NIST 2013 Open Handwriting Recognition and Translation Evaluation Plan

version 1.4

1 Introduction

The NIST Open Handwriting Recognition and Translation evaluation (OpenHaRT) is an evaluation series dedicated to support research to advance state-of-the-art in document analysis. Originated from the evaluations conducted by NIST for the DARPA Multilingual Automatic Document Classification Analysis and Translation (MADCAT)[1] Program) [1], the OpenHaRT series was designed to be accessible to all who find the tasks of interest and had its first evaluation in 2010. The 2013 OpenHaRT is the second evaluation in the series and continues to focus on core recognition and translation technologies for document images containing primary Arabic handwritten script. It is planned that future evaluations will build on these core technologies to include more complex tasks that are required to achieve document understanding capabilities.

While OpenHaRT'13 is similar to OpenHaRT'10 in many aspects, there are some minor differences. OpenHaRT'13 includes two training data tracks to understand the relationship between the amount of training data and system performance. OpenHaRT'13 does not evaluate the word segmentation condition.

To participate in the evaluation, interested parties must register by completing the registration form[2] and the data license agreement[3] available at the NIST OpenHaRT website[4]. 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

OpenHaRT'13 focuses on recognition and translation technologies of Arabic script in document images and explores the relationship between component performance and system performance. To this end, three tasks are defined to assess performance of particular algorithmic approaches. The tasks are described below and summarized in the appendix. Figure 1 which provides a pictorial summary of how the tasks are related. Participants may choose to be evaluated in any one or any combination of the three task groups.

2.1 Document Image Recognition

The **Document Image Recognition** (DIR) task measures the system capability in recognizing the text in the document images. The system is given a set of Arabic document images and text line segmentation and is asked to output the Arabic transcription found in each image. Refer to sections 7.2 and 7.3 for input and output format requirements.

2.2 Document Image Translation

The **Document Image Translation** (DIT) task measures the system capability in recognizing the foreign language text in document images and translating the recognized text into accurate and fluent English. The system is given a set of Arabic document images and reference line-level segmentation and sentence unit level segmentation and is asked to output the English translation of the text found in each image

2.3 Document Text Translation

The **Document Text Translation** (DTT) task measures the system capability in translating the foreign language text in document images into accurate and fluent English. The system is given the reference Arabic transcription and is asked to output the English translation. Refer to sections 7.2 and 7.3 for input and output format requirements.

3 Evaluation Tracks

There are two evaluation tracks in OpenHaRT'13 to facilitate cross system comparisons.

[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] http://www.nist.gov/itl/iad/mig/upload/OpenHaRT2013_Registrati
[2] http://www.nist.gov/itl/iad/mig/upload/OpenHaRT2013_RegistrationForm.pdf

[3] http://www.nist.gov/itl/iad/mig/upload/OpenHaRT2013_Eval_Agreement_Final.pdf

[4] http://www.nist.gov/itl/iad/mig/hart2013.cfm

[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.

- *Constrained* (required) – system developed using only the provided LDC data resources.
- *Unconstrained* (optional) – system developed using additional publicly available[7] non-LDC resources.

Participants are required to participate in the constrained track and encouraged to participate in the unconstrained track. For the unconstrained track, participants should take the necessary precautions to exclude newswire and web data (see Section 4) that was originally published within the evaluation epoch **June 1-30, 2008** and are required to document in their system description of the data used for system development including the information on how others can access these data sources. See section 9.2 for more information about system descriptions.

4 LDC Data Resources

A set of corpora is provided for system development. To receive this data resource, participants must register for the 2013 OpenHaRT evaluation and sign the Linguistic Data Consortium (LDC) license agreement acknowledging the terms governing the uses and rights to the data.

Data used in OpenHaRT'13 come from previous MADCAT evaluations. The data from the MADCAT Program are created in a controlled environment. Native Arabs who are proficient in Arabic produce handwritten copies of news related Arabic passages. These passages are originally in electronic format and come from a variety of newswire publications[8], web blogs and online discussion forums[9]. 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 TIF 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 10% pencil	75% unlined paper 25% lined paper	90% normal 5% fast 5% careful

[7] Publicly available data means the data can be obtained by individual researchers in the general population and not limited to very selected groups such as is true with government classified data, company proprietary data, etc.

[8] Newswire (NW) represents formal or structured text.

[9] Web text (WB) represents informal or unstructured text.

Table 2: Data profile for OpenHaRT datasets[10].

	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	~2000[11]	~100	~100
Arabic tokens per passage	125	125	125
Number of scribes per passage	1 – 15	2	3
Total num of pages	~42,000	~500	~600

5 Evaluation Metrics

This section describes the metrics used to score each evaluation task.

