BOLT Activity A Machine Translation Evaluation Plan for Phase 3

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 3 of the program.

Specifically, the BOLT MT evaluation in this third year will test the translation into English of material from three genres and for two languages for each genre:

1. Text drawn from "discussion forums" in Egyptian Arabic and Mandarin Chinese
2. Text drawn from SMS and Chat in those same two languages, Egyptian Arabic and Mandarin Chinese.
3. Speech (audio) from "conversational telephone speech (CTS) in those same two languages, Egyptian Arabic and Mandarin Chinese.

   In addition, as a *contrastive condition*, translation from careful human transcriptions of these conversations will also be evaluated, to approximate the effect of having Automatic Speech Recognition with human-level quality.

Translation from those two source languages will be evaluated separately.

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

2 EVALUATION TASK

The BOLT MT evaluation for Phase 3 will test system capabilities of translation into English of the languages and genres listed above. Translation from the two source languages will be evaluated separately.

2.1 TEXT-TO-TEXT TRANSLATION

Text-to-Text translation 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.

2.2 SPEECH-TO-TEXT TRANSLATION

Speech-to-text translation tests a combination of a system's ability to transcribe the audio, via automatic speech recognition (ASR) and the system's ability to translate the resulting transcription from Arabic or Chinese into English. The ASR component will *not* be evaluated separately, and systems are free to jointly optimize their ASR and MT components in order to produce the best translation into English.

As a *contrastive condition*, translation from careful human transcriptions of these conversations will also be evaluated, to approximate the effect of having Automatic Speech Recognition with human-level quality.

As is the case for text-to-text translation, evaluation will focus on the ability of the system to produce an English translation that completely captures the meaning conveyed by the source audio data, using easily understandable English.

Systems should assume that the translated output in the form of English text will not be accompanied by the source-language audio; thus users of the translation will not have available any prosody or other audio-only information from the source data. As a rule of thumb, the goal is for the textual translation into English to have the same meaning, and the same degree of understandability, as the human transcriptions that will be provided as input for the contrastive condition.

3 DATA

The discussion forum data for Phase 3 will be the discussion forum data from Phase 2, 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 CTS data is expected to be entire telephone conversations and conversations that are truncated at the end.

The textual source data will be in XML format (a DTD will be provided for each genre). The textual data will be UTF-8 encoded.

The MT outputs from the systems will be in the same XML format and UTF-8 encoding as the input data, with the translation substituted for the source language.

The audio data format will be SPHERE-format 8-bit mu-law, in order to give the developer teams data that is as "raw and unmodified" as possible.

The means by which MT output from audio data will acquire the format of textual source data is T.B.D.

3.1 TRAINING DATA

Discussion Forum data for both languages is already in the teams' hands from Phase 1. SMS/chat training data for Chinese and English is already in the teams' hands from Phase 2. SMS/chat training data in Arabic has been collected by the Linguistic Data Consortium (LDC) and is being incrementally distributed to the teams. Textual CTS training data is being incrementally distributed to the teams: as stated in the next paragraph, examples in Arabic and examples in Chinese are already in the teams' hands.

All the Arabic 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.

Examples of Chinese CTS have been distributed in textual form, in data release LDC2014E08, and examples of Arabic CTS have been distributed in release LDC2013E49. All textual CTS data for the program will have the same format as those two data releases.

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 September 29 and shared by October 6.

3.2 DEVELOPMENT-TEST DATA SET

A DevTest dataset (**DEV**) for each genre for each source language will be provided. This data will be drawn from the LDC data collections. The **DEV** data will be selected using the same procedures as will be used to select the evaluation data as described below (section 3.6).

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

This data is intended to be used by the teams for assessing statistical models that they have built from the training datasets. For example, teams can use these data to train model parameters or to assess the performance of their systems. NIST will not be providing any assessments that use the DevTest datasets as input.

