The 2009 NIST Open Machine Translation Evaluation Plan (MT09)

1 INTRODUCTION

The 2009 NIST Open Machine Translation evaluation (MT09) 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) 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 2009 evaluation requires the translation of text data from a given source language into a given target language. Highlights of MT09 include:

- Support for the evaluation of three language pairs.
- The inclusion of last year's *Progress* test sets.
- Evaluation by automatic metrics and coordination of volunteer human assessments.
- A new definition of system tracks, separating analysis of single systems from those representing multiple systems.

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.[1] All MT09 participants must attend the evaluation workshop and be prepared to discuss their system, results and their research findings in detail. This workshop is restricted to the group of registered MT09 participants and representatives of supporting government agencies.

To participate in the evaluation, sites must officially register with NIST[2] and agree to the terms specified in the registration form. For more information, visit the NIST Open MT web site.[3]

[1] 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. The MT09 evaluation workshop will be held in conjunction with MT Summit XII in Ottawa, Canada.

[2] MT09 registration form: http://www.nist.gov/itl/iad/mig/tests/mt/2009/MT09_RegistrationForm.pdf. Contact mt_poc@nist.gov if you have difficulties registering.

[3] http://www.nist.gov/itl/iad/mig/tests/mt

2 EVALUATION 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, MT09 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 Open MT evaluations. Both training conditions are offered for all language pairs.

Much of the data is provided by the LDC. All participants are required to sign a license agreement[4] governing the use of LDC's data resources available for system development in preparation for MT09. 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.

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:

Only data available from the LDC that is explicitly designated for the **`Constrained`** training condition may be used for core MT engine development.[5]

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.

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.[6] This ensures that research results are broadly applicable and accessible to all participants.
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*) may be used for system development. Participants may, however, continue to search the web up through the evaluation week and use data created outside of the relevant test epoch. Only data that can unequivocally be identified as published outside of the relevant test epoch

[4] LDC License agreement: http://www.nist.gov/itl/iad/mig/tests/mt/2009/2009_NIST_MTOpenEval_Agmnt_V3.pdf

[5] http://www.nist.gov/itl/iad/mig/tests/mt/2009/MT09_ConstrainedResources.pdf

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

may be used. See Table 1 for specific evaluation test epochs.

Table 1: Evaluation test epochs for MT09 language tests. Data created and/or published during these dates are off-limits for system development.[7]

Arabic-to-English	Current test set	June 2007
	Progress test set	July 2007
Chinese-to-English	Progress test set	July 2007
Urdu-to-English	Current test set	Newswire: January 2009 Web data: December 2008 – January 2009[8]

3 EVALUATION SYSTEM TRACKS

For the first time, Open MT09 will offer two different 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 11.2).

3.1 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.

3.2 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.

MT09 accommodates two variants of the **`System Combination`** track.

3.2.1 System Combination for *Formal* Evaluation

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

Systems submitted by the close of the evaluation period will be considered as official evaluation systems and results will be reported in NIST's public release of results.

3.2.2 System Combination for *Informal* Evaluation

NIST will organize the sharing of MT09 system output for the purpose of system combination experimentation. Shortly after the evaluation period, output from those sites that agree to have their output shared for system combination purposes will be distributed to those interested in submitting informal system combination results.

There will be a separate deadline for submitting such system combination results (see Schedule, section 13).

System output available for the informal System Combination track will include output for the *Current* test set for the Arabic-to-English and Urdu-to-English tasks. For each of these, the output will be partitioned into two parts. Approximately 30% of the test data will be designated as a development set for system combination; the system output for this subset will be provided with all four reference translations. The remainder of the system output will be provided as the System Combination evaluation data set. These sets will be provided separately for the Arabic-to-English and Urdu-to-English output, and System Combination submissions are to be made separately for combination output based on the two language pairs as well.

4 EVALUATION DATA SETS

The MT09 evaluation data sets will be defined as either *Current* or *Progress* test sets.

4.1 CURRENT TEST SET

A *Current* test set is newly collected data that systems process, and after submitting the system translations to NIST for scoring, the reference translations are made available for system analysis and future development testing.

A *Current* is offered for Arabic-to-English and Urdu-to-English.

4.2 PROGRESS TEST SET

A *Progress* test set is unseen data that systems process, and after submitting the system translations to NIST for scoring, all evidence of ever possessing the data is destroyed. The *Progress* test set option was first introduced in MT08 and will be reused in future NIST Open MT evaluations in an effort to better measure year-to-year improvement.

