NIST Open Machine Translation 2012
Evaluation Plan (OpenMT12)

1 Introduction

The 2012 NIST Open Machine Translation evaluation (OpenMT12) continues the ongoing series of evaluations of human language translation technology. NIST conducts these evaluations in order to support MT research and help advance the state of the art in MT technology. To do this, NIST:

- Defines a set of translation tasks,
- Collaborates with the Linguistic Data Consortium (LDC)[1] and the Defense Language Institute Foreign Language Center (DLIFLC)[2] to provide corpus resources to support research on these tasks,
- Creates and administers formal evaluations of MT technology,
- Provides evaluation utilities to the MT community, and
- Coordinates workshops to discuss MT research findings and results of task performance in the context of these evaluations.

These evaluations provide an important contribution to the direction of research efforts and the calibration of technical capabilities. They are intended to be of interest to all researchers working on the general problem of translating between human languages. To this end, the evaluations are designed to be simple, to focus on core technology issues, to be fully supported, and to be accessible to those wishing to participate.

The 2012 evaluation requires the translation of text data from a given source language into a given target language. Highlights of OpenMT12 include:

- Evaluation on prior years' Arabic-to-English and Chinese-to-English Progress test data, with the data to be made available to participants after OpenMT12,
- Evaluation on a parallel data set based on prior years' Progress tests for Arabic, Chinese, Dari, Farsi, and Korean to English, in two source data styles,
- Evaluation on Restricted Domain data for Chinese-to-English,
- Support for both a single system and a system combination track (provided sufficient interest by the participants), and
- Evaluation by automatic metrics and coordination of volunteer human assessments using a new tool to rank alternative translations.

Participation in the evaluation is invited for all researchers who find the tasks and the evaluation of interest. There is no fee for participation. However, participation in the evaluation requires participation in the follow-up workshop.[3] All OpenMT12 participants must attend the evaluation workshop and be prepared to discuss their system(s), results, and their research findings in detail. This workshop is restricted to the group of registered OpenMT12 participants, data providers and representatives of supporting government agencies.

To participate in the evaluation, sites must officially register with NIST[4] and agree to the terms specified in the registration form. For more information, visit the NIST OpenMT12 website.[5]

2 Training Conditions

MT R&D requires language data resources. System performance and R&D effort are strongly affected by the type and amount of resources used. Therefore, OpenMT12 has two different resource categories as conditions of evaluation. They differ solely by the specification of the data that may be used for system training and development. These evaluation conditions are *Constrained Training* and *Unconstrained Training*, as implemented in previous OpenMT evaluations.

Much of the data is provided by the LDC. All participants are required to sign a license agreement[6] governing the use of LDC's data resources available for system development in preparation for OpenMT12. Participants must fully comply with all requirements that are (1) stated in this evaluation plan, (2) stated on the registration form, and (3) stated on the LDC license agreement, in order to retain rights to data obtained under the LDC license agreement.

[1] http://www.ldc.edu

[2] http://www.dliflc.edu

[3] There is a registration fee associated with attending the evaluation workshop. This fee is normally between $300 and $500 and does not include travel or accommodation expenses.

[4] http://www.nist.gov/itl/iad/mig/upload/OpenMT12_Registration.pdf

[5] http://www.nist.gov/itl/iad/mig/openmt12.cfm

[6] http://www.nist.gov/itl/iad/mig/upload/OpenMT12_LDCAgreement.pdf

For Dari and Farsi to English, NIST can make available a set of parallel data from the TRANSTAC program.[7] There are only a limited number of Korean Language data resources available via the LDC license agreement.

2.1 Constrained Training

Systems entered in the Constrained Training condition allow for direct comparisons of different algorithmic approaches.

