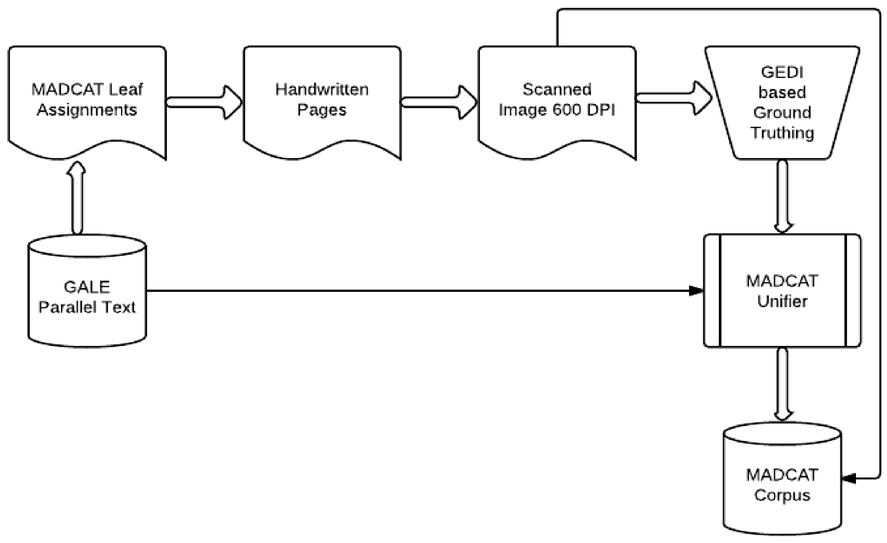
Linguistic Resources for Handwriting Recognition and Translation Evaluation

Zhiyi Song, Safa Ismael, Steven Grimes, David Doermann, Stephanie Strassel

Linguistic Data Consortium

Introduction

- LDC supports handwriting recognition and translation evaluation programs: MADCAT, OpenHaRT
  - Arabic scribe collection 2008-2010
    - Handwritten version of existing GALE parallel texts
  - Chinese scribe collection in 2011
    - Handwritten version of existing GALE parallel texts that has Treebank and/or word alignment annotation



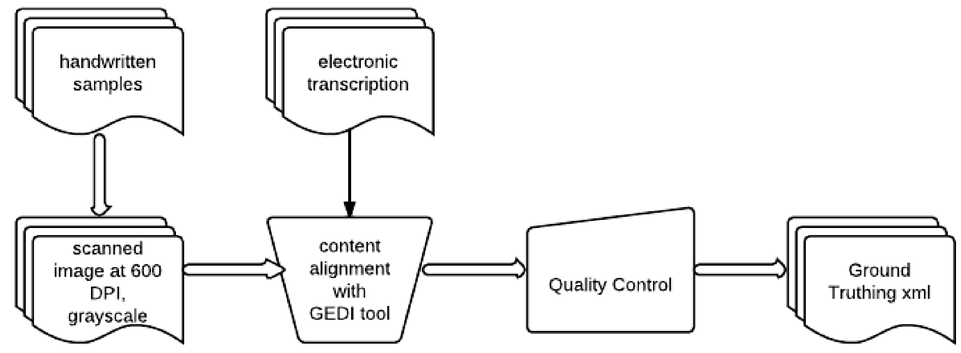
Scribe collection

- Participant recruitment, testing, training
  - Literate native speakers of Arabic/Chinese
  - All participants trained and tested
  - Vetted training assignments before production assignments
- Work flow and data management
  - Scribble: a PHP-based web application using CodeIgniter on the back end and jQuery for front end validation
    - manage scribe registration
    - handle kit creation, assignments
    - handle document validation and check-in
    - track and update kit/page status
    - manage e-text packages for ground truth annotation
  - Ground truth annotation and data delivery not handled by Scribble

