BOLT Phase 2 Activity B/C Evaluation Plan

Updated: September 27, 2013
Version: 4.5

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 conversation. The DARPA BOLT (Broad Operational Language Translation) program's Activity B/C will evaluate conversation robustness in the context of a speaker of English communicating with a speaker of Iraqi Arabic while using a translation and dialog management system. For phase 2, Activity B/C will focus on a two-way human-to-human conversation with and without clarification.

1. Tasks

Three tasks will be evaluated in phase 2 and are described below. Task 1 is the primary task that evaluates Activity C as intended – full speech-to-speech translation with error clarification capability. Tasks 2 and 3 are contrastive (or diagnostic) tests.

1.1. Task 1 - Speech-to-Speech Translation with Error Clarification

Task 1 will evaluate automatic translation capability for conversations between an English speaker and an Iraqi Arabic speaker *with* system-solicited clarification. For this task, systems are permitted to interact with a speaker to clarify input prior to translations. An evaluation trial is one such conversation guided by a scenario and is expected to last approximately 7 minutes.

Task 1 performance will be contrasted with Task 2, where the system-solicited clarification feature is not allowed. This two-part setup will allow NIST to assess the impact of the dialog clarification component on the success of the interaction.

> **INPUT** – A conversation between an English speaker and an Iraqi Arabic speaker guided by their own scenarios within tactical and non-tactical domains. The English speaker will not have knowledge of the Iraqi Arabic speaker's scenario description and visa-versa.
>
> **SYSTEM OUTPUT** – Validated log files produced by the translation systems. NIST will provide a directory structure and log file XML schema to capture the entire dialogue including audio files, timing information, transcriptions and translations. At the conclusion of each evaluation scenario, log files will be validated; log files that cannot be validated will not be evaluated. A trial will not be repeated if a corresponding log file is invalid.

ANALYSIS – Validated log files and subjective feedback will be analyzed to yield quantitative and qualitative performance reports. The impact of the system-solicited clarification will also be measured.

1.2. Task 2 – Speech-to-Speech Translation without Error Clarification

Task 2 will evaluate automatic translation capability for conversations between an English speaker and an Iraqi Arabic speaker *without* system-solicited clarification.

It is anticipated that the resulting dialogue will be mostly freeform with the speakers using their own words to communicate. The only exception is that some of the scenarios will purposefully introduce phase 1 terminology tailored to be challenging to the systems.

Input and output requirements are the same as for task #1. The analysis of task #2 will follow the same protocols as used for task #1.

1.3. Task 3 - English Only Dialog Clarification (Phase 1 B Retest)

Task 3 is a diagnostic test included to allow for a more direct comparison to phase 1 results. NIST will sequester the phase 2 Activity C systems and will implement a scaled version of the phase 1 Activity B evaluation at NIST using the phase 2 Activity C systems. Approximately 100 English starting utterances from the phase 1 B evaluation will be re-used for this test. Speakers and test moderators will be NIST staff. Some of the speakers may be the same as in phase 1.

Task 3 will follow the protocols of phase 1 to a large extent. **Notable differences from phase 1:**

- The system logging requirements have been updated and will be the same as for Activity C.
- No maximum will be imposed on the permissible number of clarifications.
- Starting utterances will be completely independent from each other; i.e. there will be no sets of starting utterances grouped together as belonging to a certain domain.
- Only the utterances themselves, no additional domain background information will be provided to the speakers.
- Speaker training will be performed by NIST moderators familiar with the systems from the preceding Activity C evaluation. Developers will not be present for training or observation for the Task 3 evaluation.

INPUT – A scripted starting utterance and potential clarifications (in response to system prompts) spoken by an English speaker. Every starting utterance will contain a target challenge of one of the types specified in section 2.2.4.

SYSTEM OUTPUT – Validated log files produced by the translation systems. NIST will provide a directory structure and log file XML schema to capture the entire trial, including audio files, timing information, transcriptions and translations. At the conclusion of each evaluation scenario, log files will be validated; log files that

cannot be validated will not be evaluated. A trial will not be repeated if a corresponding log file is invalid.

ANALYSIS – Validated log files and subjective feedback will be analyzed to yield quantitative and qualitative performance reports.

2. Data

2.1. Training Data

In phase 1, BOLT teams 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.

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, for the TRANSTAC program, "Iraqi Arabic names lexicon" containing names as well as names of places, streets, and tribes that can be used for system development.

EXCEPTION 3 - BOLT teams may use additional data resources as long as such use does not increase their systems' vocabulary. These resources must be documented in the system description. If additional data resources are used, they must be shared with other BOLT teams by the agreed date. Please refer to the schedule for the exact date.

This training data applies to ALL TASKS (1, 2, and 3).