To maximize the usage of the *Progress* test data, many restrictions are in place, and only sites that have demonstrated the ability to fully and successfully participate in past NIST Open MT evaluations are encouraged to process the *Progress* test data. First time participants will be handled on a case-by-case basis.[9]

A *Progress* test is offered for Arabic-to-English and Chinese-to-English. The *Progress* test data is identical to last year's *Progress* test data.

5 EVALUATION LANGUAGE PAIR TRACKS

There are three language pairs offered for evaluation in MT09, and there are some differences in evaluation structure between them.

Participants indicate during registration the language pair(s) that they will process. They will only receive resources under the LDC license agreement for the language pairs for which they register, and they will only receive the evaluation data for those language pairs.

[7] In the event that this requirement would restrict participation, please contact NIST at mt_poc@nist.gov to discuss alternatives.

[8] An announcement will be made as soon as the evaluation epoch dates are established.

[9] http://www.nist.gov/itl/iad/mig/tests/mt/2009/MT09_ProgressTestForm.pdf

5.1 ARABIC-TO-ENGLISH

The Arabic-to-English (A2E) track will be implemented in a very similar fashion to previous NIST Open MT evaluations. There will be two test sets, both comprised of Newswire and Web data, as seen in Table 2.

The *Current* test set will be processed by all participants and will be drawn from the GALE phase 3 evaluation data, updated to include four reference translations.

Results for submissions from GALE participants, who will have had previous access to the test data, will be reported separately from results submitted by those that are not part of the GALE program.

There is a possibility that the Arabic Current test set will be upgraded to a larger all-new data set. In the event that this upgrade becomes real, an announcement will be made via the NIST MT mailing list.

Participation in the *Progress* test is limited to those participants agreeing to the terms of usage. The *Progress* test set is identical to MT08's *Progress* test set.

The A2E test is offered for the **`Constrained`** and **`Unconstrained`** training conditions.

System translations of the *Current* test set will be evaluated using automatic metrics and volunteer human assessments. System translations of the *Progress* test will be evaluated using automatic metrics.

5.2 CHINESE-TO-ENGLISH

The Chinese-to-English (C2E) track is offered using last year's *Progress* test set consisting of Newswire and Web data, as seen in Table 2. Participation in this test is limited to those participants agreeing to the terms of usage. The *Progress* test set is identical to the *Progress* test set from MT08.

The C2E *Progress* test is offered for the **`Constrained`** and **`Unconstrained`** training conditions.

System translations will be evaluated using automatic metrics.

5.3 URDU-TO-ENGLISH

The Urdu-to-English (U2E) track is limited to an evaluation of a *Current* test set using equal amounts of Newswire and Web source data; see Table 2.

The U2E test is offered for the **`Constrained`** and **`Unconstrained`** training conditions. **`Constrained`** submissions that meet the restrictions of MT08's *strictly* **`Constrained`** track for Urdu-to-English, which constrained resources for *both* language sides and for core engine as well as supporting tool development to data on the resource DVD,[10] should be marked as such in the system description. Such submissions will be useful for gauging progress over last year's Urdu-to-English results.

MIT-LL will host a wiki for information exchange related to the U2E test.[11]

System translations will be evaluated using automatic metrics. Human assessments will be offered if enough volunteers express interest.

6 PRIMARY AND CONTRASTIVE SUBMISSIONS

MT09 allows sites to enter up to four systems for each training condition / system track / test set / source language combination that they register for.

Sites submitting more than one system for a given evaluation track must declare one entry as the primary system at the time of submission. 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 PERFORMANCE MEASUREMENT

MT09 will employ automatic metrics and human assessments of system translations.

7.1 AUTOMATIC MT 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.[12] This represents a change over recent years, in which the NIST implementation of BLEU, with a different calculation of the brevity penalty, was the official primary metric. The change was made based on discussions held at the MT08 workshop.

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.[13]

NIST updated the Open MT evaluation scoring script "**`mteval`**" such that scores are calculated using the original definition of the brevity penalty. **`mteval`** also calculates the "NIST" score and, optionally, the BLEU score using the brevity penalty as defined in recent Open MT evaluations. The latter two metrics will be reported as alternative scores. **`mteval`** is available as a downloadable software utility.[14] Research sites may use it to support their own research efforts, independent of NIST tasks/evaluations. All that is required is a set of source data, machine translation data, and at least one reference translation of high quality.

