Raytheon
BBN Technologies

Progress in the Raytheon BBN Arabic Offline Handwriting Recognition System

4 September 2014

Huaigu Cao
Raytheon BBN

Prem Natarajan
Information Sciences Institute (ISI), USC

Xujun Peng
Raytheon BBN

Krishna Subramanian
Raytheon BBN

David Belanger
former Raytheon BBN employee

Nan Li
Raytheon BBN

Raytheon
BBN Technologies

Progress in the Raytheon BBN Arabic Offline Handwriting Recognition System

4 September 2014

Huaigu Cao
Raytheon BBN

Prem Natarajan
Information Sciences Institute (ISI), USC

Xujun Peng
Raytheon BBN

Krishna Subramanian
Raytheon BBN

David Belanger
former Raytheon BBN employee

Nan Li
Raytheon BBN

Overview

- Arabic handwriting is difficult to recognize
- State of the art method primarily based on HMM+LM
- Additional methods in good systems: SVM-SSM, DNN-based tandem, DNN/HMM hybrid system,
- Virtually all speech recognition modeling improvements may apply to or is worth being tried in Arabic handwriting recognition

Effective Arabic Character Classification Techniques

HWR Technique	Research Teams
SVM	BBN CERPARMI IRISA
GMM-HMM	A2iA BBN RWTH UOB-Télécom ParisTech Siemens IRISA
Bernoulli-HMM	UPV
Neural Network	A2iA BBN TU Munich

Open Evaluations: Arabic HWRC and OpenHaRT

Evaluation	Arabic HWRC: Isolated Word Recognition		OpenHaRT: Word Sequence Recognition
Test set	IfN/ENIT Set f (easier)	IfN/ENIT Set s (more difficult)	NIST OpenHaRT
lexicon size	937 words (closed set)	937 words (closed set)	Open lexicon
Best Word Error Rate %	6.6 [TU Munich, 2009]	15.4 [UPV, 2010]	16.1 [RWTH, 2013]

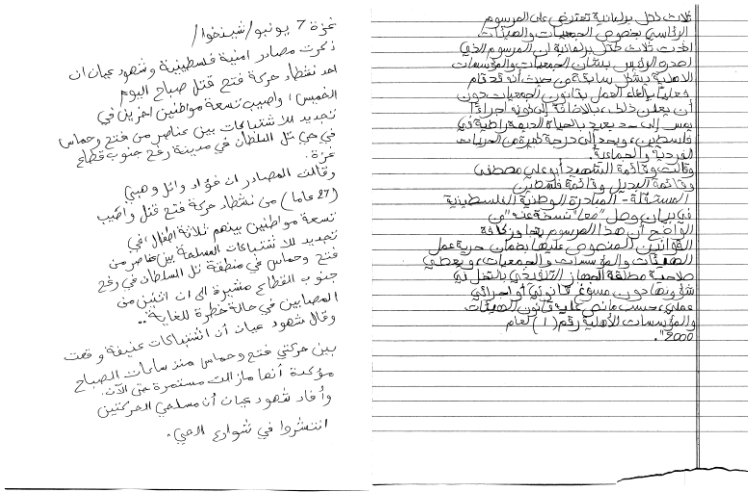
BBN OCR System at a Glance

- **HMM Decoder:** Decode text line images using HMM and n-gram LM; output can be generated as 1-best, n-best or lattice
- **N-best Rescore:** Improve 1-best accuracy by re-ranking n-best results using several different ways of re-evaluating glyph and language scores
- **Adaptive Training:** Choose appropriate model adaptation approaches for target application
- **System Combination:** Combine OCR results from more than one independently designed de-noising algorithms using a consensus network-based approach