System development must adhere to the following restrictions:

For Arabic and Chinese to English, only data available from the LDC that is explicitly designated for the Constrained Training condition may be used for core MT engine development; these data resources are listed in the LDC license agreement. **Even with the restriction to these resources, care must be taken to remove any data stemming from the test epochs specified in Table 1: Evaluation test epochs for OpenMT12 tests for all language pairs. Specifically, data stemming from July of 2007 contained in the Gigaword corpora must be excluded from system development.**

Resources that assist the core engine (such as segmenters, tokenizers, parsers, or taggers) are not subject to the same restriction. If such additional resources are used, they must be listed in the system description.

The Constrained training condition is available only for Arabic and Chinese to English. Due to the limited amount of resources we can make available for Dari, Farsi, and Korean to English, Constrained Training is not offered for those language pairs.

2.2 Unconstrained Training

Systems entered in the Unconstrained Training condition may demonstrate the gains achieved by adding data from other sources. This training condition allows for more creativity in system development.

System development must adhere to the following restrictions:

1. Data must be publicly available, at least in principle.[8] This ensures that research results are broadly applicable and accessible to all participants. Participants must specify in their system descriptions what data they used.
2. Only data that was created outside of each language pair's evaluation test epoch (the period from which the evaluation data set is drawn) is to be used for system development for that language pair. See Table 1 for test epochs. **Note that the test epoch for all language pairs includes July 2007, which coincides with the test epoch of the OpenMT08 data sets. Participants should be particularly careful to exclude all data from July 2007 that was part of the OpenMT08 Arabic-to-English and Chinese-to-English tests from training for OpenMT12 for all language pairs.**

The Unconstrained training condition is offered for all language pairs of OpenMT12.

Table 1: Evaluation test epochs for OpenMT12 tests

Language pair	**Test epoch**
Arabic-to-English	July 2007
Chinese-to-English	July 2007, January – March 2011
Dari-to-English	July 2007
Farsi-to-English	July 2007
Korean-to-English	July 2007

3 Data Sets

The OpenMT12 evaluation data will be available by language pair (see section 4). For each language pair, one or more of the following data sets will be included in the test:

The data subsets (not all of which will be available for each language pair) are:

- Original Progress test: The Progress test sets used in OpenMT08 and OpenMT09, unchanged, with data to be released after OpenMT12.
- Current test: New data collected in the same way, and for the same data genres (see section 5), as the Original Progress test data.
- Progress tests subset: A combined subset of the OpenMT08 and OpenMT09 Progress tests, with new source data generated by human translation of the English reference translation. The source data will be provided in two styles:

[7] This data will be provided as-is and without further support, may not be distributed further, and its use is strictly limited to MT research purposes related to OpenMT12. Participants may contact NIST at mt_poc@nist.gov after registering for either Dari or Farsi to English in order to obtain this data.

[8] Data limited to government use, such as the FBIS data, is deemed to be publicly available and admissible for system development.

- English-true: A more English-oriented translation; requires that the text reads well and does not use any idiomatic expressions in the foreign language to convey meaning, unless absolutely necessary.
- Foreign-true: A translation as close as possible to the foreign language, as if the text had originated in that language.

- Restricted Domain test: Data from a variety of genres, varying in formality and structured, with a common theme. The theme will be announced after the evaluation period concludes.

4 LANGUAGE PAIRS

There are five language pairs offered for evaluation in OpenMT12. Participants indicate during registration the language pair(s) that they will process. Participants will only receive resources under the LDC license agreement for the language pair(s) for which they register. They will receive one test file per registered language pair, containing all test data for that language pair. Participants must submit output for the entire language test file. Scores will be reported for the subsets identified in section 3.

The different data subsets will be available for each language pair as follows:

- Arabic-to-English:
  - Original Arabic-to-English Progress test
  - New Arabic source Progress test subset, derived from the English reference translations of the original Arabic-to-English and Chinese-to-English Progress tests
- Chinese-to-English:
  - Original Chinese-to-English Progress test
  - Chinese-to-English Current test
  - New Chinese source Progress test subset, derived from the English reference translations of the original Arabic-to-English and Chinese-to-English Progress tests
- Dari-to-English, Farsi-to-English, Korean-to-English:
  - New Dari, Farsi, Korean source subsets, derived from the English reference translations of the original Arabic-to-English and Chinese-to-English Progress tests

The approximate number of words to be processed (based on reference translation word count) for each language pair is given in Table 2: **OpenMT12 Evaluation Data Set Sizes**.

