Linguistic Resources for Handwriting Recognition and Translation Evaluation

Zhiyi Song*, Safa Ismael*, Steven Grimes*, David Doermann×, Stephanie Strassel*
*Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA USA
× University of Maryland and Applied Media Analysis, Inc, College Park MD

E-mail: {zhiyi,safa,sgrimes,strassel}@ldc.upenn.edu, doermann@appliedmediaanalysis.com

Abstract

We describe efforts to create corpora to support development and evaluation of handwriting recognition and translation technology. LDC has developed a stable pipeline and infrastructures for collecting and annotating handwriting linguistic resources to support the evaluation of MADCAT and OpenHaRT. We collect handwritten samples of pre-processed Arabic and Chinese data that has been already translated in English that is used in the GALE program. To date, LDC has recruited more than 600 scribes and collected, annotated and released more than 225,000 handwriting images. Most linguistic resources created for these programs will be made available to the larger research community by publishing in LDC's catalog. The phase 1 MADCAT corpus is now available.

Keywords: handwriting image, recognition, translation

1. Introduction

LDC[1] has been producing linguistic resources to support handwriting technology evaluation since 2008 (Strassel, 2008, 2009). The two programs that LDC is currently supporting are MADCAT[2], funded by DARPA[3] (DARPA, 2008), and OpenHaRT[4] (NIST, 2012), funded by NIST[5]. In both programs, LDC collects handwritten samples of pre-processed data that has been already translated or spontaneously produced materials. This paper is going to discuss the data collection and production that we do for MADCAT and OpenHaRT. As shown in the figure 1 below, the effort involves data pre-processing, handwriting sample collection, annotation, data post-processing, and distribution.

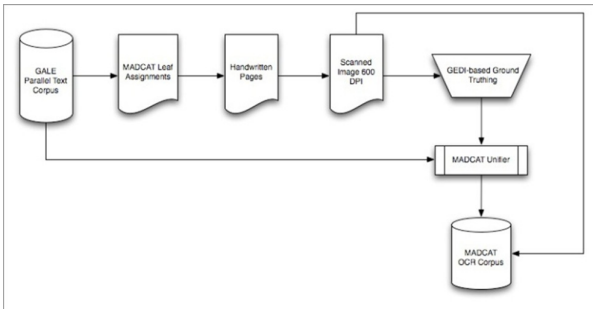

Figure 1: MADCAT data pipeline

A pilot study of Chinese was added to the MADCAT program in 2011, and the collection procedures for Chinese are similar to that for Arabic (Strassel, 2009). The Chinese portion of the project is focused on acquiring handwritten versions of existing GALE[6] parallel text data (Song et al, 2010) which also has Treebank (REF) or word alignment annotation (Li et al, 2010). The purpose of using data which have been word aligned or treebanked is to create richly annotated corpora, considering that systems are evaluated not only on OCR accuracy but on the quality of the translation of foreign language images to English text as well.

2. Scribe Collection

The GALE program collected, annotated and translated a large amount of Arabic and Chinese text. One of the major efforts of LDC to support the MADCAT program is to acquire handwritten versions of GALE data by undertaking scribe collection.

2.1 Scribe recruiting

Scribe participants are recruited either locally or remotely via our collaborators overseas. Only native speakers of the target language who can read and write that language are allowed to participate. In the case of Arabic, we recruit participants from different regional backgrounds as there may be differences in handwriting styles. One of the examples is that Arabs from the Middle Eastern region usually use the Indic numbers, vs. those from the Maghreb and North African region who use the Arabic numbers. Participants are all screened for their literacy and eligibility; as part of the registration process, participants indicate their level of education, and the main language in which they were taught in each of those levels. All participants have to have completed primary school at minimum in order to qualify for participation. All participants are trained and tested following scribing guidelines developed by LDC. Participants are given test pages, with different implements (pen vs. pencil, lined vs. unlined paper; fast, normal or slow speed), to test their attention to details. This is also to test that they actually are literate in the language of which they claim as natives. Eligible participants are registered in the study and

[1] Linguistic Data Consortium
[2] Multilingual Automatic Document Classification, Analysis and Translation
[3] Defense Advance Research Projects Agency
[4] Open Handwriting Recognition and Translation
[5] National Institute of Standards and Technology
[6] Global Autonomous Language Exploitation

receive a provisional first assignment. Once the first assignment is returned and verified, the participant will receive future assignments without waiting for additional verification. To date, more than 450 Arabic participants and 150 Chinese participants have registered in the study.

