BOLT Phase 3 Activity C Evaluation Plan

Introduction

Speech-to-speech translation systems have made great strides over the past several years, but conversations continue to break down due to the inability for a system to capture and resolve ambiguities in the input or errors in recognition prior to issuing a translation and continuing the dialog. The DARPA BOLT (Broad Operational Language Translation) program's Activity C evaluates conversation robustness in the context of an English speaking user communicating with an Iraqi Arabic speaking user via a machine translation and dialog management system. For Phase 3, Activity C focuses on a two-way human-to-human dialog with and without system-solicited clarification capability.

1. Tasks

Three tasks are evaluated in Activity C in Phase 3. Task 1 is the primary speech to speech translation task, and has two subtasks. Task 1a tests the full speech-to-speech translation with system-solicited clarification capability. Task 1b, full speech-to-speech translation without system-solicited clarification capability, is a contrastive test. In addition, Task 2, a smaller-scale trial Utility Test, investigates alternative system architectures.

1.1. Task 1: Speech-to-Speech Translation Main Test

Task 1 evaluates automatic translation capability for dialogs between an English speaking user and an Iraqi Arabic speaking user. An evaluation trial is one such dialog guided by a common scenario for which the two users receive separate background information. The resulting dialog is anticipated to be mostly freeform, with users using their own words to communicate. The only exception is that some of the scenarios purposefully introduce out-of-vocabulary names.

A trial is allowed to last at most eight minutes and is cut off by the system moderator if it has not concluded after eight minutes.

1.1.1. Task 1a: Speech-to-Speech Translation with System-Solicited Clarification

In Task 1a, system-solicited clarification is allowed.

1.1.2. Task 1b: Speech-to-Speech Translation without System-Solicited Clarification

Task 1b is identical to Task 1a, except that system-solicited clarification is not allowed. This setup allows for assessment of the impact of the dialog clarification component on the success of the interaction.

1.2. Task 2: Speech to Speech Translation Trial Utility Test

Task 2 is a smaller-scale trial utility test to be carried out after Task 1, to investigate the use of innovative approaches for advancing the performance and user interaction of speech-to-speech machine translation systems, using variations not limited by the constraints specified for Task 1. Specifically, Task 2 systems differ from Task 1 systems as follows:

- Hardware: Systems must employ mobile devices (tablet or smartphone) for user input and system output. These may be used as standalone devices or as thin clients connecting to a server laptop in the testing area. Performers may choose to have a mobile device for one or both users. Complete logs for each trial must be written to a single device that the NIST system and Subject Matter Expert moderator has access to. The hardware setup must work within the constraints of the evaluation facility.
- Input/output modalities: Systems are encouraged to experiment with input and output modalities in addition to speech, by utilizing the screen, including touch screen functionality. Thus the eyes-free constraint of Task 1 does not hold for Task 2. *Typing on a virtual keyboard is not permitted to replace speech as the primary input method.*

Input and output requirements as well as analysis protocols are the same as for Task 1.

System-solicited clarification is allowed for Task 2. There is no separate evaluation of systems with system-solicited clarification allowed vs. not allowed; each performer provides only one Task 2 system.

Performers must provide a description of their proposed Task 2 system to DARPA and NIST by the date indicated in section 7.10.

2. Data

2.1. Training Data

2.1.1. Task 1

In Phase 1, BOLT performers were given a hard drive containing the complete set of dialog data used in the TRANSTAC program. All data on this drive is available for use as training. Systems must limit their vocabulary to the data contained on this drive. This means the systems' entire vocabulary (both audio and text) is limited to the data on this hard drive. The following exceptions are permissible:

- **Exception 1**: Systems may expand the vocabulary items into morphological variants, using an algorithm that does not require access to additional data. Such algorithms should be clearly identified in the system descriptions.
- **Exception 2**: Appen-produced "Iraqi Arabic names lexicon", containing names of persons, places, streets, and tribes.
- **Exception 3**: Performers may use additional data resources that were specified and shared among all users in Phase 2, as well as new additional resources that are to be declared and

shared among all performers by the dates specified in the schedule. All such resources used must be indicated in each system description. *The use of these additional resources is not allowed to increase the systems' vocabulary*. *The BOLT Phase 2 evaluation data does not qualify under this exception; see special note below.*

Special note on BOLT Phase 2 Activity B and C evaluation data: For Task 1, the BOLT Phase 2 Activity B and C evaluation data that was shared among all performers may be used *for development testing and error analysis, as well as manual development of resources and adjustments based on such error analysis. It may not be used for automatic system training in any way. It may also not be used to expand the systems' vocabulary.*

2.1.2. Task 2

There is no restriction on permissible training data for Task 2. The data specified for Task 1 may be used. The BOLT Phase 2 Activity B and C evaluation data may be used, with no restrictions on the kind of training. Additional resources may be used. Both the BOLT Phase 2 evaluation data and other additional resources are allowed to increase the systems' vocabulary. Additional resources are not required to be disclosed to the other performers, but must be listed, along with a brief description of how they are used, in the Task 2 proposals and system descriptions submitted to DARPA and NIST.

2.2. Evaluation Data

NIST provides 30 to 40 evaluation scenarios to be used for both Task 1a and 1b. A subset of these scenarios is also used for Task 2.

About 50% of scenarios are in-domain, i.e. topics match those from the TRANSTAC training data. 30% are marginally in-domain. 20% are out-of-domain. Each scenario has a designated initiator; for one subset of scenarios this is the Subject Matter Expert, for the other, the Foreign Language Expert. A subset of the scenarios contains purposefully introduced out-of-vocabulary names. A small subset of scenarios may consist of clusters of two or three scenarios, to be tested consecutively, in which the later scenarios address plausible follow-up situations to the first one.

A small set of representative sample scenarios are distributed to the performers ahead of the dry run.

A scenario consists of:

- For the initiator:
    - A brief description of the background and scene from the initiator's perspective.
    - An overall goal for the interaction
- For the respondent:

- A brief description of the background and scene from the respondent's side that contains embedded information that may be useful in response to the initiator's utterances

Topic areas for the evaluation data may include, but not be limited to:

- TRANSTAC
  - Civil Affairs
  - Combined Operations
  - Combined Training
  - Facilities Inspection
  - Medical
  - Traffic Control
- HADR
  - Disaster Relief
  - Humanitarian Aid
- General (out-of-domain) topics

2.2.1. Clarification Challenges

In Phase 3, challenges explicitly incorporated into the evaluation scenarios in an attempt to trigger the systems' dialog clarification components on both the English and Arabic side are limited to out-of-vocabulary (OOV) named entities, i.e. names of persons, locations, and organizations that are not in the limited vocabulary available for system training. Only a subset of scenarios contains such challenges. It is expected that other challenges – non-name OOVs, word sense ambiguities, idiomatic expressions – occur naturally in a free-flowing dialog.

2.3. Evaluation Users

NIST recruits 9-12 Subject Matter Experts and 9-12 Foreign Language Experts to use the systems during the evaluation for Tasks 1a and 1b. For Task 2, a subset of these users is used. The users must meet the following requirements:

- The primary language is US English for Subject Matter Experts and Iraqi Arabic for Foreign Language Experts.
- They are required to have some post high school education.
- They are free of heavy regional accents.
- Male and female Subject Matter Experts and Foreign Language Experts are used at an approximate ratio of 80% male to 20% female, to match the gender ratio of the TRANSTAC training data.

