BOLT Activity A Machine Translation Evaluation Plan for Phase 2

1 INTRODUCTION

The goal of the Broad Operational Language Translation (BOLT) program is to create technology capable of translating multiple foreign languages in all genres, retrieve information from the translated material, and enable bilingual communication via speech or text. NIST is managing the evaluations for the various activities in BOLT. This evaluation plan is for Machine Translation (MT) of BOLT Activity A for Phase 2 of the program.

Specifically, the BOLT MT evaluation in this second year will test the translation into English of text drawn from "discussion forums" in Egyptian Arabic and Mandarin Chinese and from SMS/chat in those same two languages. Translation from those two source languages will be evaluated separately.

The evaluation will be limited to the two research teams funded to participate in BOLT.

2 EVALUATION TASK

The BOLT MT evaluation for Phase 2 will test system translation capabilities into English of text drawn from "discussion forums" and from SMS/chat, in Egyptian Arabic and Mandarin Chinese. Translation from those two source languages will be evaluated separately.

2.1 TEXT-TO-TEXT TRANSLATION

Text-to-Text translation will be the only translation technology mode evaluated in Phase 2. Translation of text tests a system's ability to translate foreign text data into understandable and accurate English text.

Systems must produce English text that completely captures the meaning conveyed by the source data, using easily understandable English.

3 DATA

The discussion form data for Phase 2 will be the discussion forum data from Phase 1, and reads as entire threads.[1] It is drawn from Egyptian Arabic and Mandarin Chinese language data from a variety of discussion forums.

The SMS/chat data is expected to be entire text messages.

The source data will be UTF-8 encoded.

3.1 TRAINING DATA

Discussion Forum data is already in the teams' hands from Phase 1. The primary source of SMS/chat training data will be data collected or accepted by the Linguistic Data Consortium (LDC) and incrementally distributed to the teams throughout the BOLT program. LDC will be accepting donated data as well as collecting data. The SMS/chat data will be collected separately in the three languages (Arabic, Chinese, and English), and the source data will be distributed to the BOLT MT developer teams. The Arabic data and Chinese data will be translated into English, and the translations will be distributed to the BOLT MT developer teams. None of the English data will be translated into Arabic or Chinese.

BOLT teams may use training data outside of the resources distributed by the LDC if that data is specifically authorized by DARPA and shared with all BOLT developer teams. All such data must be declared by August 9 and shared by August 16.

3.2 DEVELOPMENT DATA SET

A development set (**DEV**) of approximately 40k words of SMS/chat from each source language will be provided. This data will be drawn from the LDC training releases (at least R1 and R2). The **DEV** will be selected using the same procedures as will be used to select the evaluation data as described below (section 3.4).

The development dataset will be accompanied by a first-pass reference translation.

The two research teams will jointly select 5kw per language of SMS/chat data, from the development set, as a validation set to be evaluated by HTER. The edits on this 5kw will be released along with the HTER scores after the evaluation, so as to enable the research teams to do error analysis. The 5kw per language validation set for forums data will remain unchanged from Phase 1.

[1] A thread is data from a single discussion forum with an explicit initial topic. It consists of an *initial post* and zero or more *follow-up posts*.

3.3 EVALUATION DATA SET

The SMS/chat evaluation dataset (**EVAL**) will contain approximately 200k source words from each language. The discussion forum data is approximately 200k source words from each language. The discussion forum data will read as entire threads. Special steps will be followed to protect the **EVAL**, keeping its contents sequestered (or blind) throughout all phases of the BOLT program.

NIST will sub-select approximately 20k source words of SMS/chat from each language for the primary HTER scoring. That data will be used in addition to the approximately 20k source words of discussion forum data from each language (exactly the same discussion forum data that was used in phase 1).

The entire **EVAL** will be accompanied by a first-pass reference translation. The sub-selection for HTER scoring will have careful translations that include alternatives, referred to as "gold standard" references. In cases where the original source language is ambiguous, the reference data will contain allowable alternatives for words or phrases. Idioms will typically receive a literal translation and a translation that captures the intended meaning

3.4 DATA SELECTION PROCEDURES

The evaluation data will typically represent informal language. The discussion forum data is from threads with a focus on current or dynamic events. The SMS/chat data will have no restrictions on topic content.

Data will be chosen for the development and evaluation datasets in a way that reasonably resembles how the training data is chosen. The **DEV** and **EVAL** datasets will be chosen by parallel procedures so that they match each other reasonably well.

