Portable Language-Independent Adaptive Translation from OCR

Quarterly R&D Status Report No. 6

April 15, 2009

Contractor:	**BBN Technologies** 10 Moulton Street, Cambridge, MA 02138
Principal Investigator:	Prem Natarajan Tel: 617-873-5472 Fax: 617-873-2473 Email: pnataraj@bbn.com
Reporting Period:	1 January 2009 – 31 March 2009

This material is based upon work supported by the Defense Advanced Research Projects Agency
DARPA/IPTO
Portable Language-Independent Adaptive Translation from OCR
MADCAT Program
ARPA Order No: X 103
Program Code: 7M30
Issues by DARPA/CMO under Contract #HR001-08-C-0004

Executive Summary

This is the sixth R&D quarterly progress report (QPR) of the BBN-led team under DARPA's MADCAT program. This report is organized by technical task area.

1.1. Pre-Processing and Page Segmentation [BBN, Argon, Lehigh, Polar Rain, UMD, SUNY]

Text Segmentation and Verification [Polar Rain]: This quarter, we developed an algorithm to verify and detect text characters in different sizes (e.g. large versus small font/text sizes). For this technique, we first compute a tiered multi-scale image pyramid consisting: the first tier is composed of the original image; the second tier is computed by down-sampling the image by half along *x* and *y* axes, and the third tier of the pyramid is computed by down-sampling the original image by a factor of four along each the *x* and *y* axes. For text segmentation, the Shape-DNA models are applied on a multi-scale image pyramid and segmentation maps from each tier are integrated into a single map. Similarly, text verification is repeated on each level of the pyramid and verification results are integrated for determining the final segmentation.

Shape-DNA based Handwritten Text Line Detection [Polar Rain]: We designed a text line detection algorithm on handwritten text images in the MADCAT corpus using Shape-DNA models. First, we apply our ruled-line detection and cleaning technique to the image. Next, we project the binary input image onto the Shape-DNA text database and apply morphological operators to obtain a segmentation map. Finally, using the segmentation map and the binary input image, we group the text characters into their corresponding text lines. Initial results showed that Shape-DNA based text line detection for handwritten documents holds promise for degraded documents with image noise and ruled lines.

Text line detection and separation [SUNY]: In this quarter, we implemented a new approach for text detection to deal with documents with varying degree of skew and inadequate line spacing that result in touching/overlapping lines. Our approach is based on "steerable directional" filter, which finds the local orientation of a text line by scanning in multiple directions. The maximum response from a convolution of the filter with the image is the estimated direction of the line of text. Specifically, our approach has the following key steps: First, a stroke segment that crosses a text line is automatically detected. Next, a reference line for splitting touching lines is estimated based on center of gravities of the contours from the detected lines. Finally, we split touching components at the contour level and reconstruct the character images. Figure 1 shows a crowded page and the results from the text line finding. All lines are accurately segmented. The highlighted areas are bounding boxes of touching characters crossing different text lines. The split characters are reconstructed and accurately grouped with the text lines that they belong to.

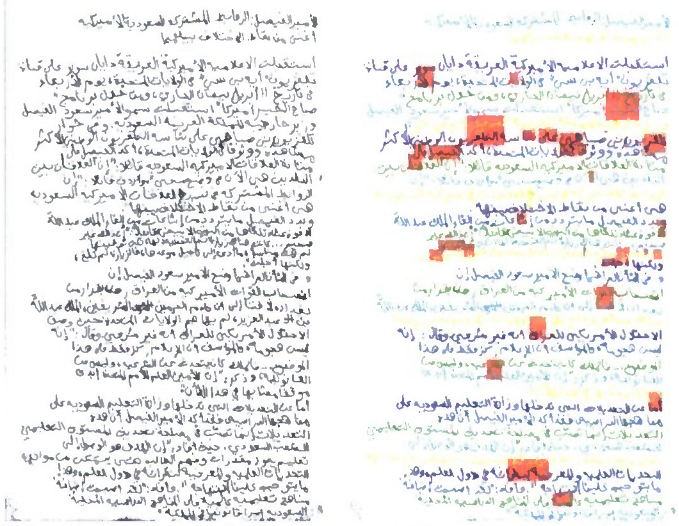

Figure 1: Example of text-line separation using Steerable filter

1.2. Text Recognition [BBN, Argon, Columbia, SUNY]

Error Analysis [BBN]: This quarter, we continued our analysis of causes of errors for text recognition. Our error analysis methodology is based on a novel approach designed for assigning "blame" to error types for machine translation. Based on a subjective analysis of images, we defined seven categories of causes of errors. Next, we randomly chose 300 images from the test set and had three bilingual annotators classify each image into one or more of the seven categories. The final error categories assigned to each page were based on a majority vote of each of the three sets of annotations. We then estimated the best set of weights for each error category such that the weighted sum of error categories in a page approximates the overall word error rate (WER) for that page. The results of our analysis are shown in Table 1. As can be inferred from the numbers in the table, poor legibility and overlapping words/lines contribute the most to overall WER while the presence of slant and non-Arabic characters has the least impact.

Cause of Error	Fraction of Total Error (%)
Poor Legibility	39.0
Overlapping Words/Lines	24.8
Ruled Page Lines	19.5
Skew	10.5
Short Lines	4.5
Non-Arabic Characters	1.1
Slant	0.6

Table 1: Results of Error Analysis.

Training with Phase 2 Data [BBN]: In this quarter, we received an additional 7500 images written by 70 different scribes and their corresponding ground truth annotations from LDC. Approximately half of these images were from scribes who were never seen in previous training sets. With the addition of this set, the total amount of available training data for text recognition is 16K images written by 101 unique scribes. We trained the text recognition system on the entire training set using the GC+PACE (Gradient-Concavity and Percentile stream) features. Next, we decoded the MADCAT Phase 1 DevTest Part2 with unsupervised page-wise adaptation. A 5% relative improvement in WER was obtained over the Phase 1 model trained with 9K pages.

