As we can notice, the rightmost word is successfully corrected with our correction model, and the translation result is “I'd like <unk> I”.

Our noisy channel model can not solve the segmentation error in the above case, and sometimes it even makes such errors worse. For example, as Figure 11 shows, the first line means “how about a drink”. The OCR segments the leftmost word on the first line into 2 separate strings, and there is no character recognition error on them. Since these 2 separate strings are out of our vocabulary, the translation is “how about <unk> <unk>”. Our correction model converts the 2nd leftmost string on the 2nd line into another word which means “hair”, and the translation

turns into “how about hair <unk>”, which is misleading in our end-to-end system.
From analyzing the OCR results in our system, we that word segmentation errors dominate in our application. If we explicitly fix all word segmentation errors in OCR result, the word error rate drops to less than 2%, as the 4th row in Table 3 shows. Correspondingly, all BLEU scores rise to above 42, which is close to the upper bound of translation, as represented in the 4th row of Table 3.

Enhance Correction with Bigrams

To overcome segmentation errors, we enhance our noisy channel model for segmentation correction. Given a string of the OCR output, we explore all of its bigrams, if the edit distance between a bigram and a word in our dictionary is less than a threshold, we replace the bigram with the new word. Iterating this replacement processing yields a new string. In the decoding process, we use the Viterbi algorithm to decode the most possible word sequence given both the output of OCR and its corresponding replacement strings. Figure 12 explains the algorithm in detail. In our experiment, we set n as 2.

BigramCorrection(OcrOutput $\hat{A}'$){

A). For all segmentions in $\hat{A}'$, construct a set of bigram strings using the adjacent segmentation.

B). Iterate the replacement process, and generate a new string $\hat{A}''$: If the edit distance between a bigram and a dictionary word is less than n, replace the bigram with the word.

C). Decode both $\hat{A}'$ and $\hat{A}''$, and choose the most probable sequence as the correction of $\hat{A}'$.

}

Figure 12: The Bigram Correction Algorithm

The 3rd rows in Table 3 and Table 4 present the result of our enhanced correction model. In Arabic Transparent Font, we improve the BLEU score from 18.70 to 33.47 with our enhanced correction model, while the most accurate translation result reaches 34.01. In our experiment, we also tried trigram combination in our enhanced noisy channel model, but there was no further improvement on it.
The enhanced correction model is designed to improve the error of segmenting one single word in an image into multiple words. In cases of merging multiple adjacent words in an image into one single word, the capability of our enhanced correction model is limited. As we can see, the improvement in Traditional Arabic Font with our enhanced correction model is marginal.

The Related Work

The research presented in this paper is related to some previous works reported in the literature. Hong corrected the OCR result through passage-level post-processing using visual constraints and linguistic constraints (Hong, 1994). Kolak and Resnik modeled a noisy channel in OCR Error Correction with syntactic information (Kolak & Resnik, 2002). Kanungo et al. compared the recognition accuracy of the two most commonly used Arabic OCR products, Sakhr and OmniPage (Kanungo et al., 1999). Taghva et al. built an expert system for automatically correcting OCR errors to post-process the OCR result text in preparation for a subsequent retrieval system (Taghva et al., 1994). The system only focuses on words that will likely be used for retrieval, and claimed 87% of the errors were corrected. Doermann and Yao presented a system for modeling the OCR output errors, and they described some symbol and page models to simulate the degraded images during scanning, decomposing and recognition (Doermann & Yao, 1995). Sato et al. implement some video OCR techniques to solve the low resolution characters and extremely complex background problems in digital video data. They post-process the OCR result by mapping the OCR result into a dictionary by a self-defined word similarity (Sato et al., 1999).
In this paper, we focus on correcting OCR errors to improve translation quality for an image-based document translation task. Similarly, the OCR errors might impair the accuracy of other applications that rely on text processing techniques. Croft et al. examined the information retrieval performance on OCR output, and showed that low quality OCR output can result in significant degradation on the accuracy of retrieval (Croft et al., 1993). Instead of correcting OCR errors, Harding et al. used *n*-gram formulations with a probabilistic retrieval system and improved retrieval performance over standard queries on the same data when a level of 10 percent degradation or worse was achieved (Harding et al., 1997). Similarly, Mittendorf et al. showed that recognition errors can be ignored in retrieval if the number of documents and their lengths are sufficiently large (Mittendorf et al., 1995).