Table 2: OpenMT12 Evaluation Data Set Sizes

Language pair	Approx. reference word count
Arabic-to-English	81,000
Chinese-to-English	135,000
Dari-to-English	46,000
Farsi-to-English	46,000
Korean-to-English	46,000

5 DATA GENRES

Each language pair test set consists of data from two or more genres. Systems will be required to process the entire test set for a specific language pair for the submission to be considered valid.

5.1 NEWSWIRE TEXT

A portion of each language pair test set will consist of Newswire texts, similar to those used in past OpenMT evaluations. These Newswire documents may be drawn from several types of sources, including newswire releases and the web. Newswire data will be specified as such in the test sets.

5.2 WEB DATA TEXT

A portion of each language pair test set will contain Web data, similar to what was referred to as Newsgroup data in previous OpenMT evaluations. These Web data documents may be drawn from user forums, discussion groups, and blogs. Web data will be specified as such in the test sets.

5.3 UNSPECIFIED GENRES

A portion of the Restricted Domain data subset of the Chinese-to-English test will consist of other genres of varying formality and structure; these genres will ***not*** be specified in the test set (genre will be set to “xx”).

6 PRIMARY AND CONTRASTIVE SUBMISSIONS

For each language pair registered for, OpenMT12 allows participants to submit exactly one primary single system submission, up to three contrastive single system submissions, up to three Original SysCombo submissions, and up to three Post-Eval SysCombo submissions (see section 7.2) per training condition.

At the time of submission exactly one system must be identified as the primary system, for each given language pair/training condition combination. Only primary systems will be compared and contrasted across sites in NIST's reporting of results.

Contrastive systems are encouraged to test significant alternatives to the primary system. NIST discourages contrastive entries that represent mere tweaks and minor parameter setting differences.

7 SYSTEM TRACKS

Open OpenMT12 will offer two system tracks for evaluation, the Single System track and the System Combination track. The system track is to be identified in the system description (see section 11).

7.1 SINGLE SYSTEM TRACK

Single System track systems enable research focused on the core issues of specific algorithmic approaches needed to advance machine translation technology. They allow the strengths and weaknesses of particular algorithms to be more clearly analyzed. Single System track systems exhibit the following key characteristic:

- The submitted translations result from a single core engine producing a translation using *primarily* one algorithmic approach. Note the use of *primarily* here; a predominantly SMT system that has a set of rules to handle certain types of data, often referred to as a "hybrid" system, is considered a single system.

7.2 SYSTEM COMBINATION TRACK

System combination research has intensified over the past several years as significant performance gains have been achieved through various combination techniques. Systems entered in the System Combination track exhibit one or more of the following characteristics:

- The submitted translations are the result of a core engine producing more than one translation, each using a different algorithmic approach before an internal combination process.
- The submitted translations are the result of comparing two or more alternative translations that were produced by different systems, possibly on different CPUs.

OpenMT12 features two variants of the System Combination track.

7.2.1 Original SysCombo

The output based on system combination is produced during the initial evaluation period, either site-internally or by cross-site collaboration.

Systems in the Original SysCombo track will be considered as official evaluation systems, and results will be reported in NIST's public release of results.

7.2.2 Post-Eval SysCombo

As a non-official, diagnostic test, NIST will organize the sharing of OpenMT12 system output for the purpose of system combination experimentation. During registration, participants will indicate whether they agree to have their primary system output be made available, anonymized, for system combination purposes. For those sites that agree to share, we ask that they also submit output of their primary system for an older OpenMT test set, to be made available, also anonymized, as a development set for the system combination track.

Shortly after the evaluation period, output from those sites that agreed to share will be distributed to those registered for the Post-Eval SysCombo track. There will be a separate deadline for submitting such system combination results.