2.2. Evaluation Data

2.2.1. Task 1

NIST will develop up to several hundred scenarios where approximately 70% of the scenarios will be designed to be initiated by the English speaker, and 30% will be initiated by the Iraqi speaker. About 70% will be in-domain, i.e. topics will match those from the TRANSTAC training data, 20% will be marginally in-domain, and 10% will cover out-of-domain topics. The male/female ratio will attempt to match that of the TRANSTAC training data (approximately 80% male and 20% female). Approximately 25% of these scenarios will introduce terminology tailored to challenge the systems. Approximately twenty representative scenarios will be distributed to the research teams. These scenarios will have characteristics that are representative of the evaluation scenarios including a mix of tactical v. non-tactical and initiated by the English speaker v. Iraqi Arabic speaker. An

evaluation trial is one such conversation guided by a scenario and is expected to last approximately 7 minutes.

A structured scenario will consist of:

- For the Scenario Driver:
    - A brief description of the background and scene from the driver's perspective.
    - A set of approximately five critical concepts to be communicated to, acquired from, or resolved with the other speaker
    - An overall goal for the interaction
- For the Scenario 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

An unstructured scenario will consist of:

- For the Scenario Driver:
    - A brief description of the background and scene from the driver's perspective.
    - An overall goal for the interaction
- For the Scenario 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 **will not** be limited to:

- Humanitarian aid
    - Food distribution
    - Neighborhood construction
    - Vaccine coordination
- Disaster relief
    - Shelter mitigation
    - Task planning
    - Medical triage
- Check point operations
    - Car search
    - Identification validation
- General Out of Domain topics

2.2.2. Task 2

Same as Task 1.

2.2.3. Task 3

A subset of approximately 100 starting utterances from phase 1 will be re-used for the Task 3 evaluation. Each of these starting utterances will contain a target challenge to be clarified from one of the categories listed in section 2.2.4

2.2.4. Categories of Clarification Types

This section outlines the categories of ambiguity that will be included in the evaluation data in an explicit attempt to trigger the systems' dialog clarification components for English and Arabic.

2.2.4.1. Out-of-Vocabulary (OOV)

The same limited vocabulary will be available for system training. An OOV is a word or set of words that is not part of this vocabulary.

- Noun – named entity (name of person, location, or organization)
- Noun – common noun
- Verb
- Modifier – adjective, adverb

For the Iraqi Arabic side, the only possible error types will be OOV.

2.2.4.2. Word Sense Ambiguities

The target word can have multiple meanings while the pronunciation is the same.

- Homophone-heterograph: different meaning, same pronunciation, *different* spelling
  - Example: *wait* vs. *weight*
- Homophone-homograph: different meaning, same pronunciation, *same* spelling
  - Example: *plane* as in *airplane* vs. *flat surface*

2.2.4.3. Idioms

The target item is a phrase with figurative meaning, i.e. a meaning that cannot be deduced from the literal meanings of the words it consists of. Idioms cannot be translated literally from one language to another without losing the figurative meaning.

- Example: *drop the ball* (meaning *make a mistake*)

3. Metrics

Prior to implementing the metrics identified in this section, bi-lingual judges will make a determination for each evaluation trial as to whether or not the SPEAKERS performed as expected. Trials where the speaker behavior caused a fatal trial will not be scored. Note, if it is determined that the system may have elicited the behavior, the trial will be scored.

3.1. Task 1 Speech-to-Speech Translation with Error Clarification

3.1.1. Human Assessment – Goal Accomplishment (primary)

The primary metric for this task will be a new form of human assessment that evaluates the dialog as a whole. NIST will report the percentage of evaluation trials that a bilingual judge rates in the following categories:

Q: The driver of the conversation achieved his goal(s)?

__ Strongly Agree __Agree __Slightly Agree

__Slightly Disagree __Disagree __ Strongly Disagree

Each evaluation trial will receive three independent judgments. The judgments will be compared to the user feedback answering the same question.

3.1.2. Human Assessment – Critical Concept Transfer (primary)

Both structured and unstructured scenario trials will be annotated for Critical Concept Transfer (CCT). However, the unstructured scenario will have the CCT generated by judges after the data-collection, whereas the structured scenario will have pre-designed CCT which exist before the data collection.

For annotation of unstructured scenarios, at least two judges will be assigned to each trial. Each judge will annotate all utterances for one side of the conversation and then they will swap sides for a review of each other's annotations.