[10] See section 2.1 of http://nist.gov/itl/iad/mig/tests/mt/2008/doc/MT08_EvalPlan.v2.4.pdf.

[11] Registered Urdu participants will receive the URL, username, and password for the wiki.

[12] 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).

[13] http://www.nist.gov/itl/iad/mig/tests/mt/2009/ngram-study.pdf

[14] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13.pl

While BLEU is the official evaluation metric for MT09, NIST will run a suite of MT evaluation metrics as time and resources permit.[15] The results of alternative scoring techniques will be included as part of the public release of results.

7.2 HUMAN ASSESSMENTS OF SYSTEM TRANSLATIONS

As in previous Open MT evaluations, human assessments are part of MT09. All human assessments will be performed using a participant-volunteer model.

If a site wants its system output to be included in the assessments, it will be required to perform some assessments of MT09 system translations. NIST will provide an updated web-based assessment tool.[16] There is a separate registration form[17] for participating in the human assessments.

If a site wants to enter more than one system per language track in the human assessments, one of them must be the primary system in either training condition.

There are three types of human assessment. Judges will work on only one type of assessment at a time, i.e. they will not go through the different types of assessment for the same data at the same time.

Guidelines, a tutorial, and a training session for the human assessments are available on the MT09 human assessment web site.

7.2.1 Comprehensibility

Human assessment of comprehensibility will be performed using the following model: An assessor views the first translated segment of a document (MT only) and answers the question, "How comprehensible is this translation for you?" Answers will be given using a 7-point scale with an optional "Notes" box for comments. Immediately following the comprehensibility question, a single reference translation for the segment is shown and a second question is asked: "Does the reference change your understanding of this sentence?" Answers to this second question are given using a 5-point scale.

This process of assessing comprehensibility continues for the same document and the same system, showing the next MT translated segment while preserving all previously assessed reference segments to provide context.

7.2.2 SEMANTIC ADEQUACY

7.2.2.1 SEMANTIC ADEQUACY – SEGMENT LEVEL

Human assessment of semantic adequacy at the *segment* level will be performed using the following model: An assessor views the first translated segment of a document (MT only) and is asked to "Predict the range of semantic adequacy scores you and others will give this machine translation once the reference translation is shown." Answers are given using a 7-point scale. Immediately following this predictive question, a single reference translation for the segment is shown and a second question is asked: "Now provide your actual semantic adequacy score for the machine translation of this segment compared to the reference translation." Answers are given using the same 7-point scale as for the first question.

Judges can indicate their confidence in their assessment by selecting a specific scale point or a range of adjacent scale points for both questions.

This process of assessing semantic adequacy continues for the same document and the same system, showing the next MT translated segment while preserving all previously assessed segments to provide context.

7.2.2.2 SEMANTIC ADEQUACY – DOCUMENT LEVEL

Human assessments of semantic adequacy at the *document* level will be performed using the following model: An assessor views a machine translation of a complete document together with a single reference translation and is asked to "Provide your semantic adequacy score for the machine translation of this document compared to the reference translation." Answers are given using the same 7-point scale as for the segment-level semantic adequacy question.

Judges can indicate their confidence in their assessment by selecting a specific scale point or a range of adjacent scale points for both questions.

8 NIST MT 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 Open MT source, reference, and translation files have an "xml" extension; their format is defined by the current XML DTD.[18] NIST requires that all submitted translation files are well-formed and valid XML (against the above-mentioned DTD).

8.1 DIFFERENCES FROM PREVIOUS DATA FORMAT

Starting with the 2009 evaluation, NIST provides XML-only files. Some differences between the former SGML file format and the new XML file format are listed below:

- The root element of each file must be "`mteval`".
- Element and attribute names are case-sensitive; they must match their definition in the DTD; e.g.:
  - Invalid: `<DOC>…</DOC>`
  - Invalid: `<Doc>…</Doc>`
  - Valid: `<doc>…</doc>`
- The "less-than" sign and the "ampersand" sign must be properly escaped (using "<" and "&", respectively), e.g.:
  - Invalid: `<seg id="1">A & B</seg>`
  - Valid: `<seg id="1"> A &amp; B<seg>`
- Attribute values must be enclosed using double or single quotes, e.g.:
  - Invalid: `<seg id=12>…</seg>`
  - Valid: `<seg id="12">…</seg>`