The test set for the Post-Eval SysCombo track will be primary system output for the Original Progress test set for Arabic-to-English and Chinese-to-English (see Data Sets, section 3). These sets will be provided separately for the Arabic-to-English and Chinese-to-English output, and System Combination output must be submitted separately for combination output based on the two language pairs as well.

8 PERFORMANCE MEASUREMENT

OpenMT12 will employ automatic metrics and human assessments of system translations.

8.1 Automatic Metrics

The official primary metric for measuring performance will be the automatic N-gram co-occurrence scoring technique called BLEU, as originally developed by IBM.[9]

The N-gram co-occurrence scoring technique evaluates translations one "segment" at a time. (A segment is a cohesive span of text, typically one sentence, sometimes more.) Segments are delimited in the source text, and this organization must be preserved in the translation. An N-gram, in this context, is simply a *case sensitive* sequence of N tokens. (Words and punctuation are counted as separate tokens.) The N-gram co-occurrence technique scores a translation according to the N-grams that it shares with one or more reference translations of high quality. In essence, the more co-occurrences, the better the translation.

The N-gram co-occurrence technique BLEU employs provides stable estimates of a system's performance with scores that correlate well with human judgments of translation quality. Details of a study of the N-gram co-occurrence technique as a performance measure of translation quality may be accessed on the MT web site.[10]

For the official scores, NIST will compute case-sensitive BLEU scores using NIST's publicly available *mteval* software.[11] Researchers may use it to support their own research efforts, independent of NIST evaluations. All that is required is a set of source data, machine translation data, and at least one reference translation of high quality.

While BLEU is the official evaluation metric for OpenMT12, NIST will run a suite of MT evaluation metrics as time and resources permit.[12] The results of alternative scoring techniques will be included as part of the public release of results.

8.2 Human Assessments

As in previous OpenMT evaluations, human assessments are part of OpenMT12. Human assessments will be performed using a participant-volunteer model.

If a site wishes for its system output to be included in the assessments, it will be required to perform some assessments of OpenMT12 system translations. NIST will provide an assessment tool for download; judges will return their assessments to NIST. Participation in the human assessments must be indicated on the OpenMT12 registration form. Only primary submissions will be eligible for inclusion in the human assessments.

The human assessments for OpenMT12 will consist of five-way rankings of system outputs at the segment level. The data to be included in the human assessments will be drawn from the Combined Progress tests new source subset, which is available for all five language pairs.

Participants who would also like to participate in the human assessments must submit a separate registration form for human assessments by the registration deadline of February 3 2012.[13]

9 Schedule

- (See Table 1): Training data off-limits periods
- October 28 2011: Evaluation plan available
- October 28 2011 – February 3 2012: Registration period (early registration highly encouraged)
- November 4 2011: Training data available from LDC
- January 16 – February 24 2012: Dry run period (early submission highly encouraged); output due at NIST February 24 11.59am EST.
- April 2 – 6 2012: Main evaluation period; output due at NIST April 6 11.59am EDT.
- April 16 – 20 2012: Post-eval system combination evaluation period; output due at NIST April 20 11.59am EDT.
- April 20 2012: Preliminary release of main evaluation results to participants
- April 27 2012: System descriptions due
- April 30 – June 1 2012: Human assessment period
- June 27 – 28 2012: Workshop in the Washington DC area, co-located with the NIST OpenHaRT12 evaluation workshop
- August 31 2012: Official public release of results

10 Evaluation Procedures

The OpenMT12 evaluation process includes a number of mandatory steps; please see the schedule in section 9 for the dates for each of these:

1 Register to participate. Each site electing to participate in the evaluation must register with NIST.

[9] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu (2001). "Bleu: a Method for Automatic Evaluation of Machine Translation". This report may be downloaded from URL http://domino.watson.ibm.com/library/CyberDig.nsf/home (keyword RC22176).

[10] http://www.nist.gov/itl/iad/mig/tests/mt/2012/ngram-study.pdf

[11] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13a-20091001.tar.gz

[12] Contact NIST at mt_poc@nist.gov if you would like to recommend the inclusion of an (already published) MT metric.

