NIST 2012 Open Handwriting Recognition and Translation Evaluation Plan

version 1.5

1 Introduction

The National Institute of Standards and Technology (NIST) is pleased to announce the second evaluation in the Open Handwriting Recognition and Translation (OpenHaRT) evaluation series. The OpenHaRT evaluation series is similar to evaluations conducted by NIST for the DARPA Multilingual Automatic Document Classification Analysis and Translation (MADCAT)[1] Program [1]; however, there are some notable differences between the two. First the OpenHaRT evaluation series is open to the public. And second, while the 2012 OpenHaRT evaluation uses data from MADCAT Phase 3, it also includes new data collected from Amazon Mechanical Turk[2].

The 2012 OpenHaRT evaluation again focuses on recognition and machine translation technologies for document images containing primary Arabic handwritten script as initial steps toward achieving document understanding. The evaluation series will build the critical mass required to solve challenges posed in these areas.

To participate in the evaluation, interested parties must register by completing the registration form[3] and the data license agreement[4] available at the NIST OpenHaRT website. There is no fee to participate, but participants are required to attend the post-evaluation workshop[5] to present their systems. While the evaluation is open to all who wish to participate, workshop attendance is limited to evaluation participants, data providers, (potential) evaluation sponsors, and interested U.S. government personnel.

2 Evaluation Tasks

The 2012 OpenHaRT evaluation seeks to advance the current state-of-the-art in recognizing and translating Arabic script in document images. The goal of OpenHaRT is to assess system performance and to understand the strengths and weaknesses of particular algorithmic approaches. It is planned that future evaluations will build on these technologies to include more complex tasks that are required to achieve document understanding capabilities.

Segmentation plays a vital role in deconstructing the document images. Translation and recognition tasks are paired with segmentation conditions to explore the relationship between system performance and the system's ability to segment the data. Segmentation is represented as a series of polygon coordinates indicating the locations of the text segments within the image. The 2012 OpenHaRT evaluation focuses on line segmentation. *Line segmentation* is defined as a bounding box that surrounds a line of text and is derived algorithmically from the word segmentations[6] by creating polygons that minimize the amount of text overlap between the lines.

There are two tasks defined for the 2012 OpenHaRT evaluation. Each task is designed to measure components within the overall system. Tasks are described below and summarized in the appendix (Table 4). Participants may choose to be evaluated in either one or both evaluation tasks.

2.1 Document Image Translation Task

Document image translation (DIT) task measures the overall performance of the system in translating text in foreign language document images into accurate and fluent English. The system is given a document image and the line level segmentation information and is required to output the English translation. Refer to sections 6.2 and 6.3 for input and output format requirements. Re-rendering the image is not required and is not a focus for OpenHaRT at this time.

2.2 Document Image Recognition Task

Document image recognition (DIR) task measures the text recognition component (Optical Character Recognition, or OCR) of the system transcribing the text in the document image.

[1] The NIST OpenHaRT evaluation is closely related to the DARPA MADCAT evaluation. Thus, there will be many references to "MADCAT" throughout this document.

[2] https://www.mturk.com/mturk/welcome

[3] http://www.nist.gov/itl/iad/mig/upload/OpenHaRT2012_RegistrationForm.pdf

[4] Data license agreement will be made available pending negotiation with data provider.

[5] There is a small workshop registration fee which does not include travel and accommodation. The workshop location is tentatively planned to be in the Washington DC Metro area. NIST reserves the right to cancel the workshop if there are too few participants.

[6] Human annotators mark the word boundaries using the GEDI [2] tool. The input to GEDI is a document image.

3 Data Resources

A set of corpora is provided for system development. To receive this data resource, participants must register for the 2012 OpenHaRT evaluation and sign the Linguistic Data Consortium (LDC) license agreement acknowledging the terms governing the uses and rights to the data.

