The 2008 NIST Open Machine Translation Evaluation Plan (MT08)

1 Introduction

The 2008 NIST Open Machine Translation evaluation (MT08) 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 task performance,
- Provides evaluation tools and 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 2008 evaluation requires the translation of text data from a given source language into a given target language. Highlights of MT08 include:

- Support for the evaluation of four language pairs,
- The introduction of Progress test sets,
- Inclusion of automatic metrics, two types of human assessments, and MT comprehension tests.

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 participants must attend this workshop and be prepared to discuss their system, results and their research findings in detail. This workshop is restricted to the group of registered 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]

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, two different resource categories have been defined as conditions of evaluation for MT08. The categories differ solely by the specification of the data that may be used for system training and development. The evaluation conditions are Constrained training and Unconstrained training.

Much of the data is provided by the LDC. Participants who are not current members of the LDC will be required to sign a license agreement[4] which governs the use of LDC's data resources available for system development in preparation for the MT08 evaluation.

2.1 Constrained Training

Systems entered in the Constrained training condition allow for direct comparisons of differing algorithmic approaches.

System development must adhere to the following restrictions:

For all language pairs **except Urdu to English**, only data that is available from the Linguistic Data Consortium's public catalog and is 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, or taggers) are not subject to the same restriction. If such additional resources are used, they must be listed in the system description.

For **Urdu to English**, only the LDC resource DVD designated specifically for this task may be used.[6] Resources that assist the core engine (such as segmenters, tokenizers, or taggers), as well as any rule development, are subject to the same restriction. These restrictions holds for both the Urdu and the English side of the Urdu to English task.

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 may allow for more creativity in system development.

System development must adhere to the following restrictions:

1. Data must be publicly available, at least in principle.[7] This ensures that research results are broadly applicable and accessible to all participants.
2. Only data that was created before July 1st, 2007 may be used for system development (*this is the month from which the evaluation data set will be drawn*). Participants may, however, continue to search the web

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

[2] The 2008 Machine Translation Registration form is online at: http://www.nist.gov/speech/tests/mt/2008/doc/MT08_RegistrationForm.pdf. Contact mt_poc@nist.gov if you have difficulties registering.

[3] http://www.nist.gov/speech/tests/mt

[4] http://www.nist.gov/speech/tests/mt/2008/doc/2008_NIST_MTOpenEval_Agmnt_StandardV3.pdf

[5] http://www.nist.gov/speech/tests/mt/2008/doc/mt08_constrained.html

[6] http://www.nist.gov/speech/tests/mt/2008/doc/mt08_constrained.html

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

up through the evaluation week and use data that had existed on or before June 30th 2007.

The Unconstrained training condition applies to tests all language pairs except *Urdu-to-English*.

3 Evaluation Data Sets

The NIST MT08 evaluation data sets will be defined as either Progress or Current test sets.

3.1 Progress Test Set

A Progress test set is newly collected data that systems process, and after turning in the system translations to NIST for scoring, all evidence of ever possessing the data is destroyed. The Progress test set data will be used again in future NIST Open MT evaluations in an effort to more directly calibrate year-to-year improvement.

To maximize the usage of the Progress test set data, many restrictions will be put in to place and only sites that have demonstrated the ability to fully and successfully participate in past NIST Open MT evaluations will be encouraged to process the Progress test set. First time participants will be handled on a case-by-case basis8.

3.2 Current Test Set

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

4 Evaluation Language Pairs (Tracks)

There are four language pairs offered for evaluation in MT08, and there are some subtle (and not so subtle) differences in the evaluation structure between them.

4.1 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 of equal size, both comprised of Newswire and Web source data, as seen in Table 1. The Current test set will be processed by all participants, and the Progress test set will be processed by those participants agreeing to the terms of usage.

The A2E test will be offered for both training conditions, Constrained and Unconstrained. Participants will be able to run systems from both training tracks for both data sets.

System translations will be evaluated using automatic metrics, human assessments, and the DLPT-star test.

4.2 Chinese-to-English

The Chinese-to-English (C2E) track will have the same evaluation structure as A2E. There will be two tests sets of equal size, both comprised of Newswire and Web source data, as shown in Table 1. The Current test set will be processed by all participants, and the Progress test set be processed by those participants agreeing to the terms of usage.

The C2E test will be offered for both training conditions, Constrained and Unconstrained.

System translations will be evaluated using automatic metrics, human assessments, and the DLPT-star test.

4.3 English-to-Chinese

The English-to-Chinese (E2C) track will be limited to Newswire data and will only be offered with a Current test set.

