The 2005 NIST Machine Translation Evaluation Plan (MT-05)

1 INTRODUCTION

The 2005 NIST Machine Translation evaluation (MT-05) is part of an 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 2005 evaluation may be viewed as two tasks. Each task requires performing translation 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.

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, participants are required to attend the follow-up evaluation workshop and are expected to discuss their research findings in detail. For more information, visit the MT web site.[1] To participate in the evaluation, sites must officially register with NIST.[2]

2 PERFORMANCE MEASUREMENT

Performance will be measured using both human assessments and automatic N-gram co-occurrence scoring techniques for MT-05.[3] Both of these techniques evaluate 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.

2.1 HUMAN ASSESSMENTS

Human judges will assess translation quality with respect to both the "adequacy" of the translation and its "fluency". This technique was used by DARPA in its MT evaluations during the early 1990's and has been adapted and refined by the LDC for the current series of evaluations. The assessments will be performed by native (monolingual) speakers of American English.

Adequacy is judged by comparing each translated segment with the corresponding segment of a high quality reference translation. A segment's adequacy is scored according to how well the meaning of the test translation matches the meaning of the reference translation. Fluency is scored independent of the source or any reference translation. Additional information on the human assessment technique is available from the LDC's web site.[4]

2.2 N-GRAM CO-OCCURRENCE SCORING

Translation quality will be measured automatically using N-gram co-occurrence statistics. 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.

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 amount of data that may be used for system training

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

[2] The 2005 Machine Translation Registration form is online at: http://www.nist.gov/speech/tests/mt/doc/RegistrationForm-mt05.pdf
Contact Mark Przybocki (Mark.Przybocki@nist.gov) if you have difficulties registering.

[3] Subsequent evaluations may use only automated scoring if that proves adequate.

[4] http://www.ldc.upenn.edu/Projects/TIDES/Translation/TranAssessSpec.pdf

[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.pdf

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

and development. The evaluation conditions are called "Unlimited Data", and "Large Data".[9]

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.[10] This ensures that research results are broadly applicable and accessible to all participants. Second, participants may not use data that was created (or posted on the web) after December 1st 2004 to develop their system. Participants may, however, continue to search the web up through the evaluation week and use data that had existed before December 1st 2004. This is the basic condition that applies to both tasks.

3.2 LARGE DATA

In addition to the restrictions of the Unlimited Data condition, the Large Data condition limits the use of bilingual resources to the parallel corpora and bilingual dictionaries available from the LDC. The Large Data condition is a contrastive test that applies to both tasks.

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.[11] 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.

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">
<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 "<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" sysid="NIST-primary-
large">
<seg id=1> TRANSLATED ENGLISH TEXT </seg>
<seg id=2> TRANSLATED ENGLISH TEXT </seg>
…
</DOC>
<DOC docid="NYT-doc2" sysid="NIST-primary-
large">
…
</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.

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

Evaluation test sets will be created separately for each language under test.

Systems will be required to process the entire test set for each source language attempted.

Source documents will be UTF-8 encoded for Arabic and will be GB-encoded for Chinese.

Participants in the evaluation may submit translations for both MT tasks. Participants may also submit translations for one or both of the training data conditions. Each submission must be complete, however, in order to be acceptable.

Evaluation participants may submit one or more sets of translations for each such test.

The first submission of test results from a participating site will be treated as the site's primary submission, and the term 'primary' must be part of the name of the system identified in the 'sysid' field. All subsequent submissions will be deemed to be 'contrast' submissions."

Systems will be evaluated separately on each language and for each training condition. System performance will be reported over the entire test set. Unlike the MT-04 test set, this year's test set will contain only "news" stories. These "news" stories are similar to those used in past MT evaluations. These stories may

[9] Previous evaluations have supported a "Small Data" task. This condition is no longer part of the NIST MT evaluations.

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

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

be drawn from several kinds of sources including newswire, broadcast news, and the web. There will be 100 stories for each source language. The overall size of the newswire test data will be approximately 25,000 words (of English translation), which is essentially the same as in past MT evaluations.

6 EVALUATION PROCEDURES

There are eight steps in the MT-05 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 MT-05 Registration form.

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

4 *Upload the translations.* The translations are uploaded via email according to instructions in *section 7*. Translations for each language must be submitted separately. Each submission file will be validated against the MT DTD.[11] Submission files that fail this validation check will not be scored, and the submitting site will be notified of the failure by e-mail.