3. System Learning and Adaption

Each performer provides three physically separate systems for Tasks 1a, 1b, and 2, respectively. Learning and adaptation are permitted, separately for the Task 1a, 1b, and 2 systems, following these rules and restrictions:

- Learning of language content (e.g. OOV names) is permitted across scenarios, users, and system sessions for both Subject Matter Expert and Foreign Language Expert side.
- Learning of and adaptation to Subject Matter Expert user characteristics is permitted across scenarios and sessions; systems are allowed to make use of the Subject Matter Expert ID for this purpose.
- Learning of and adaption to Foreign Language Expert user characteristics is permitted within a scenario, but not across scenarios and sessions.
- Learning and adaptation is not permitted to make use of scenario IDs and scenario order.

4. Hardware

4.1. Task 1

Task 1 systems must operate on one of these laptop platforms:

- http://nist.gov/itl/iad/mig/upload/BOLT-P3-C-LaptopSpec_DellM4800.pdf
- http://nist.gov/itl/iad/mig/upload/BOLT-P2-BC-LaptopSpec_Dell-M4700.pdf
- http://nist.gov/itl/iad/mig/upload/BOLT-P1-B-LaptopSpec_DellM4600.pdf

There are no constraints on the type of microphone except that push-to-talk technology must be used, and that the technology must work within the constraints of the evaluation facility, which provides a soundproof barrier between the two users.

4.2. Task 2

Task 2 systems must employ mobile platforms (one or two tablets or smartphones); there are no constraints on the models to be used. These mobile platforms may be used in a standalone manner, or as thin clients connecting to a laptop server in the testing area. There are no other constraints on the hardware, except that the technology must work within the constraints of the evaluation facility, which provides a soundproof barrier between the two users.

5. Experimental Design

Systems from all three performers are tested simultaneously.

NIST employs a Latin-square design for the assignment of user pairs, scenarios, and systems.

Subject Matter Experts and Foreign Language Experts are paired, and the pairings remain constant throughout the evaluation. For Task 1, sets of three user pairs are assigned an identical subset of scenarios, such that the same scenarios are tested on the three systems in parallel and in the same order by different user pairs. For Task 2, fewer scenarios are used. By default, this is a

subset of the Task 1 scenarios that approximates the domain and initiator distribution from Task 1. Alternatively, each performer may specify, in the Task 2 system proposal, a focus on a particular domain or subset of domains, or a focus on one user side being the initiator, for their Task 2 system, in which case the subset of scenarios from Task 1 is selected accordingly.

Users complete their entire set of scenarios on one system before moving on to the next system, reducing the complexity of the evaluation and confusion for the users, and allows them to become expert users to the extent possible.

For Task 1, all systems are tested on all scenarios and with clarification on and clarification off for each scenario, thus allowing comparison of how clarification helps or hinders the dialog. For Task 2, the distinction between clarification on and clarification off does not apply.

6. Metrics

Validated log files and subjective feedback are analyzed to yield quantitative and qualitative performance reports. The impact of system-solicited clarification is also measured.

Prior to implementing the metrics identified in this section, a determination is made for each evaluation trial as to whether or not the users performed as expected. Trials which are significantly flawed due to user (rather than system) are excluded from assessment.

6.1. Bilingual Human Assessment – Goal Accomplishment (primary)

The primary metric evaluates the dialog as a whole for achievement of the initiator's goal. For each trial, bilingual assessors review the backgrounds and goal, then listen to the audio logs of the dialog (including the users' speech and the system translations and clarifications) and assess the statement:

The initiator of the dialog achieved his/her goal.

by selecting an answer from the following scale:

_ Strongly disagree _ Disagree _ Slightly disagree _ Neutral _ Slightly Agree _ Agree _ Strongly Agree

Each evaluation trial receives several independent assessments. NIST reports the percentage of times that each answer category was chosen. The assessments may be compared to the initiating users' answers to a similar question.

6.2. Secondary Bilingual Human Assessment

The bilingual assessors may also assess the overall helpfulness of clarifications and adequacy of the translation for each trial.