2.2 Workflow and Data Management

LDC's handwriting collection web application is locally known as Scribble. Scribble, derived from another in-house application known as Scribe, is a PHP-based web application using CodeIgniter on the back end and jQuery for front end validation. This framework was chosen because it is lightweight and allows for rapid application development. Standoff annotation of individual handwriting documents is file-based, but annotation progress and workflow management is handled by Scribble, and details are stored in a MySQL database.

The Scribble application is tasked with many duties related to MADCAT handwriting collection. It is used to manage scribe registration, handle kit creation, serve annotation assignments, and handle document validation and check-in. Additional features include functionality to track and update kit/page status and manage e-text packages for ground truth annotation.

At this time Scribble handles nearly all aspects of the project except two critical functions – annotation and data processing. Ground truth annotation is handled using GEDI, which is described in more detail in Section 4. GEDI is written in Java though functions best on the Windows platform. Many of the polygon drawing functions of GEDI are gradually reimplemented in Scribble. Data processing for release is handled by standalone Perl and Python scripts used to manipulate the data format from the native format of the annotation tools, and this process is described further in Section 5.

3. Data Processing for handwriting collection

The processing procedures include creation of kits and annotation preparation after collection.

3.1 Kit creation

Before participants are recruited and register for collection, all documents are processed and divided into pages which are then grouped into kits for assigning to participants. The process takes the input of a set of segmented GALE source document and outputs a series of kits which are generated using partition parameters from the MADCAT database along with the corresponding kit objects in the database used for workflow management.

The kit generation process can be divided into three steps. The first step tokenizes the text, word and line wraps and paginates GALE source text into kit pages. To ensure that each scribe will not experience word/line wrapping problems, each Arabic page is limited to no more than 20 lines and a line no more than 5 words. A Chinese page has no more than 15 lines and a line no more than 15 characters. The second step, which is optional, is to have annotators manually review the MADCAT pages for content and formatting issues. This step can eventually be phased out once the team is confident there are no display issues for the kit pages.

The last step is to generate alternate kits given a set of MADCAT pages and kit parameters preselected by the team and entered in the database. Each version contains the exact same pages with different ordering and writing conditions. For an Arabic kit, there are 2-7 versions of the same kit. For the Chinese pilot study, a kit has 15 versions of each page. For both the Arabic and Chinese collections, the writing conditions stipulated 90% pen, 10% pencil; 75% unlined paper, 25% lined paper; 90% normal speed, 5% careful speed and 5% fast speed. By dividing writing, paper, and speed conditions across scribes while holding the content constant, we were able to obtain a variety of handwriting samples from different scribes under a range of conditions.

3.2 Annotation preparation

Once all handwritten pages of a kit are collected and checked for quality and completeness, the kit is ready to be processed for downstream annotation. The management and bookkeeping of kit selection is coordinated through Scribble. Project coordinator may click a button to process the data which includes identifying the corresponding tokenized text of each page, generating the kit and page profiles which include ID, writing condition, scribe ID. Hence each scanned image is associated with the attributes that were assigned when the kit was created.

4. Annotation

The annotation process includes high resolution scanning, content alignment and refinement of the markup with the GEDI tool (Doermann et al, 2010) originally developed at and currently available for download from the University of Maryland.

4.1 Scanning of documents

Collected handwritten samples are scanned at high resolution (600 dpi, grayscale). Since the number handwritten samples is quite large, this process is handled by our external vendor, Applied Media Analysis (AMA), using a sheet fed scanner that handles 40-60 documents per minute. Document images are integrated into AMAs workflow for binarization and alignment with the original transcription.

4.2 Content Alignment

Using the GEDI tool extended by AMA, a rectangle or polygon bounding box is first drawn around each line of text. Each line zone is assigned a unique ID and line starts are marked using the tool to preserve the parallel alignment of the content with its transcription. The system then aligns the original transcription files with each physical line on the image, and the annotator uses this content as a guide to divide the lines into tokens that

correspond to the original transcription. A token's physical coordinates on the page are recorded as the "ground truth" in XML format along with a set of other attributes. In Arabic, a token is a word, while in Chinese, a token is a Chinese character. Once a token is drawn, it is automatically assigned an ordinal number token ID in the order the tokens are annotated. Figure 2 indicates the polygon bounding box in the GEDI tool and how it preserves the alignment.