3.3 VALIDATION DATA SET

The two research teams will jointly select additional data from the DevTest dataset, which NIST will score as part of each evaluation and for which NIST will provide the complete scoring details, so that the teams can see how the scoring was done; for example, how the data was post-edited for 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.

It is expected that the Validation dataset will have approximately 5kw for each language for each genre.

Selections already exist for Chinese and Arabic discussion forums and for Chinese SMS/chat, and those selections will remain unchanged. During phase 3, teams will select validation dataset data for Arabic SMS/chat as well as for both Arabic and Chinese CTS.

3.4 MAIN EVALUATION DATA SET

The main evaluation dataset (**MAIN EVAL**) will contain approximately 200k source words from each language for discussion forums and a large number of files in both languages for SMS/Chat and for CTS. The discussion forum data reads as entire threads. Special steps will be followed to protect the **MAIN EVAL** dataset, keeping its contents sequestered (or blind) throughout all phases of the BOLT program.

3.5 HTER EVALUATION DATA SET

From the available pool of data from which the MainEval dataset is chosen, NIST will sub-select approximately 20k source words for each genre for

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

each language (Chinese and Arabic), to be used for the primary HTER scoring. The existing HTER Evaluation dataset for discussion forums and for Chinese SMS/chat will be unchanged. In phase 3, NIST will choose data for Arabic SMS/chat and for both Arabic and Chinese CTS.

In the evaluation, the HTER dataset will be scored for HTER but will also be scored for the automated MT metrics (BLEU, METEOR, and TER) so that NIST and the teams can examine the relationships between HTER scores and the scores on the automated MT metrics.

The entire **MAIN EVAL** dataset will be accompanied by a first-pass reference translation.

The **HTER EVALUATION** dataset (for HTER scoring) will have careful translations that include alternatives (translations 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.

3.6 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. The CTS data will be informal telephone conversations.

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 focused on avoiding personally-identifiable information, but were otherwise minimally selective. Current events at the time the data was collected and expressions of personal opinions were typical topics.

It is anticipated that topical coverage of CTS data may be somewhat similar to the SMS/chat data.

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 is in a somewhat similar format. All textual data will be UTF-8 encoded. Genre (forums, SMS/chat, CTS) will be made known to the systems during BOLT phase 3, implicitly from file formats and directory structure.

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 and CTS data will be in somewhat similar formats, matching the training data that has been distributed to the developer teams. In all cases, the MT outputs will replace the source-language inputs. System MT 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 **<DATE T.B.D.>**, 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

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 for discussion forums, at the

[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_p3.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.nist.gov/itl/iad/mig/bolt_p3.cfm
The previous GALE documents are at:
http://www.itl.nist.gov/iad/mig/tests/gale/2008/

file level for SMS/Chat, and at the conversation level for Conversational Telephone Speech.

6.2 AUTOMATED MT METRICS

NIST will run BLEU, METEOR, and TER over the entire **EVAL** consisting of about 200k words per source 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 *a subset of* the same data on which HTER is performed.

In phase 3, given the volume of data with the addition of CTS, it will no longer be possible for the judges to score *all* the data in the available time.

7 EVALUATION PROCEDURES

7.1 DELIVERING SYSTEM OUTPUT TO NIST

This section is not yet final.

Create a directory that names your 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/smschat/primary
./arabic/smschat/contrastive_1
./arabic/smschat/contrastive_2

./chinese/smschat/primary
./chinese/smschat/contrastive_1
./chinese/smschat/contrastive_2

./arabic/ctsfromaudio/primary
./arabic/ctsfromaudio/contrastive_1
./arabic/ctsfromaudio/contrastive_2

./chinese/ctsfromaudio/primary
./chinese/ctsfromaudio/contrastive_1
./chinese/ctsfromaudio/contrastive_2

/arabic/ctsfromtranscript/primary
./arabic/ctsfromtranscript/contrastive_1
./arabic/ctsfromtranscript/contrastive_2