Participants are free to use additional non-LDC collected data for system development. However, participants should take the necessary precautions to exclude newswire and web data (see below) that was originally published within the evaluation epoch **June 1-30, 2008.** We acknowledge that this exclusion may not always be practical. Participants are required to document data used for system development in their system description. See 8.2 for more information about system descriptions.

Data used in OpenHaRT 2012 come from two sources: data from previous MADCAT evaluations and data collected from the Amazon Mechanical Turk.

3.1 MADCAT Data

The data from the MADCAT Program are created in a more controlled environment. Literate native Arabic writers produce handwritten copies of news related Arabic passages. These passages are originally in electronic format and come from a variety of newswire publications[7], web blogs and online discussion forums[8]. Each passage is copied by three scribes. The handwritten copies are then scanned at 600 dpi to create the corresponding document images. The document images are in TIFF format. Table 1 lists the target distribution of the various writing factors, and Table 2 lists statistics for the datasets.

Table 1: Target distribution of various writing factors

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

[7] Newswire (NW) represents formal or structured text.
[8] Web text (WB) represents informal or unstructured text.

Table 2: Data profile for OpenHaRT datasets[9].

	Training Set	Dev Set	Eval Set
Source	MADCAT P1/P2/P3 training sets MADCAT P1/P2 devsets	MADCAT P1 pilot eval set	MADCAT P3 eval set
Genres	Newswire & Web text		
Num. of passages	~1500	~100	~100
Arabic tokens per passage	125	125	125
Number of scribes per passage	1 – 15	2	3
Total num of pages	~28,000	~500	~600

3.2 Mechanical Turk Data

The data from Amazon Mechanical Turk are created in a less controlled environment. Documents are created by having Turkers[10] create natural occurring data types and topics (i.e., shopping list). The Turkers then created the corresponding images using whatever method available to them (i.e., digital camera, scanner, etc.). No constraints or requirement are placed on the writers such as the writing utensils, the writing surfaces, the dpi of their uploaded images, the length of the document, etc.

No training or development data are provided for this data set.

4 Evaluation Metrics

This section describes the metrics used to score each of the evaluation tasks.

4.1 TER

The system performance on the **document image translation** tasks is measured automatically using the official evaluation metric Translation Error Rate (TER) [4]. TER is an edit distance metric which calculates the exact match distance between the system translation and the reference translation.

[9] "P1", "P2", and "P3" refer to phase 1, phase 2, and phase 3 MADCAT evaluations, respectively.
[10] "Turkers" is the term used to refer to people who perform tasks for money via Amazon's Mechanical Turk – an online marketplace for work.

$$TER = \frac{(\#insertions + \#deletions + \#substitutions + \#shifts)}{\#reference\_translated\_words}$$

In addition, system performance will be measured using a set of alternative automatic metrics such as BLEU and METEOR.

4.2 WER

The system performance on the **document image transcription** task is measured automatically using the Word Error Rate (WER) [5] metric. WER is an edit distance metric which calculates the errors (insertions, deletions, and substitutions) in the system transcription. (*TER and its use of shifts are not applicable to measuring transcription errors.*)

$$WER = \frac{(\#insertions + \#deletions + \#substitutions)}{\#reference\_transcribed\_words}$$

5 Scoring Package

A scoring package to facilitate the calculation of the OpenHaRT metrics will be made available to registered participants. The package utilizes the software "tercom" developed by UMD-BBN [4] as well as those developed internally at NIST [5]. The availability of the package will be announced on the OpenHaRT mailing list hart_list@nist.gov.

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 (e.g., typos, word omissions) are to be included as-is for scoring. A stand-off annotation file will identify such segments allowing them to be analyzed separately.

All translation and transcription scoring preserves the casing information.

6 Data File Format

OpenHaRT data use an XML format that defines storage elements which capture the various annotation layers in a document image. The format is described in version (v6) of the MADCAT Format Specifications document[11] and is extendable to future planned evaluation tracks. All training, development, and evaluation data will adhere to this XML format. System output will be validated using the DTD version 1.1.1 before being scored.