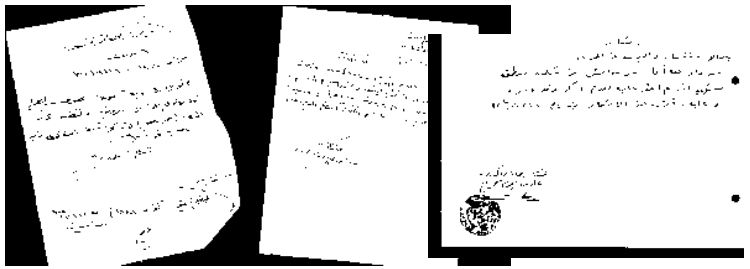
Data for Model Training and Evaluation

- Controlled Data
  - People hired to hand copy prepared Arabic web and news text
  - 42K pages, by 308 writers, ~20M characters for training
  - High quality 600DPI scanned, mostly clean, 10% noisy
  - Evaluation set selected by NIST and used in OpenHaRT Evaluation
  - Publicly available
- Uncontrolled Data
  - 18K pages of real documents collected from field, ~ 8M characters
  - Low quality noisily scanned 200DPI, many documents are noisy with clutters and cursive writing
  - Multiple genres (memos, letters, forms, tables)
- Text-only Data Corpora
  - LDC Arabic Gigaword v. 3 and v. 4

Controlled data

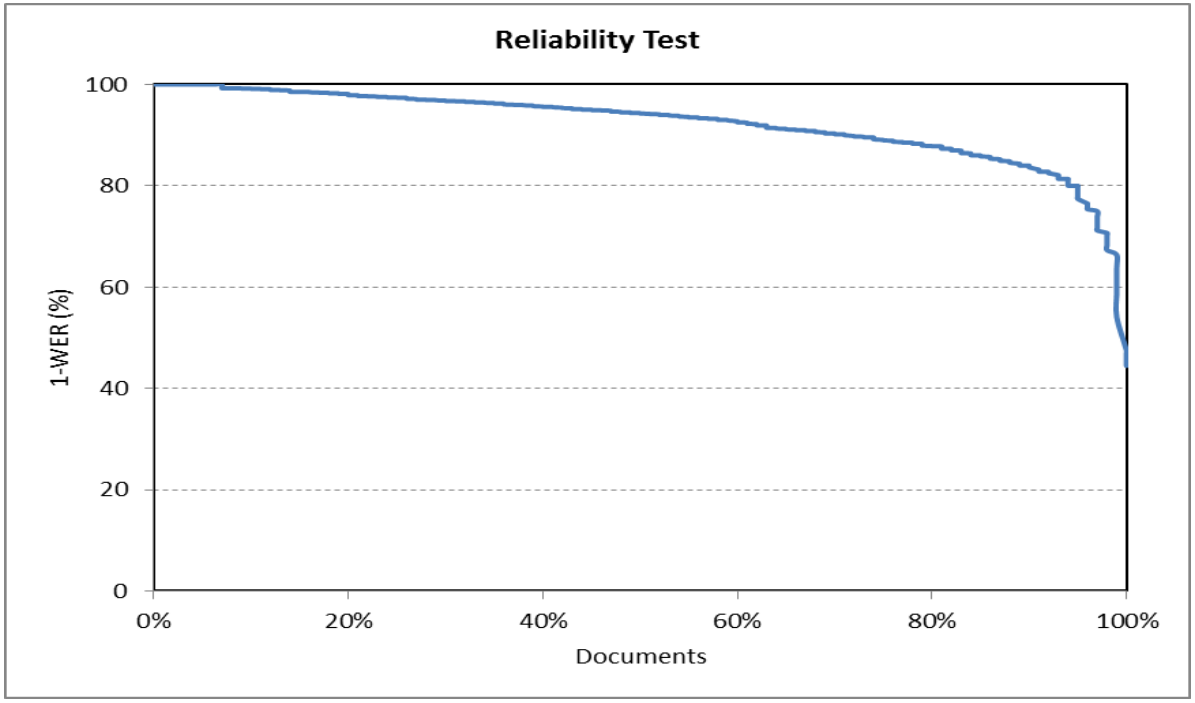


Uncontrolled data



System Performance

System	Test Set	Word Error Rate (WER) %
BBN	Controlled	7.9
Best Team of OpenHaRT'13	Controlled	16.1