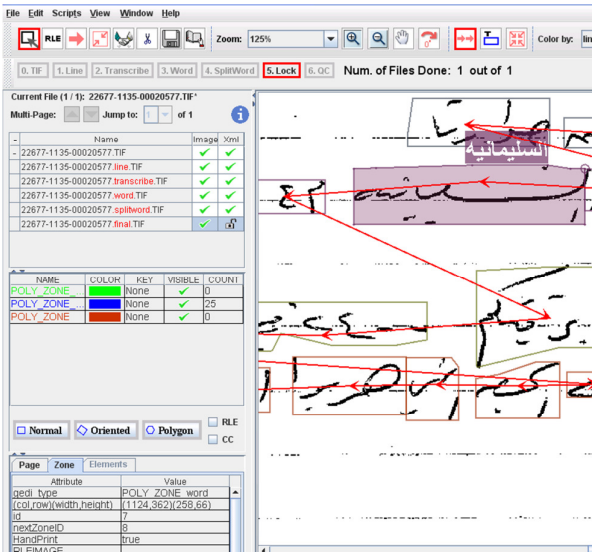

Figure 2: GEDI tool and content alignment

The reading order is then automatically added and refined if necessary to indicate the natural flow of writing (i.e. in case of Arabic from right to left, and from left to write in Chinese), as shown in figure 3.. Reading order can be difficult to determine in more complex documents which may contain other physical properties such as signatures, dates, images, or logos and for which there is no one obvious order to read the tokens on the page.

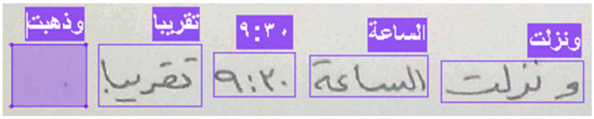

Figure 3: Reading order in GEDI tool

As part of quality control process, each token is reviewed and additional "status" tag is used to indicate scribe mistakes in creating the original document such as extra tokens, missing tokens, and typos.

In GALE-style data, electronic transcriptions exist for each document. Content of each document is imported into the tool. The number of tokens in the digital text should match that of the tokens drawn in the tool. Should there be a missing or an extra token in a given line, they are handled manually; in the case of a missing token, an empty box is drawn and its content is added to indicate what the missing token should have been, as shown in figure 4.

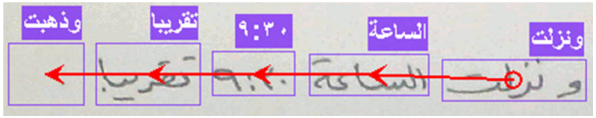

Figure 4: Missing token mark-up in GEDI tool

4.3 Quality Control

AMA has developed an extensive set of tools to provide quality control. The GEDI tool itself enforces various constraints on reading order, alignment of text with original content, and consistency of toke attributes. It also provide mechanisms for visualization of content aligned with the annotation, pseudo coloring of zones by attribute and a "listener" feature that lets outside processes automatically load and control GEDI functions. For example after all annotation is complete, the workflow clips and organizes all tokens by content, color coding the attributes so a reviewer can easily pick out errors. Clicking on individual token clips the document containing this token will automatically load into GEDI and the zone will be highlighted.

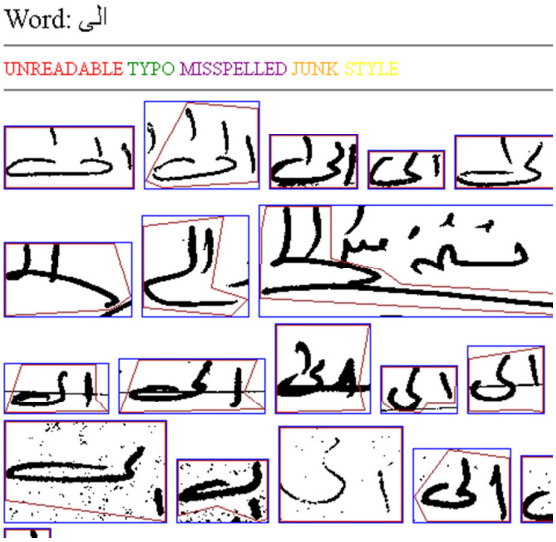

Figure 5: Quality control in GEDI tool

4.4 Challenges

In Arabic data, it is common to find mismatches in the way numbers are written (Arabic vs. Indic) which is due to each region's convention in writing numbers. This is a potential challenge to OCR. The GEDI tool supports all Unicode input and Arabic and Indic numbers as needed. Email, links and other English content can often be found amongst Arabic or Chinese data, and this poses challenges for annotating the data.

In Chinese data, a polygon bounding box is drawn around each character to achieve high annotation consistency and flexibility of alignment with both the source text and English translation. There are some minor challenges in Chinese writing. Dot on the upper right corner or right side of a character block sometimes float away from the block. It is not very common, but once it occurs, it may pose some challenge in drawing the bonding box, with the right dot be confused as 、 (a coma like punctuation).