./chinese/ctsfromtranscript/primary
./chinese/ctsfromtranscript/contrastive_1
./chinese/ctsfromtranscript/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 its identity 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 (p3)
- team name (astral or delphi)
- submission
  (dryrun, validation, primary, contrastive1, or contrastive2),
- source language (arabic or chinese),
- genre (forums, smschat, ctsfromaudio, or ctsfromtranscript),
  with no hyphen or underscore in the genre
- date/time when the submission was *assembled* by your team (a date/time meaningful to you)
  If it is 2014-December-19 at 17:30 edt, this should appear as `2014-12-19-1730edt`

A possible experiment ID is as follows
`bolt_p3_delphi_primary_arabic_smschat_2014-12-19-1730edt`

7.2 SYSTEM DESCRIPTION

A separate system description will be required for the BOLT evaluation, due at the end of January 2015.

7.3 DELIVERING COTS SYSTEMS TO NIST

As of September 2014, it appears to NIST that evaluating COTS systems will NOT be possible before the end of the BOLT program funding.

8 DATA FORMATS FOR TEXTUAL INPUT

The examples provided here are intended to show exactly what text in the textual input files you should translate. In all cases, the target language MT output should **replace** the Arabic or Chinese source language input. The XML markup and should remain unchanged.

The directory structure of your output should be the same as the directory structure of the input except that in the filenames the `.arz.` or `.cmn.` of the input filenames should become `.eng.` in the MT output filenames.

8.1 DISCUSSION FORUMS

```
<?xml version="1.0" encoding="UTF-8"?>
<multipost id="bolt-arz-DF-123_123456_12345678" language="Arabic">
<post id="bolt-arz-DF-123-123456-12345678_p1">
<su id="bolt-arz-DF-123-123456-12345678_p1_su1">
ابن موسى        ← Translate this
</su>
</post>
<post id="bolt-arz-DF-123-123456-12345678_p2">
<su id="bolt-arz-DF-123-123456-12345678_p2_all1">
مرحروسة ي ا مصر        ← Translate this too
</su>
</post>
</multipost>
```

Note carefully that SUs identified as `_all1` `_all2` `_all3` and so forth are needed for context by the HTER posteditors and therefore **must** also be translated.

8.2 SMS / CHAT

The rule here is simple: for each <su> element, the MT system should translate the text that is between <body> and </body> tags.

Explanation: For SMS/Chat, there can be splits or merges, so the <body> may differ from the <message>. For example, the LDC may break a lengthy message into multiple pieces (that is, multiple <su>'s), in which case the entire <message> (and message id) will repeat in successive <su>'s whose <body> elements contain the successive pieces of the message. Inversely, a text-messaging provider may limit a message to some maximum number of characters, and therefore may break what the sender intended to be one message into multiple messages (perhaps even breaking in the middle of a word), and the LDC may merge them into one <su> (with one <body>), in which case there may be multiple <message> elements inside the <messages> element of an <su>.

```
<?xml version="1.0" encoding="UTF-8"?>
  <conversation id="SMS_CMN_20130303.0001" medium="SMS" donated="true">
    <su id="s0">
      <messages>
        <message id="m0000" medium="SMS" time="2013-03-03 22:27:08 UTC" participant="131671">回来给我带碗小米粥买个卷饼吧，谢了</message>        ← DO NOT translate the <message>
      </messages>
```

```
    <body>回来给我带碗小米粥买个卷饼吧，谢了</body>        ← Instead, translate the <body>
  </su>
  <su id="s1">
    <messages>
      <message id="m0001" medium="SMS" time="2013-03-03 22:27:44 UTC" participant="131623">嗯
</message>
    </messages>
    <body>嗯</body>        ← Translate this
   </su>
</conversation>
```