System Performance

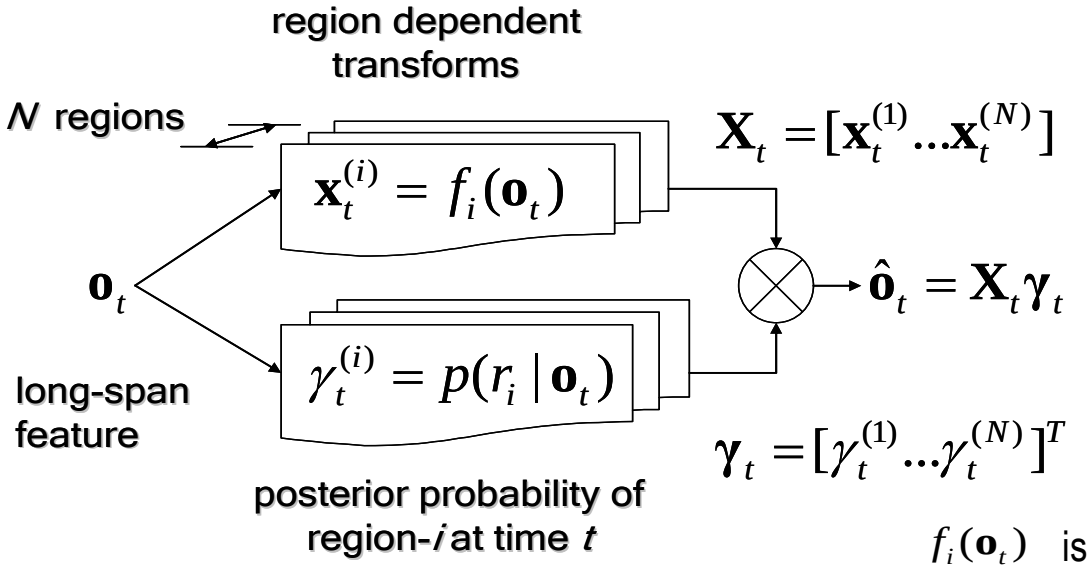
System	Test Set	WER %
BBN	Uncontrolled	22.1

Sliding Window Features

- Estimate text size within each line and determine sample rate accordingly
- Moving interval is 1/60 height of largest text in the line
- Dynamic window top and bottom selection according to change of text position
- Compute script-independent features on each slice as input to HMM: intensity percentile, total window intensity, stroke orientation angle, correlation value for computing the angle, gradient, concavity, Gabor filter response (177 dim/frame)

Nonlinear Feature Transform [Chen, ICFHR2012]

- Region Dependent feature Transform (RDT) [Zhang, INTERSPEECH06]
  - Non-linear feature transform
  - Originally developed for HMM-based ASR
- Major steps
  - Divide feature space into multiple regions using a GMM
  - Estimate linear feature transform for each regions
  - Interpolate locally transformed features with region posteriors
  - Optimization based on MPE criteria
- 5% relative improvement in WER over Linear Discriminant Analysis (LDA)



$f_i(\mathbf{o}_t)$ is a region-dependent transform

Usually $f_i(\mathbf{o}_t) = \mathbf{A}_i \mathbf{o}_t + \mathbf{b}_i$

HMM Recognizer

- 14-state character HMM (Arabic, English letters, digits, and punctuations)
  - Triphone is used; each triphone state has a unique set of GMM weights
  - Each character state has a unique set of GMM means/cov
- Minimum Character Error (MCE) training [Povey, ICASSP2002]

Language Model

- N-gram LM: 2-gram & 3-gram LM w/ Kneser-Ney smoothing
  - 2-gram for quickly generating sufficient candidates
  - 3-gram for searching for more accurate result
- External training data: LDC Arabic Gigaword 3 & 4, text within the same epochs as the newswire test data is taken out of the training data
- Dictionary keeps 300 thousand most frequent words

