The UPV Handwriting Recognition and Translation System for OpenHaRT 2013

DSIC/ITI
Universitat Politècnica de València
E-46022 València (Spain)

I. Khoury, A. Giménez, J. Andrés, A. Juan, J.A. Sánchez
[ialkhoury,agimenez,jandres,ajuan,jandreu]@dsic.upv.es

August 23, 2013
Outline

Site Introduction

Transcription System

Translation System

Submissions

Tools and Means

Results

Conclusion

References
Site Introduction

- Pattern Recognition and Human Language Technology research group (PRHLT)

- From the Universitat Politècnica de València (UPV)
  - DSIC and DISCA of the Universitat Politècnica de València (UPV)
  - Instituto Tecnológico de Informática (ITI) from UPV

- Interests:
  - Multimodal Interaction
  - Machine Translation
  - Handwritten Text Recognition (HTR) and Document Analysis
  - Automatic Speech Recognition and Understanding
  - Image Analysis and Computer Vision
  - Transcription and Translation of Video lectures (transLectures) [1]
Related work and current research in HTR:

- HTR using Bernoulli and Gaussian HMMs applied to:
  - Arabic IFN/ENIT database [9]
  - Arabic APTI database for Printed Arabic [10]
  - NIST OpenHaRT 2010 and 2013 (LDC) corpus
  - IAM database [7]
- BHMMs using discriminative training
Transcription System

- **Image Processing**
  - Scaling to a given height (30 pixels)
  - Image Binarization using Otsu method

- **Text Processing**
  - Adding shape information to Arabic transcripts

- **Feature extraction**
  - Window extraction to a given width (9 pixels)
  - Window repositioning to its center of mass
    - Vertical, Horizontal, and Both directions (Vertical)

- **HMM system using Bernoulli mixtures (BHMM)**
  - Fixed number of states (6 states per character)
  - Mixture components per state (128)
  - Tri-character approach
  - EM algorithm for training and recognition
  - 5-grams Language Model (LM) for recognition
  - Grammar Scale Factor (GSF) on LM (30)
Figure: Example of transformation of a $4 \times 5$ binary image (top) into a sequence of 4 15-dimensional binary feature vectors $O = (o_1, o_2, o_3, o_4)$ using a window of width 3. After window extraction, the standard method is compared with the vertical repositioning. Mass centers of extracted windows are also indicated.
Translation System

- Our system is based on a state-of-the-art log-linear translation system (Moses toolkit)
- Standard moses features
  - Phrased-based model
    - Phrase translation probabilities (both directions)
    - Lexical weights (both directions)
  - Language Model (5-grams trained with SRILM)
  - Distance-based reordering model
  - Word penalty
  - Lexicalized reordering model
Translation System (Cont.)

- Text processing
  - tokenization:
    - English was tokenized with Moses tokenization tools
    - Arabic was tokenized with MADA+TOKAN tools
    - Removing long sentences (longer than 150 words)

- Standard Moses training
  - Alignment extraction
  - Phrase extraction
  - MERT
Submissions

- Document Image Recognition (DIR)
  - Two systems followed the constrained training condition
  - Trained with our BHMMs approach
  - The contrastive system was trained using the complete data
  - The primary system was trained with less data

- Document Text Translation (DTT)
  - Two systems: Different training conditions
  - Trained with Moses toolkit
  - For the constrained training condition:
    - We used only the LDC resources for the OpenHaRT’13
  - For the unconstrained training condition:
    - We used the MultiUN and TED corpus (IWSLT 2011)
    - Aligned on sentence level using the Champollion tool
    - Sentences were selected according to the infrequent $n$-grams score
Document Image Translation (DIT) Given a handwritten image \( f \), it can be expressed as follows

\[
y^* = \arg\max_{y \in Y} p(y | f) = \arg\max_{y \in Y} \sum_x p(x | f) p(y | x) \tag{1}
\]

where,

- \( f \): input image
- \( x \): candidate recognized source (Arabic) text
- \( y \): candidate translated sentence (in English) corresponding to \( f \).

- Three systems followed the constrained training condition
- The probability \( p(x | f) \) in Eq. (1) was approximated by the primary DIR transcription system
- The key difference among systems lay in the translation subsystems
Translation subsystems for the DIT task (Three Systems)

- The first DIT system (DIT1), Eq. (1) was approximated as follows,

\[
y^* \approx \arg\max_{y \in Y} \left[ \max_{x} \left\{ p(x|f) \cdot p(y|x) \right\} \right]
\]

\[
\approx \arg\max_{y \in Y} \left[ p(y|\max_{x} \left\{ p(x|f) \right\}) \right]
\]

*The \( p(y|x^*) \) was approximated by the primary DTT translation system

- The input image was recognized by the primary DIR transcription system, and the recognized text was fed into the primary DTT translation system.
The second DIT system (DIT2):