[13] http://www.nist.gov/itl/iad/mig/upload/OpenMT12_HumanAssessmentRegistration.pdf

2 Sign LDC's data license agreement and return it to LDC. **Even if not selecting any training data, participants must sign the agreement to receive the evaluation data, which are listed on the agreement.**

3 Receive the dry run source data from NIST. Dry run source data will be sent to evaluation participants via email at the beginning of the dry run period.

4 Perform the dry run translation. Each site must run its translation system(s) on the entire dry run set for each language pair attempted.

5 Verify the compliance of the dry run translation output with the OpenMT12 submission requirements, using the OpenMT12 submission checker tool as described in section 12.2.

6 Return the dry run translations to NIST via FTP according to the instructions in section 12.3.

7 Receive the evaluation source data from NIST. Source data will be sent to evaluation participants via email at the beginning of the evaluation period. Inspection and manipulation of the evaluation data before the end of the evaluation period are prohibited.

8 Perform the translation. Each site must run its translation system(s) on the entire test set for each language pair attempted.

9 Verify the compliance of the translation output with the OpenMT12 requirements, using the OpenMT12 submission checker tool as described in section 12.2.

10 Return the translations to NIST via FTP according to the instructions in section 12.3.

11 Repeat steps 7-10 for the Post-Eval SysCombo track, if participating.

12 Submit a system description (see section 14).

13 Complete assigned human assessments, if participating.

14 Receive the evaluation results. NIST will score the submitted system translations and distribute the evaluation results to the participants.

15 Receive the complete set of reference translations for the language pair(s) attempted. Once the evaluation is complete, the set of reference translations used for evaluation will be made available to the evaluation participants to support error analysis and further research and to allow preparation for the evaluation workshop.

16 Be ready to prepare an oral or poster presentation for the workshop; NIST will contact selected presenters in ample time before the workshop.

17 Attend the evaluation workshop. NIST sponsors a follow-up evaluation workshop where evaluation participants and government sponsors meet to review evaluation results, share knowledge gained, and plan for the next evaluation. A knowledgeable representative from each participating site is required to attend this workshop and be ready to describe their technology, research, and findings. Attendance at the workshop is restricted to evaluation participants and government sponsors of MT research.

It is imperative that participants successfully complete all steps for all tasks they registered for. Failure to comply at any point will jeopardize chances of being permitted to participate in future NIST OpenMT evaluations and will require immediate return and removal of any data obtained under the LDC OpenMT12 license agreement.

Both the submission checker[14] (used to verify compliance of submissions) and the *mteval* utility[15] (used for scoring) are available for download from NIST. All those participating must complete the dry run, which includes performing the translations on the dry run set, using the submission checker to verify the compliance of their dry run output, and returning the dry run output to NIST before the main evaluation period (see section 13). The dry run provides an opportunity to uncover any potential formatting problems with submissions.

Participants are required to verify their evaluation submissions' compliance using the submission checker before submitting them to NIST. NIST will not accept any submissions that do not comply with the submission requirements (see section 12).

Handling of late and debugged submissions: Scores on submissions received at NIST after the submission deadline, as well as submissions that were debugged beyond formatting errors after an initial submission, will not be listed in the official public release of results. The respective sites will be listed in the release as having participated with a late and/or debugged submission. Such submissions will be scored as time permits and may be reported at the evaluation workshop.

11 NIST OpenMT Data Format

NIST has defined a set of XML tags that are used to format MT source, reference, and translation files for evaluation. Translation systems must be able to input the source documents and output translations that meet these formatting standards. All NIST OpenMT source, reference, and translation files have an *xml* extension; their format is defined by the current XML DTD.[16] NIST requires that all submitted translation files are well-formed and valid against the above-mentioned DTD.