Participants are required to participate in a dry run to exercise their evaluation pipeline and avoid unnecessary delay due to data format and submission procedure. See section 9 for further information.

6.1 Reference Data

Each reference file contains two main layers of information along with a pointer to an accompanying document image. The first layer contains the physical segmentation of the image. The second layer contains semantic information in the image. The reference files are identified with the extension ".ref.madcat.xml".

For example: <FILENAME>.ref.madcat.xml

6.2 Input Data

The input to the system under test consists of document images and their corresponding XML files identifying the segmentation of interest.

The XML input files are derived from the reference files. Depending on the task, certain information will be removed from the reference files to create the input files. For the document image translation and the document image transcription tasks, the translation and transcription information is removed. For the document text translation task, translation information is removed. If a task excludes some segmentation information, the corresponding segmentation sub-layer is also removed.

Table 5 in the appendix summarizes the information content in the input for each task-condition pairing. Note the filename extensions also indicate the information content in the input.

6.3 Output Data

The output from the system under test consists of input XML files with the missing information added by the system.

Depending on the task, certain information will be added to the input files to create the output files. For the document image translation and the document image transcription tasks, the OpenHaRT system is to add the missing translation and transcription information, respectively. For the document text translation task, the OpenHaRT system is to output the translation information.

The name of the output file is to follow a similar naming convention as its corresponding input file. Note the difference between input and output

[11] ftp://jaguar.ncsl.nist.gov/madcat/resources/MADCATDataFormatSpec_V6.tgz.

filenames is highlighted below. Refer to Table 5 in the appendix for the output filename convention.

For example:

Input: <FILENAME>.in.madcat.xml

Output: <FILENAME>.out.madcat.xml

7 Evaluation Rules

The following rules must be observed when participating in the OpenHaRT evaluations:

- All tasks must be processed independently. That is, data files provided to complete other evaluation tasks may only be used for their designated task.
- Language model adaptation across pages is not allowed.
- Investigation of the evaluation data prior to submission of system output is not allowed. Both human and automatic probing is prohibited.
- To the extent possible, participants must exclude data that overlaps the evaluation epoch of June 1 – 30, 2008 from system development.
- Participation in the post-evaluation workshop is required. Each participating organization is to be represented by at least one technical individual who has the knowledge required to discuss system details (algorithmic approaches, data, issues ...) in the workshop's open forum.

8 Submission of Results

Participants may submit output from multiple systems and multiple output versions of the same system[12] for a given task and segmentation condition pairing. One system and version of that system must be declared as primary[13] at the time of submission and all other as contrastive. Participants must also include a single system description describing the system(s) submitted for evaluation.

Each configuration (task, segmentation condition, system, and system version) is considered as a single experiment and is identified by an experiment identifier (EXP-ID). See section 8.1 for the format of the EXP-ID. All experiments are to reside in a single submission file. See section 8.3 for the format of the submission file.

Submission will be made via FTP. If more than one submission is made, the last submission as indicated by the submission number replaces all previous submissions. Submissions that fail validation[14] will be returned to participants for correction. Late and/or debugged submissions will be documented and scored but will not be compared to other on-time submissions in NIST's reports.

8.1 System Output

System outputs are organized by experiment identifiers (EXP-ID). EXP-ID has the format:

EXP-ID =
HART12_<TEAM>_<TYPE>_<TASK>_<COND>_
<SYSID>_<VER>_<DATE>
where,

<TEAM>: is a participant-specified string (that does not contain underscores) indicating the name of the participating organization.