5. Data processing for release

After ground truth annotation has taken place, files are spot-checked for accurate transcriptions. At this time a reading order is also assigned to tokens. In cases where GALE documents have been scribed in-house, this is trivial because token reading order is simply left-to-right, top-to-bottom on the page (or right-to-left in the case of Arabic). A reading order is assigned by the annotator, and this allows for sentence segmentation. The reading order, sentence segmentation, and often an English translation are not included in the GEDI XML format used during ground truth annotation but are added to the MADCAT XML delivery release format.

As the MADCAT data have many layers of annotation (ground truth, transcription, sentence segmentation, reading order, translation), a unified data format was defined by LDC early in the MADCAT program to consolidate information from GALE source text and translation text and ground truth files (Strassel 2009). This format creates a single XML output which contains multiple layers of information: a text layer for source text with word/character tokenization and sentence segmentation, an image layer with all the bounding boxes, a document metadata layer, and a translation layer. There are plans to further annotate MADCAT data for word alignment and treebank, thereby creating a single data set rich in linguistic markup.

Once the ground truth annotation is completed, LDC processes the data to generate the single MADCAT XML output. The input GEDI XML is validated to ensure the annotation tool hasn't changed during annotation by creating new valid attributes or values in the XML. The XML is further validated with one script comparing corresponding MADCAT XML and GEDI XML files to make sure that no tokens were lost in the conversion process. (The conversion process renumbers tokens and reorders them to a natural numerical order.) Another script reviews a MADCAT XML file to ensure that there is a one-to-one correspondence between token polygons and source tokens. At this point summary statistics for the data release are prepared and we are prepared to send our handwriting annotations to our data users!

6. Resulting Data

To date, more than 42,000 Arabic handwritten pages and 223,600 Chinese handwritten pages have been collected, annotated and released to the MADCAT program participants, as shown table 1.

The linguistic resources described in this paper have been distributed to MADCAT performers and sponsors. Most linguistic resources developed by LDC for MADCAT will also be published in LDC's catalog, making them generally available to the larger research community; this includes all MADCAT data based on GALE sources. The following corpus is scheduled to be added in LDC's catalog in the spring of 2012:

	Training			
	phase 1	phase2	phase3	Chinese
Genre	NW, WB	NW, WB	NW, WB	NW, WB
Number of pages	2000	5000	621	1491
tokens/ page	unconstrained	unconstrained	<=125	<=225
scribes/ page	5	up to 5	up to 7	15
Total handwritten pages	10000	27915	4540	223600
number of unique scribes	100	152	53	150

Table 1: Data profile of MADCAT training data

LDC2012TXX: MADCAT Phase 1 Arabic Handwriting Training Corpus

7. Conclusion and future work

We have described the pipeline and infrastructures that LDC adopts and builds to collect and annotate handwriting linguistic resources to support recognition and translation evaluation. We have also annotated handwriting which we did not collect. While handwriting collection is in many ways ideal because we may vary the conditions as described above, collected samples represent idealized writing. Found handwriting in documents is more natural because there may be writing skew, mix of machine type and handwritten text, and smudges or stamps. While such training data may be noisier and pose unique annotation challenges, such found data are more representative of the goal of MADCAT which seeks to recognize and translate a variety of text samples.

8. Acknowledgements

We acknowledge and appreciate the work of David Lee on 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.

9. References

DARPA. 2008. Multilingual Automatic Document Classification, Analysis and Translation (MADCAT). (http://www.darpa.mil/Our_Work/I2O/Programs/Multilingual_Automatic_Document_Classification,_Analys

is_and_Translation_%28MADCAT%29.aspx)
DARPA. 2007. Global Autonomous Language Exploitation (GALE). (http://www.darpa.mil/Our_Work/I2O/Programs/Global_Autonomous_Language_Exploitation_%28GALE%29.aspx)
Doermann, D., Zotkina, E., Li H. (2010). GEDI - A Groundtruthing Environment for Document Images. *Ninth IAPR International Workshop on Document Analysis Systems (DAS 2010)*.
Li, X., Ge, N., Grimes, S., Strassel, S. M., and Maeda, K. (2010). Enriching Word Alignment with Linguistic Tags. LREC 2010.
NIST. 2012. NIST 2012 Open Handwriting Recognition and Translation Evaluation Plan. (http://www.nist.gov/itl/iad/mig/upload/OpenHaRT2012_EvalPlan_v1-5.pdf)
Song, Z., Strassel, S., Krug, G., Maeda, K. Enhanced Infrastructure for Creation and Collection of Translation Resources. LREC 2010.
Strassel, S., Friedman, L., Ismael, S., Brandschain, L. (2008). New Resources for Document Classification, Analysis and Translation Technologies. LREC 2008.
Strassel, S. (2009). Linguistic Resources for Arabic Handwriting Recognition. Second International Conference on Arabic Language Resources and Tools.