The LDC identified discussion forum data by a combination of hand-selection and automatic selection. The hand-selection process identified posts with the desired characteristics (such as Egyptian Arabic dialect and current events as the topic). Forums in which desired data had been identified were considered "promising" and data selection focused on such forums. An appreciable fraction of the **DEV** and **EVAL** datasets was chosen by automatic selection that was informed by the hand-selections.

The procedures for choosing the SMS/chat data are still under development.

4 DATA FORMATS

For discussion forum data, both the source language input and the target language output will be in the LDC's "multipost" XML format. For SMS/chat data, the format is not yet final but will closely resemble the multipost format, and that format will be used for both source-language input and MT output.

For discussion forum data, the source language data includes markup that identifies each post in the thread. Within each post, there is markup identifying the sentence-like units (SUs). BOLT systems will be required to include corresponding post and SU markup in their MT output, and that markup will be used to align the MT output with the reference translation for the purposes of HTER editing. Posts and SUs should appear in the target-language MT output in the same order as in the source-language inputs.

The markup for SMS/chat data has not yet been determined.

NIST will identify the data that is to be translated by the MT systems.

4.1 INPUT FORMATS

The MT discussion forum source-language data will be distributed in the LDC multipost data format. The SMS/chat source-language data will be in a similar format, still to be determined. All data will be UTF-8 encoded. Genre (forums vs. SMS/chat) will be made known to the systems during BOLT phase 2.

4.2 OUTPUT FORMATS

The system MT outputs for discussion forum data will be in the LDC multipost data format. The system MT outputs for SMS/chat data will be in a similar format, still to be determined. System output should be UTF-8 encoded.

5 SYSTEM SUBMISSIONS

5.1 DRY RUN

Teams are required to participate in a dry run. A single system submission for the **DEV** is to be submitted to NIST before July 15, using the same submission procedures and formats as for the actual evaluation.

No system scores for the dry run will be reported, and the quality of the MT will not be assessed. The purpose of the dry run is to validate that systems are producing output in the valid data formats and also to verify the submission procedures and the evaluation tool-chain.

5.2 EVALUATION SYSTEMS

5.2.1 Primary Systems

Each team is required to submit a single primary system. The primary system must be the first system submission and is the system that will be evaluated using the primary evaluation metric HTER.

5.2.2 Contrastive Systems

Each team is permitted to submit up to two additional (contrastive) systems. Scoring of contrastive systems will be limited to automated MT metrics. Reporting of contrastive system scores will be limited to the overall **EVAL** score. The intent of accepting contrastive systems is to evaluate alternate approaches, not to evaluate additional, later development efforts.

Late and/or debugged contrastive systems will not be accepted.

6 METRICS

6.1 PRIMARY EVALUATION METRIC

BOLT will use HTER, an edit-distance metric, to evaluate system translation quality. This will be accomplished by having a team of trained human editor(s) make changes to the MT output so that the resulting edited-MT output contains understandable English that conveys exactly the same information as the reference data. The editors will do so using as few edits as they can.

6.1.1 Post Editing Process

This section will be updated later to reflect plans for handling SMS/chat data. The final post-editing guidelines for SMS/chat are due to be finalized by August 23rd. The remainder of this section addresses the handling of discussion forum data.

NIST has developed an editing interface[2] designed for the post editing task. An editor will have access to the entire contents of the thread for full context of the post being edited.

The editor's focus will be on a single sentence-like unit (SU) at a time, and the editors will edit complete posts (all SUs in each selected post). The aligned reference and system translations will be displayed in two separate columns. Alternative words and phrases will be given to the editor in instances when the original source language data was ambiguous or if independent human translators did not agree on the exact meaning.

The editors will be given specific guidelines[3] to follow while performing the edits. The post editor will modify the SU under focus until the editor believes that the MT output completely captures the meaning conveyed in the reference data. The editors are instructed to make modifications using as few edits as possible. Although the editor will be looking at the aligned SUs, they will be free to use context before and after the SU currently in focus. See the post editing guidelines for more details.

Each translated document, by each system, will be post-edited by 2 editors. Both edited documents will be reviewed in a second pass. There will be quality control measures in place to verify that the post editors are performing their job in an acceptable manner.

The official HTER score will use the minimum HTER (at the SU level) between the two versions of the post edited document.