6.3. User Feedback

At the conclusion of a dialog, the initiator may be asked to rate if he/she achieved the predefined goal, and the respondent may be asked to rate if he/she was able to communicate successfully using the system. Both users may be asked to rate the helpfulness of clarifications, if applicable. At the end of a session on one system, users may be asked provide to feedback regarding the success of the interactions, as well as usability of the systems without and with clarification.

There may also be opportunity to leave informal written feedback after each system session on a voluntary basis. NIST removes any personally identifiable information from this feedback.

6.4. Additional Metrics

The duration of time needed for each dialog is noted and used as an additional metric.

The number of clarifications, if applicable, may be noted and used as an additional metric. For this purpose, each time a user, prompted by the system, interacts with the system again before the other user interacts with the system counts as one clarification turn.

The number of trials cut short due to system error may be noted as an additional metric.

7. Evaluation Logistics

7.1. Input

An evaluation trial begins by the system moderator entering, using the system's keyboard, a unique evaluation trial ID. In case a trial is cut short due to system, operator, user, or facility problems, a trial ID may be re-entered, in which case the evaluation trial starts anew, distinguished by a new time stamp.

All other input to the system is direct user input (speech for Task 1, potentially other modalities for Task 2).

7.2. Output

System output to be evaluated consists of validated log files produced by the translation systems. NIST provides a directory structure and log file XML schema to capture each entire dialog, including audio files, automatic transcriptions, system clarification requests, automatic translations, and timing information. At the conclusion of each evaluation scenario, log files are validated.

This section describes the directory structure, file naming requirements, and encoding requirements that the Phase 3 BOLT C submissions must adhere to.

7.2.1. Directory Structure

There is no restriction on the directory structure where log files are stored. A flat directory structure (all files in one directory) is preferred.

7.2.2. File Naming Convention

Log-file names must comply with the following naming convention:

```
<system>_<userpair>_<systemtype>_<scenarioid>-<epoch>.xml
```

Audio file names must comply with the following naming convention:

```
<system>_<userpair>_<systemtype>_<scenario-id>-<type-tag>-
<epoch>.wav
```

Note:

- Audio filenames have the same epoch as the log file epoch.
- All information except <epoch> and <type-tag> is being entered by the moderator.
- Audio file names only differ from the log file name by <type-tag>.

The following convention is used:

- A represents an alphabetic character.
- N represents a numeric character.
- Other letters represent themselves.
- | (pipe character) represents logical OR

<scenario-id> is defined as: ANNNN

<system> is defined as: (S|B|I)

- Identifies the system participating, using the first letter of the performing organization's name.

<userpair> is defined as: (SNNFNN) or (FNNSNN)

- S is the Subject Matter Expert, F is the Foreign Language Expert; each is further identified by a two-digit number. The first user entered is the initiator of the dialog. The system needs to preserve the order in which Subject Matter Expert and Foreign Language Expert are entered (e.g. not change S01F01 to F01S01).
- Two digits must be used.

<systemtype> is defined as: (CLA|NOC|UTL)

- Identifies Task 1a (CLA) vs. Task 1b (NOC) vs. Task 2 (UTL)

<epoch> is defined as (N)*10 or (N)*13:

- POSIX-epoch (see http://en.wikipedia.org/wiki/Unix_time) **or** Windows-FILETIME (see http://msdn.microsoft.com/enus/library/windows/desktop/ms724290%28v=vs.85%29.aspx)

- POSIX-Epoch and Windows-FILETIME are monotonically increasing numbers. As consequence of this, a distinct time is being used for each trial run.
- **Since it is possible to re-run the same trial multiple times with the same trial ID, only the log file with the latest epoch is assumed to be correct and is used for assessment.**
- POSIX-epoch and Windows-FILETIME have different lengths. POSIX uses 10, Windows uses 13 digits.

<type-tag> is defined as: AANNN

- First two characters should represent what segment of the trial the audio file refers to. The actual naming is undefined, and up to the systems.
- The last three digits represent the segments index.
- Three digits must be used. The numbering does not have to be in sequence.
- AANNN must uniquely identify each audio file used.

