Linguistic Resources for Arabic Handwriting Recognition

Stephanie M. Strassel

Linguistic Data Consortium
3600 Market Street, Suite 810
Philadelphia, PA 19104 USA
strassel@ldc.upenn.edu

Abstract

MADCAT (Multilingual Automatic Document Classification Analysis and Translation) is a five year DARPA program that will produce systems to automatically convert foreign language text images into English transcripts for use by humans and downstream processes including summarization and information extraction. The first two phases of MADCAT focus on handwritten Arabic. Linguistic Data Consortium (LDC) creates and distributes linguistic resources for MADCAT, including data, annotations, specifications and tools for system training and evaluation. To date LDC has recruited over 300 scribes from around the Arabic speaking world to produce handwritten text for MADCAT. A web-based collection toolkit supports scribe recruitment, registration, data assignment and tracking, progress reporting, quality control and compensation both at LDC and at remote collection sites. Handwritten pages are scanned at high resolution and manually annotated with information including bounding boxes for each line and word on the page. Corresponding digital text and English translations are generated, and the multiple data layers are unified into a single xml output file containing: a text layer consisting of source text, tokenization and sentence segmentation; an image layer consisting of bounding boxes; a scribe demographic layer consisting of scribe ID and partition (train/dev/test); and a document metadata layer. LDC has collected, annotated and distributed over 38,000 handwritten pages thus far, and collection continues at a rapid pace. Most linguistic resources developed for the program will also be published in LDC's catalog making them generally available to the larger research community; the MADCAT Phase 1 Training Corpus is expected to be published in late 2009.

Introduction

MADCAT (Multilingual Automatic Document Classification Analysis and Translation) is a five year technology evaluation program sponsored by the US Defense Advanced Research Projects Agency (DARPA). MADCAT's goal is to produce systems that can automatically convert foreign language text images into English transcripts for use by humans and downstream processes including summarization and information extraction. To meet this goal, system developers must tackle a number of challenges including the integration and optimization of disparate technologies such as page segmentation, metadata extraction, OCR and translation technologies. The ultimate goal is to create a highly accurate end-to-end system for deployment at the program's end. While baseline performance for such a system was estimated at approximately 2% at the program's outset (Olive 2007), the target performance in year 5 is 90% accuracy. The core evaluation task thus far in MADCAT has been translation of handwritten Arabic documents. Linguistic Data Consortium (LDC) is creating publicly available linguistic resources for MADCAT technologies on a scale and richness not previously available. MADCAT corpora consist of new annotations of existing data, data donations from program performers, and substantial new data collection to create high quality material for evaluation and to address strategic gaps (for genre, dialect, image quality, etc.) in existing training resources.

Data Requirements

Existing linguistic resources for MADCAT technology development are limited. Beyond data developed by MADCAT performers themselves, two prior corpora were identified as particularly relevant, and were acquired by LDC for redistribution to MADCAT program participants.

First, the AMA Arabic Dataset developed by Applied Media Analysis (AMA 2007) consists of 5000 handwritten pages, derived from a unique set of 200 Arabic documents transcribed by 49 different writers from six different origins. The handwritten collection contains various document types including forms, memos, poems, diagrams, and number lists in both English and Indic digits. The texts were produced in various utensils including pencil, thick markers, fine point pen, and ball point pens. The dataset is in binary and grayscale settings and additional settings are available. Ground truth annotations are also incorporated, at both the PAW (Partial Arabic Word) and word level. Second, LDC acquired 3000 pages of handwritten Arabic images collected by Sakhr. Sakhr's corpus consists of 15 Arabic newswire documents each transcribed by 200 unique writers. LDC added line and word level ground truth annotations to each handwritten image, and distributed these along with English translations for each document to MADCAT performers.

Beyond existing corpora, MADCAT performers requested additional new training data totaling at least 10,000 handwritten pages in the first year and 20,000 pages in the second year of the program, plus ground truth annotations for each page. Although the primary evaluation metric for MADCAT measures machine translation performance, the program is not funding research in MT and MADCAT performers are expecting to embed existing MT technology into their MADCAT systems. Given this constraint, it is important for training and evaluation data, at least in early phases of the program, to remain consistent with the types of data used to train the teams' MT systems. The primary MADCAT performer BBN is also part of the DARPA GALE Program (Strassel 2006), and so the MT system used by BBN for MADCAT is the one developed under GALE. This scenario heavily influenced the approach to training and evaluation data creation for the first phase of MADCAT and a decision

was made to take advantage of existing GALE infrastructure and linguistic resources. GALE data characteristics are well understood, and cost and time factors for handling this data are reasonably well known. Annotation costs are controlled since translations already exist. More importantly, utilizing GALE data eliminates a potential domain mismatch between BBN's GALE MT models and MADCAT test sets, and it provides controlled test sets for evaluation. For all these reasons, source data from two GALE domains, one formal (newswire) and one informal (weblogs and newsgroups), was targeted as the primary MADCAT Phase 1 training, devtest and evaluation corpora, with minimum data requirements specified as follows:

Phase 1	Training	DevTest	Eval
Genre	Newswire & Web Text	Newswire & Web Text	Newswire & Web Text
Origin of source docs	GALE Parallel Text Training Corpora	GALE Phase 1-2 Eval Data	GALE Phase 3 Eval Data
Number of pages	2000	320	160
Arabic tokens/page	unconstrained	125	125
Scribes/page	5	2	2
Total handwritten pages	10000	640	320
Number of unique scribes	100	50	24
Scribe exposure	all exposed	half unexposed	10 unexposed

Figure 1: Phase 1 Data Requirements

Handwriting Collection

Overview

High quality English translations already exist for the GALE documents used in MADCAT, but the Arabic source data is electronic text, not handwritten. Therefore the majority of LDC's MADCAT effort focuses on acquisition of handwritten versions of the GALE data, and the annotation and processing of that data. To acquire handwritten versions of the GALE documents, LDC has undertaken a new human subject collection. Native, literate Arabic speakers from a variety of geographic and demographic backgrounds have been recruited to produce and donate writing samples in person and via a web application created and maintained by LDC for this project. Each potential participant is first screened for literacy and their ability to perform the writing task. After screening, new scribes register for the study via LDC's website, providing salient demographic information (geographic origin, education, right or left-handedness). To date, nearly 400 scribes have participated in the collection, with more being added daily. Scribes are compensated for their participation based on the number of handwritten pages they produce.

Collection Infrastructure

To support supplemental handwriting collection by partner sites around the world and ensure consistency and efficiency in our local collection, LDC has developed a web application to control all aspects of the collection process from scribe registration and assignment tracking to quality control and compensation. The entire web application is password protected and every user is given a unique account which allows LDC to track every action made through the website. Users can be assigned different roles to restrain their level of access to different functions in the system, and a user may only view and manage scribes that have been registered at that coordinator's location. For instance, a remote collection coordinator working in Morocco would only be allowed to view/manage scribes that were registered at that site in Morocco.

Upon enrollment, the web application assigns each scribe a unique subject ID for the duration of the collection and assigns them to a data partition (exposed/training versus unexposed/test). Prior to assignment to scribes, documents are assembled into "kits", which consist of a set of source documents to be handwritten along with the writing conditions required for each document. Writing conditions for the collection as a whole are established as follows: Implement: 90% ballpoint pen, 10% pencil; Paper: 75% unlined white paper, 25% lined paper; Writing speed: 90% normal, 5% fast, 5% careful. Scribes are provided with detailed guidelines describing requirements for the writing task, with examples illustrating each rule and condition.

Data Pre-Processing

Prior to assignment, documents are processed to optimize their appearance for the handwriting task. Starting with a full document from the GALE newswire or web collection, files are first manually segmented into sentences. Those sentence segments may be re-ordered and additional formatting (line breaks and spacing) added to create optimal pages for handwriting assignment. To avoid word and line wrapping problems, we target roughly 5 words per line and no more than 25 lines per page. This conservative standard ensures that different scribes writing the same page will all have the same line and page breaks. Additional processing is required to handle issues like lines containing bi-directional text (for instance, an email address embedded in a line of Arabic text). A document header and line numbers are added to help scribes keep track of their assignment.

AAW_ARB_20070102.0045_1

لا تنسخ أي شيء مكتوب فوق الخط

<1> فيما عقدت مجموعة من حركة
<2> العدل والمساواة المسلحة المعارضة في
<3> دارفور مؤتمرا في العاصمة الإثيوبية
<4> أديس أبابا وناقشت خلاله العديد
<5> من القضايا واختارت خلاله ادريس
<6> آدم ازرق رئيسا جديداً للحركة ،
<7> اعتبرت حركة العدل والمساواة التي
<8> يتزعمها خليل ابراهيم ان اعلان
<9> ميلاد حركة باسمها وعزل رئيسها
<10> « تم بتخطيط وتمويل من
<11> جهاز الامن والمخابرات السوداني » ،
<12> واصفة المجموعة الجديدة بان لا
<13> قيمة لها ونفت وجود أي
<14> انقسام داخلها .
<15> وقال مصدر ديبلوماسي أفريقي مطلع في
<16> اديس ابابا ، ان هذا المؤتمر
<17> للمجموعة المناوئة لرئيس الحركة خليل
<18> إبراهيم يعتبر انقلابا عليه .

