New Resources for Document Classification, Analysis and Translation Technologies

Stephanie Strassel, Lauren Friedman, Safa Ismael, Linda Brandschain

Linguistic Data Consortium

3600 Market Street, Suite 810

Philadelphia, PA 19104 USA

{strassel, lf, safa, brndschn}@ldc.upenn.edu

Abstract

The goal of the DARPA MADCAT (Multilingual Automatic Document Classification Analysis and Translation) Program is to automatically convert foreign language text images into English transcripts, for use by humans and downstream applications. The first phase the program focuses on 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. Corpora will consist of existing LDC corpora and data donations from MADCAT partners, plus new data collection to provide high quality material for evaluation and to address strategic gaps (for genre, dialect, image quality, etc.) in the existing resources. Training and test data properties will expand over time to encompass a wide range of topics and genres: letters, diaries, training manuals, brochures, signs, ledgers, memos, instructions, postcards and forms among others. Data will be ground truthed, with line, word and token segmentation and zoning, and translations and word alignments will be produced for a subset. Evaluation data will be carefully selected from the available data pools and high quality references will be produced, which can be used to compare MADCAT system performance against the human-produced gold standard.

.

1. Introduction

The goal of the DARPA MADCAT (Multilingual Automatic Document Classification Analysis and Translation) Program is to automatically convert foreign language text images into English transcripts for use by humans and downstream engines like Distillation. In the first phase, the program will focus on handwritten Arabic documents. In response to these technology challenges, Linguistic Data Consortium (LDC) is undertaking an integrated set of activities to create publicly available language resources on a scale and richness not currently available. Source data will include images of handwritten and mixed Arabic text from all over the Arabic speaking world, emphasizing source and data type variety including: maps, training manuals, brochures, graffiti, signs, letters, memos, instructions, postcards, forms, with and without marginalia. LDC and its principle partner, Applied Media Analysis [1] (AMA), will create ground truth annotation, zone analysis, segmentation and transcription/digitization of a large subset of the collected data, and translation of a large subset of the annotated data, using LDC and AMA staff as well as an extensive network of skilled vendors from around the world. LDC will also create evaluation resources for MADCAT, including high quality gold standard translations and post-editing of translation system output.

LDC is working with program participants, sponsors and NIST to create and implement a data plan for MADCAT that achieves the desired balance of volume, variety and complexity while remaining responsive, flexible and targeted to the goals of the program. This paper reports on our current and planned MADCAT data activities.

2. Data Roadmap

Ultimately, MADCAT systems will be required to process images that span a wide variety of challenging conditions. LDC has developed a data roadmap defining four salient features for MADCAT data and comparing current and required technology capabilities for handling each condition.

First, ***genre*** refers to the structural features of a given data type. State of the art translation technology from the DARPA GALE Program handles four genres: structured text (Newswire), structured audio (broadcast news), unstructured text (weblogs and newsgroups), and unstructured audio (talk shows). MADCAT adds a large number of genres, including how-to manuals, letters, signs, memos, postcards, forms, diaries and ledgers, just to name a few. These genres will be added gradually to the data pool over the life of the program, starting with the genres (e.g. letters, diaries) that are structurally closest to existing GALE genres.

Second, ***topic*** refers to the subject matter of the data. GALE topics target primarily news, current events and commentary, although the unstructured text genres (weblogs and newsgroups) provide supplemental data on an almost unlimited set of topics. MADCAT expands the topic focus to include scientific topics (e.g. civil engineering plans), topics of military interest (e.g. handwritten manuals for repairing weapons), and highly personal materials appearing in diaries and letters. Again, we will phase in new topic areas over time.

Third, ***medium*** refers to the technique or method of communication in the source document, and can be