Example:

Repeated trial (System I, trial T0001, using clarification, S01 is initiator, Date: Wed, 09 May 2012 13:30:31 GMT)

- I_S01F01_CLA_T0001-1336569922.xml
- I_S01F01_CLA_T0001-1336570231.xml

Corresponding audio files for the trial at time 1336570231:

- I_S01F01_CLA_T0001_UT000-1336570231.wav
- I_S01F01_CLA_T0001-XY012-1336570231.wav

7.2.3. Encoding

Log file must contain UTF-8-encoded content only. Audio files must use the RIFF WAVE (WAV/LPCM) format.

7.2.4. XML Validity and XSD Compliance

Each log file must contain only valid XML data and comply (validate) with the current BOLT-BC XSD-Schema that is available online at http://www.nist.gov/itl/iad/mig/bolt_p3.cfm.

Log files which do not validate against the schema are excluded from assessment.

7.2.5. Consistency Checker

NIST provides a cross-platform consistency checker tool which can validate individual log syntax as well as check log file content for consistency (in terms of structure, labels, file naming, etc.). The tool is applied to detect problems after each block of scenarios during the evaluation, and it

is applied after the evaluation to validate all log files before scoring. The tool is available at http://www.nist.gov/itl/iad/mig/bolt_p3.cfm.

7.3. System Description

In addition to providing the log files capturing the system behavior, performers are required to submit a system description. Refer to the schedule in section 7.10 for the exact date.

The system description must provide the following information:

- Modifications to the hardware, if any
- New methods/techniques developed in Phase 3, if any
- Data used for system training (limited as specified in section 2.1)
- System start-up instructions
- Log file extraction instructions
- System shut-down instructions

7.4. Training of Subject Matter Experts and Foreign Language Experts

- Each user pair receives training from the performers regarding the functional capabilities of each system being tested before the user pair uses that system. Training should cover using the system both with system-solicited clarification on and off. Performers must make every effort to keep training consistent between user pairs by using a dedicated script to follow for training purposes. Training time for each user pair on each system includes opportunity for hands-on training on the system and question/answer time with the performers.
- Each user receives descriptions of their assigned scenarios, training on how to appropriately conduct these dialogs using a translation system, and scenario rehearsal time, as deemed appropriate to meet the test objectives.
- Subject Matter Experts and Foreign Language Experts rehearse their scenarios separately.
- Immediately before each testing session on a system, the testing user pair completes one trial scenario on the system to be tested in the testing room. This trial scenario is the same across all user pairs, systems, and sessions.

7.5. NIST Moderators

7.5.1. System and Subject Matter Expert Moderators

A NIST system and Subject Matter Expert moderator is in the room with the Subject Matter Expert. The moderator is responsible for entering the correct information for logging of trials on the system and interacts with the Subject Matter Expert as needed. Prior to the evaluation week, the moderators are provided with written guidelines that explain the evaluation goals and procedures as well as guidelines that explain the basic operations of the systems. They are instructed to be cooperative but not enabling. System and Subject Matter Expert moderators are responsible for:

- Cross-checking upcoming scenario ID with users and the Foreign Language Expert moderator
- Entering trial ID into system
- Starting trial after ensuring system and users are ready
- Deciding on need to restart due to technical or other issues
- Ending trial if not finished after eight minutes
- Resolving any problems with the Subject Matter Expert, if necessary
- Recording any irregularities

7.5.2. Foreign Language Expert Moderators

A bilingual Iraqi/English moderator is in the room with the Foreign Language Expert. Prior to the evaluation week, the moderators are provided with written guidelines that explain the evaluation goals and procedures as well as guidelines that explain the basic operations of the systems. Foreign Language Expert moderators are responsible for:

- Resolving any problems with the Foreign Language Expert, if necessary
- Recording any Foreign Language Expert-side irregularities