The E2C test will be offered for both training tracks, Constrained and Unconstrained. System translations will be evaluated using automatic metrics.

4.4 Urdu-to-English

The Urdu-to-English (U2E) track will be limited to an evaluation of a Current test set using equal amounts of Newswire and Web source data, see Table 1.

The U2E test will be offered for one training track, Constrained. For this test, system development is to be limited to the Urdu-resource-DVD that NIST will supply to participants registered for this track.

MIT-LL will host a WIKI site[9] for information exchange, related to the U2E test. While additional collected data is not allowed for system development, tools derived from the resource DVD, alternative mark-up or alignment of the provided data may be shared using the WIKI.

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

5 Performance Measurement

MT08 will make use of automatic metrics as well as human assessments of system translations.

5.1 Automatic MT Metrics

As in previous NIST Open MT evaluations, the primary metric for measuring performance will be the automatic N-gram co-occurrence scoring technique called BLEU-4.

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, originally developed by IBM[10], 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.[11]

NIST provides an N-gram co-occurrence evaluation tool as a downloadable software utility.[12] Research sites may use this utility to support their own research efforts, independent of NIST tasks/evaluations. All that is required, in addition to the source

[8] http://www.nist.gov/speech/tests/mt/2008/doc/MT08_ProgressTestForms.pdf

[9] Registered participants will receive the username, password and URL to the MIT-LL hosted WIKI.

[10] *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).

[11] http://www.nist.gov/speech/tests/mt/2008/doc/ngram-study.pdf

[12] http://www.nist.gov/speech/tests/mt/2008/scoring.html

language data, is a set of one (or more) reference translations of high quality.

Although BLEU-4 will be the official evaluation metric for MT08, NIST will run a suite of MT evaluation metrics as time and resources permit[13]. The results of alternate scoring techniques will be included as part of the public release of results.

5.2 Human Assessments of System Translations

As in previous MT evaluations, human assessments will be part of MT08, however, this year all human assessments will be performed using a participant-volunteer model.

Human assessments will be limited to one system per participant, which must be their primary system entered in either the Constrained or Unconstrained training task.

If a participant would like their system output to be included in what is assessed, the site will be required to perform some assessments of MT08 system translations. NIST will provide a web-based assessment tool[14]. There is a separate registration form[15] for participating in the human assessments.

There will be two types of human assessments.

5.2.1 Human Assessment – Adequacy

Human assessment of adequacy will be performed using the following model: An assessor will be presented with one reference translation and one system translation at a time. The assessor will decide on a scale how *adequate* the MT output is by judging how much pertinent information is preserved. For segments that the assessor gives one of the higher scores, the assessor will also provide a more global yes/no judgment of the adequacy of the MT. Visual clues will be given to the assessor to identify matches between the system translations and the reference translation. The assessor will continue the assessments segment-by-segment for a given document, although each successive segment will come from a randomly chosen system. Guidelines, a tutorial, and a training session are available on the NIST Open MT human assessment web site.

5.2.2 Human Assessment – Preferred Translation

The human assessments will include a second type of decision. For the preferred translation mode, an assessor will view one reference translation and two system translations. The assessor will select the MT output that they feel is more appropriate given the reference.

Preferred translation assessments will begin after all of the adequacy assessments are completed.

5.3 DLPT-star - MT Comprehension Tests

The Defense Language Institute (DLI) and MIT Lincoln Laboratory (MIT-LL) have collaborated to develop a MT Comprehension test protocol based on the Defense Language Proficiency Test (DLPT)[16] , a broadly accepted measure of effectiveness for evaluating foreign language proficiency. The adaptation of the DLPT for machine translation evaluation is known as the DLPT-star[17]. The NIST Open MT evaluation will include such a test, working in collaboration with DLI and MIT-LL.

Since these tests are labor intensive to create, the test materials will be based on the Progress test for Arabic-to-English and English-to-Chinese. This will allow for future rerunning of the same DLPT-star tests in conjunction with future NIST Open MT evaluations.

It is planned that up to six system translations will be chosen for inclusion in DLPT-star. System translations will be chosen based on the automatic scores and by the different algorithmic approaches.

6 NIST MT Data Format

NIST has defined a set of SGML tags that are used to format MT source, translation, and reference files for evaluation. Translation systems must be able to input the source documents and output translations that meet these formatting standards. All NIST MT source, translation, and reference files have a ".sgm" extension.

Evaluation data is packaged in SGML format as defined by the current MT DTD.[18] Translation output data must include the system designator. This system ID should contain site identification information and also provide unique identification of the system used to produce the output data. See section 6.2 for additional information regarding the required format for a system ID label.