All dialogues (structured and unstructured) will then have at least three judgments by bilingual assessors that determine for each annotated critical concept whether or not it was satisfactory handled (question and response) by the translation system using the following scales, respectively:

For structured scenarios:

__ Concept addressed and all relevant information was conveyed

__ Concept addressed and most relevant information was conveyed

__ Concept addressed and some relevant information was conveyed

__ Concept addressed and no relevant information was conveyed or concept was not addressed

__ Concept addressed and misleading information was conveyed

For unstructured scenarios:

__ All relevant information of concept was conveyed

__ Most relevant information of concept was conveyed

__ Some relevant information of concept was conveyed

__ No relevant information of concept was conveyed

__ Misleading information was conveyed in attempt to address concept

NIST will report the percentage of critical concepts conveyed overall and separately for each conversation side. Per system analysis will be provided, and for comparison across systems, NIST will normalize by the number of turns.

3.1.3. Speaker User Feedback

At the conclusion of each evaluation trial the speakers and respondents may provide their personal feedback of the system's capabilities. This will be done informally and is not required. Paper and pen will be available. NIST IET will review any comments to purge any PPI from the feedback.

3.1.4. Minor Metrics

Time will be noted and used as an additional minor metric.

3.2. Task 2

The same metrics as described above will be applied to this task (except for the user feedback regarding clarification).

3.3. Task 3

Similar metrics as used in phase 1 will be used for Task 3 to allow for more direct comparison. The phase 1 trials of the subset of starting utterances selected for phase 2 may be re-assessed by the same phase 2 assessors and using the phase same protocols to allow more maximum comparability.

4. Rules and Restrictions

4.1. Tasks 1 and 2

System learning will be permitted within a scenario trial, but not across scenario trials.

4.2. Task 3

System learning across starting utterances will not be permitted.

5. Evaluation Details and Logistics

5.1. Hardware

Teams will determine the laptop platform by the agreed date. Please refer to the schedule in section 5.11 for the exact date. If a decision could not be converged by the agreed date, all teams will use the same platform that was used in phase 1.

5.2. Input

An evaluation trial will begin by the system moderator entering, using the system's keyboard, a unique evaluation trial ID. In the case of a system failure or unintended input by the system operator, a trial ID may be re-entered in which case the evaluation trial will start anew and only the final log file for the particular trial ID will be evaluated.

All other input to the system will be speech.

5.3. Output

NIST has defined a set of XML tags that are used to format the log file output for evaluation. NIST requires that all submitted log files meet these formatting standards.

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

5.3.1. 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.

5.3.2. File Naming Convention

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

```
<system>_<sme1|fle1><sme2|fle2>_<structure>_<scenarioid>-
<epoch>.xml
```

Audio file names must comply with the following naming convention:

```
<system>_<sme1|fle1><sme2|fle2>_<structure>_<scenario-id>-
<type-tag>-<epoch>.wav
```

Note:

- Audio filenames should have the same epoch as the log file epoch.
- All information, but <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 being used:

- *A (or any other character but N) represents an alphabetic character*
- *N represents a numeric character*
- *| (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 systems company name.

<sme|fle> is defined as: (S|F)NN

- The first SME/FLE will be the driver of the conversation as entered by the moderator. The systems GUI needs to respect the order in which SME/FLE1 and SME/FLE2 are entered (e.g. do not swap F01S01 with S01F01).
- Two digits must be used

<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/en-us/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 will be automatically used in the evaluation !
- POSIX-epoch and Windows-FILETIME have different lengths. POSIX uses 10, Windows uses 13 digits.

<structure> is defined as: (S|U)

This entry marks structured (S) or un-structured (U) trials.

<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.

Examples:

Repeated trial (System I, structured trial T0001, S02 is driver, Date: Wed, 09 May 2012 13:30:31 GMT)

I_S02F05_S_T0001-1336569922.xml

I_S02F05_S_T0001-1336570231.xml

Corresponding audio files for the trial at time 1336570231

I_S02F05_S_T0001_UT000-1336570231.wav

I_S02F05_S_T0001-XY012-1336570231.wav

5.3.3. Encoding

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

5.3.4. XML Validity and XSD Compliance

Each log file must contain only valid XML data and comply (validate) with the current BOLT-BC XSD-Schema available online at:
http://www.nist.gov/itl/iad/mig/bolt_p2.cfm

Log files which do not validate against the schema will not be accepted in the evaluation !

5.3.5. Consistency Checker

NIST will develop and provide the systems with a cross-platform consistency checker tool which can validate an individual log syntax as well as check log-file content for consistency (in terms of structure, labels, file-naming, etc.). The tool will be applied to detect problems after each block of scenarios during the evaluation, and it will be applied after the evaluation to validate all log-files before scoring. The tool will be made available online at:
http://www.nist.gov/itl/iad/mig/bolt_p2.cfm

5.4. System Description

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

The system description must provide the following information:

- Modifications to the hardware (if any)
- New methods/techniques developed in phase 2 (if any)
- Data used in system training

- Where obtained, how available to others
- System start-up instructions
- Log file extraction instructions
- System shut-down instructions

5.5. Evaluation Speakers

NIST will recruit 10-15 Subject Matter Expert (SME) and 10-15 Foreign Language Expert (FLE) to use the systems during the evaluation. These speakers will meet at least the following requirements:

- Their primary language will be English for Subject Matter Expert or Iraqi Arabic for Foreign Language Expert.
- They will be required to have some post high school education.
- They will be free of heavy regional accents as determined by NIST language experts.
- Both male and female Subject Matter Experts and Foreign Language Experts will be utilized at an approximate ratio of 80% to 20%, respectively.