6.1.2 The HTER Edit Distance Metric

Software will be used to compute HTER scores by comparing the resulting edited-MT with the original MT and counting the number of edits. An edit is an insertion of a word, deletion of a word, replacement of a word, or a block move of a string (possibly of multiple words) from one location to another. Each edit is weighted equally. The number reported will be the ratio of the number of edits to the number of words in the gold standard reference data. In the case of alternative words and phrases, only the first choice listed will be counted as part of the reference.

HTER will be automatically calculated using BBN created software called `tercom.0.7.25.jar`[4].

NIST will report the mean HTER scores over the first-pass and second-pass edited data. The official HTER score is found by taking the lowest HTER segment score when comparing the two edited versions.

For the official evaluation, NIST will report HTER scores at the post level.

6.2 AUTOMATED MT METRICS

NIST will run BLEU, METEOR, and TER over the entire **EVAL** consisting of about 200k words per source

[2] The JAVA based post editing interface maybe accessed via the NIST GALE website at: http://www.itl.nist.gov/iad/mig/tests/gale/2008

[3] The post-editing guidelines may be accessed via the NIST BOLT website at: http://www.nist.gov/itl/iad/mig/bolt_p1.cfm
The previous GALE documents are at the URL http://www.itl.nist.gov/iad/mig/tests/gale/2008/

[4] The BBN supplied evaluation script is available via the NIST GALE website at: http://www.itl.nist.gov/iad/mig/tests/gale/2008/

language. Automatic metrics will use the first-pass translations as reference.

6.3 SENTENCE-LEVEL JUDGMENTS OF SEMANTIC ACCURACY

NIST expects to have a panel of bilingual judges perform sentence-level judgments of MT semantic accuracy over the same data on which HTER is performed. NIST may need to create software for this purpose.

7 EVALUATION PROCEDURES

7.1 DELIVERING SYSTEM OUTPUT TO NIST

Create a directory that names the BOLT team, all in lower-case:

"astral" or "delphi"

Under your team directory create any/all parts of the following structure that are relevant to your submission (this is *required*):

./arabic/forums/primary
./arabic/forums/contrastive_1
./arabic/forums/contrastive_2

./chinese/forums/primary
./chinese/forums/contrastive_1
./chinese/forums/contrastive_2

./arabic/sms_chat/primary
./arabic/sms_chat/contrastive_1
./arabic/sms_chat/contrastive_2

./chinese/sms_chat/primary
./chinese/sms_chat/contrastive_1
./chinese/sms_chat/contrastive_2

Place the system translations in their proper directory.

System translation files should have names that match the input file. For example a source file named

`bolt-arz-DF-123-200912-12345678.arz.su.xml`

should result in a target-language file named

`bolt-arz-DF-123-200912-12345678.eng.su.xml`

Within the MT output, the identity of each SU should be *exactly* the same as in the source-language file.

The submission file is to be assembled with tar and gzip. The submission file name will be an experiment ID that includes

- bolt
- phase (p2)
- team name (astral or delphi)
- submission
  (dryrun, validation, primary, contrastive1, or contrastive2),
- source language (arabic or chinese),
- genre (forums or smschat),
  with no hyphen or underscore here
- date/time when the submission was assembled by your team
  If it is 2013-August-27 at 17:30 edt, this would appear as `2013-08-27-1730edt`

A possible experiment ID is as follows

`bolt_p2_delphi_primary_arabic_smschat_2013-08-27-1730edt`

7.2 SYSTEM DESCRIPTION

A separate system description will be required for the BOLT evaluation. It will due at the end of October.

7.3 DELIVERING SYSTEMS TO NIST

A copy of the 2012 BOLT systems (used to generate the primary submissions) is to be delivered to NIST by 6 January 2014. During the week of January 6–10, after the evaluations are complete, the systems will be installed and verified to run at NIST. The system should accept inputs in the same directory structures and so forth as the data for the 2012 evaluation. The system should generate outputs with the same file formats as the MT outputs that were delivered for the 2012 evaluation.