[1] http://appliedmediaanalysis.com

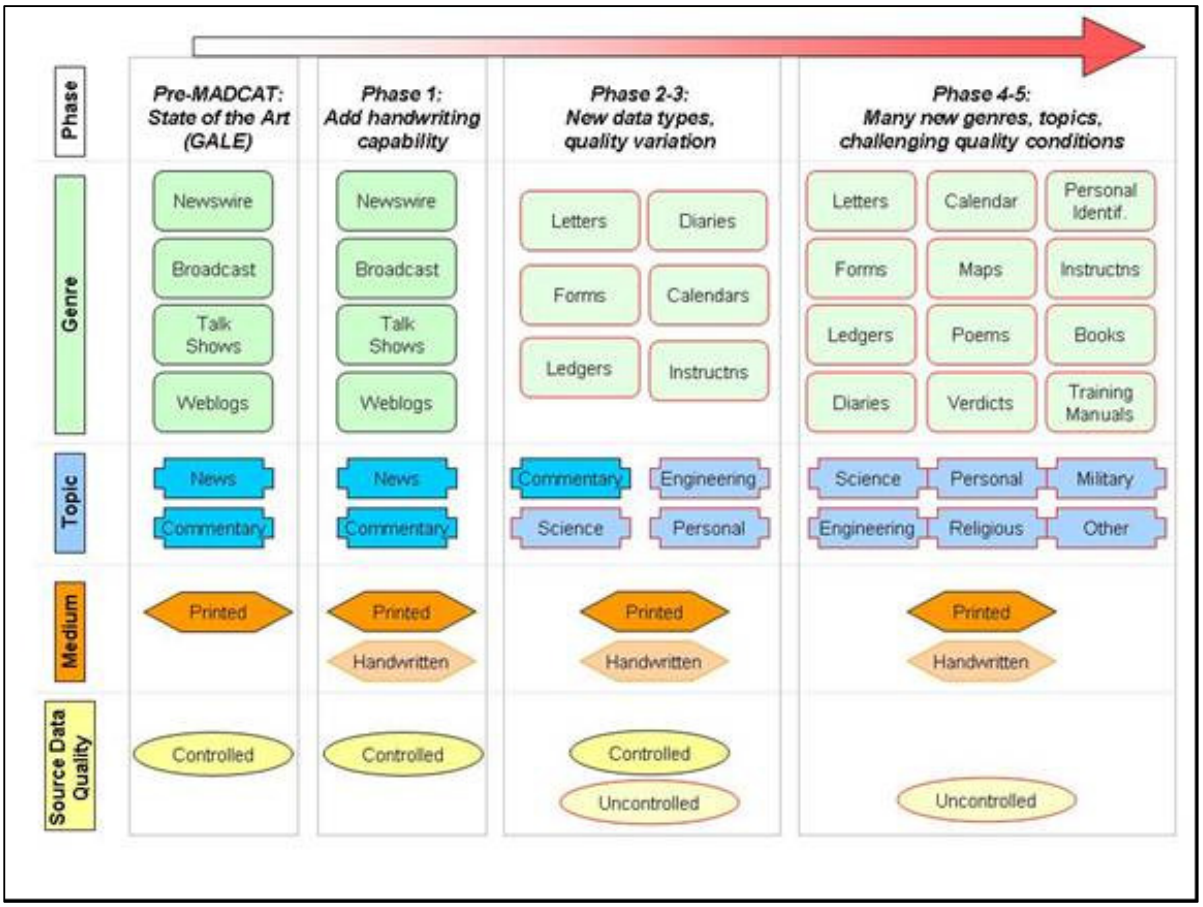

Figure 1: MADCAT Data Roadmap

technology goal is a dramatic improvement in the performance of Arabic handwriting recognition systems, so Phase 1 of MADCAT adds handwritten data as a primary condition.

Finally, ***source data quality*** is used here to refer to the conditions under which the source data is collected, and how much variation can be expected to exist in the quality of the data. Data collected under carefully controlled conditions (good handwriting with no overlapping lines, scanned at high resolution, limited "noise" like image artifacts from a copier or dirt on the page) is obviously easier to process than found data, which may be collected and processed under extreme conditions (e.g. in the military theater) where quality is difficult to control. Early phases of MADCAT will include more carefully controlled source data conditions, but over time more noise will be introduced to push the technology toward the goal of being able to perform on real world data.

Figure 1 depicts the overall data roadmap for the MADCAT Program.

3. Collection

LDC has adopted multiple strategies to address the program's training data needs. A large portion of the targeted data will come from existing corpora of interest contributed by sponsoring agencies and affiliated researchers. Related technology programs like DARPA GALE (Global Autonomous Language Exploitation) will also contribute data, particularly in the first phase of the program (Strassel et. al. 2006). The Internet is also a valuable source of Arabic data of interest to MADCAT, in particular images of graffiti and signage. The recent explosion of photo sharing sites and photoblogs provides a treasure trove of data, much of it freely available under Creative Commons[2] or similar licenses.