N-best Rescore

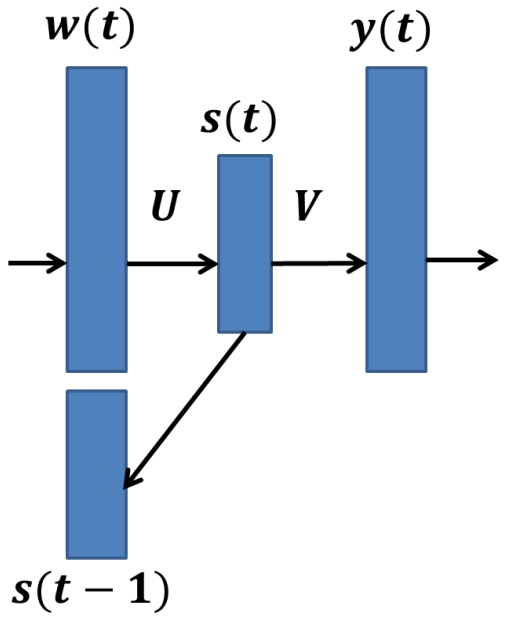
- N-best rescore: a classic way to leverage additional classifiers and analysis tools
- Generate n-bests using HMM decoder
- Basic scores: HMM loglikelihood, LM probability
- Additional scores from
  - Discriminative segment classifiers (SVM, DNN)
  - Transformed images
  - Stronger LM (e.g., higher order, RNNLM)

Segment Classifier Score

- In a character classifier, each character has a class label
- In a character state classifier, each character HMM's state has a class label
- Examples:
  - Stochastic Segment Model (SSM): character classifier using SVM [Prasad, ICPR2010]
  - Hybrid DNN-HMM for speech: frame classifier using deep network [Yu, IEEE TASLP2013]
- Log-likelihood score generated by normalization of segment classifier decision scores

Recurrent Neural Network Language Model (RNNLM)

- RNNLM [Mikolov, INTERSPEECH 2010]
  - Modeled as a recurrent neural network with input layer w, hidden layer s (context layer) and output layer y
  - Input is an concatenation of the vector representing the current word and the output of context layer
  - Similar to n-gram, an RNNLM joint probability can be derived for each n-best

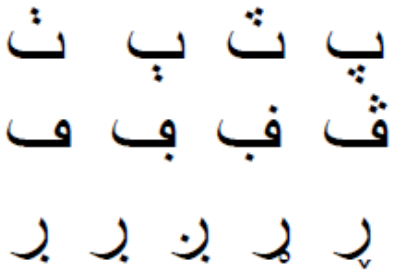


$$x(t) = w(t) + s(t-1)$$
$$s(t) = f\left(U \cdot x(t)\right)$$
$$y(t) = g\left(V \cdot s(t)\right)$$
$$f(z) = \frac{1}{1+e^{-z}}, g(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}}$$

HMM using Reduced Symbol Set (RSS)
(for Arabic Only)

Raytheon
BBN Technologies

- Similar Arabic characters



- Rescore using reduced symbol set recognizer
  - Remove dots from Arabic text line images
  - Generate a reduced symbol set consisting of normalized Arabic letters and all other non-Arabic symbols
  - Train an HMM recognizer using reduced symbol set

LM in Translated Foreign Language [Devlin, ICFHR2012]

- MT score/ foreign language LM:

  If we translate each n-best into English, the machine translation probability, after normalization, can be used in n-best rescore

- Yielded significant improvements on NIST Arabic data [Devlin, ICFHR2012]

N-best Rescore Results

Scores	Rel. WER Reduction %	Note
HMM, n-gram, char-SVM	3.6	Verified on controlled set *i*
HMM, n-gram, reduced symbol score	2.9	Verified on controlled set *i*
HMM, n-gram, frame-DNN	0.9	Verified on uncontrolled set
HMM, n-gram, RNNLM	3.5	Verified on uncontrolled set
HMM, n-gram, MT-LM	2.5	Verified on controlled set *i*

