

Multilingual Automatic Document Classification Analysis and Translation Phase 2 (MADCAT-P2) Evaluation Plan

1 Introduction

The Multilingual Automatic Document Classification Analysis and Translation (MADCAT) program is a five-year DARPA research program whose purpose is to explore and develop technologies that convert non-English language document images into English transcripts so that the information can be readily used [1]. The National Institute of Standards and Technology (NIST) conducted the first evaluation of MADCAT in 2008 (Phase 1) to measure the performance of these developed technologies. The Phase 2 (P2) evaluation is a continuation of that work.

P2 consists of two evaluation tracks: a Go/No-Go and a challenge track. The Go/No-Go track is similar to what was done in Phase 1 (P1), and focuses on monitoring progress using controlled data sets. The challenge track, added in P2, focuses on system performance using real-life data. While the data used for training, development, and evaluation of the Go/No-Go and challenge tracks are separate, both tracks have the same evaluation tasks in P2. In addition the system's ability to segment data plays a more significant role in P2. The rest of this document describes the evaluation protocols employed in the second phase of the MADCAT program.

2 Evaluation Tasks

The technologies that the MADCAT program seeks to develop are multidisciplinary. These technologies include optical character recognition (OCR), machine translation (MT), and document structure extraction. The three evaluation tasks defined in P1 remain unchanged for P2. Each task was designed to measure various aspects within the system. The goal of the evaluation is to assess system performance and to understand the strengths and weaknesses of the system. The three tasks are described in detail in the subsections below and are summarized in Table 1.

2.1 Document Image Translation

Document image translation is the primary evaluation task. It measures the overall performance of the MADCAT system in translating foreign language document images into accurate and fluent English documents. This task is to be run in both image segmentation conditions: (1) when line segmentation is given and (2) when word segmentation is given. However, the line segmentation condition is the primary condition which will be post-edited.

2.2 Document Image Recognition

Document image recognition is a contrastive task measuring the text recognition (OCR) component of the MADCAT system. This task is to be run in both image segmentation conditions.

2.3 Document Text Translation

Document text translation is a contrastive task measuring the translation (MT) component of the MADCAT system. Image segmentation condition is not applicable in this task.

Table 1: Evaluation tasks for MADCAT P2

Task	Primary	Input	Output	Metric
Document image translation	Yes	Arabic document images <ul style="list-style-type: none">• with line segmentation	Segmented English translation	HTER
	<i>No</i>	<ul style="list-style-type: none">• with word segmentation	Segmented English translation	TER
Document image recognition	<i>No</i>	Arabic document images <ul style="list-style-type: none">• with line segmentation• with word segmentation	Segmented Arabic transcription	WER
Document text translation	<i>No</i>	Segmented manual transcription of Arabic document images	Segmented English translation	TER

3 Data

3.1 Data for Go/No-Go Evaluation Track

The data used in the P2 Go/No-Go track is again drawn from the data used in the Global Autonomous Language Exploitation (GALE) program. Utilizing the same data between the two programs eliminates the domain mismatch and allows the incorporation of MT models developed under GALE for MADCAT use. See [2] for details regarding GALE data suitability for MADCAT use.

Two GALE data genres – newswire and web text – are used in MADCAT as formal text and informal text, respectively.

3.1.1 Data Creation for MADCAT

Literate, native Arabic writers are recruited by the Linguistic Data Consortium (LDC) to act as scribes. These scribes are instructed to produce handwritten copies of the chosen GALE passages using various writing implements. Table 2 lists the target distribution of the various writing implements in the MADCAT P2 data.

Table 2: Target distribution of various writing factors for the data used in MADCAT

Writing Instrument	Writing Surface	Writing Speed
90% ballpoint pen	75% unlined white paper	90% normal
10% pencil	25% lined paper	5% fast
		5% careful

3.1.2 Segmentation Creation

Two levels of image segmentation are created for the MADCAT P2 data to test the system's ability to partition the data in a meaningful way. Image word segmentation is the first layer and is created manually. A human annotator takes a given document image along with its corresponding printed transcript and marks the boundary for each word in the image.

Image line segmentation is the second layer and is derived algorithmically from the word segmentations. The algorithm used creates polygons that minimize bleeds between the lines [3].

3.1.3 Data Sets

3.1.3.1 Evaluation Data Set

The evaluation data will be chosen from the GALE Phase 3 (P3) evaluation candidate epoch, selecting passages that do not overlap with those used for the MADCAT P1 pilot evaluation. To limit test set variations (difficulty), the passages will be chosen such that their TER score distribution matches the TER score distribution of P1 pilot¹.

Each passage will be copied by three scribes. The passages are assigned to the scribes in a pre-determined order to maximize scribe overlap to study scribe handwriting effect [4].

Table 3: Target size for MADCAT P1 evaluation data set

Genre	Newswire	Web Text
Source	GALE P3 Evaluation Epoch	GALE P3 Evaluation Epoch
Number of passages	80	80
Arabic tokens per passage	125	125
Number of scribe per passage	3	3
Scribe distribution	50% exposed, 50% unexposed	
Total Number of Scribes	24 (12 in training, 12 new)	

3.1.3.2 Development Data Set

The MADCAT P2 development set will come from the GALE P2 development data set. The only selection requirement is that the chosen passage met the required number of Arabic tokens per passage. Table 4 lists the target size for the development set.