7.6. System Malfunctions and Incomplete Trials

A trial may end before completion due to a variety of reasons. The number of such trials is anticipated to be small, and thus any effects on system learning and adaptation are anticipated to be small. Incomplete trials are handled differently depending on the reason:

- **System error**: A trial cannot be completed due to a system malfunction. Only one redo of this trial is allowed before moving on to the next trial. The later attempt at the trial is used in the assessments. A maximum is set for the number of trials that may be reattempted over the course of the entire evaluation. After this maximum is met for a system, no more redos are allowed for that system, and any subsequent incomplete trial is used as-is in the assessments. This per-system maximum is set to 15 for each of Task 1a and 1b, and 10 for Task 2.
- **User/moderator/facility error**: A trial cannot be completed due to a problem that is not the system's fault. This trial can be redone several times if necessary, until complete, until a system malfunction occurs twice, or until the allotted time ends. The latest version of the trial is used in the assessments.

NIST keeps track of the number of incomplete trials due to system malfunction. If a system encounters malfunctions numerous times during a session, performers are called in to troubleshoot. Permissible changes made during such troubleshooting are limited to those that enable the system to run without fatal malfunctions, and must be documented. Depending on how long troubleshooting takes, it is possible that the system in question misses trials.

As a general rule, the latest time stamp for any trial is used for the subsequent assessments, unless semi-automatic inspection by NIST reveals that the latest time stamp version is clearly not the correct one to use (e.g. if that log file is mostly empty).

7.7. Evaluation Facility

The evaluation occurs at a suitable facility in the Washington DC metro area. Users are separated by a soundproof barrier so that they can see, but not hear each other.

7.8. Evaluation Rules

The systems for all tasks (1a, 1b, 2) must be in testing-ready stage and turned over to NIST on the morning of the first evaluation day for Task 1, before user training and testing begins.

Performers must be present to start and shutdown their systems on each evaluation day.

Performers are asked to sign a written statement that their systems are in proper working order prior to each day's testing.

NIST maintains possession of the systems throughout the testing week, taking them off-site and recharging if necessary, each evening.

Performers may be present in their own testing room prior to a testing session for training purposes.

Performers may view and hear the interaction with the system from a separate room during evaluation sessions; however they may not interact with the system or the users. Only if technical difficulties are encountered during an evaluation session that cannot be resolved by the NIST team and NIST requests technical assistance from the performers, performers may enter the testing room for the sole purpose of providing such technical assistance.

The systems and logs are released to the respective performer teams immediately after live system testing concludes.

7.9. Dry Run

A dry run is required of the performers. The dry run is conducted with two Subject Matter Experts and two Foreign Language Experts, using a small set of training scenarios to interact with the systems for all three tasks. Performers must provide three physically distinct systems for Tasks 1a, 1b, and 2, respectively. The system must be in an evaluation-ready state and accept input and log output in accordance with section 7.2.

The log files for the dry-run scenarios are retrieved and analyzed by NIST to ensure proper formatting, and are used to exercise the scoring pipeline.

7.10. Schedule

Date(s)	Event
March 26 2014	Draft evaluation plan sent to performers
June 13 2014	Evaluation plan published
August 1 2014	Data to be shared (as described in section 2.1.1 Exception 3) specified
August 1 2014	Task 2 system proposals due
October 17 2014	Data (as described in section 2.1.1 Exception 3) shared
December 3-5 2014	Dry run at Omega Recording Studios, Rockville MD (one day per performer)
January 26 – February 1 2014	Evaluation at Omega Recording Studios, Rockville MD (days 1-5: Task 1a and 1b, days 6-7: Task 2)
March 2015	Bilingual Human Assessments
February 27 2015	System descriptions due
March 2014	Results to DARPA
(At DARPA's discretion)	Results to performers
May 5 – 6 2015	DARPA BOLT PI Meeting