5.1 WER

The system performance on the **document image transcription** task is measured using Word Error Rate (WER) as described in [4]. WER is an edit distance metric which calculates the errors (insertions, deletions, and substitutions) in the system transcription.

$$WER = \frac{\#\ insertions + \#\ deletions + \#substitutions}{\#\ reference\ transcribed\ words}$$

5.2 TER

The system performance on the **document image translation** and **document text translation** tasks is measured using Translation Error Rate (TER) as described in [5]. TER is an edit distance metric which calculates the exact match distance between the system translation and the reference translation.

$$TER = \frac{\#\ insertions + \#\ deletions + \#substitutionsn + \#shifts}{\#\ reference\ transcribed\ words}$$

If time permits, the system performance will also be measured with other alternative automatic metrics such as BLEU [6] and METEOR [7].

[10] "P1", "P2", and "P3" refer to phase 1, phase 2, and phase 3 MADCAT evaluations, respectively.

[11] A passage can be more than one page.

6 Scoring Package

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

Normalization (e.g., punctuation tokenization) is to be performed on the reference and system output prior to scoring. For a complete list of normalization rules, refer to the scoring package.

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 scoring preserves the casing information.

7 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 the MADCAT Format Specifications document[12] and is designed to be 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 10 for further information.

7.1 Reference Files

Each reference file contains two main layers of information along with a pointer to the accompanying image. The first layer contains the physical segmentation of the image. The second layer contains semantic information in the image.

Data participants received from the LDC are the reference files from which the input files are derived. The reference files have the extension ".madcat.xml".

For example: <FILENAME>.madcat.xml

In the scoring pipeline, the reference files are linked to ".ref.madcat.xml" to facilitate parallelism among the file naming convention. This mechanism should be transparent to the users.

7.2 System Input Files

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

The input XML 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 line segmentation task, only the image is provided. For the document image transcription and translation tasks, the transcription and translation 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. The input files have the extension ".in.madcat.xml".

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

7.3 System Output Files

The output from the system under test consists of the 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 line segmentation task, since no input XML files are provided, the system is to produce the output XML files containing the line segmentation information. For the document image transcription and the document image translation tasks, the system is to add the missing transcription and translation information, respectively. For the document text translation task, the system is to output the translation information. The output files have the extension ".out.madcat.xml".

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

8 Evaluation Rules

The following rules must be observed by participants in the OpenHaRT evaluation:

- Language model adaptation across pages is not allowed when processing the OpenHaRT evaluation data. The rule is only included because the evaluation data contains duplicate passages (that is, the same passage text copied by different scribes), this rule is to prevent the system from leveraging information gained from "easier"

[12] ftp://jaguar.ncsl.nist.gov/madcat/resources/MADCATDataFormatSpec_V6.3.tgz.

scribes (e.g., neat handwriting) to assist its performance on "harder" scribes (e.g. messy handwriting).

- Investigation of the evaluation data prior to submission of all system output is not allowed. Both human and automatic probing is prohibited to ensure that all participating systems have the same amount of information on the evaluation data.

- To the extent possible, participants must exclude data that overlaps the evaluation epoch of June 1 – 30, 2008 from system development to avoid possible overlaps of the development data with the evaluation data.

- 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 ...) to be able to discuss his/her system in the workshop's open forum.

9 Submission of Results

Participants can submit output from multiple systems (i.e. different algorithmic approaches) and output from multiple versions (i.e., different tuning parameters) of the same system[13] for each track/task they have registered. One system and version of that system must be declared as primary[14] 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 9.1 for the format of the EXP-ID. All experiments are to reside in a single submission file. See section 9.3 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 will be returned to participants for correction. A validation script will be made available in the near future to help participants check their submission prior to sending it to NIST. Late and/or debugged submissions will be documented and scored but will not be compared to other on-time submissions in NIST's reports.

9.1 System Output

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

EXP-ID =
HART13_<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:
- DIR – document image recognition
-
- DIT – document image translation
- DTT – document text translation

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

<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).

<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, 2013. The participant then decided to

[13] 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.

[14] 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.

submit a bug-fix for one of the experiments and a new (different) experiment at a later date on May 6, 2013. 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, 2013 while the bug-fixed submission and the new experiment will contain the new submission date of May 6, 2013.

9.2 System Description

In addition to the system output, participants are to include a description describing in detail all of the system(s) they submit for evaluation. The system description consists of, but is not limited to, the algorithm approaches employed, the training data used and how others can access this data if the data is outside of the LDC provided training data, and/or any other pertinent information. A template for the system description[16] can be obtained from the NIST OpenHaRT website. There should be only one description comprising of all the systems submitted from each participating team with the name:

HART13_<TEAM>_<TYPE>_sysdesc.txt
where,

<TEAM> and <TYPE> are same as in 9.1.

