BOLT Activity B Evaluation Plan

Version 3.0 1 June 2012

1. Introduction

Speech-to-speech translation systems have made great strides over the past 5 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.

Activity B will advance 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 1, Activity B will focus on a single side of a conversation (an initial utterance issued by the English speaker through translation of the meaning of that utterance into Iraqi Arabic text). The system will have the opportunity to seek clarification through audio for one or more parts of the utterance that it determines to be ambiguous, or in error, prior to issuing the translation. This clarification process is the focus of Activity B.

2. Task

The task is dialog clarification for speech-to-speech translation in the presence of ambiguous or errorful input. A translation system will receive a starting statement or question through spoken English input and must issue a textual translation that conveys the ***intended meaning*** of the English input. Each system will be evaluated using the identical set of starting utterances, recited by the same system operator.

The system will have up to three opportunities to seek clarification with the operator before issuing the translation. The system must use audio to communicate the clarification requests to the operator. Responses to the system clarification questions will be free-form from the operators, although these operators will be provided basic instructions as to be "cooperative" but not "enabling" (see Section 6.2).

The starting utterances will be designed to evoke certain situations that are known to be problematic for speech-to-speech translation systems. The Phase 1 Activity B evaluation will probe several of these factors, many of which will be evaluated separately. The system operator will be provided side-information that identifies the true intent of the communication.

2.1. Primary Categories of Conversation Clarification Types

The following subsections list the error categories that will be of primary interest in the evaluation.

2.1.1. English OOV

The general category of English OOV will be evaluated separately from the other clarification types. A known, and limited, set of English words will be available to train each system. Several evaluation trials will include *one or more* English words that the system does not have in its ASR language model. Note, that some utterances may contain OOV's that are applicable to two or more sub-categories listed below (for example, verb and synonym).

For Activity B evaluation, a change in the form of a known word in the systems vocabulary to express a grammatical function or attribute such as tense, mood, person, number, case or gender (inflections or morphological word differences), **will not** be considered for OOV.

2.1.1.1. Nouns - Named Entities

For the purpose of this evaluation, the category of named entities will be limited to names of people, names of locations, and names of organizations. There will be several evaluation trials that include named entities in the starting utterance that are not present in the system's English vocabulary.

2.1.1.2. Nouns – Non-Named Entities

There will be several evaluation trials that will include ***non***-named-entity nouns that are not present in the system's English vocabulary. These additional noun types often relay important content of the instruction.

2.1.1.3. Verbs

There will be several evaluation trials that will include verbs that are not present in the system's English vocabulary. The verb of an utterance often relays the important action of the instruction.

2.1.1.4. Synonyms

There will be several evaluation trials where the starting utterance will include an OOV synonym for a word known to be in the system's English lexicon.

2.1.2. Homophones

The general category of Homophones will be evaluated separately from the other Conversation Clarification Types. Several evaluation trials will include a starting utterance where it is unclear as to which meaning of a given homophone is intended.

> *Example*: "*How much was the weight/wait.*"

With limited context it is unclear when spoken by the subject if they are asking "how long" or "how heavy".

Arabic names are often also common nouns, and such names may be among the English inputs that will be tested.

2.1.3. Word Sense

For the purpose of this evaluation, Word Sense differs from homophones in exactly one way. Under Homophones, we consider words that sound the same, have multiple meanings, ***and*** have multiple spellings. Word Sense will be limited to words that have multiple meanings but are spelled ***in only one way***.

Example: The word “grave” meaning a burial plot is in vocabulary

“*We have grave concerns.*”

Example: The word “plane” meaning airplane is in vocabulary

“*I need a plane to fix this sticking door*.”

2.2. Additional Categories of Conversation Clarification Types

In addition to the categories defined in Section 2.1.1, 2.1.2, and 2.1.3, this evaluation will probe many other possible sources of error. Many of these types of ambiguities or errors are difficult to elicit in a systematic fashion across all systems. Evaluation trials that attempt to evoke these types of errors will be analyzed separately for each team and will be used to provide the BOLT program manager with information covering trouble spots as the program moves forward.