For a submission to be valid there must be an output translation for each source document. Further, each output translation must have the same number of segments as the corresponding source document and these segments must appear in the same order as in the source document. Translation is to be performed only for data within the span of each segment tag. These segments contain only source language data.

6.1 Source File Format

Each evaluation source file is defined using a set of SGML tags. A source set begins with the tag (**srcset**) which is followed by several documents each defined by a (**doc**) tag. Each document consists of a series of segments that are defined with a (**seg**) tag. Each (**seg**) tag has an id attribute, which sequentially identifies the segments. Each tag has a corresponding closing tag. An example of a source file:

```
<srcset setid="mt-arab-v0" srclang="Arabic">
<DOC docid="NYT-doc1" genre="text">
<seg id="1"> ARABIC LANGUAGE TEXT </seg>
<seg id="2"> ARABIC LANGUAGE TEXT </seg>
...
</DOC>
<DOC docid="NYT-doc2">
...
</DOC>
</srcset>
```

Note: Test data may contain other SGML tags such as (but not limited to) "<h1>" or "<p>". For the purpose of evaluation, only the native language text that is surrounded by a (**seg**) tag is to be translated. Details regarding the "source" file format may be found in Appendix A.

6.2 Translation (Test) File Format

Each set of translations must adhere to the NIST MT data format. A single translation file may have results for several systems, but they must all be translations of the same source set. A translation

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

[14] For a description, see http://www.nist.gov/speech/tests/mt/2008/ha.

[15] http://www.nist.gov/speech/tests/mt/2008/doc/MT08_HumanAssessmentsForm.pdf

[16] http://en.wikipedia.org/wiki/Defense_Language_Proficiency_Tests

[17] http://www.ll.mit.edu/IST/pubs/0510_Jones.pdf

[18] http://www.nist.gov/speech/tests/mt/2008/doc/mteval.dtd

set begins with the tag (**tstset**) which is followed by one or more systems' translations. The translation test set file format is very similar to the source set file format. An example follows:

```
<tstset setid="mt-arab-v0" srclang="Arabic"
trglang="English">
<DOC docid="NYT-doc1"  genre="text"
sysid="NIST_arabic_constrained_primary">
<seg id="1"> TRANSLATED ENGLISH TEXT </seg>
<seg id="2"> TRANSLATED ENGLISH TEXT </seg>
…
</DOC>
<DOC docid="NYT-doc2"
sysid="NIST_arabic_constrained_primary">
…
</DOC>
</tstset>
```

Note: this translation file may contain results for more than one system simply by adding the additional translations between the (**tstset**) tags (identified by a different "**sysid**"). Details regarding the translation file format may be found in Appendix A. The "**sysid**" attribute for the translation file must conform to the following format:

sysid=<site-id>_<language>_<condition>_<type>

where

<site-id> is a short name identifying the site.

<language> is "arabic", "chinese", "urdu", or "english".

<condition> is either "constrained" or "unconstrained".

<type> is either "primary" or "contrastX" where "X" is an integer from 1..N uniquely identifying the contrastive system that produces the translation. There can only be one primary system per language and condition combination.

6.3 REFERENCE FILE FORMAT

The format of MT reference files is exactly the same as is used for the translation files, except that reference files use a (**refset**) tag in place of the (**tstset**) tag. Details regarding the reference file format may be found in Appendix A.

7 EVALUATION DATA GENRES

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

Source documents will be UTF-8 encoded.

Systems will be evaluated separately on each language and for each training condition. 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 1 provides details on how the evaluation data will be divided into selected subsets for reporting results.

Systems will not have prior knowledge of the data type (Newswire or Web data) of each source document.

Data Statistics for MT08 (approximate word counts)			
Source Language	Genre	Current set	Progress set
Arabic	Newswire	20,000	20,000
	Web	15,000	15,000
Chinese	Newswire	20,000	20,000
	Web	15,000	15,000
English	Newswire	40,000	N/A
	Web	N/A	N/A
Urdu	Newswire	20,000	N/A
	Web	20,000	N/A

Table 1: MT08 Evaluation Data Sets

7.1 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 1, above.

7.2 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 1.

8 EVALUATION PROCEDURES

There are ten steps in the MT08 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 11 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 are to be provided to NIST on the MT08 Registration form.

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 instructions in section 9. 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 data set.*

6 *Send in the system description (following the template provided) by the deadline.* See section 11 for details.

7 *Receive the evaluation results.* NIST will score the submitted system translations and distribute the evaluation results to the participants. See section 11 for more details.