Unsupervised Scribe Adaptation [BBN]: In Phase 1, we performed unsupervised page-wise maximum likelihood linear regression (MLLR) adaptation using the 1-best hypothesis produced by the decoder. Since many MADCAT images contain only a small number of words, a page-wise adaptation scheme is likely to over-fit to the errors in the 1-best decoder hypothesis. Therefore, we modified the adaptation procedure to adapt to automatically determined scribe clusters from training instead of just adapting to the 1-best hypothesis for the page. During training, we first perform unsupervised clustering of pages. For clustering, we compute the following features: (1) average width and height of the connected components in the image, (2) pen pressure, (3) text density, (4) slope distribution, and (5) ratio between the external and internal contours. Next, K-means algorithm is used to cluster the pages. In the second stage, we adapt the "global" glyph HMM to each cluster. During recognition, we first assign the page to a cluster based on Euclidean distance. Next, we decode the test page using the models for the cluster it was assigned to. Finally, the models corresponding to the cluster to which the test page belongs are further adapted to the 1-best hypothesis obtained from the earlier decoding.

We performed experiments to evaluate both the efficacy of the features used for clustering pages and the recognition performance using a combination of cluster-based and page-wise adaptation. For evaluating the efficacy of the features for clustering, we performed a supervised closed-set scribe identification experiment with 58 scribes. Using the features described above with a nearest neighbor classifier, we obtained 77.8% top-1 accuracy. Next, for evaluating our adaptation approach, we performed three sets of experiments using PACE features on MADCAT DevTest Part1 and Part2: (1) un-adapted recognition, (2) page-wise adaptation as implemented in our Phase 1 system, and (3) the combination of page-cluster and page-wise adaptation approach described above. The WER of un-adapted decoding is 44.0% and the WER of page-level adaptation is 41.2%. In contrast the lowest overall WER of combining page cluster adaptation and page-level adaptation is 39.8%, obtained from

MLLR page-cluster adaptation on 1024 page clusters. We plan to incorporate this new adaptation architecture in our Phase 2 evaluation system.

Named Entity Detection using Lattices [BBN]: Named-entities (NEs) are difficult to recognize and translate, especially for Arabic, because, the vast majority of Arabic names are also commonly used words. For example, the word Salim could be the name of a person or it could mean "blessed" – the *sense* in which the word is used can only be disambiguated by considering the surrounding context.

This quarter, we performed experiments to improve recognition of named entities using lattices. First, we annotated the MADCAT DevTest Part1b for three types of named entities: Person, Organization, and Geographical Location. A total of 1246 occurrences of 428 unique NEs were annotated. We then created a list of NEs to be detected in the decoder hypotheses of the test set. We searched for an exact match of the NE string in two forms of decoder hypotheses: (1) the 1-best hypothesis, and (2) the word lattice. As one would expect, the search for NEs in lattices improves recall at the expense of increasing the number of false positives. Therefore, we applied a threshold on word confidence posteriors to reject the less likely hypothesis. As illustrated in the ROC curve shown in Figure 2, searching for NEs in the lattice results in a larger number of true detects for the same number of false accepts as the 1-best search (7% relative improvement in %True Detects at the same %False Accepts). In addition, searching for NEs in the lattice provides the option of being able to detect a larger number of NEs if we are willing to tolerate a larger number of false accepts.

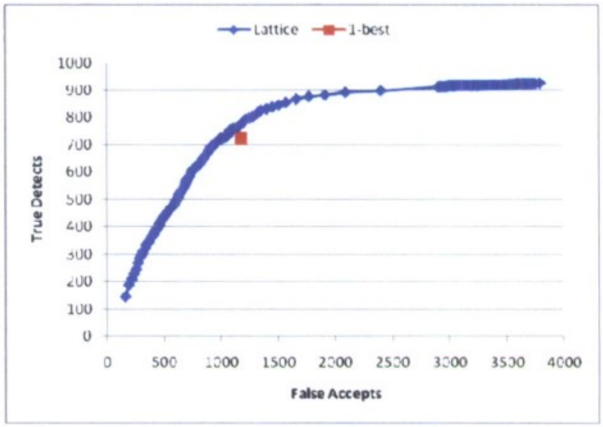

Figure 2: Comparison of named entity detection performance using 1-best search and lattice search

1.4. Integration with GALE MT [BBN]

Recognition Lattices for MT [BBN]: Since a word lattice is more likely to contain a better answer than the 1-best recognition hypothesis, this quarter we setup the infrastructure for using lattices from text recognition as input for machine translation (MT). Specifically, we updated our experiment infrastructure to generate lattices from text recognition. Next period, we will perform MT experiments with lattice input and compare it to translating the 1-best text recognition hypothesis.

1.5.Metadata Extraction [BBN, BAE, Lehigh, SUNY, UMD]

Logo Recognition [BAE]: During this period, we investigated performance of the alpha-rooted phase correlation (ARPC) based matcher on the augmented tobacco logo database and the ANFAL database. The augmented tobacco logo database contains 56 logo classes with 291 logo exemplars spread unevenly across the logo class set. The ANFAL dataset consists of logos extracted from Arabic documents. The ANFAL logo data is a challenging data set due to the degraded quality of the logo data. From roughly 100,000 document images, we culled a set of logos for the augmented ANFAL database. The augmented ANFAL logo database contains 44 logo classes with 300 test logo realizations spread across the different classes. On the tobacco database, we got a performance of 97% correct recognition. On the ANFAL database, we obtained a recognition accuracy of 90.3%.