<TYPE>: can be one of the following values:
- DRYRUN – practice run on some sample data to validate the evaluation pipeline
- EVAL – official evaluation run on the official evaluation test set

<TASK>: is the evaluation task and can be one of the following values:
- DIT – document image translation
- DIR – document image recognition

<COND>: is the evaluation condition:
- LINE – line segmentation given

<SYSID>: is a participant-specified string (that does not contain underscores) designating the system used. The string *must begin* with *p-* for a primary system and with *c-* for any contrastive systems. For example: *p-baseline*. Note that there can only be one primary system per task/condition.

<VER>: is an integer (1 to n) indicating the version number. Values greater than 1 indicating multiple versions of the same system (i.e., the same system is run with a different set of parameters).

[12] A "system" is defined as a set of technology components interacting with each other to produce some output. For example, different noise removal algorithms would be labeled as different systems but different tuning parameters would be considered as different versions of the same system.

[13] The "primary" run is expected to yield the best performance on the blind test set. Only "primary" runs are used in cross-site analysis. "Contrastive" runs are compared only against their corresponding primary run for the same task/condition pairing.

[14] A submission validation script will be made available to participants in the near future.

<DATE>: is an 8-digit submission date of the format YYYYMMDD where YYYY is a 4-digit year, MM is a 2-digit month, and DD is a 2-digit day. This date will be used to distinguish experiments in different submission files.

For example, a participant submitted four experiments on April 27, 2012. The participant then decided to submit a bug-fix for one of the experiments and a new (different) experiment at a later date on May 6, 2012. The participant must include the first three unchanged experiments, the bug-fixed experiment, and the new experiment in the second submission file; but the three unchanged experiments[15] will contain the original submission date of April 27, 2012 while the bug-fixed submission and the new experiment will contain the new submission date of May 6, 2012.

8.2 System Description

Participants are to include a system description describing the system(s) submitted for evaluation in addition to the system output. The system description consists of, but is not limited to, the algorithm approaches employed, the training data used, and/or any other pertinent information. A template for the system description[16] can be obtained from the NIST OpenHaRT website. The system description will be distributed to other participants before the workshop. There should be only one system description per participating team with the name:

HART12_<TEAM>_<TYPE>_sysdesc.txt
where,

<TEAM> and <TYPE> are same as in 8.1.

8.3 Submission Instructions

Participants are to follow the steps outlined below when packaging and submitting their results.

1) Create an experiment directory for each experiment (see 8.1).

2) Place the system output in the corresponding experiment directory.

3) Create a submission directory with the format:
HART12_<TEAM>_<TYPE>_<SUB-NUM>
where,
<TEAM> and <TYPE> are same in 8.1.
<SUB-NUM>[17] is an integer (1 to n) where 1 identifies your first submission, 2 your second, etc.

4) Place all the experiment directories in the submission directory.

5) Place the system description in the submission directory (see Section 8.2 for more information on the system description).

6) Tar and gzip the submission directory.

7) FTP the compressed tar file to jaguar.ncsl.nist.gov/openhart/incoming using anonymous ftp.

8) Send an email to hart_poc@nist.gov to notify the submission was made.

For example:

- mkdir HART12_NIST_DRYRUN_DIT_LINE_p-baseline_1_20120427
- cp *.out.madcat.xml HART12_NIST_DRYRUN_DIT_LINE_p-baseline_1_20120427
- mkdir HART12_NIST_DRYRUN _1
- mv HART12_NIST_DRYRUN_DIT_LINE_p-baseline_1_20120427 HART12_NIST_DRYRUN _1
- cp HART12_NIST_DRYRUN_sysdesc.tx HART12_NIST_DRYRUN _1
- tar zcvf HART12_NIST_DRYRUN _1.tgz HART12_NIST_DRYRUN _1
- ftp jaguar.ncsl.nist.gov (anonymous login with email as password)
- binary
- cd openhart/incoming
- put HART12_NIST_DRYRUN _1.tgz
- bye
- send an email to hart_poc@nist.gov

An example submission directory content is given below:

HART12_NIST_DRYRUN _1 /
./HART12_NIST_DRYRUN_sysdesc.txt
./HART12_NIST_DRYRUN_DIT_LINE_p-baseline_1_20120427

[15] If the unchanged experiment submitted in the second submission is different from the first submission, it will be flagged and the entire submission will be returned to participant for correction.
[16] http://www.nist.gov/itl/iad/mig/upload/OpenHaRT2012_SystemDescription.txt
[17] Do not confuse submission number and version number. The submission number indicates the submission sent to NIST. The version number indicates a run of a specific system with a certain set of parameters.