[14] ftp://jaguar.ncsl.nist.gov/mt/resources/OpenMT12.tar.gz

[15] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13a-20091001.tar.gz

[16] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.6.dtd

11.1 Source File Format

A source file contains one single `srcset` element, immediately beneath the root `mteval` element. The `srcset` element has the following attributes:

- `setid`: The dataset.
- `srclang`: The source language. One of: Arabic, Chinese, Dari, Farsi, Korean.

The *srcset* element contains one or more *doc* elements, which have the following attributes:

- `docid`: The document name.
- `genre`: The data genre. One of: nw, wb, xx.

Each `doc` element contains several segments (`seg` elements). Each segment has a single attribute, `id`, which must be enclosed using double quotes.

One or more segments may be encapsulated inside additional elements, such as (but not limited to) `hl`, `p`, or `poster`. Only the native language text that is surrounded by a *seg* start-tag and its corresponding end-tag is to be translated.

Sample source file:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE mteval SYSTEM "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.6.dtd">
<mteval>
        <srcset setid="sample_set" srclang="Arabic">
                <doc docid="sample_document_1" genre="nw">
                        <seg id="1">ARABIC SENTENCE #1</seg>
                        <seg id=”2”>ARABIC SENTENCE #2</seg>
                        ...
                </doc>
                <doc docid="sample_document_2" genre="nw">
                        <seg id="1">ARABIC SENTENCE #1</seg>
                        ...
                </doc>
                ...
        </srcset>
</mteval>
```

11.2 Reference File Format

A reference file contains one or more `refset` elements, immediately beneath the root `mteval` element. Each `refset` element contains the following attributes:

- `setid`: The dataset.
- `srclang`: The source language. One of: Arabic, Chinese, Dari, Farsi, Korean.
- `trglang`: The target language, English.
- `refid`: The current reference.

Each `refset` element contains one or more documents, which, in turn, contain segments. The format of the document elements and their subsequent child elements is exactly the same as described in section 0 above for the source file.

Sample reference file:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE mteval SYSTEM "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.6.dtd">
<mteval>
        <refset setid="sample_set" srclang="Arabic" trglang="English" refid="reference01">
                <doc docid="sample_document_1" genre="nw">
                        <seg id="1">ENGLISH REFERENCE TRANSLATION #1</seg>
                        <seg id="2">ENGLISH REFERENCE TRANSLATION #2</seg>
                        ...
                </doc>
                <doc docid="sample_document_2" genre="nw">
                        <seg id="1">ENGLISH REFERENCE TRANSLATION #1</seg>
                        ...
                </doc>
                ...
        </refset>
        ...
</mteval>
```

11.3 Translation (Test) File Format

A translation file contains one or more tstset elements, immediately beneath the root mteval element. Each tstset element contains the following attributes:

- setid: The dataset.
- srclang: The source language. One of: Arabic, Chinese, Dari, Farsi, Korean.
- trglang: The target language, English.
- sysid: A name identifying site and system (see section 12.1.2 for requirements).

The content of each tstset element is exactly the same as described previously for the source file format and the reference file format.

