The 2004 NIST Machine Translation Evaluation Plan (MT-04)

1 INTRODUCTION

The 2004 NIST Machine Translation evaluation (MT-04) 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 research community, and
- Coordinates workshops to discuss MT research findings and results 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 2004 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 participants that find the tasks and the evaluation of interest. There is no fee for participation. However, participants are required to attend a follow-up 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-04.[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. Details of the human assessment technique may be accessed on 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, tokens include words and punctuation.) 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 (target) quality.

3 EVALUATION CONDITIONS

MT R&D requires language data resources, and system performance and R&D effort are strongly affected by the type and amount of resources used. Therefore three different resource categories have been defined as conditions of evaluation. These categories limit the data that may be used for system training and development. The evaluation conditions are called "Unlimited Data", "Large Data", and "Small 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.[9] This

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

[2] The 2004 Machine Translation Registration form is online at: http://www.nist.gov/speech/tests/mt/doc/RegistrationForm-mt04.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: For MT-04, systems will ***only*** be scored using case sensitive reference translations. Systems will be penalized if they do not output case sensitive translations.

[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

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

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 January 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 January 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. For the Large Data condition, parallel corpora and bilingual dictionaries are limited to those available from the LDC. The Large Data condition applies to both the Chinese-to-English and Arabic-to-English translation tasks for MT-04.

3.3 SMALL DATA

In addition to the restrictions of the Large Data condition, the Small Data condition limits the use of source language resources to those specified in Table 1. Further, no indirect use of data, including the use of corpus-trained tools such as LDC-provided tokenizers or parsers, is allowed. There are no limitations on the use of English data, however. The Small Data condition applies to the Chinese-to-English task, only. There is no Small Data condition for the Arabic-to-English task.

Table 1: Resources allowable for the Small Data condition

Chinese-to-English
The bilingual texts from the 100k-word UPenn Chinese treebank. (But the trees are not allowed to be used.)
The 10k-word dictionary from CMU (S. Vogel)

4 NIST MT DATA FORMAT

NIST has defined a set of SGML 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 MT source, reference, and translation files have a ".sgm" extension.

Evaluation source data is packaged in the SGML format for source data, according to the current MT DTD[10]. Translation output data must be packaged in the SGML format for translation output data. The output format includes the requirement for a system designator. This system ID should contain site identification information and also provide unique identification of the system used to produce the output data.

Note that 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 the segments. 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 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-ara-set-5" 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: There may exist 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.

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-ara-set-5" 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.

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.

5 EVALUATION DATA

This year, each system[11] will be required to process two separate source sets for each source language attempted. The resulting translations must be submitted separately for scoring (see below).

Chinese source documents will be GB-encoded. Arabic source documents will be in UTF-8.

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

Systems will be evaluated separately on each language, for each training condition, and on each source set. Evaluation participants may submit one or more sets of translations for each such test.

> **If a site submits more than one system output for any one test, one system must be declared in advance as the primary (best) submission. Furthermore, the first**

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

[11] A single system submitted for each of the three data conditions (unlimited, large and small) is to be considered three submissions and each should process both test sets for the given source language.

submission to the NIST automatic email scorer must contain the primary submission.

5.1 SOURCE SET I – NEWS STORIES

The first source 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, broadcast news, and the web. There will be approximately 100 stories for each source language, with approximately 420 Chinese characters per Chinese story, and about 160 Arabic words per Arabic story.

5.2 SOURCE SET II – NEW GENRE

The second source set will contain a genre of data that is new to NIST MT evaluations. The new genre data set may contain editorials, speeches, or technical documents. The intent is to challenge systems by supplying data that is unlike the data used for system development. The overall size of ***source set II*** will be about the same as that of ***source set I*** in character and/or word counts, and may be different in the number of overall documents.

6 EVALUATION PROCEDURES

There are seven steps in the MT-04 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 8*.

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-04 Registration form.

3 *Perform the translation.* Each site must run its translation system(s) on both source data sets for each language attempted.

4 *Upload the translations.* The translations are uploaded via email according to instructions in *section 7*. Translations for each source set must be submitted separately.

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 autoscorer 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 at the evaluation workshop, or shortly thereafter.

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 workshop. Some participants however, have refused to accept the reference translations so that they could continue to make blind submissions with the automatic email scorer.

7 *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 where they are expected 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 evaluation is open to all interested contributors, but NIST does not publish results outside the list of participants and government sponsors. Further, while participants are allowed to discuss their own results without restriction, disclosure of the results of other sites is not allowed. This restriction is to ensure the scientific focus of the evaluation, to make participation as collaborative as possible, and to encourage new participants and new approaches.

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-04 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[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.

8 SCHEDULE

Date (2004)	Event
01 January	Cut-off date for training data (section 3.1) Chinese-to-English and Arabic-to-English tasks.
30 April	Registration Deadline for the Chinese-to-English and Arabic-to-English tasks.
10 May 9 am EDT	Registered participants will receive the Chinese and Arabic evaluation source data via Email.
14 May 12 noon EDT	Deadline for ON-TIME results submitted to NIST for Email scoring.
20 May	Composite results released.
22-23 June	Workshop for evaluation participants and government sponsors of MT research, to be held in the Baltimore/Washington D.C. area.
TBA	Human Assessments will be distributed to the participants.

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