./*.out.madcat.xml
./HART12_NIST_DRYRUN_DIT_LINE_ c-red_1_20120427
./*.out.madcat.xml
./HART12_NIST_DRYRUN_DIT_LINE_ c-white_1_20120427
./*. out.madcat.xml
./HART12_NIST_DRYRUN_DIT_LINE _c-white_2_20120427
./*. out.madcat.xml
./HART12_NIST_DRYRUN_DIT_LINE_ c-blue_1_20120427
./*. out.madcat.xml

9 Dry Run Evaluation

Participants are required to take part in a dry run exercise of their system prior to the official evaluation. The purpose of the dry run is to demonstrate readiness and to resolve any issues in the evaluation pipeline before the official evaluation starts. The dry run follows the exact protocol as the official evaluation (i.e., tasks, conditions, input/output file format, submission instructions). A small data set taken from training is used as the test data. The dry run is available from February 8, 2012 to March 30, 2012. Participants can contact NIST any time during that period to begin the dry run.

10 Publication of Results

NIST will release an official scoring report following the evaluation workshop. The report will be made public on the NIST website. Participants are free to publish and discuss their own results. However, participants must not publicly compare their results to that of other participants but can point to the NIST report for the results of the other participants. Participants must reference the NIST report when publishing their results.

11 Schedule

Table 3 lists important dates of the evaluation. Participating sites will receive training data after they have sent the completed and signed the registration form to NIST and data license agreement to LDC.

Table 3: OpenHaRT'10 evaluation schedule

Event	Date
Evaluation epoch (training and development data cannot overlap this epoch)	June 1-30, 2008
Evaluation registration period	September 1, 2011 – March 16, 2012
Training data availability	10 working days from the receipt of the signed registration and data license agreement (tentative)
Dry run evaluation availability	Feb 8, 2012 – Mar 30, 2012
Dry run submission	10 working days from the receipt of the dry run data
Evaluation period	April 16 - 27, 2012
DIT/DIR data distributed to participants	*April 16 (~09:00am US Eastern)*
DIT/DIR results and system description due to NIST	*April 27 (~11:59pm US Eastern)*
Preliminary results released to participants	May 15, 2012
System description distributed to all participants	June 15, 2012
Post-evaluation workshop	June 28, 2012 (tentative)
Official results published	July 30, 2012

12 References

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

[2] E. Zotkina, H. Suri, D. Doermann, "GEDI: Groundtruthing Environment for Document Images (Software)", http://lampsrv02.umiacs.umd.edu/projdb/project.php?id=53.

[3] GALE_p3_evalplan-v1f.pdf at http://www.nist.gov/itl/iad/mig/upload

[4] 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.

[5] 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.

Appendix

Table 4: Overview of OpenHaRT'12

Task	Primary Metric	Input	Output
Document Image Translation (DIT)	TER	Arabic document image • with line segmentation	Segmented English translation
Document Image Recognition (DIR)	WER	Arabic document image • with line segmentation	Segmented Arabic transcription

Table 5: Information content for the evaluation tasks

Task	Condition	Annotation Layer Removed	Input/Output File Extension
Document Image Translation (DIT)	Line Segmentation	• transcription • translation • word-level segmentation	<BASE>.in.madcat.xml <BASE>.out.madcat.xml
Document Image Recognition (DIR)	Line Segmentation	• transcription • translation • word-level segmentation	<BASE>.in.madcat.xml <BASE>.out.madcat.xml