8 SCHEDULE (NOTE: MOST FUTURE DATA RELEASE DATES ARE *NOT* FIRM OR CLEAR)

Late January: Draft MT Evaluation Plan available for comment (discussion expected to occur in the telecons)

- - -

Early February: Sample SMS/chat was released by LDC

Feb. 15: MT Evaluation Plan was released

- - -

- - -

Apr. 22: First-pass translations were scheduled to be available for 20-k word per language of SMS/chat DevTest

- - -

May 9: Translation Release 1
120k-words per language of first-pass translations were scheduled to be available for SMS/chat training data (Chinese and Egyptian)

May 27: Memorial Day (Decoration Day)

May 30: 5k-words per language of SMS/chat scheduled to be identified from DevTest as phase-two validation set

May 31: First-pass translations were scheduled to be available for the 20k-words per language of SMS/chat DevTest

- - -

June 4: SMS/chat training data source release 1 was published by LDC
200k-words per language (English and Chinese)
A portion was held back for dev/eval data

June 28: 10k-words of Chinese SMS/chat were identified for DevTest
(from the June 4th source release)

Jun. 28: Translation Release 2
300k-words per language of first-pass translations were scheduled to be available for SMS/chat training data (Chinese and Egyptian)

Jun. 28: Gold-standard translations scheduled to be available for the 5k-words per language validation set for SMS/chat data

- - -

July 5: Gold-standard translations were released for the 5k-words per language validation set for forums data

July 16: SMS/chat training data source release 2 was published (English and Chinese)
A portion was held back for dev/eval data

July 25: 30k-words of Chinese SMS/chat was identified for DevTest (2nd batch) — from the July 16th source release

July 25: 60k-words of Chinese SMS/chat was identified as main eval data
(Note: no downstream annotations are planned for this dataset)

July 31: SMS/chat training data source release 3 is scheduled to be released
800k-words per language (English, Chinese)

- - -

Aug. 7: Translation Release 3
200k-words per language of first-pass translations is scheduled to be available for SMS/chat training data
Chinese *(and Egyptian?)*

Aug. 9: Developer teams to publicly identify any private training data to be shared with other teams (see end of Section 3.1)

Aug. 14: 20k-words per language of SMS/chat identified as HTER Eval Set (subset of the data identified on Jun. 28)

Aug. 16: Any private training data must be shared with other teams by this date (see end of Section 3.1)

Aug. 19 – 23: MT Dry Run (**No** *scores will be distributed, and quality of the MT will* **not** *be assessed.*)
The dry run is intended to make sure all submission procedures, data formats, and scoring procedures work without problems.

Aug. 23: MT outputs submissions for Dry Run (using the DevTest data) due at NIST

Aug. 23: MT post-editing guidelines for SMS/chat finalized

Aug. 26 – 30: Crunch MT Dry Run

- - -

Sept. 2: Labor Day

Sept. 2 – 6: MT Summit (Nice, France)

Sept. 13: NIST re-provides **forums** eval source data (progress set) to participants

Sept. 15: Translation Release 4
280–300 k-words per language of first-pass translations available for SMS/chat training data (Chinese and Egyptian)

Sept. 16 – 20: ***** MT EVALUATION PERIOD for forums data *******

Sept. 20: NIST provides **SMS/chat** eval source data (progress set) to participants

Sept. 23 – 27: ***** MT EVALUATION PERIOD for SMS/chat data *****

Sept. 27 (3:00 p.m.): primary submission MT outputs due at NIST via ftp, both for forum data and for SMS/chat data

Sept. 30: SMS/chat training data source release 4
(*possibly earlier*) 500k-words per language (English, Chinese, Egyptian)
A portion will be held back for dev/eval data

- - -

Oct. 4 (3:00 p.m.): Any contrastive submissions of MT outputs on main eval set due at NIST via ftp

Oct. 7: Post-editing of HTER subset (from the Sept. 27 submissions) begins (20k-words per language, per genre)

Oct. 9 (3:00 p.m.): Validation set MT outputs (both forum and SMS/chat) due at NIST, via ftp (will be post-edited for HTER)

Oct. 28: SMS/chat training data first-pass translation release 5
300k-words per language of first-pass translations available for SMS/chat training data (Chinese and Egyptian)

Oct. 31: System description for the 2013 primary submission system due at NIST

- - -

Nov. 1: Tentative date for results of automated MT metrics to be sent to developers
(for primary submissions, contrastive submissions, and validation set submissions)

Nov. 18 – 22: TAC, TREC, and TRECVID

- - -

Dec. 11: Post-edits of HTER subset delivered to NIST

Dec. 20: All final MT results (including HTER) to DARPA and performers

- - -

Jan. 6, 2014: Systems to be delivered to NIST, on COTS hardware

Jan. 6 – 10: Systems installed and verified to run at NIST