Conclusions and Future Work

In this paper, we have proposed an approach to correct Arabic OCR errors in an image-based Arabic document translation system. The correction model is trained with synthetic images with different fonts and sizes. We have further enhanced our correction model with bigrams to improve the word segmentation correction. We achieved substantial improvement in both word correction and the translation accuracy.
However, the correction models we proposed are limited with the cases of recognizing multiple adjacent words on image into one single word. Furthermore, those more complicated conditions with the mixture of character recognition errors and word segmentation errors are even challenging for us. In order to address these challenges, we will work on modeling the context information in the training data in our future work.

Acknowledgements

This work was partially supported by DARPA through the ASSIST program. We would like to thank Justin Betteridge for his valuable help on the final version of the paper, and we are also grateful to comments from the anonymous reviewers.

References

Chen, X., Yang, J., Zhang, J., and Waibel, A. (2004). Automatic detection and recognition of signs from

natural scenes, IEEE Transactions on Image Processing, 13(1), 87-99.

Croft, W.B., Harding, S., Taghva, K. and Borsack, J. (1993). An Evaluation of Information Retrieval Accuracy with Simulated OCR Output. Technical Report: UM-CS-1993-076.

Doermann, D., and Yao, S. (1995). Generating synthetic data for text analysis systems. Symposium on Document Analysis and Information Retrieval, (pp. 449-467).

Eck, M. and Hori, C. (2005). Overview of the IWSLT 2005 evaluation campaign. In Proceedings of International Workshop on Spoken Language Translation, (pp. 11-17), Lisbon, Portugal.

Eikvil, L. (1993). OCR Optical Character Recognition. Technical Report, Report No. 876.

Harding, S.M., Croft, W.B. and Weir, C. (1997). Probabilistic Retrieval of OCR Degraded Text Using N-Grams. Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries (pp. 345-359), Pisa, Italy.

Hong, T. (1995). Degraded Text Recognition Using Visual and Linguistic Context. Ph.D. thesis, Computer Science Department, SUNY Buffalo.

Huang, X., Acero, A. and Hon, H. (2001). Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice-Hall.

Kanungo, T., Marton, G. and Bulbul, O. (1999). OmniPage vs. Sakhr: Paired model evaluation of two Arabic OCR products. Proceedings of SPIE Conf. on Document Recognition, (pp. 109-120), San Jose, CA.

Kolak, O. and Resnik, P. (2002). OCR error correction using a noisy channel model. In Human Language Technology Conference (HLT 2002), San Diego, CA.

Mittendorf, E., Schäuble, P. and Sheridan, P. (1995). Applying Probabilistic Term Weighting to OCR Text in the Case of a Large Alphabetic Library Catalogue. Proceedings of the 18th ACM SIGIR Conference on Research and Development in Information Retrieval, (pp. 328-335) Seattle, WA.

Myka, A. and Güntzer, U. (1997). Measuring the Effects of OCR Errors on Similarity Linking, Proceeding of ICDAR, (pp. 968-973), Ulm, Germany.

Papineni, K. and Roukos, S. Ward, T. and Zhu, W. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. Proceedings of the 20th Annual Meeting of the Association for Computational Linguistics, (pp. 311-318), Philadelphia, Pennsylvania.

Rice, S; Nagy, G.; Nartker, T. (1999). Optical Character Recognition: An Illustrated Guide to the Frontier. Kluwer, Boston,

Sato, T., Kanade, T., Hughes, E.K., Smith, M.A. and Satoh, S. (1999). Video OCR: Indexing Digital News Libraries by Recognition of Superimposed Captions. ACM Multimedia Systems, vol.7, 385-394.

Taghva, K., Borsack, J. and Condit, A. (1994). An Expert System for Automatically Correcting OCR Output. Proceedings of the SPIE-Document Recognition, (pp. 270-278), San Jose, CA.

Vogel, S., Zhang, Y., Huang, F., Tribble, A., Venogupal, A., Zhao, B. and Waibel, A. (2003). The CMU Statistical Translation System. Proceedings of MT Summit IX, New Orleans.

Yang, J., Gao, J., Zhang, Y. and Waibel, A. (2001). Toward Automatic Sign Translation, Proceedings of the Human Language Technology Conference (HLT-2001).

Yang, J., Chen, X., Zhang, J., Zhang Y. and Waibel, A. (2002). Automatic detection and translation of text from natural scenes, Proceedings of ICASSP 2002, (pp. 2101 – 2104), Orlando, Florida.

Zhang, Y., Vogel, S. (2007). PanDoRA: A Large-scale Two-way Statistical Machine Translation System for Hand-held Devices. In Proceedings of MT Summit XI, Copenhagen, Denmark.