¹ TER scores will be computed based on the MT output and the first-pass reference translations for the two sets using the production MT engine ACM.

Table 4: Target size for MADCAT development data set

Genre	Newswire	Web Text
Source	GALE P2 Development Set	GALE P2 Development Set
Number of passages	100	100
Arabic tokens per passage	125	125
Number of scribe per passage	3	3
Number of unique scribes	50% exposed, 50% unexposed	

3.1.3.3 Training Data Set

The MADCAT P2 training data will come from a subset of the GALE P1 and P4 parallel text training data releases. It contains a mix of newswire and web text with no more than five scribes per passage.

Table 5 lists the target size for the training set.

Table 5: Target size for MADCAT training data set

Genre	Newswire / Web Text
Source	GALE P1 & P4 parallel text
Number of passages	4000
Arabic tokens per passage	125
Number of scribe per passage	Max 5
Scribe distribution	50% exposed, 50% unexposed

3.1.3.4 Progress Data Set

To track progress, a portion of the P1 pilot evaluation data set was sequestered to be included as part of the P2 evaluation data set. The handling of this data follows the same rules and restrictions as the evaluation data set. See section 4 for the complete list governing the handling of evaluation data. The sequestered data was made at the passage level and selected such that its HTER distribution matched the HTER distribution of the P1 pilot evaluation data.

3.1.3.5 Control Data Set

To track inter-editor agreement, a portion (10%) of the system output for the P1 pilot evaluation data set was selected to be included for repost-editing. The selection was made at the passage level and selected such that its HTER distribution matched the HTER distribution of the P1 pilot evaluation data set using ranking statistics.

3.2 Data for Challenge Evaluation Track

The details of the MADCAT P2 challenge evaluation track *{task(s), data, and schedule}* are currently being planned. This section of the evaluation plan will be updated and all appropriate parties will be notified.

4 Evaluation Rules and Restrictions

The following list of rules and restrictions must be observed during the MADCAT P2 evaluation:

- Language model adaptation across pages is not allowed.
- Interaction with the evaluation test data before submission of system results is not allowed. This includes both human interaction and automatic probing of the data.
- Passages in P1 pilot evaluation that were identified as sequestered are not to be used in training and/or development. Manual or automatic interaction with this data is also not allowed.

5 Data File Format

All data created for MADCAT use an XML format that defines storage elements which capture the various annotation layers in a document image. The format is defined in version v4h2 of the MADCAT Format Specifications document [5].

5.1 Reference Data

Each reference file contains two main layers of information. The first layer contains the word and line level segmentation for the document image, and the second layer contains the transcription and translation of the text in the image. See section 3 of [5]. The reference files are identified with the extension “.madcat.xml”.

For example: <BASE>.madcat.xml

5.2 Input Data

Each input file is derived from the corresponding reference file. Depending on the evaluation tasks, certain information is removed from the reference file. For the document image translation and document image transcription tasks, information from the translation and transcription layers is removed. For the document text translation task, information from the translation layer is removed. If a task excludes some segmentation information, the corresponding segmentation sub-layer is also removed. Table 6 summarizes the information content in the input for each task-condition pairing.

Table 6: Information content

Task	Condition	Annotation Layer Removed	File Extension
Document image translation	Line segmentation	<ul style="list-style-type: none">• transcription• translation• word-level segmentation	<BASE>.lineseg.madcat.xml
	Word segmentation	<ul style="list-style-type: none">• transcription• translation	<BASE>.wordseg.madcat.xml

Document image recognition	Line segmentation	<ul style="list-style-type: none"> • transcription • translation • word-level segmentation 	<BASE>.lineseg.madcat.xml
	Word segmentation	<ul style="list-style-type: none"> • transcription • translation 	<BASE>.wordseg.madcat.xml
Document text translation	N/A	<ul style="list-style-type: none"> • translation 	<BASE>.textseg.madcat.xml

5.3 Output Data

For the document image translation and document image transcription tasks, the MADCAT system is to output the transcription and translation information. These output files are to be identified with the “. [word|line]sys.madcat.xml” extension.

For example: <BASENAME>.wordsys.madcat.xml

For the document text translation task, the MADCAT system is to output the translation information. These output files are to be identified with the “transys.madcat.xml” extension.

For example: <BASENAME>.transys.madcat.xml

6 Post-Editing Protocol

Post-editing is the act of manually modifying system output such that it contains the same meaning as the reference translation using understandable English.

Each system output will be edited for correctness by two independent teams of editors. A team consists of a pair of editors with one editor making edits in a first pass and a second editor acting as a reviewer. The reviewer checks the first pass edits for correctness while making additional modifications if needed. The output of the reviewer is the final version from the team. The output of the two reviewers is compared at the segment level, choosing the segment that has the lower HTER score (see section 7.2). The final document level HTER score is the resulting HTER when choosing the lower segment across the both sets of post-edited MT.