8 *Receive the complete set of reference translations for the Current test sets.* Once the evaluation is complete, the set of reference translations used for evaluation will be available to the evaluation participants. This is intended to support error analysis and further research and to prepare 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, to share knowledge gained, and to plan for the next evaluation. A knowledgeable representative from each participating site is **required** to attend this workshop and to describe their technology and research and present their research findings. Attendance at this 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.

We stress to new participants 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.[12] Further, NIST strongly encourages sample submissions for previous NIST Dry Run, Evaluation, and Development test sets to verify proper formatting, following the procedure outlined in the next section. It is vitally important that all those planning to participate in MT08 verify that they are prepared for the formal evaluation by making successful submissions of practice data sets.

9 Submitting Translations to NIST

Participants in the evaluation may submit translations for one or all of the MT test languages. Participants may also submit translations for one or both of the training conditions. Furthermore, evaluation participants may submit up to three sets of translations for each language/condition. Each submission must be complete, however, in order to be acceptable.

9.1 System Translations

E-mail is the preferred method[19] for sites to submit their system translations to NIST. MT08 participants are to send the translations to mt_poc@nist.gov.

To properly package a translation file for submission to NIST, follow these 4-steps:

1. Create a directory that identifies your site (i.e., ./NIST)
2. Create a sub-directory for each source language attempted (i.e., ./NIST/Arabic, ./NIST/Chinese, ./NIST/Urdu, ./NIST/English)
3. Put the properly formatted translation files in the corresponding sub-directories. Each translation file must have a ".sgm" extension (e.g., ./NIST/Arabic/NIST-primary.sgm).
4. Create the compressed tar file using the Unix **tar** and **gzip** commands. (tar –cf NIST.tar ./NIST; gzip NIST.tar)
5. Send the file as an attachment to mt_poc@nist.gov.

9.2 System Descriptions

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

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

10 Guidelines For Publication of Results

NIST Speech Group's HLT evaluations are moving towards an open model which promotes interchange with the outside world. The rules governing the publication of MT08 evaluation results are the same as were used the previous year.

10.1 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 (and the alternative metric scores) achieved for each task, and for each training condition. Scores will be reported for the overall test set for each source language processed, and also for the following subsets of the evaluation test data:

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

When available, results from the DLPT-star test and results from the participant-based human assessments will also be posted.

10.2 Participants' Reporting of Results in Publications

Participants will be free to publish results for their own system, but will not be allowed to cite another site's results without permission from the other site. Publications should not identify the other participating sites, but may point to the NIST paper as a reference.

11 Schedule

Date	Event
June 27 2007	Evaluation Plan released.
July 1 2007	Training data cutoff date. All data created, posted, or published during or after this date is off-limits for system training and development.
Approx. August 31 2007	Urdu resources DVD available.
January 10 2008	Registration Deadline.
January 28 2008 9:00am EST	Evaluation test data e-mailed to participants.
February 1 2008 12 noon EST	Deadline for on-time submission of results to NIST.
February 8 2008	Preliminary release of results to participants.
February 15 2008	Deadline for submitting system description.
March 27-28 2008	Evaluation workshop, open to participants and government sponsors only, held in the Baltimore/Washington DC area.
May 23 2008	Official public release of results.

[19] If sending translations as an e-mail attachment is not possible, contact mt_poc@nist.gov to make other arrangements.

[20] http://www.nist.gov/speech/tests/mt/2008/doc/template_sys_desc.txt

Appendix A: NIST MT Data Format

1. Source File Format

The source file contains the source documents to be translated. The format of the source file is defined by the current MT DTD.[18] The source file begins with a **<srcset>** tag which contains a set of documents. Each document, defined by the **<doc>** tag, contains a set of segments. Each segment, defined by the **<seg>** tag, contains the source text to be translated.

The **<srcset>** tag has two required attributes—**setid** and **srclang—** and one implied attribute **trglang**. The **setid** attribute contains the name of the set of documents to be translated. This name is globally unique, meaning that no other source files will have that same name. The **srclang** attribute identifies the language of the source set, and for MT08 it can be one of these two values: **Arabic** or **Chinese**. The **trglang** attribute identifies the language of the system translations and is usually not specified in the source file.

The **<doc>** tag has two required attributes **docid** and **genre**, and one implied attribute **sysid**. The **docid** attribute contains the name identifying the document within the given source set. The genre attribute indicates the type of data for a given documents. The **sysid** attribute is usually not specified in the source file.

The **<seg>** tag has an implied attribute called **id**. The **id** attribute contains a number identifying the segment within the given document.

