The 2006 NIST Machine Translation Evaluation Plan (MT06)

1 INTRODUCTION

The 2006 NIST Machine Translation evaluation (MT06) 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 2006 evaluation requires the translation of text data from a given source language into the target language. The source languages under test are Arabic and Chinese, and the target language under test is English. The text data will consist of newswire text documents, web-based newsgroup documents, human transcription of broadcast news, and human transcription of broadcast conversations.

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 present their results. Participants are expected to discuss 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 MT web site.[3]

[1] There is a nominal registration fee associated with attending the evaluation workshop. This fee is normally between $200 and $400, and does not include travel or accommodation expenses.

[2] The 2006 Machine Translation Registration form is online at: http://www.nist.gov/speech/tests/mt/doc/RegistrationForm-mt06.pdf. Contact NIST (mt_poc@nist.gov) if you have difficulties registering.

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

2 PERFORMANCE MEASUREMENT

As in previous NIST MT evaluations, performance will be measured using an 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[4], and this organization must be preserved in the translation. An N-gram, in this context, is simply a ***case sensitive***[5] 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[6], 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.[7]

NIST provides an N-gram co-occurrence evaluation tool as a downloadable software utility.[8] 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 language data, is a set of one (or more) reference translations of high quality.

Although BLEU-4 will be the official evaluation metric for MT06, NIST will run a suite of MT evaluation metrics as time and resources permit[9]. The results of alternate scoring techniques will not be released as part of the public release of results, but will be available to the participants and discussed at the NIST evaluation workshop.

As in previous MT evaluations, human assessments[10,11] will be part of MT06. Up to six participating systems will be assessed by

[4] Note to GALE contractors: This is a key difference between the source data for the GALE program and the source data for the NIST open MT evaluation. The source data provided for the GALE Translation evaluation is **not** segmented.

[5] Note: Beginning with MT-04, systems are ***only*** scored using case sensitive reference translations. Systems are penalized if their output words do not have the correct case.

[6] *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)

[7] http://www.nist.gov/speech/tests/mt/doc/ngram-study.v26.pdf

[8] http://www.nist.gov/speech/tests/mt/resources/scoring.htm

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

[10] In the past, the quality of MT system outputs was assessed by human judges with respect to both "adequacy" and "fluency". BLEU scores have been observed to correlate well with these

human judges for "adequacy". NIST will select four systems based on the relative rankings of BLEU scores, while two systems will be hand selected based on information provided in the system descriptions.

The assessments will be performed for about 10,000 words of newswire data. For comparisons, several human translations will be included in what is to be assessed.

3 EVALUATION 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. The categories differ solely by the specification of the data that may be used for system training and development. The evaluation conditions are called "Unlimited Data", and "Large Data".

3.1 (ALMOST) UNLIMITED DATA

For the Unlimited Data condition there are only two restrictions on the data that may be used for system development. First, the data must be publicly available, at least in principle.[12] This ensures that research results are broadly applicable and accessible to all participants. Second, participants may not use data that was created on or after February 1st, 2006 to develop their system (*this is the month from which the evaluation data set will be drawn*). Participants may, however, continue to search the web up through the evaluation week and use data that had existed on or before January 31st 2006. This is the basic test condition that applies to tests in both languages.

3.2 LARGE DATA

In addition to the time restriction of the Unlimited Data condition, the Large Data condition limits the use of resources to those available from the LDC. The Large Data condition applies to tests in both languages.

Participants who are not current members of the LDC will be required to sign a license agreement[13] which governs the use of LDC's data resources available for system development in preparation for the MT06 evaluation.

4 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 the SGML format as defined by the current MT DTD.[14] 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 4.2 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.

4.1 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.

4.2 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 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_large_primary">
<seg id="1"> TRANSLATED ENGLISH TEXT </seg>
<seg id="2"> TRANSLATED ENGLISH TEXT </seg>
…
</DOC>
<DOC docid="NYT-doc2"
sysid="NIST_arabic_large_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 either "arabic" or "chinese".

<condition> is either "large" or "unlimited".

<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.

---

human *quality* measures. MT *utility* is being addressed in the GALE program (http://www.nist.gov/speech/tests/gale/).

[11] http://projects.ldc.upenn.edu/TIDES/Translation/TransAssess04.pdf

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

[13] http://www.nist.gov/speech/tests/mt/doc/LDCLicense-mt06.pdf

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

4.3 REFERENCE FILE FORMAT

The format of MT reference files are 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.

5 EVALUATION DATA

The NIST MT06 evaluation will leverage evaluation data resources with DARPA's Global Autonomous Language Exploitation (GALE) program[15], allowing for a much larger and more diverse MT06 evaluation test set than would otherwise occur. NIST will use all the data being used for GALE Translation evaluation and add to it greater or equal amounts of non-GALE data from the same sources and epoch. Evaluation test sets will be created separately for each language under test.

Participants who are not GALE-funded contractors will be required to sign a data usage license agreement[16] that governs the use of the GALE portion of the evaluation data.

Each test set will be drawn from four types of data (newswire texts, newsgroup texts, human transcripts of broadcast news, and human transcripts of broadcast conversations – or talk shows). 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 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, newsgroup, broadcast news, or broadcast conversations) of each source document. They will however, have knowledge of the original genre, "text" or "audio". This will be defined as an attribute of the "<DOC" tag in the source data.

