BOLT Activity A Machine Translation Evaluation Plan for Phase 1

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 1 of the program.

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

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

2 EVALUATION TASK

The BOLT MT evaluation for Phase 1 will test system translation capabilities into English of text drawn from "discussion forums" 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 1. 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 input data for the BOLT MT evaluation will consist of entire threads[1] drawn from Egyptian Arabic and Mandarin Chinese language data from a variety of discussion forums. The source data will be UTF-8 encoded.

3.1 TRAINING DATA

The primary source of training data will be collected by the Linguistic Data Consortium (LDC) and incrementally distributed to the teams throughout the BOLT program.

BOLT teams may use 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 May 15, 2012 and shared with the other teams by **May 30, 2012**.

3.2 DEVELOPMENT DATA SET

A development set (**DEV**) of approximately 40k words from each source language will be provided. This data will be drawn from the LDC training releases (R2 and R3). The **DEV** will be selected using the same procedures as will be used to select the evaluation data as described below (section 3.5).

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

The four research teams will jointly select 5kw per language of data, from the development 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.

3.3 EVALUATION DATA SET

The evaluation dataset (**EVAL**) will contain approximately 200k source words from each language. The 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 from each language for the primary HTER scoring. The sub-selection will consist of complete posts chosen from a large number of distinct threads. This selection is to be representative of the overall evaluation dataset. The targets are 200 threads and 400 posts, for inclusion in HTER scoring.

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.

[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*.

[2] The JAVA based post editing interface maybe accessed via the

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 and will be from threads with a focus on current or dynamic events.

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

The LDC identifies data by a combination of hand-selection and automatic selection. The hand-selection process identifies posts that have the desired characteristics (such as Egyptian Arabic dialect and current events as the topic). Forums in which desired data has been identified are considered "promising" and data selection will focus on such forums. An appreciable fraction of the **DEV** and **EVAL** datasets will be chosen by automatic selection that is informed by the hand-selections.

4 DATA FORMATS

Both the source language input and the target language output will be in the LDC's "multipost" XML format. The source language data will include markup that identifies each post in the thread. Within each post, there will be 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. NIST will identify the data that is to be translated.

4.1 INPUT FORMATS

The MT source-language data will be distributed in the LDC multipost data format. All data will be UTF-8 encoded.

4.2 OUTPUT FORMATS

The system MT outputs will be in the LDC multipost data format. 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 prior to June 8th using the same submission procedures and formats as for the actual evaluation.

System scores for the dry run will not be reported. The purpose of the dry run is to validate that systems are producing output in the valid data formats and to verify 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.

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

[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

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

6.3 Sentence-Level Judgments of Semantic Accuracy

NIST will have a panel of bilingual judges perform sentence-level judgments of MT semantic accuracy over the same data on which HTER is performed. Due to resource constraints, these judgments will only be performed for the Arabic source-language data.

7 Evaluation Procedures

7.1 Delivering System Output to NIST

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

"astral", "delphi", thunderbolt", or "fletcher"

Under your team directory create the following structure (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

Place the system translations in their proper directory.

System translation files should have the same name as the input file.

A system description will be required for the BOLT evaluation.

7.2 Delivering Systems to NIST

A copy of the BOLT systems (used to generate the primary submissions) is to be delivered and installed at NIST the week of September 24–28, after the evaluations are complete. The system should accept inputs in the same directory structures and so forth as the data for the evaluation. The system should generate outputs with the same file formats as the MT outputs that were delivered for evaluation.

8 Schedule

25 April 2012

MT Eval dataset: 100k selection per language from R2 to LDC (Note, the volume of Chinese fell a bit short)

[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/

1 – 3 May 2012

R2 of Chinese and Arabic training data available

14 May 2012

MT Eval dataset: 100k selection per language from R3 from NIST to LDC (shortfall of Chinese in the R2 selections was made up in the R3 selections).

14 May 2012

Additional 20 k-word per language DevTest data from R3, selected by NIST and selections sent to LDC. LDC has done SU annotations, but may not have budget to do translations (possibility to do translations is T.B.D.)

18 May (or before)

HTER experiment results and lessons learned

18 May 2012

Arabic DevTest available (approx. 40 k-word)

18 May 2012

Chinese DevTest available (approx. 60 k-character)

5 June 2012

10k per language selection for HTER from R2 (NIST to LDC)

5 June 2012

NIST re-distributes the DevTest data to teams (for teams to use for the dry run).

8 June 2012

Teams are requested to run a system over the DevTest data and submit the outputs to NIST as a dry run.

12 June 2012

10k per language selection for HTER from R3: (NIST to LDC)

22 June 2012

200k per language MT Evaluation dataset sent to participants (to be in their hands by 26 June 2012).

The extra 5k (from DevSet) for HTER will be selected by NIST and identified for performers as part of this release.

26 June – 3 July 2012

Performers run eval data through MT systems

3 July 2012 / 3 p.m. EDT

MT outputs due at NIST (via ftp)

3 July 2012

Full final references for the 20k per language

HTER datasets in NIST's hands (from LDC)

12 July 2012 / 3 p.m. EDT

Outputs from up to two contrastive MT systems due at NIST (via ftp).

12 July 2012

HTER editing kits etc. in LDC's hands (post-editing begins 13 July 2012)

30 July – 3 August 2012

NIST runs automated MT metrics on primary and contrastive MT system outputs

6 – 10 August

NIST analyzes results of automated MT metrics

10 August 2012

Results of automated MT metrics available

23 August 2012

HTER post-editing complete

24–28 August 2012

rolling releases of HTER data from LDC to NIST

29–31 August 2012

crunching/analysis of HTER results

7 September 2012 (or before)

HTER results due to DARPA and performers

7 September 2012

All MT scoring results due to DARPA

24 – 28 September

MT systems (to be sequestered/mothballed) delivered to NIST and verified to run (teams help if necessary).