Sample translation (test) file:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE mteval SYSTEM "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.6.dtd">
<mteval>
        <tstset setid="sample_set" srclang="Arabic" trglang="English" sysid="
NIST_ara2eng_primary_cn">
                <doc docid="sample_document_1" genre="nw">
                        <seg id="1">ENGLISH SYSTEM TRANSLATION #1</seg>
                        <seg id="2">ENGLISH SYSTEM TRANSLATION #2</seg>
                        ...
                </doc>
                <doc docid="sample_document_2" genre="nw">
                        <seg id="1">ENGLISH SYSTEM TRANSLATION #1</seg>
                        ...
                </doc>
                ...
        </tstset>
        ...
</mteval>
```

12 Formal Submission Requirements

12.1 Output Formatting

This section describes the directory structure, file naming, formatting, and encoding requirements that OpenMT12 submissions must adhere to. The specifications outlined in this section will be checked by the submission checker described in section 12.2.

12.1.1 Directory structure

The directory tree of each submission must comply with the following structure:
output/<site>/<langpair>
Where:

- site: The unique ID ***assigned by NIST*** to the site upon registration
- langpair: The language pair attempted in this submission (one of ara2eng, chi2eng, dar2eng, far2eng, kor2eng)
- version: The initial submission version must be 01. Potential later (debugged) submissions are to be numbered consecutively (02, 03, etc.). The first submission uploaded to NIST must contain the primary system. Additional submissions must use incremental versions. In the case of submitting debugged systems, only the debugged files must be present in a later submission. NIST will accept additional submissions through the end of the main evaluation period.

Example of a well-formed directory structure: output/NIST/ara2eng

12.1.2 File naming

Test file names must comply with the following naming convention:
<site>_<langpair>_<systype>_<train>_<evaltype>_<datestamp>.xml
Where:

- site: The unique ID ***assigned by NIST*** to the site upon registration
- langpair: The language pair attempted in this submission (one of ara2eng, chi2eng, dar2eng, far2eng, kor2eng)
- systype: The type of system of the particular submission. A primary submission must always be present. Up to three contrastive submissions are allowed. In addition, up to three submissions featuring a system combination are allowed.
    - One of: primary, contrast1, contrast2, contrast3, combo1, combo2, combo3
- train: The training condition, referring to Constrained or Unconstrained training
    - One of: cn, un
- evaltype: The evaluation being performed, referring to DryRun , Main or Post Evaluation System Combination.
    - One of: dryrun, eval, combo
- datestamp: The submission date.

- Format is: yyyymmdd. For OpenMT12 yyyy must be 2012.

The base file name is <site>_<langpair>_<systype>_<train>, and must match the sysid defined in section 0.

Example of a well-formed file name: NIST_ara2eng_primary_cn_dryrun_20120105.xml

12.1.3 Encoding

Submitted test files must contain UTF-8 encoded content only.

12.1.4 XML validity and DTD Compliance

Each test file must contain only valid XML data and comply with the current MT XML DTD (mteval-xml-v1.6.dtd).

12.1.5 File Content

A submitted test file must match the respective source data file in:

- setid value
- srclang value
- number of documents
- docid values
- genre value for each document
- number of segments for each document
- seg id values for each document

In addition, the sysid value must match the base file name as defined in section 12.1.2.

12.1.6 Compression

The submission for each language pair must be compressed as follows:

- Change directory to the parent directory of output directory; the output directory should contain all output for one language pair.
- For a dry run submission, issue the command:
  `tar cfvz OpenMT12_DryRun_<site>_<langpair>_<version>.tgz output.`
- For a main evaluation submission, issue the command:
  `tar cfvz OpenMT12_Eval_<site>_<langpair>_<version>.tgz output.` This archive contains all of the participant's primary, contrastive, and original SysCombo system output for the Constrained and/or Unconstrained Training condition for the given language pair.
- For a Post-eval SysCombo submission, issue the command:
  `tar cfvz OpenMT12_Combo_<site>_<langpair>_combo_<version>.tgz output.`

However zip files will be accepted as long as their content follows the naming and structure rules.

12.2 Output Validation

OpenMT12 will make use of a submission checker utility that is available for download[17]. NIST requires all participants to use it to validate their submission(s) prior to providing them to NIST. The submission checker will check all the requirements outlined earlier in this section and provide detailed error messages on file formatting, file completeness, file naming, and directory structure of a submission. Participants must fix all errors and re-check their submission(s) before submitting to NIST. NIST will reject any submissions that do not pass the submission checker; such submissions can be re-submitted as a new version after participants fix the reported errors. NIST requires the participation in a dry run (see section 13) to minimize the chance of submission problems during the main evaluation.