5.5.1. Training of Subject Matter Experts and Foreign Language Experts

- Speakers will receive training from a research team representative regarding the functional capabilities of each system being tested. Additional time will be provided for hands-on demonstration of each system and one-on-one question/answer time with members of the research teams.
- Each speaker will receive descriptions of their assigned scenarios, training on how to appropriately conduct these conversations using a translation system, and scenario practice time – as deemed appropriate to meet the test objectives.
- Subject Matter Experts and Foreign Language Experts will be rehearsed separately.

5.6. NIST Moderators

NIST will recruit approximately ten of their employees to moderate the evaluation. Prior to the evaluation week, the moderators will be provided with written guidelines that explain the evaluation goals and procedures as well as guidelines that explain the basic operations of the systems. They will be instructed to be cooperative but not enabling. There are two types of moderators:

Speaker moderator manages the speakers and his/her responsibilities include but not limited to:

- Reviews scripted utterance
- Decides if restart is necessary
- Supplies additional information about the scenario – true intent

System moderator manages the systems and his/her responsibilities include but not limited to:

- Enters trial ID and other information
- Confirms ready state
- Gives indication to start the scenario
- Records any oddities

5.7. Evaluation Rooms

The evaluation of task 1 and 2 will occur at a suitable facility in the DC metro area. Speakers will be isolated in a soundproof room or booth. Speakers will be able to see each other however they will not hear the other language; this may accomplished by renting a sound studio or purchasing of soundproof booth.

Task 3 will be held at NIST.

5.8. Researcher Accessibility

Researchers will be allowed to view/hear the interaction with the system, however they will be not able to interact with the system during the evaluation trail.

5.9. Evaluation Procedures

BOLT teams will be required to be present to start and shutdown their systems on each evaluation day.

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

On each day of the evaluation week, team representatives may be present in their own testing room to review and assist in the training of system moderators. Teams will have up to 30 minutes to demonstrate how to operate their system with groups of 3-4 system moderators.

NIST will employ a Latin-square assignment for the system operators, test trials, and systems.

SME's will be assigned a set of evaluation trials. Over the course of the evaluation each system operator will implement their entire set of evaluation trials for each system.

All systems will be tested simultaneously.

All systems will be tested with 50% of the scenarios with clarification off and again using different SMEs/FLEs with clarification on, thus allowing comparison on how clarification helps or hinders the conversation.

There will be one system moderator, one speaker moderator present in each room during a testing session. Each moderator will have specific responsibilities.

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

At the conclusion of the B Retest, NIST will return the systems to the developers within 6months. These systems may be utilized to assist in planning future phases of BOLT.

5.10. Dry Run

A dry run will be required of the teams. The dry run will consist of approximately 5 scenarios using 2-3 SME/FLEs to interact with the system. The system must be in an evaluation-ready-state and accept the trial ID.

The log file for 5 scenarios will be retrieved and analyzed by NIST to ensure proper formatting, and will be used to exercise the scoring pipeline.

The dry run will be conducted as closely to the evaluation as possible.

5.11. Schedule

Date(s)	Event
February 15	Draft Eval Plan sent to Teams
February 22	Small set of Scenarios to Teams
March 1	Vocabulary locked down
May 1	Shared data deadline
May 1	Platform agreed upon
June 17-18	Dry Run (@ Omega Studio)
August 21	Shared monolingual data lists between teams (the Exception 3 lists)

October 21-25	Evaluation (@ Omega Studio)
	Evaluation Day 1
	Evaluation Day 2
	Evaluation Day 3
October 28- November 1	*Live Evaluation B*
November 1	System Description Due
November 12- 22	Human Assessment
December 20	Results to DARPA
At DARPA's discretion	Results to Teams
January 14-16, 2014	DARPA BOLT PI Meeting

6. Glossary of Terms

Scenario – Is brief synopsis of an event or situation. The description should be constructed such that it sets the scene for the speaker to gain specific information.

Topic - A topic is a combination of the Scenario and domain. It is possible for topics to be related to each other but there will be some specificity to a difference between them.

Topic variability is a goal to reduce the impact caused by limited training data.

Trial - An evaluation trial is a conversation guided by a scenario and is expected to last approximately 7 minutes.