Figure 2: Processed document for assignment

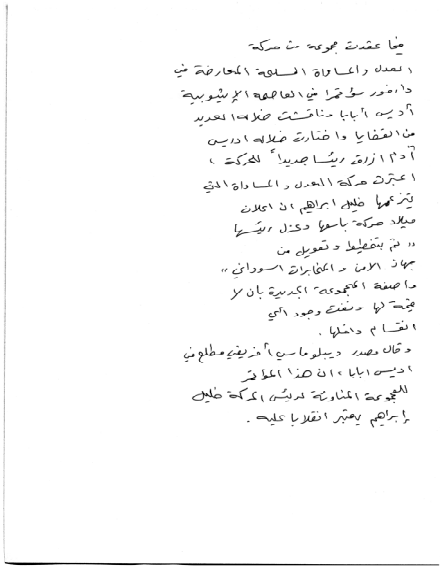

Figure 3: Handwritten version of document

Each scribe is assigned between one and ten kits at a time, and assignment information, due dates, writing conditions and other relevant data is tracked via the web application. Collection managers can use the web application to assign and download new kits, review/update previously assigned kits, view the scribe's online history, and view the scribe's registration information. The application can also be used to track collection progress and report statistics about various aspects of the collection.

Quality Control

Typically scribes are not assigned additional kits until their current assignment has been validated. LDC staff review each collected kit for completeness and accuracy on a number of conditions including consistency of line breaks, accuracy of writing conditions and presence of typographical errors or missing text. Depending on the severity of any errors discovered, scribes may be asked to redo portions of a kit, they may receive a compensation penalty or they may be removed from the study entirely.

Ground Truthing and Annotation

Once handwritten data has been collected from scribes, it is checked for quality and completeness, then each page is scanned at a high resolution (600 dpi, greyscale) to create a digital version of the handwritten document. Data is then annotated with several important features to create ground truth references that can be used to model desired MADCAT system output. The current ground truthing task consists of drawing a bounding box drawn around each line, word or other targeted element on the handwritten page. Each bounding box, or zone, consists of a unique ID, the contents and coordinates indicating the location of the zone on the page. Explicit reading order is preserved by applying a nextZoneID tag. Finally, scribe mistakes are noted by assigning status tags to zones as required -- for instance, denoting a particular word or region on the page as extra, missing or typo. Ground truthing is performed using the GEDI (GroundTruth - Editor and Document Interface) tool created by Applied Media Analysis (AMA) and modified to support MADCAT.

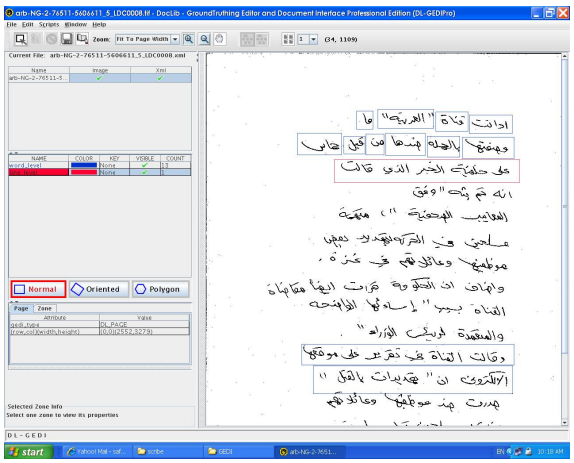

Figure 4: GEDI Annotation Toolkit

A subset of collected handwritten, annotated data is additionally labeled for a set of features designed to support error analysis on MADCAT system performance. For each line of handwritten text, a binary decision (present/absent) is made regarding the presence of the various features. Such characteristics are not necessarily mistakes or errors on the part of scribes, but are instead expected features of natural handwriting. Some features are quite frequent in the collected data; in the Phase 1 pilot evaluation corpus up to 39% of pages were judged to have at least one line of text with poor legibility, and up to 25% of pages contained some overlapping lines/words.

Overlapping word or line boundaries	A part of one word bleeds into another due to writer's style or inadequate spacing
Poor legibility	Some characters or even whole words are not clearly written; careful consideration of details and/or context is required for disambiguation
Presence of non-Arabic characters	The presence of English, Roman numerals, etc. in Arabic images
Ruled lines	Text is written on pages with ruled lines
Short text lines	Only one or two words are on a line
Skew	The "baseline" for the text line (or even the word) is not a horizontal line with zero slope
Slant	The words or characters "lean" to the right or left

Figure 5: Handwriting Attribute Categories

Data Format

A major technical challenge in producing linguistic resources for MADCAT is the logical storage of many layers of information across multiple versions of the same data. To tackle this problem, LDC defined a unified data format early in the MADCAT program, along with supporting software that takes multiple data streams and generates single xml output file which contains all required information. The xml file contains distinct components: a text layer that consists of the source text, tokenization and sentence segmentation; an image layer that consist of bounding boxes; a scribe demographic layer that consists of scribe ID and partition (train/test); and a document metadata layer.