It is also likely that some of these errors will naturally occur in the categories defined above.

2.2.1. Speech Recognition Errors

Each system will be presented with several evaluation trials that are designed to evoke speech recognition errors. It is unlikely each utterance will produce similar recognition errors across all systems in every case. Therefore, this set of evaluation trials will be analyzed separately for each system and NIST will report the percentage of trials evoking errors, and the level of success each system achieves in coping with their set of errors.

Note, the intention for this category is to evoke speech recognition errors that are not caused by the presence of an OOV (those types of errors will be reported separately).

2.2.2. Missing concept (target side)

There will be an attempt to develop evaluation trials that require the translation of a concept that is not understood by the Iraqi Arabic language training. That is, a valid unambiguous English utterance cannot be successfully translated without being reworded. Idioms may fall into this category. Word Sense errors (Section 2.1.3) will not be analyzed as falling in this category.

Example: The adjacent unit must move.

2.2.3. User Error

The general category of User Error will be evaluated separately from the other error types. Several evaluation trials will fall into one or more of the categories listed below. They are designed to evaluate how well a system can cope with bad input. Note, that some utterances may have the potential fit into two or more sub-categories listed below.

2.2.3.1. Mispronunciations and Poor Grammar

Several evaluation trials may instruct the user to mispronounce one or more words of the utterance. These trials will include enough evidence to understand the intended word(s). A phonetic representation of the word to be mispronounced will be provided. The test moderator may ask the system operator to recite the phonetically represented word prior to starting the evaluation trial.

Example: "*Where are the exclosives located?*" or "Who drove the caaa?"

Several evaluation trials may instruct the user to recite an initial utterance that is not proper English, but which conveys a question or statement that is reasonably understood by others; in some instances they will use the wrong word.

Example: "*People often come here to conjugate.*"

2.2.3.2. Incomplete utterance

Several evaluation trials may instruct the user to stop mid-sentence or begin in the middle of the utterance. These trials will not include enough context to form a valid statement or questions. An incomplete sentence will be provided for the system operator to recite.

Example: "*Did the car in the ...*" or "*… westside of the curb.*"

3. Data

3.1. TRANSTAC Data

All BOLT teams received a hard drive containing the complete set of **Iraqi Arabic** data as used in the TRANSTAC program. Activity B systems must limit (*with the exception as stated in section 3.2*) their source side (**English**) vocabulary to the English data contained on this drive. ***All data on this drive is available for use***. Systems may expand the vocabulary items into morphological variants, using an algorithm that does not require access to additional English data. Such algorithms should be clearly identified in the system descriptions.

For the TRANSTAC Program, APPEN produced a "names lexicon" containing male and female names, as well as names of places, streets, and tribes. This data **may** be used for system development.

3.2. Other Data Resources

The list of 69 additional words, as proposed by IBM (www.nist.gov/itl/iad/mig/bolg_p1.cfg), is available for all teams to extend their English ASR vocabulary. In addition, each team is permitted to create a private list of up to 50 words to be used for the same purpose. These private lists must be shared with NIST no later than June 6, 2012. These private lists will be shared with each team after the phase 1 evaluation period.

No other sources of data may be used to increase the systems' English vocabulary.

BOLT teams may use other sources of target side (**Iraqi Arabic**) data to increase the system's translation capability. Such sources of data must be identified prior to the evaluation and its use must be

documented in the submitted system description (see Appendix A). The additional data sources must be identified and shared with the other BOLT teams by May 16th 2012.

Allowed are data resources that are required to train tools that parse and annotate available training data as long as:

1. The use of this data does not increase the English vocabulary of the system
2. The data is identified in the system description

It is permissible to use open-source tools where it is unknown what data resources were used in training the tool as long as:

1. The use of the tool does not increase the English vocabulary of the system
2. The tool is identified prior to May 16th 2012
3. The tool is identified in the system description

3.3. Evaluation Scripts

Starting utterances will be scripted. This section defines the generic method used to create the starting utterances.