Domain	Approximate size by English reference	Date of test epoch
Data used in the GALE evaluation		
Newswire text	10,000 words	Feb. 2006
Newsgroup text	10,000 words	Feb. 2006
Broadcast News transcripts	10,000 words	Feb. 2006
Broadcast Conversation transcripts	10,000 words	Feb. 2006
Additional Data for NIST open MT evaluation		
Newswire text	20,000 words	Feb. 2006
Newsgroup text	10,000 words	Feb. 2006
Broadcast News transcripts	10,000 words	Feb. 2006

Table 1: MT06 Evaluation Data sets

5.1 NEWSWIRE TEXT

A portion of the test set will contain "news" 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 overall size of the newswire test data will be approximately 30,000 words (of English translation), which is essentially the same as in past MT evaluations.

5.2 NEWSGROUP TEXT

A portion of the test set will contain "newsgroup" data, which is a new data source being used in the NIST MT06 evaluation. There will be approximately 20,000 words from several sources. Newsgroup data is world-wide-web data from user forums and discussion groups.

5.3 BROADCAST NEWS TRANSCRIPTS

The portion of the test set will include manually created transcripts from audio "broadcast news" reports. There will be approximately 20,000 words of broadcast news transcript data. Broadcast news transcripts will include verbalized hesitations and repetitions, and punctuation is added to make the transcript grammatical[17].

5.4 BROADCAST CONVERSATION TRANSCRIPTS

The portion of the test set will include manually created transcripts from audio "broadcast conversation" shows. There will be approximately 10,000 words of broadcast conversation data. Broadcast conversation transcripts will include verbalized hesitations and repetitions, and punctuation is added to make the transcript grammatical.

6 EVALUATION PROCEDURES

There are eight steps in the MT06 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 the schedule in *section 9*.

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 address(es) to receive the evaluation source sets is provided to NIST on the MT06 Registration form.

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

4 *Return the translations.* The translations are to be returned to NIST via email according to instructions in *section 7*. Translations for each language must be submitted separately.

5 *Receive the evaluation results.* NIST will score the submitted system translations and distribute the evaluation results to the participants. See *section 9* "Schedule" for more details.

6 *Receive the complete set of reference translations.* 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.

7 *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.

8 *Attend the evaluation workshop.* NIST sponsors a follow-up evaluation workshop where evaluation participants and government sponsors meet to review evaluation results, to

[15] There is a machine translation component to the GALE program, but the evaluation is limited to GALE funded sites.

[16] http://www.nist.gov/speech/tests/mt/doc/LDC_GALE_mt06.pdf

[17] The additional broadcast news transcripts will follow the guidelines posted at: URL

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.

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.[8] Further, an email evaluation facility[18] is continuously available and is accepting submissions for all previous NIST Dry Run, Evaluation, and Development test sets. It is vitally important that all those planning to participate in the MT06 evaluation verify that they are prepared for the formal evaluation by making successful submissions of these practice data sets.

7 SUBMITTING TRANSLATIONS TO NIST

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

7.1 SYSTEM TRANSLATIONS

E-mail is the preferred method[19] for sites to submit their system translations to NIST. Unlike previous MT evaluations where participants sent the translations to the automatic scorer, MT06 participants are to send the translations to mt_poc@nist.gov. The automatic scorer will be ready to handle MT06 evaluation data immediately following the first preliminary release of results to the participants. This will allow for convenient pre-workshop analysis by the sites without being exposed to the reference data.

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. Put the properly formatted translation files in the directory. Each translation file must have a ".sgm" extension (i.e., ./NIST/NIST-primary.sgm).
3. Create the compressed tar file using the Unix **tar** and **gzip** commands. (tar –cf NIST.tar ./NIST; gzip NIST.tar)
4. Send the file as an attachment to mt_poc@nist.gov.

7.2 SYSTEM DESCRIPTIONS

Sites are required to prepare a system description for each system submitted for evaluation. g. As a minimum, the system description must include:

- The identity of the system
- Evaluation condition (Unlimited vs. Large)
- A description of the algorithmic approach
- Key differences between multiple submissions

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

8 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 MT06 evaluation results are the same as were used for MT05.

8.1 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 achieved for each task, and for each 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 "Newsgroup" text documents
3. The "Broadcast News" transcript documents
4. The "Broadcast Conversation" transcript documents

8.2 PARTICIPANT'S 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.

9 SCHEDULE

Date	Event
01 February '06	Training data cut-off date. All data created, posted, or published during or after this month is "off-limits" for system training and development.
31 May '06	Registration Deadline for the Chinese-to-English and Arabic-to-English tasks.
24 July '06 9 am EDT	Registered participants will receive the Chinese and Arabic evaluation source data via Email.
28 July '06 12 noon EDT	Deadline for ON-TIME results submitted to NIST for Email scoring.
02 August '06	Preliminary release of results to the participants.
Tentative: 6-7 September '06	Workshop for evaluation participants and government sponsors of MT research, to be held in the Baltimore/Washington D.C. area.
October 1st, 2006	Official public release of results.

[18] http://www.nist.gov/speech/tests/mt/doc/autoscore.htm

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

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.[14] 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 MT06 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.[14] 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 MT06 it can be one of these two values: **Arabic** or **Chinese**. The **trglang** attribute indicates the language of the translated set, and for MT06 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_large_primary">
              <seg id="1"> TRANSLATED ENGLISH TEXT </seg>
              <seg id="2"> TRANSLATED ENGLISH TEXT </seg>
              …
```

```
        </doc>
        <doc docid="NYT-doc2" sysid="NIST_arabic_large_primary">
        …
        </doc>
        <doc docid="NYT-doc1" sysid="NIST_arabic_large_contrast1">
        …
        </doc>
        <doc docid="NYT-doc2" sysid="NIST_arabic_large_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[14]. 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 MT06 it can be one of these two values: **Arabic** or **Chinese**. The **trglang** attribute indicates the language of the translated set, and for MT06 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>
```