NIST requires that the following two lines are added at the beginning of each XML file to facilitate validation against the DTD:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE mteval SYSTEM
"ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.3.dtd">
```

In the following file format descriptions, the possible values for the attributes `srclang`, `trglang`, and `genre` are

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

[16] For a description, see http://www.nist.gov/itl/iad/mig/tests/mt/2009/ha.

[17] http://www.nist.gov/itl/iad/mig/tests/mt/2009/MT09_HumanAssessmentForm.pdf

[18] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.3.dtd

limited to the values used in MT09; the DTD itself may have more values for these attributes.

8.2 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, Urdu.

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

- "docid": The document.
- "genre": The data genre. One of: nw, wb.

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

One or more segments may be encapsulated inside other 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 ("</seg>") is to be translated.

A sample XML source file can be found in Appendix A.

8.3 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, Urdu.
- "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.

A sample XML reference file is provided in Appendix A.

8.4 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, Urdu.
- "trglang": The target language, English.
- "sysid": A name identifying site and system (see section 11.1 item 4).

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

A sample XML translation file is provided in Appendix A.

9 EVALUATION DATA GENRES

Each test set is drawn from two types of data (Newswire, Web data). Systems will be required to process the entire test set for each language pair attempted.

Source documents are UTF-8 encoded.

Systems will be evaluated separately for each language pair and for each training condition, and results will be reported separately for each system track. System performance will be evaluated separately for the *Progress* and *Current* test sets, where available. System performance will be reported over the entire test set and over selected subsets of the test set. Table 2 provides details on how the evaluation data will be divided into selected subsets for reporting results.

Table 2: MT09 Evaluation Data Sets.

Data Statistics for MT09 (approximate *reference* word counts)			
Source Language	Genre	Current set	Progress set
Arabic	Newswire	15,000	20,000
	Web data	15,000	15,000
Chinese	Newswire	N/A	20,000
	Web data	N/A	15,000
Urdu	Newswire	20,000	N/A
	Web data	20,000	N/A

9.1 NEWSWIRE TEXT

A portion of each test set will contain Newswire stories, similar to those used in past MT evaluations. These stories may be drawn from several kinds of sources, including newswire releases and the web. The expected sizes of the Newswire data sets are described in Table 2 above.

9.2 WEB DATA TEXT

A portion of the test set will contain Web data, similar to what was referred to as Newsgroup data in previous NIST Open MT evaluations. Web data is world-wide-web data from user forums, discussion groups, and blogs. The expected amounts of Web data are listed in Table 2.

10 EVALUATION PROCEDURES

There are ten steps in the MT09 evaluation process:

1 *Register to participate.* Each site electing to participate in the evaluation must register with NIST no later than the deadline for registration.[2] See section 13 for more details.

2 *Receive the evaluation source data from NIST.* Source data will be sent to evaluation participants via email at the beginning of the evaluation period. The email addresses to receive the evaluation source sets must be provided to NIST on the MT09 Registration form. **Inspection and manipulation of the evaluation data before the end of the evaluation period are prohibited.**

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

4 *Return the translations.* The translations are to be returned to NIST via email according to the instructions in section 11. Translations for each language must be submitted separately.

5 *If applicable, delete all Progress test set source material and any derivative files generated in the course of processing the Progress test set(s).*

6 *Submit a system description (see section 11.2) by the deadline indicated in section 13.*

7 *Receive the evaluation results.* NIST will score the submitted system translations and distribute the evaluation results to the participants. See section 13 for more details.

8 *Receive the complete set of reference translations for the Current test set(s).* 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.

9 *Prepare a presentation including a description of your system and your research findings*. Participants will be asked to send a soft copy of their talk to NIST about a week before the evaluation workshop so that workshop material may be prepared in advance.

10 *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 describe their technology and research and present their research findings. Attendance at the workshop is restricted to evaluation participants and government sponsors of MT research.

It is imperative that participants successfully complete all 10 steps. Failure to meet any one of the requirements will jeopardize the chances of your organization being invited to participate in future NIST Open MT evaluations. Also, failure to comply will require immediate return and removal of any data obtained under the LDC MT09 license agreement.

We stress that the **`mteval`** utility to be used for the evaluation is available for download from NIST for all those who are interested in using the tool.**Error! Bookmark not defined.** It is vitally important that all those planning to participate in MT09 verify that they are fully prepared for the formal evaluation prior to the evaluation period; we therefore encourage submitting output on practice data sets, following the procedures outlined in section 11.1. NIST will require all first-time participants to submit such a practice run to verify proper formatting.