Evaluation

A pilot evaluation for the first phase of MADCAT was held in late 2008. The primary evaluation task is document image translation, where systems take Arabic handwritten images as input, and output translated English text. For the pilot evaluation systems were also given reference segmentation (bounding boxes for each line of text); this information is not provided in future evaluations. The primary metric for measuring document image translation performance is HTER, or human translation edit rate (Snover 2006). HTER calculates the minimum number of changes required for highly-trained human editors to correct MT output so that it has the same meaning as the reference translation, through a process called post-editing (NIST 2007). While there is no separate evaluation of OCR or processing components, in addition to the primary evaluation task, there are two secondary component evaluations. Document image recognition assesses the ability of MADCAT systems to accurately transcribe Arabic text from a segmented Arabic image. Performance is measured by Word Error Rate (WER). Document text translation measures the ability of systems to translate Arabic text into English text, performance is measured by Translation Edit Rate (TER).

MADCAT Phase 1 "go/no-go" evaluation targets for the primary task were set at no less than 40% accuracy for at least 70% of the documents, as measured by TER. By the program's end systems are required to achieve at least 95% accuracy for 90% of the documents. MADCAT system performance in the Phase 1 pilot evaluation met the evaluation targets, and the second phase evaluation is scheduled for early 2010 (NIST 2008).

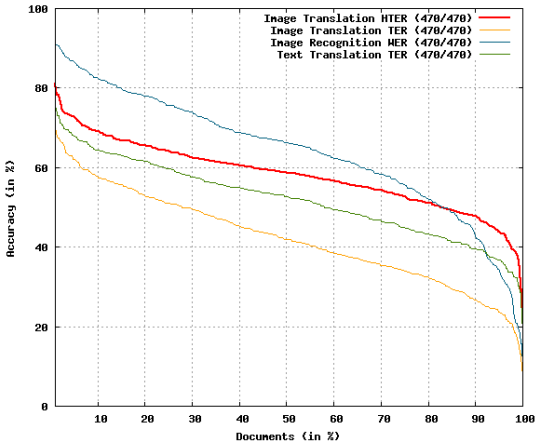

Figure 6: Phase 1 Pilot Evaluation Results

Future Plans

The first two phases of data collection have concentrated on acquisition of handwritten versions of printed Arabic newswire and web text. As the program continues new domains and challenge scenarios will be added to support ongoing system development and to better approximate real world scenarios. LDC is currently in the planning stages for a new handwriting collection that will focus on spontaneously produced materials, including journal entries, lists, ledgers, letters, memos, and instructions written on paper, chalkboards and white boards. To satisfy the need for mixed printed and handwritten material, scribes will also fill out postcards and printed forms, hand-label maps and add marginalia commenting on printed materials.

Data Distribution

LDC has collected, annotated and distributed over 38,000 handwritten pages for MADCAT thus far, and collection continues at a rapid pace. Annotated corpora are delivered to MADCAT performers on a rolling basis to meet program milestones. 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 first general MADCAT publication is the Phase 1 Training corpus, expected to appear in LDC's catalog during the fourth quarter of 2009. This release comprises 9693 collected handwritten pages and will include all elements released to MADCAT performers: original Arabic source documents (tokenized and sentence-segmented); high-resolution scanned images of the corresponding handwritten page for each document; ground truth annotations for each handwritten page; and English translations for each document.

Acknowledgements

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.

References

Applied Media Analysis (2007). Arabic-Handwritten-1.0. http://appliedmediaanalysis.com/Datasets.htm.

National Institute of Standards and Technology (NIST) (2007). Post Editing Guidelines Version 3.0.2 http://www.nist.gov/speech/tests/gale/2007/doc/GALEpostedit_guidelines-3.0.2.pdf

National Institute of Standards and Technology (NIST) (2008). MADCAT Phase 1 Evaluation Report. http://www.itl.nist.gov/iad/mig//tests/madcat/2008/pilot/MADCAT_Reports_20081114-1239/index.htm

Olive, Joseph (2007). Multilingual Automatic Document Classification Analysis and Translation (MADCAT) SOL BAA 07-38 Proposer Information Pamplet, DARPA/IPTO.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul (2006). "A Study of Translation Edit Rate with Targeted Human Annotation," Proceedings of Association for Machine Translation in the Americas.

Stephanie Strassel, Christopher Cieri, Andrew Cole, Denise DiPersio, Mark Liberman, Xiaoyi Ma, Mohamed Maamouri, Kazuaki Maeda (2006). Integrated Linguistic Resources for Language Exploitation Technologies. Proceedings of LREC 2006: Fifth International Conference on Language Resources and Evaluation.