To supplement existing, donated and found resources, some amount of controlled data collection is also necessary, for instance to provide data for evaluation and devtest sets that is of known quality, and to address strategic gaps (for genre, dialect, image quality, etc.) in the training corpus. A targeted human subject collection is currently underway, recruiting native, literate Arabic speakers from a variety of geographic and demographic backgrounds to produce and donate writing samples in person and via a website that we create and maintain. In the initial stages, scribes will produce handwritten versions of training data from GALE newswire and web parallel text collections. Because high-quality translations and keyboarded (digital) copies of the documents already exist, the collection can be streamlined to reduce startup costs. Scribes will be recruited in the US and from Arabic speaking countries around the world in partnership with LDC's network of vendors and other collaborators. Scribes will provide salient demographic information (geographic origin, education, right- or left-handedness)

[2] creativecommons.org

and will be assigned a unique subject ID for the duration of the collection. Scribes will further be assigned to training or evaluation data sets. GALE documents will be assembled into "kits", which consist of the set of source documents to be handwritten along with the writing conditions required for each document. Writing conditions vary in terms of paper (lined or unlined, border or no border), writing instrument (pen, marker, pencil) and writing speed (hurried, natural, neat). Each scribe is assigned one or more kits for completion. In later stages of the collection, scribes will produce writing samples in a wider variety of genres and writing conditions, for instance producing journal entries, letters, memos, instructions written on paper, chalkboards and white boards and lists. 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.

4. Annotation: Ground Truthing and Translation

Collected data must be annotated with several important features to create ground truth references that can be used to model desired MADCAT system output. In order to provide training data in a format that can be readily annotated with additional features and easily incorporated by MADCAT technology developers, we will first create a digitized version of every image, PDF file or paper document that has been collected. Because the challenges of Arabic handwriting and the quality of image files targeted by MADCAT will confound available OCR technology to the point that no usable output file can be generated, most data will be digitized through manual keyboarding by human editors. Similarly, automatic performance on character segmentation and zone analysis of handwritten Arabic documents has not reached a level of maturity that would allow automated solutions to be easily incorporated into an annotation pipeline, so we must largely rely on human annotators to perform and correct document region classification. LDC and AMA have developed annotation guidelines and adapted existing tools to ground truth data at the document, page, zone, line, word and the PAW (part of Arabic word) level.

One of the key shortcomings of the existing image processing community is a lack of a uniform standard for document representation. Although several have been proposed (OASIS Open Document, Dublin Core Metadata Initiative, hOCR) they have not provided the optimal combination of completeness and simplicity, and thus have not been widely accepted. A new data format definition has been developed for MADCAT, with AMA providing critical expertise on imaging, ground truthing and document descriptions and representations, LDC contributing insights into document metadata and annotation format considerations, and NIST providing perspective on evaluation requirements.

While all collected materials will be digitized, transcribed, segmented and zoned, MADCAT sites also require some data to be translated into English. Because manual translation of the entire collection would be prohibitively expensive, LDC employs semiautomatic methods to identify a subset of the data pool for manual translation. Translated material is selected to be representative of the variables important to MADCAT: data type, medium, geographical region of origin and quality of the original, to provide a representative, heterogeneous pool of richly annotated training data. Initial pools of training data are automatically sorted into these categories and human annotators then verify each document as being suitable for translation.

For each selected document, LDC produces a high-quality English reference translation, using the training data translation process developed for the DARPA TIDES program, modified for GALE, and further refined to support MADCAT (LDC 2006). In addition to manual creation of translations for selected data, LDC pursues several other methods to obtain translations of collected material. First, wherever possible we harvest existing translations from the web, with LDC software like Bilingual Internet Text Search (BITS) and Champollion alignment tools providing a starting point for this task (Ma & Liberman 1999, Ma 2006). We also exploit the fact that some collected data already has English translations, captions or summaries as part of the source document. For instance, images of Arabic signs and graffiti on photo sharing websites like Flickr often provide English translations as part of the comment stream associated with the image.

5. Evaluation Resources

A subset of the regular MADCAT data collection will be earmarked to serve as the evaluation test set. This data will be carefully selected to meet the requirements of the MADCAT program. Documents will be digitized, transcribed and processed to produce a translator-friendly format, substantially similar to the one used by LDC in GALE, NIST Open MT evaluations and related programs. Once evaluation documents have been selected and processed using the techniques described above, "raw" translations are created by professional translators under contract to LDC. We then execute a multi-stage process to produce a single gold standard reference translation for each document:

1. First pass quality control (QC) at LDC will be performed by Arabic-dominant bilingual junior annotators. First pass QC reviewers will focus on correcting obvious errors in the translations and flagging any indications of problems with upstream processes.
2. Second pass QC at LDC will be performed by Arabic-dominant fluent bilingual senior or lead annotators. Second pass QC reviewers will focus on improving fluency and correcting subtle translation mistakes. QC reviewers at this stage will have access to scans of the original documents so that translations can be verified against the source when necessary.
3. Third pass QC at LDC will be performed by English-dominant fluent bilingual annotators. Third

pass QC reviewers will focus on fine-tuning fluency, standardizing proper nouns, and adding translation variants where necessary.
4. Fourth pass QC at LDC will be performed by monolingual English senior annotators. Fourth pass QC reviewers will check for understandability and fluency, expected use of translation alternatives, and correct spelling of proper nouns and technical terms. QC reviewers at this stage will also flag formatting problems, conflicts with the translation guidelines, inconsistencies, and questionable regions. Any problems discovered at this phase will be resolved with input from Arabic linguists.
5. Corpus-wide scans will be performed by LDC's programming team to standardize and validate data format and identify any lingering errors.
6. Arabic team leaders will perform final spot checks on a subset of the data, verifying all versions of a file to ensure that all problems have been resolved.

After the evaluation, LDC will manage post-editing of MADCAT translation system output. Post-editing refers to the task of modifying MT output such that the resulting text completely captures the meaning of the gold standard reference translation (Przybocki et. al. 2006). Post-editing requires editors to achieve this end while making the minimum number of edits to the MT output. The resulting edited text is used to measure the edit distance between a system output and the gold standard reference translation.

6. Document Management

MADCAT data tracking system is under development to manage all of the data collected, processed, annotated and distributed for the program. Each document in the collection is assigned a unique document ID, and derivative files incorporate the same file stem to provide straightforward reinforcement of the connection between the collected document and its associated files. The resulting MADCAT database tracks information such as document ID; language; data type (map, graffiti, etc.); data partition (training, devtest, eval); data medium (print, handwriting, mixed); author information; acquired date; acquired location; source type (paper, electronic); offsite storage location ID; orthographic information; file format (jpeg, tiff, etc.); file size; IPR status; and other fields as required.

7. Resource Integration and Coordination

The effort involved in coordinating linguistic resources for a program like MADCAT is partly administrative, partly consultative, and partly technical. The coordinator role requires ongoing interaction with collection and annotation project leaders to make sure that all deliverables are accounted for; standards for data selection, annotation, quality control and data formatting are normalized across tasks wherever possible; both common and task-specific QC and formatting issues are fully documented. The role also requires ongoing interaction with other MADCAT participants and sponsors, to make sure that collection and annotation standards and guidelines are consistent with (supportive of) research goals; delivery schedules and data formatting standards are appropriate to program requirements; data distribution is as efficient as possible; quality control processes are updated and maintained in a manner that is fully responsive to all members of the research community. LDC will maintain a central point of reference that can serve sponsors, contractors and researchers with consistent, up-to-date and comprehensive information about targets, status, standards and issues involving the various other tasks for collection, creation and release of linguistic resources for MADCAT.

8. Resource Distribution

LDC will use techniques developed in programs like TIDES and refined for GALE to provide timely deliveries of new resources to program participants. As the linguistic resources described above are distributed to the program, LDC will wherever possible distribute the data more broadly, for example to our members and licensees, through the usual mechanisms. Standards, tools, software and best practices developed by LDC will also be made freely available to the research community under an "open source" model for use in education, research and technology development.

In addition to serving the primary goal of improved performance for MADCAT engines, our efforts will lead to substantial corpora with durable value to the worldwide Human Language Technology community and the technology users who benefit from HLT. .

9. 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.

10. References

Linguistic Data Consortium (2006.) GALE Arabic Translation Guidelines V2.3 http://projects.ldc.upenn.edu/gale/Translation/specs/GALE_Arabic_translation_guidelines_v2.3.pdf

Ma, X. and Liberman, M. (1999.) BITS: A Method for Bilingual Text Search over the Web. Machine Translation Summit VII, September 13th, 1999, Kent Ridge Digital Labs, National University of Singapore.

Ma, X. (2006.) Champollion: A Robust Parallel Text Sentence Aligner. LREC 2006: Fifth International Conference on Language Resources and Evaluation

Przybocki, M., Sanders, G., Le, A. (2006.) Edit Distance: A Metric for Machine Translation. LREC 2006: Fifth International Conference on Language Resources and Evaluation

Strassel, S., Cieri, C., Cole, A., DiPersio, D., Liberman, M., Ma, X., Maamouri, M., Maeda K. (2006.) Integrated Linguistic Resources for Language Exploitation Technologies. LREC 2006: Fifth International Conference on Language Resources and Evaluation