To prepare the submissions, first create the proper directory structure containing properly formatted and named files and compress them properly, as outlined earlier in this section.

The submission checker takes as input a compressed file of a submission for one language pair as described in section 12.1.5. A submission must pass all of the requirements outlined earlier in this section in order to pass the submission checker.

- Change directory to the directory of the submission checker tool
- Run `MT12SubmissionChecker.sh <SUBMISSION-ARCHIVE>.tgz`, where SUBMISSION-ARCHIVE is the file name generated in the compression step above

The tool only validates a single compressed file at a time.

If the archive passes the validation, proceed to upload your submission to NIST as described below. Otherwise, please review the submission checker output and fix any reported problems, then re-compress your output and repeat the validation process.

The submission checker is available for Linux, Mac and Windows (using Cygwin [18]) Operating Systems. Instructions for setup are available on the README document provided with the tool package. The README file contains the directory structure of the package.

The submission checker primer document is a step-by-step example of use of the tool and is provided as part of the submission checker package.

[17] ftp://jaguar.ncsl.nist.gov/mt/resources/OpenMT12.tar.gz

[18] http://www.cygwin.com/

12.3 Output Submission to NIST

System translations are to be submitted to NIST via anonymous ftp upload to ftp://jaguar.ncsl.nist.gov/OpenMT12/incoming.

13 Dry Run

In order to ensure smooth and timely processing of the OpenMT12 results, all participants are required to participate in a dry run ahead of the main evaluation period. For each language pair registered, participants will receive a very small source data set and must submit translation output following the requirements and procedures outlined in section 12. See section 9 for the schedule of the dry run. During the dry run period, NIST will assist participants in resolving any formal issues with their submissions. Early submission of dry run output is highly encouraged so that all issues can be resolved well before the main evaluation period.

14 System Descriptions

Participants are required to submit system descriptions of the MT systems used for their submissions. Please use NIST's template[19] for system descriptions. System descriptions should be submitted in text format, and the file name should reflect the site ID.

15 Guidelines For Publication of Results

NIST Multimodal Information Group's MT evaluations follow an open model to promotes interchange with the outside world. The rules governing the publication of NIST Open OpenMT12 evaluation results are the same as were used the previous year.

15.1 NIST Publication of Results

At the conclusion of the evaluation cycle, NIST will create a report which documents the evaluation. The report will be posted on the NIST web space and will identify the participants and the official BLEU-4 scores achieved for each language pair, training condition, and system track. Alternative metrics scores will be included, but it will be made clear that system tuning can affect rank ordering and that BLEU-4 scores are the only official scores for system comparison. Scores will be reported by language pair for the data subsets described in section 3, both across genres and separately by genre.

Results from the participant-based human assessments may also be posted.

The report that NIST creates should not be construed or represented as endorsements for any participant's system or commercial product, or as official findings on the part of NIST or the U.S. Government.

15.2 Participants' Reporting of Results in Publications

Participants must refrain from publishing results and/or releasing statements of performance until the official OpenMT12 results are posted by NIST.

Participants may not compare their results with the results of other participants, such as stating rank ordering or score difference. Participants will be free to publish results for their own system, but, participants will not be allowed to name other participants, or cite another site's results without permission from the other site. Publications should point to the NIST report as a reference.[20]

All publications must contain the following NIST disclaimer:

NIST serves to coordinate the NIST OpenMT evaluations in order to support machine translation research and to help advance the state-of-the-art in machine translation technologies. NIST OpenMT evaluations are not viewed as a competition, as such results reported by NIST are not to be construed, or represented, as endorsements of any participant's system, or as official findings on the part of NIST or the U.S. Government.

Linguistic resources used in building systems for OpenMT12 should be referenced in the system description. Corpora should be given a formal citation, like any other information source. LDC corpus references should adopt the following citation format:

Author(s), Year. Catalog Title (Catalog Number). Linguistic Data Consortium, Philadelphia PA.

For example:

Xiaoyi Ma et al, 2005. Arabic News Translation Text Part 1 (LDC2004T17). Linguistic Data Consortium, Philadelphia PA.

[19] http://www.nist.gov/itl/iad/mig/upload/OpenMT12_SysDescTemplate.txt

[20] This restriction exists to ensure that readers concerned with a particular system's performance will see the entire set of participants and tasks attempted by all researchers.