Model Adaptation

- Supervised adaptation
- Unsupervised adaptation
- HMM adaptation methods
  - Applied *Maximum A Posteriori* (*MAP*) [Gauvain, IEEE TSAP1994] to adapt HMM mean/cov when there are more training data
  - Applied *Maximum Likelihood Linear Regression* (*MLLR*) [Leggetter, Computer Speech & Language 1995] or *Constrained Maximum Likelihood Linear Regression* (*CMLLR*) to adapt HMM mean/cov when there are less training data
  - Applied character-tied *Duration Adaptation* [Cao, ICFHR2010] to adapt HMM transition probabilities

Writer Identification – Use Cases

- Use case 1: select writer dependent HMM
- Use case 2: group test data that are written by the same writer to get more adaptation data

Writer Identification Methods

- Text-independent: GMM, supervector, *i* –vector, Fisher vector: normally applied in text-independent, bag-of-words style, suitable when there are fewer training data

- Text-dependent: OCR-driven ensemble of character-level SVM writer classifiers [Cao, ICDAR2011]: generates character boundary using HMM-based forced alignment, suitable when there are more training data

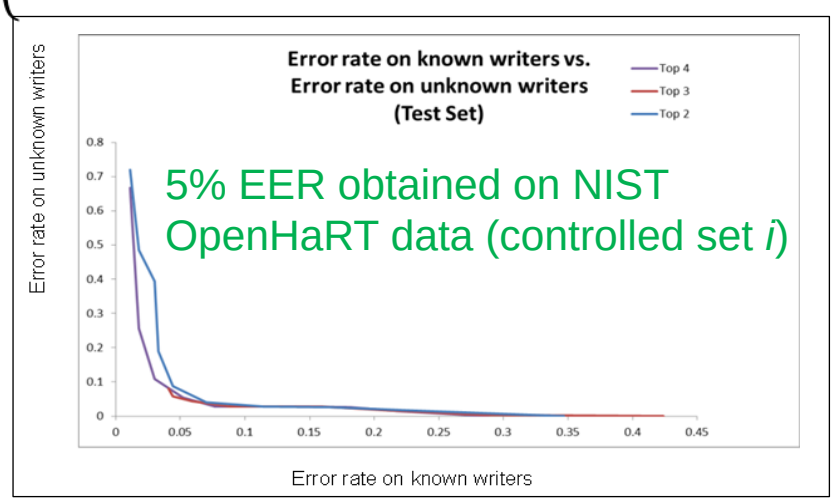
Writer Verification

- Motivation: writer-independent HMM slightly outperforms mismatched writer-dependent HMM
- Added writer verification stage based on OCR-driven writer ID (prior work: [Cao, ICDAR2011])

$w_1, w_2, \ldots w_n$ top-n writer ID candidates

Criteria to verify w_1:

$$\begin{cases} \textbf{accept} \text{ if } \dfrac{\mathcal{L}(w_1)}{\mathcal{L}(w_1)+\mathcal{L}(w_2)+\cdots+\mathcal{L}(w_n)} > T \\ \textbf{reject} \text{ otherwise} \end{cases}$$