5 *Receive the evaluation results.* The system output submissions are evaluated using NIST's automatic scoring utility and the results of this evaluation are returned to the submitter's email address. (The MT auto-scorer uses the email address in the *FROM* field.) This process is automatic and the site usually receives results within a few minutes of submission. Human judgments obviously take much longer and those results will be presented shortly after the evaluation workshop.

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, share knowledge gained, and 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, the email evaluation facility 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 MT-05 evaluation verify that they are prepared for the formal evaluation by making successful submissions of these practice data sets.

7 SUBMITTING TRANSLATIONS TO NIST

E-mail is the preferred method for sites to submit their system translations and system descriptions to NIST

7.1 SYSTEM TRANSLATIONS

E-mail is the preferred method[12] for sites to submit their system translations to NIST. Translations should be sent to the automatic e-mail scorer, mteval@nist.gov.

To properly package a translation file for the automatic e-mail scorer, follow these 4-steps:

1. Create a directory that identifies your site (i.e., ./NIST)
2. Put the properly formatted translation file in the directory. Only one file with a ".sgm" extension should be placed in this directory. (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 mteval@nist.gov.

The e-mail scorer accepts compressed tar files that have the extension *.tar.gz.

7.2 SYSTEM DESCRIPTIONS

Sites should prepare a system description for each system submitted for evaluation. These system descriptions must be sent to NIST separately from the translations that are submitted for automatic scoring. 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.

8 GUIDELINES FOR PUBLICATION OF RESULTS

THIS IS A NEW SECTION OF THE EVALUATION PLAN, PLEASE READ CAREFULLY BEFORE REGISTERING YOUR PARTICIPATION IN THE NIST 2005 MACHINE TRANSLATION EVALUATION.

NIST Speech Group's HLT evaluations are moving towards an open model which promotes interchange with the outside world. Therefore, the rules governing the publication of MT-05 evaluation results have been modified.

8.1 NIST PUBLICATION OF RESULTS

1. 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.

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

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

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 December 2004	Cut-off date for training data Chinese-to-English and Arabic-to-English tasks.
30 April '05	Registration Deadline for the Chinese-to-English and Arabic-to-English tasks.
09 May '05 9 am EDT	Registered participants will receive the Chinese and Arabic evaluation source data via Email.
13 May '05 12 noon EDT	Deadline for ON-TIME results submitted to NIST for Email scoring.
18 May '05	Composite results released.
27 May '05	Hotel Registration Deadline (to get a special room rate).
13 June '05	Workshop Registration Deadline.
20-21 June '05	Workshop for evaluation participants and government sponsors of MT research, to be held at the North Bethesda Marriott Hotel and Conference Center.
TBA	Human Assessments will be distributed to the participants.

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.[11] 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 MT-05 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 one required attribute **docid** and one implied attribute **sysid**. The **docid** attribute contains the name identifying the document within the given source set. 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">
              <seg id=1> ARABIC LANGUAGE TEXT </seg>
              <seg id=2> ARABIC LANGUAGE TEXT </seg>
              …
       </doc>
       <doc docid="NYT-doc2">
       …
       </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.[11] 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 MT-05 it can be one of these two values: **Arabic** or **Chinese**. The **trglang** attribute indicates the language of the translated set, and for MT-05 it is **English**.

The <**doc**> tag has one required attribute **docid** and one implied attribute **sysid**. The **docid** attribute contains the name identifying the document within the given source set. 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-primary">
              <seg id=1> TRANSLATED ENGLISH TEXT </seg>
              <seg id=2> TRANSLATED ENGLISH TEXT </seg>
              …
       </doc>
       <doc docid="NYT-doc2" sysid="NIST-primary">
```

```
        …
        </doc>
        <doc docid="NYT-doc1" sysid="NIST-contrast">
        …
        </doc>
        <doc docid="NYT-doc2" sysid="NIST-contrast">
        …
        </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[11]. 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 MT-05 it can be one of these two values: **Arabic** or **Chinese**. The **trglang** attribute indicates the language of the translated set, and for MT-05 it is **English**.

The **<doc>** tag has one required attribute **docid** and one implied attribute **sysid**. The **docid** attribute contains the name identifying the document within the given source set. 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" sysid="LDC-trans1">
                <seg id=1> TRANSLATED ENGLISH TEXT </seg>
                <seg id=2> TRANSLATED ENGLISH TEXT </seg>
                …
        </doc>
        <doc docid="NYT-doc2" sysid="LDC-trans1">
        …
        </doc>
        <doc docid="NYT-doc1" sysid="LDC-trans2">
        …
        </doc>
        <doc docid="NYT-doc2" sysid="LDC-trans2">
        …
        </doc>
        …
</refset>
```