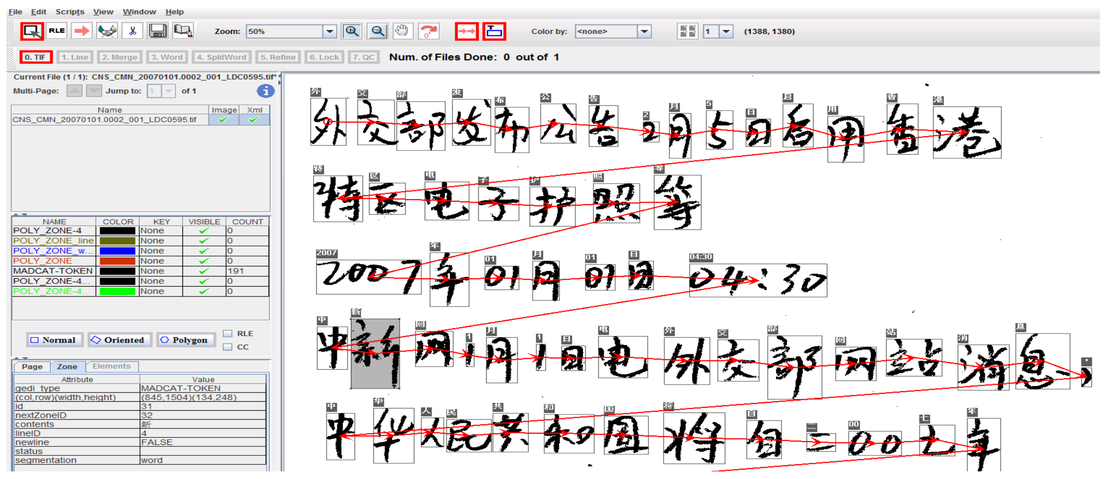
Data processing for scribe collection

- Scribe kits creation from a set of segmented GALE source documents. Three steps:
  - **Step 1**: tokenize the text, execute word and line wraps, paginate GALE source text into kit pages
    - Arabic: maximums: 20 lines/page, 5 words/line
    - Chinese: maximums: 15 lines/page, 15 characters/line
  - **Step 2**: manually review MADCAT kit pages for content and formatting
  - **Step 3**: generate alternate kits given a set of MADCAT pages and preselected kit parameters
    - Arabic: 2-7 versions of the same kit
    - Chinese: 15 versions of the same kit
    - Writing conditions: 90% pen, 10% pencil; 75% unlined paper, 25% lined paper; 90% normal speed, 5% careful speed and 5% fast speed
- Annotation preparation using Scribble to:
  - coordinate management and bookkeeping of kit selection
  - generate corresponding tokenized text of each scribe page
  - provide kit and page profiles which include ID, writing condition, scribe ID

Annotation



- Content alignment
  - using the GEDI tool to draw polygon bounding box around each line, word/character token with unique ID assigned
  - Each token's physical coordinates on the page are recorded as the "ground truth"
  - Reading order is automatically added (Chinese L>R, Arabic R>L)
  - Each token is reviewed, additional features are added to indicate status of extra token, typo, etc.
  - Missing tokens in handwritten image are aligned with empty boxes in GEDI



- QC procedures:
  - GEDI tool enforces various constraints on reading order, text alignment and consistency of token attributes
  - GEDI tool provides mechanisms for extra QC procedures

Data Processing

- Unified data format consolidates GALE source text, translation text and ground truth annotation. Output: a single XML file with multiple layers of information
  - text layer for source text with word/character tokenization and sentence segmentation
  - image layer with bounding boxes
  - document metadata layer
  - translation layer

Results

- 42,000+ Arabic handwritten pages, 223,600 Chinese handwritten pages
- All collected, annotated and released to MADCAT program participants
- Most will be made generally available to the larger research community through LDC's catalog

Acknowledgments

We acknowledge and appreciate the work of David Lee on the technical infrastructure of MADCAT. This work was supported in part by the Defense Advanced Research Projects Agency, MADCAT Program Grant No. HR0011-08-1-004. The content of this paper does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

Contacts

Linguistic Data consortium: Zhiyi Song zhiyi@ldc.upenn.edu; Steven Grimes sgrimes@ldc.upenn.edu; Stephanie Strassel strassel@ldc.upenn.edu

Applied Media Analysis: David Doermann doermann@appliedmediaanalysis.com

Science Applications International Corporation: Safa Ismael safa.s.ismael@saic.com