6.1 Post-Editing Kit and Editor Team Assignment

NIST defines “kits” as a collection of system output to be post edited by two post editing teams.

Each kit is to have around 600 words, a reasonable amount of data for an editor to edit in one session. Each kit contains a mixture of the two genres, newswire and web text

Each kit is assigned to two teams of editors. There are 10 teams. To the extent possible kits are assigned to editor pairs as to maximize their overlap for comparing inter-editor statistics.

7 Evaluation Metrics

This section describes the metrics used to score each of the three evaluation tasks. All scoring described below conserve case.

7.1 TER

The system performance on the document text translation task is measured by TER [6]. Short for Translation Edit Rate, TER is an edit distance metric which calculates the exact match distance between the system translation and the reference translation.

$$TER = \frac{(\#insertions+\#deletions+\#substitutions+\#shifts)}{\#reference_translated_words}$$

7.2 HTER

The system performance on the document image translation task is measured by HTER. Short for Human-mediated Translation Edit Rate, HTER is an edit distance metric which calculates the distance between the original system output and the post-edited system output. HTER is the primary metric of translation quality for MADCAT.

7.3 WER

The system performance on the document image transcription task is measured by WER. Short for Word Error Rate, WER is an edit distance metric which calculates the minimum number of steps required to transform the system transcript to exactly match the reference transcript. Unlike TER, shifts are not used in the calculation of WER.

$$WER = \frac{(\#insertions+\#deletions+\#substitutions)}{\#reference_transcribed_words}$$

8 Scoring Package

NIST developed a scoring package to facilitate the calculation of the above metrics. The package utilizes the software tercom-0.7.25 developed by UMD-BBN [6] as well as those developed internally at NIST [7].

Normalization is to be performed on the system output prior to scoring. For the translation tasks, punctuations in the reference and system translations are tokenized. For the transcription task, if any diacritic information is present in the reference and system transcripts, it is removed.

Segments containing scribe errors are to be included as-is for post editing. A stand-off annotation file will identify such segments allowing them to be analyzed separately.

9 Submission of Results

Submission of the evaluation results will be done via FTP:

- Create a directory where the system output will reside
- Place the output in that directory
- Tar and compress the directory
- FTP the tar and compressed file to jaguar.ncsl.nist.gov/madcat/incoming
- Send an email to madcat_poc@nist.gov to notify the submission was made

For example:

- `mkdir plato_1`
- `cp *.{word|line|trans}sys.madcat.xml plato_1`
- `tar zcvf plato_1.tgz plato_1`
- `ftp jaguar.ncsl.nist.gov` (anonymous login with email as password)
 - `binary`
 - `cd madcat/incoming`
 - `put plato_1.tgz`
 - `bye`
- send an email to madcat_poc@nist.gov

10 Schedule

Go/No-Go Track Evaluation Schedule	
Training & Development Data	
Training data release 1	February 9, 2009
Training data release 2	May 11, 2009
Training data release 3	July 9, 2009
Development data release	May 11, 2009
Evaluation	
Evaluation data for task 1 (document image translation) and task 2 (document image recognition) release to MADCAT team	October 28, 2009
Evaluation results for tasks 1 & 2 due to NIST at 10am	November 18, 2009
Evaluation data for task 3 (document text translation) release to MADCAT team at 3pm	November 18, 2009
Evaluation results for task 3 due to NIST	November 30, 2009
PE (post-editing) starts	November 30, 2009
PE ends	January 27, 2010
Final results to DARPA	January 29, 2010
Challenge Track Evaluation Schedule	TBD

11 Glossary of Terms

- Document – a naturally occurring unit of original source data of variable length
- Passage – a sub-section within a document chosen for evaluation

- Manuscript – a copy of a passage created by a scribe
- Page – one of the leaves in a manuscript created by a scribe; the basic unit of evaluation
- Scribe – a person who creates a handwritten copy of one or more passages

12 References

- [1] J. Olive, "Multilingual Automatic Document Classification Analysis and Translation (MADCAT) SOL BAA 07-38 Proposer Information Pamphlet", DARPA/IPTO, 2007.
- [2] MADCAT_Data_Planning_4_Feb_2008v11.ppt at https://madcatwiki ldc.upenn.edu/madcatwiki/index.php/Meetings/Phase1/DARPA_Brief
- [3] BBN Line Annotation Proposal.ppt presented to NIST on March 25, 2009.
- [4] MADCAT Phase 2 Scribe Assignment.docx
- [5] MADCATDataFormatSpec_V4h2.zip at https://madcatwiki ldc.upenn.edu/madcatwiki/index.php/Data_Format
- [6] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," *Proceedings of Association for Machine Translation in the Americas*, 2006.
- [7] J. Fiscus, J. Ajot, N. Radde, and C. Laprun, "Multiple Dimension Levenshtein Edit Distance Calculations for Evaluating Automatic Speech Recognition Systems During Simultaneous Speech", *Proceedings of LREC*, 2006.