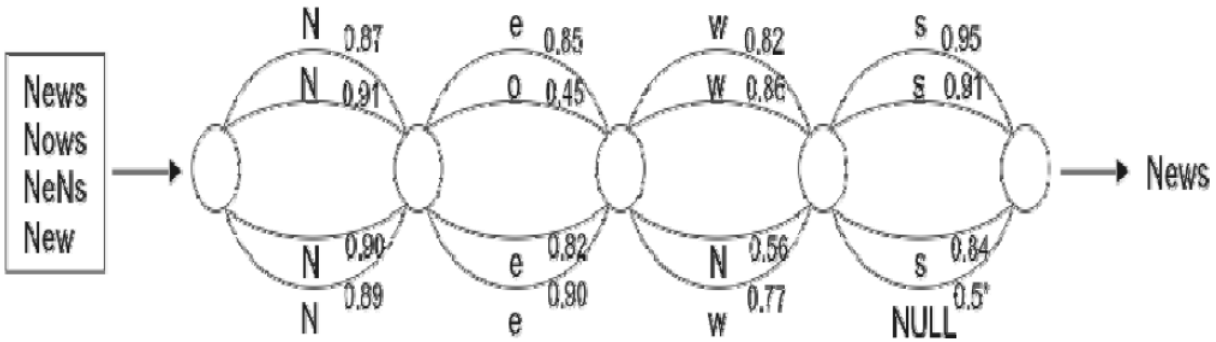
Writer Type	Relative WER Improvement % Resulted from Writer Verification
Known *	5.6
Unknown	6.4

Data: LDC controlled *i*, 50% by known writers

* Known implies presence in training data

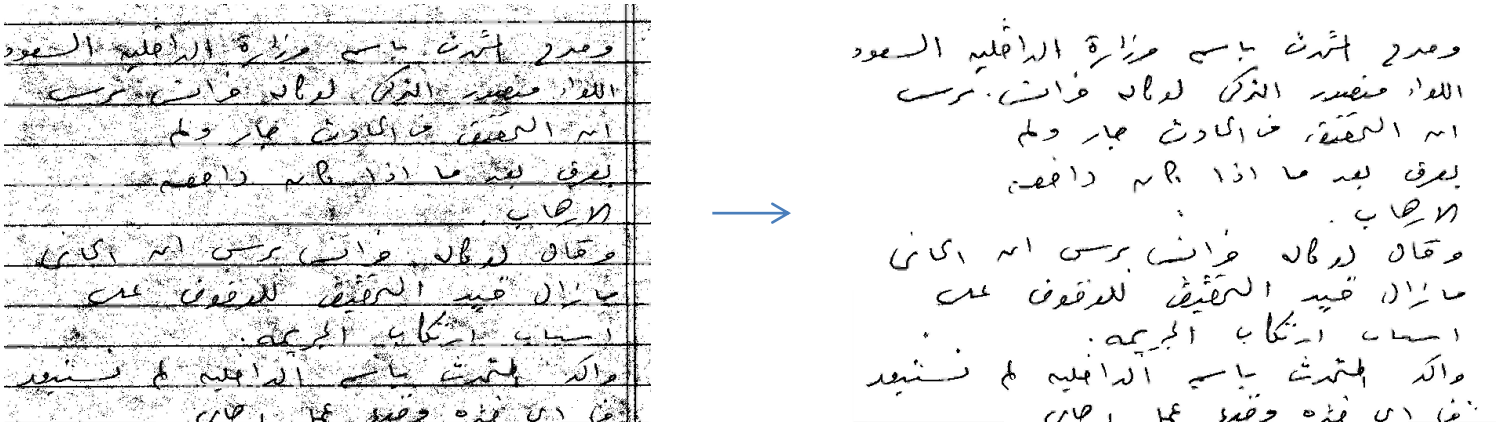
System Combination

- Build a confusion network from N-best of multiple systems
  - Arcs represent hypothesized words
  - Nodes represent transition points
- Use word posterior probability calculated based on the frequency of the word in the N-best of each system as features
- Optimize feature weights on a held-out tuning set



Selection of Single Systems

- Prior work [Prasad, ICPR2010] combined systems of different feature sets and modeling methods (MLE vs. MPE, GMM sizes, adapted vs. unadapted) but improvement diminishes as features and models become better
- Current method combines systems of different image preprocessing algorithms:
  - De-noising [Cao, MOCR2009], [Shi, ICDAR2011], [Caner, ICPR2010]



  - With or without slant correction
  - Train one model using noisy images and another model using cleaned image
  - Some systems recognize noisy images and others recognize cleaned image
  - Leads to 6.3% relative reduction of WER

Q & A