The system description will be distributed to the workshop attendees and archived on the OpenHaRT'13 website but will not be published in the official ICDAR'13 proceedings. Evaluation participants are encouraged to submit a short paper about their work for OpenHaRT to the main ICDAR'13 conference, following the ICDAR'13 guidelines for short papers and relevant dates for submissions[17].

9.3 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 9.1).

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

3) Create a submission directory with the format: HART13_<TEAM>_<TYPE>_<SUB-NUM> where,

   <TEAM> and <TYPE> are same in 9.1.
   <SUB-NUM>[18] 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 9.2 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 HART13_NIST_DRYRUN_DIT_LINE_p-baseline_1_20130427
- cp *.out.madcat.xml HART13_NIST_DRYRUN_DIT_LINE_p-baseline_1_20130427
- mkdir HART13_NIST_DRYRUN _1
- mv HART13_NIST_DRYRUN_DIT_LINE_p-baseline_1_20130427 HART13_NIST_DRYRUN _1
- cp HART13_NIST_DRYRUN_sysdesc.tx HART13_NIST_DRYRUN _1
- tar zcvf HART13_NIST_DRYRUN _1.tgz HART13_NIST_DRYRUN _1
- ftp jaguar.ncsl.nist.gov (anonymous login with email as password)
- binary
- cd openhart/incoming
- put HART13_NIST_DRYRUN _1.tgz
- bye
- send an email to hart_poc@nist.gov

An example submission directory content is given below:

[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/OpenHaRT2013_SystemDescription.txt

[17] Participation in OpenHaRT'13 does not guarantee that the paper submitted to ICDAR'13 will get accepted. The submitted papers will go through the same rigorous peer review process.

[18] 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.

HART13_NIST_DRYRUN _1 /
./HART13_NIST_DRYRUN_sysdesc.txt
./HART13_NIST_DRYRUN_DIT_LINE_p-baseline_1_20130427
./*.out.madcat.xml
./HART13_NIST_DRYRUN_DIT_LINE_ c-red_1_20130427
./*.out.madcat.xml
./HART13_NIST_DRYRUN_DIT_LINE_ c-white_1_20130427
./*. out.madcat.xml
./HART13_NIST_DRYRUN_DIT_LINE _c-white_2_20130427
./*. out.madcat.xml
./HART13_NIST_DRYRUN_DIT_LINE_ c-blue_1_20130427
./*. out.madcat.xml

10 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 begins on February 14, 2013 and ends on February 28, 2013.

11 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.

12 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)	Jun 1 – 30, 2008
Evaluation registration period	Jun 1 – Dec 30, 2012
Training data availability	10 working days from the receipt of the signed registration and data license agreement (tentative)
Dry run evaluation period	Feb 14 – Feb 28, 2013
Dry run data sent to participants (all tasks)	*Feb 14*
Dry run submission due to NIST	*Feb 28,*
Formal evaluation period	Mar 19 – Apr 23, 2013
DIR/DIT evaluation data distributed to participants for processing	*Mar 19*
DIR/DIT system results due to NIST	*Apr 9*
DTT evaluation data distributed to participants for processing	*Apr 16*
DTT system results due to NIST	*Apr 23*
System description due to NIST	Apr 30
Preliminary results released to participants	May 7, 2013
System description distributed to all participants	May 14, 2013
Post-evaluation workshop to co-locate with ICDAR'13	Aug 23, 2013 (tentative)
Official results published	Aug 23, 2013

13 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] 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.

[5] 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.

[6] K. Papineni, S. Roukos, T. Ward, W. J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation", *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics*.

[7] A. Lavie and A. Agarwal, "METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments"**,** *Proceedings of the ACL 2007 Workshop on Statistical Machine Translation, 2007*

Appendix

Table 4: Summary of OpenHaRT'13

Task	Primary Metric	Input	Output	Input/Output File Extension
Document Image Recognition (DIR)	WER	Arabic document image • with reference line segmentation • with reference sentence unit segmentation	Sentence unit segmented Arabic transcription	Input:<BASE>.tif Input:<BASE>.in.madcat.xml Output:<BASE>.out.madcat.xml
Document Image Translation (DIT)	TER	Arabic document image • with reference line segmentation • with reference sentence unit segmentation	Sentence unit segmented English translation	Input:<BASE>.tif Input:<BASE>.in.madcat.xml Output:<BASE>.out.madcat.xml
Document Text Translation (DTT)	TER	Arabic document image • with reference line segmentation • with reference sentence unit segmentation • with reference Arabic transcription	Sentence unit segmented English translation	Input:<BASE>.tif Input:<BASE>.in.madcat.xml Output:<BASE>.out.madcat.xml

Figure 1: OpenHaRT'13 evaluation tasks.