The NIST 2002 Machine Translation Evaluation Plan (MT-02)

1 INTRODUCTION

The NIST year 2002 Machine Translation evaluation (MT-02) is the first in a series of evaluations of human language translation technology. NIST is conducting 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 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 sponsors 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 2002 evaluation will evaluate translation from Chinese to English and from Arabic to 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 expected to attend a follow-up workshop and to discuss their research findings in detail at the workshop. For more information, visit the MT-02 web site.[1] To participate in the evaluation, please register with Mark Przybocki.[2]

2 PERFORMANCE MEASUREMENT

Performance will be measured using both human assessments and automatic N-gram co-occurrence scoring techniques for MT-02.[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 sequence of N words.) 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[5], 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-02 web site.[6]

NIST provides an N-gram co-occurrence evaluation tool as a downloadable software utility.[7] 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 categories 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.[8] This ensures that research results are broadly applicable and accessible to all participants. Second, March 15th 2002 is the cut-off date for collection of training data. Use of data published after this date and web crawling after this date are disallowed.

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 and listed on the MT-02 resource web page.[9]

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

---

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

[2] Mark Przybocki may be contacted either through email (Mark.Przybocki@nist.gov), or telephone (301/975-3169).

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

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

[5] *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)

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

[7] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-kit-v09.tar.gz

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

[9] http://www.nist.gov/speech/tests/mt/resources/index.htm

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 only to the Chinese-to-English translation task for MT-02. There is no Small Data condition for Arabic-to-English.

Table 1 Chinese language resources allowable for the Chinese-to-English Small Data condition[9]

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)

Source data for MT-02 will be news stories in Chinese and Arabic. These stories will be drawn from three kinds of sources – newswire, broadcast news, and the web. There will be approximately 100 stories for each source language, with approximately 250 words per story.

4 EVALUATION PROCEDURES

There are seven steps in the MT-02 evaluation process:

1 *Register to participate.* Each site desiring to participate in the evaluation must register with Mark Przybocki no later than the deadline for registration.[2]

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, according to the evaluation schedule. The appropriate email address to receive this data needs to be provided to NIST when registering to participate.

3 *Perform the translation.* Each site must then run its translation system(s) on the source data to produce the translated output data.

4 *Upload the translations.* The translations are uploaded via email according to instructions on the MT-02 web site.[10]

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 reply address. This process is automatic and the site usually receives results within minutes of submission. Human judgments obviously take much longer and those results will be presented at 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 supplied to the evaluation participants. This is intended to support error analysis and further research and to prepare for the workshop.

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. Each evaluation site is expected to describe their technology and research and present their research findings at this workshop. 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 evaluation results outside the community 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 imposed to ensure the scientific focus of the evaluation, to make participation as collaborative as possible, and to encourage new participants and new approaches.

Evaluation source data is packaged in the SGML format for source data, according to the MT-02 DTD[11][12]. 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.

Participants in the evaluation may submit translations for either the Chinese source data or the Arabic source data or both. 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 and each training condition. Evaluation participants may submit one or more sets of translations for each such test. If a site submits more than one system output for one test, however, one of these outputs must be declared in advance as the primary (best) submission.

The mteval utility to be used for the evaluation is available now for download from NIST for all those who are interested in using the tool.[7] Further, the email evaluation facility is continuously available and is accepting submissions for the December Chinese dry run source data (set_ID = "mt2001_devset_v0"[13]) and for an Arabic practice data set (set_ID = "mt2002_Arabic_p0"[14]). It is vitally important that all those planning to participate in the June evaluation verify that they are prepared for the formal evaluation by making successful submissions of these practice data sets.

5 SCHEDULE

Date (2002)	Event
7 June	Deadline for registration to participate
10 June **8 am** EDT	Evaluation data sent to registered evaluation sites via email
14 June **8 am** EDT	Deadline for email submission of results **(note the time)**
21 June	Reference data and official results released to evaluation participants
22-23 July	Workshop for evaluation participants and government sponsors of MT research, to be held in Santa Monica prior to the TIDES meeting

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

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

[12] Note to previous participants: The DTD has been changed since the last evaluation to make the MT evaluation tags and attributes conform to default SGML conventions. These changes must be accommodated in your processing.

[13] ftp://jaguar.ncsl.nist.gov/mt/data/mt2001_devset_v0.tar.gz

[14] ftp://jaguar.ncsl.nist.gov/mt/data/mt2002_arabic_p0.tar.gz