8.3 Conversational Telephone Speech

```
<?xml version="1.0" encoding="UTF-8"?>
<conversation>
  <su id="1" speaker="B" begin="117.88" end="120.14">
    <body>فيه عندنا احنا ال security كده بعجل</body>        ← Translate this
  </su>
  <su id="2" speaker="A" begin="119.57" end="123.74">
    <body>عندنا احنا ما already زيادة ده ده بس >campus</bodyال حدود على إتضرب هو يعنى        ← Translate
  </su>
  <su id="4" speaker="A" begin="124.89" end="126.81">        ← Note absence of su 3
    <body>%fp %pw فالعملية</body>        ← Translate
  </su>
  <su id="5" speaker="B" begin="126.89" end="128.12">
    <body>سهلا و اهلا</body>        ← Translate
  </su>
</conversation>
```

9 SCHEDULE FOR BOLT MT (AND CONFERENCES PEOPLE MAY BE ATTENDING)

May 4–9: ICASSP (Florence, Italy)

May 26: Memorial Day / Decoration Day (Federal holiday)

May 26–31: LREC (Reykjavik, Iceland)

- - -

June 22–27: Association for Computational Linguistics meeting (Baltimore, MD)

- - -

July 4: Independence Day (Federal holiday)

- - -

- - -

Sept. 1: Labor Day (Federal holiday)

Sept 14–18: Interspeech conference (Singapore)

Sept 29: Developer teams to publicly identify any private training data to be shared with other teams (see end of Section 3.1)

- - -

Oct. 6: Any private training data must be shared with other teams by this date (see end of Section 3.1)

Oct. 13: Columbus Day (Federal holiday)

- - -

Nov. 3: MT outputs submissions for Dry Run (using the DevTest data) due at NIST
(**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.

Nov. 4 – 21: NIST to crunch MT Dry Run and fix any problems that turn up

Nov. 10–12: TRECVID workshop (2.5 days) at University of Central Florida (Orlando, FL)

Nov. 11: Veterans Day (Federal holiday)

Nov. 17–18: TAC workshop at NIST

Nov. 18–21: TREC workshop at NIST

Nov. 27: Thanksgiving (Federal holiday)

As a practical matter, NIST will not be available Nov. 27–30
and NIST is reluctant to promise that our servers will be up and running then.

- - -

Dec. 1–19: BOLT MT Evaluation period

Notionally:

CTS from audio in the first week

Discussion forums and SMS/Chat in the second week, including Arabic SMS/Chat

CTS from source-language textual transcriptions in third week

noon Dec. 1: audio data for CTS, plus the textual data for Discussion forums and SMS/Chat made available to teams

noon Dec. 11: MT outputs on CTS from audio due at NIST from teams

2 p.m. Dec. 11: NIST releases source-language textual transcriptions of CTS to teams

5 p.m. Dec. 19: All remaining MT outputs due at NIST from teams

Dec. 25: Christmas Day (Federal holiday)

- - -

Jan. 1, 2015: New Year's Day (Federal holiday)

Jan. 8: Post-editing begins, with staggered delivery of editing kits to LDC over Jan. 8–16

Jan. 8 – March 13: Post-editing

Jan. 9: All contrastive submissions due at NIST

Jan. 16: Validation dataset submissions due at NIST (primary system only)

Jan. 19: Martin Luther King Day (Federal holiday)

Jan 30: Automated metric scores (BLEU, METEOR, TER) from NIST to DARPA and developer teams

Jan 30: System description due, to DARPA and NIST, from each developer team

Jan. 30: 1st rolling release of non-QA'd post-editing results from LDC to NIST

- - -

Feb. 13: 2nd rolling release of non-QA'd post-editing results from LDC to NIST

Feb. 16: Presidents' Day (Federal holiday)

Feb. 23–27: *Tentative week for human judgments of semantic adequacy*

Feb. 27: 3rd rolling release of non-QA'd post-editing results from LDC to NIST

- - -

March 20: Final QA'd post-editing results from LDC to NIST

March 31: Initial HTER scores from NIST to DARPA and developer teams

- - -

mid-to-late April: PI meeting

- - - - - - - - - - - -