- Followed a similar approach to the first DIT system
- The source part of each bilingual training pair was substituted by the transcription obtained by the primary DIR system
- The new training data set produced in this way was used to train the translation system
- It was expected to better handle the noisy output of the DIR system
- Better performance than the primary DTT in development set but worse performance in the test set
The third DIT system (DIT3):

- Different approximation of Eq. (1) was used

\[ y^* = \arg\max_{x \in \text{NBest}(f)} \left\{ \arg\max_{y \in \text{NBest}(f|x)} \left\{ p(x|f) [p(y|x)]^\theta \right\} \right\} \]  \hspace{1cm} (3)

- Introducing a scaling factor \( \theta \)
- The search space was approximated by \( N \)-best lists
- Each input image was first recognized using the primary DIR system into 100-Best transcriptions, and then each transcription was translated using the primary DTT system into 100-Best translations
## Data statistics

### Table: Data (lines) used for training each system and its training conditions.

<table>
<thead>
<tr>
<th>System/Condition</th>
<th>Constrained</th>
<th>Unconstrained</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIR1</td>
<td>779, 100</td>
<td>-</td>
</tr>
<tr>
<td>DIR2</td>
<td>789, 874</td>
<td>-</td>
</tr>
<tr>
<td>DIT (recognition part)</td>
<td>779, 100</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table: Data (segments) used for training each system and its training conditions.

<table>
<thead>
<tr>
<th>System/Condition</th>
<th>Constrained</th>
<th>Unconstrained</th>
<th>LDC</th>
<th>MultiUN</th>
<th>TED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DTT</td>
<td>40, 580</td>
<td>19, 956</td>
<td>2, 205</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIT (translation part)</td>
<td>40, 580</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

The UPV HTR and MT System for OpenHaRT’13

August 23, 2013

14 / 18
Tools and Means

For Text Processing:
- Moses tokenization tools [5]
- MADA+TOKAN [8] toolkit
- Champollion Toolkit (CTK) [6]

For Handwritten Text Recognition:
- TLK toolkit [2]

For Machine Translation:
- Moses toolkit [5]
Results (Line Condition)

Table: Submitted systems for DIR and line segmentation condition together with their Word Error Rate (WER\%)

<table>
<thead>
<tr>
<th>System</th>
<th>Reference</th>
<th>WER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Eval’10</td>
<td>Eval’13</td>
</tr>
<tr>
<td>DIR1</td>
<td>p-1_1_20130425</td>
<td>29.08</td>
</tr>
<tr>
<td>DIR2</td>
<td>c-1_2_20130425</td>
<td>-</td>
</tr>
<tr>
<td>UPV PRHLT</td>
<td>OpenHaRT’10</td>
<td>47.45</td>
</tr>
</tbody>
</table>

- The DIR2 system slightly outperforms the DIR1 system
  - Expected improvement: DIR2 was trained with more data
- Both DIR1 and DIR2 systems outperform the DIR system of the 2010 evaluation (UPV PRHLT)
  - Trained with more mixture components (128) per state
  - We used a bigger language model for recognition.
Results

Results (Line Condition)

Table: Submitted systems for (DTT and DIT) and line segmentation condition together with their BLEU score

<table>
<thead>
<tr>
<th>System</th>
<th>Reference</th>
<th>BLEU [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Eval’10</td>
</tr>
<tr>
<td>DTT Constrained</td>
<td>p-1_1_20130425</td>
<td>22.53</td>
</tr>
<tr>
<td>DTT Unconstrained</td>
<td>p-1_1_20130425</td>
<td>25.18</td>
</tr>
<tr>
<td>DIT1</td>
<td>p-1_1_20130425</td>
<td>16.51</td>
</tr>
<tr>
<td>DIT2</td>
<td>c-1_2_20130425</td>
<td>16.58</td>
</tr>
<tr>
<td>DIT3</td>
<td>c-1_3_20130425</td>
<td>18.13</td>
</tr>
</tbody>
</table>

► The Unconstrained DTT system significantly outperforms the Constrained DTT system.
  ► The usage of an additional data (around 20\(K\)) significantly improved the translation accuracy in the DTT system.
  ► Sentence selection according to the infrequent \(n\)-grams score [4]
► The DIT3 shows better performance over DIT1 and DIT2
Conclusion

- The UPV Recognition and Translation System for the NIST OpenHaRT’13 evaluation.
- Submissions:
  - Two systems for the DIR task (constrained training condition)
  - One system for the DTT task (both training conditions)
  - Three systems for the DIT task (constrained training condition)
- Results for the DIR task outperform previous results in OpenHaRT 2010 evaluation
- Results for DTT and DIT tasks are very promising


References


Owen Rambow Nizar Habash and Ryan Roth.
Mada+tokan: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization.


M. Pechwitz et al.
IFN/ENIT - database of handwritten Arabic words.

Fouad Slimane, Rolf Ingold, Slim Kanoun, Adel M. Alimi, and Jean Hennebert.
A new arabic printed text image database and evaluation protocols.