For example,

```
<srcset setid=”mt04-arab-evalset-v0” srclang=”Arabic”>
       <doc docid=”NYT-doc1” genre=”text”>
              <seg id=”1”> ARABIC LANGUAGE TEXT </seg>
              <seg id=”2”> ARABIC LANGUAGE TEXT </seg>
              …
       </doc>
       <doc docid=”NYT-doc2” genre=”text”>
       …
       </doc>
       …
</srcset>
```

2. Translation File Format

The translation file contains the system (or systems) output translations to be evaluated. The translation file format is also defined by the current MT DTD.[18] The translation file begins with a **<tstset>** tag which contains a set of documents. Each document, defined by the **<doc>** tag, contains a set of segments. Each segment, defined by the **<seg>** tag, contains the translated text.

The **<tstset>** tag has two required attributes—**setid** and **srclang—**and one implied attribute **trglang**. The **setid** attribute contains the name of the set of documents that has been translated. This name must match the **setid** of the source file for which system performed the translation. The **srclang** attribute indicates the language of the source set, and for MT08 it can be one of these two values: **Arabic** or **Chinese**. The **trglang** attribute indicates the language of the translated set, and for MT08 it is **English**.

The **<doc>** tag has two required attributes **docid** and **genre,** and one implied attribute **sysid**. The **docid** attribute contains the name identifying the document within the given source set. The genre attribute indicates the type of data for a given documents. The **sysid** attribute contains the name of the system that performed the translation. This attribute allows outputs from multiple systems to exist in the same translation file.

The **<seg>** tag has an implied attribute called **id**. The **id** attribute contains a number identifying the segment within the given document.

Note that the translation file must contain the same number of segments as that of the source file and that these segments must appear in the same order as the order they appear in the source file.

For example,

```
<tstset setid=”mt04-arab-evalset-v0” srclang=”Arabic” trglang=”English”>
       <doc docid=”NYT-doc1” sysid=”NIST_arabic_constrained_primary”>
              <seg id=”1”> TRANSLATED ENGLISH TEXT </seg>
              <seg id=”2”> TRANSLATED ENGLISH TEXT </seg>
              …
       </doc>
       <doc docid=”NYT-doc2” sysid=”NIST_arabic_constrained_primary”>
```

```
        …
        </doc>
        <doc docid=”NYT-doc1” sysid=”NIST_arabic_constrained_contrast1”>
        …
        </doc>
        <doc docid=”NYT-doc2” sysid=”NIST_arabic_constrained_contrast1”>
        …
        </doc>
        …
</tstset>
```

3. Reference File Format

The reference file contains high quality human output translations that NIST uses to evaluate the system output translations. The reference file format is also defined by the current MT DTD[18]. The reference file begins with a **<refset>** tag which contains a set of documents that has been translated by human translators. Each document, defined by the **<doc>** tag, contains a set of segments. Each segment, defined by the **<seg>** tag, contains the translated text.

The **<refset>** tag has two required attributes—**setid** and **srclang—**and one implied attribute **trglang**. The **setid** attribute contains the name of the set of documents that has been translated. This name must match the **setid** of the source file for which human translators performed the translation. The **srclang** attribute indicates the language of the source set, and for MT08 it can be one of these two values: **Arabic** or **Chinese**. The **trglang** attribute indicates the language of the translated set, and for MT08 it is **English**.

The **<doc>** tag has two required attributes **docid** and **genre,** and one implied attribute **sysid**. The **docid** attribute contains the name identifying the document within the given source set. The genre attribute indicates the type of data for a given documents. The **sysid** attribute contains the name of the human translator who performed the translation. This attribute allows outputs from multiple human translators to exist in the same reference file.

The **<seg>** tag has an implied attribute called **id**. The **id** attribute contains a number identifying the segment within the given document.

For example,

```
<refset setid=”mt04-arab-evalset-v0” srclang=”Arabic” trglang=”English”>
        <doc docid=”NYT-doc1” genre=”text” sysid=”LDC-trans1”>
                <seg id=”1”> TRANSLATED ENGLISH TEXT </seg>
                <seg id=”2”> TRANSLATED ENGLISH TEXT </seg>
                …
        </doc>
        <doc docid=”NYT-doc2” genre=”text” sysid=”LDC-trans1”>
        …
        </doc>
        <doc docid=”NYT-doc1” genre=”text” sysid=”LDC-trans2”>
        …
        </doc>
        <doc docid=”NYT-doc2” genre=”text” sysid=”LDC-trans2”>
        …
        </doc>
        …
</refset>
```