11 Submitting Translations to NIST

Participants are required to submit results for all language pairs that they registered for. Translations may be submitted for one or more of the training conditions. Each submission must be complete in order to be acceptable.

11.1 System Translations

System translations are to be submitted to NIST via ftp upload to ftp://jaguar.ncsl.nist.gov/MT09/incoming.[19]

To properly package translation files for submission of your *Current* test output to NIST, the following steps must be followed exactly:

1. Create a directory that identifies your site (e.g., ./NIST).
2. Create a subdirectory for each language pair you will submit: ./NIST/a2e, ./NIST/c2e, ./NIST/u2e.
3. Encode each translation file in UTF-8.
4. Name each translation file with the following elements, in order, separated by underscores:
    a. Site and system ID
    b. Language pair: a2e, c2e, u2e
    c. Training condition: cn, un for **`Constrained`**, **`Unconstrained`**, respectively
    d. Primary vs. contrastive system: primary, contrast1, contrast2, contrast3

    and with an "xml" extension (e.g., NIST_a2e_cn_primary.xml).
5. Put the properly formatted translation files in the corresponding subdirectories (e.g., ./NIST/a2e/NIST_a2e_cn_primary.xml).
6. Create the compressed tar file using the UNIX **tar** and **gzip** commands. (tar –cf NIST.tar ./NIST; gzip NIST.tar)
7. Upload the file to ftp://jaguar.ncsl.nist.gov/MT09/incoming.

Follow the same steps for submitting *Progress* test output, creating a separate tar file.

11.2 System Descriptions

Participants are required to prepare a system description for each system submitted for evaluation. Please use NIST's template[20] for system descriptions to make it easier for researchers to directly compare and contrast different algorithmic approaches.

The preferred submission format of the system descriptions is ASCII text or PDF. System descriptions will be distributed as part of the workshop materials.

12 Guidelines For Publication of Results

NIST Multimodal Information Group's HLT evaluations are moving towards an open model which promotes interchange with the outside world. The rules governing the publication of NIST Open MT09 evaluation results are the same as were used the previous year.

12.1 NIST Publication of Results

At the conclusion of the upcoming 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 (along with alternative metric 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 for the overall test set for each language pair processed, and also for the following subsets of the evaluation test data:

1. The Newswire text documents
2. The Web data text documents

Where available, results from the participant-based human assessments will 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.

[19] If submission by ftp is not possible for a participant, please contact mt_poc@nist.gov to make alternative submission arrangements.

[20] http://www.nist.gov/itl/iad/mig/tests/mt/2009/MT09_SysDescTemplate.txt

12.2 PARTICIPANTS' REPORTING OF RESULTS IN PUBLICATIONS

Participants must refrain from publishing results and/or releasing statements of performance until the official MT09 results are posted by NIST on approximately October 30th, 2009.

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.[21]

All publications must contain the following NIST disclaimer:

NIST serves to coordinate the NIST Open MT evaluations in order to support machine translation research and to help advance the state-of-the-art in machine translation technologies. NIST Open MT 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 MT09 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.

For example:

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

13 SCHEDULE

Date	Event
(See Table 1.)	Training data off-limits periods. All data created, posted, or published during these periods is off-limits for system training and development.
February 24 2009	Evaluation plan released
May 14 2009	Registration Deadline.
June 8 2009 9:00am EDT	Evaluation test data e-mailed to participants.
June 12 2009 12 noon EDT	Deadline for submission of results to NIST.
June 17 2009	Distribution of system translations to Informal System Combination track participants
June 19 2009	Preliminary release of results to participants.
June 26 2009	Deadline for submission of Informal System Combination track results to NIST.
June 26 2009	Deadline for submitting system descriptions to NIST.
July 6 2009	Human assessment tasks available.
August 15 2009	Deadline for finishing human assessments.
August 31 – September 1 2009	Evaluation workshop, open to participants and government sponsors only, held in Ottawa, ON, Canada (co-located with MT Summit XII). Separate registration is required for attending the workshop. Announcements are made to registered participants only.
October 30st, 2009	Official public release of results.

[21] 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.

Appendix A: NIST MT Data Format

Example of XML source file:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE mteval SYSTEM "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.3.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>
```

Example of XML reference file:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE mteval SYSTEM "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.3.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>
```

Example of XML translation file:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE mteval SYSTEM "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.3.dtd">
<mteval>
        <tstset setid="sample_set" srclang="Arabic" trglang="English" sysid="NIST">
                <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>
```