3.3.1. Scenario

NIST will define 100-125 testing scenarios for evaluation. Each scenario will include the motivation (a statement describing the peripheral details of what the user is trying to accomplish, in general terms); identification of the domain; and a sub-domain that specifically identifies the situation. Scenarios are to have a wide coverage and will include general topics. The number of starting utterances per scenario will be limited to reduce the affect a particular topic might have on a specific system.

Valid topic areas for the scenarios **<u>may</u>** 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
- Local Restaurants
  - Hours of operation
  - Recommendations

Table 1 shows an example of the type of information that will be provided to the system operator.

Motivation	You are meeting with a member of the Local Army/Police to plan a raid. You need to make sure that the local officer is clear on their role in the operation, what options there are for entry into the target building, and what vehicle support can be expected from the Americans.
Domain	Combined Operations
Sub-Domain	Planning a Raid
trial ID	***utterance***
TR001-01	The target area of this raid will be Al Mouhmodaya, correct?
TR001-02	The area here around these buildings will be a potential bottleneck.
TR001-03	What do we need to do to deal with this dead space here?
TR001-04	As an alternative, we may need to use the breach here.
TR001-05	We will not be able to provide any tank support for this mission.

Table 1: System Operator Information Sheet

3.3.2. Starting Utterance

Exactly 5, scripted, starting utterances per scenario will be defined. Each starting utterance is considered a single trial for the evaluation. Successive starting utterances do not build off of each other; context carry-over is not necessary and is not permitted.

A starting utterance will be scripted and the system operator is to recite the utterance word-for-word. It is the job of the test moderator to ensure proper input. If the moderator determines that the initial utterance deviated from the script, the moderator will choose, at their sole discretion, to abort and restart the evaluation trial.

In general, each of the 5 starting utterances per scenario will be designed to evoke one of the different conversation errors, as defined in Section 2.1.

4. Metrics

For Activity B, NIST will employ many metrics and report the results separately for each of the metrics. This approach will support the use of the findings from Activity B to influence the development of metrics for future phases of Activity C.

4.1. High Level Concept Transfer (HLCT)

High Level Concepts are clauses OR items in a list that are usually separated by words “and” or “or”. Such concepts are annotated for each starting utterance to be used in the evaluation. Human judges will compare each starting utterance (and the intended meaning) against the resulting translation and count how many of the concepts were successfully transferred into the translation.

Examples of English Speech (all one concept):

- My name is Sergeant Smith
- How is your family today?

- We are here today to provide security.
- How many people live in the house with you?
- How many children do you have?

Examples of Translated Foreign Language Speech:

- I sell fruit and vegetables (2 concepts)
- My family works with me (1 concept)
- One of my daughters is going to school and one of my sons is going to school (2 concepts)
- They have material but still need notebooks and still need pens (3 concepts)
- Islamic subjects and a little English (2 concepts)
- The police came around once a week or two times a week (2 concepts)
- Almost six kilometers (1 concept)
- We don't have clean water; we boil it. (2 concepts)

4.2. Timing

The system output format requires that each begin/end activity point in the evaluation trial be time-stamped. NIST will report time requirement information in regards to each system activity.

4.3. Clarification turns required

For Activity B, systems may use 0-3 exchanges with the user to clarify the input. Successful translations with fewer exchanges are preferred. NIST will report statistics in regards to the number of clarification turns required.

4.4. Human Assessment

Bilingual human judges will view the original utterance (with all provided side information) and the resulting translation and will provide Likert-type judgments of the semantic adequacy for the translation. For each translated utterance, the judge will assign one of the following scores.

+3 Completely adequate

+2

+1 Tending towards adequate

0

–1 Tending towards inadequate

–2

–3 Inadequate

Judges will be provided a set of exemplars for each of the four choices that have anchor values. They will be asked to regard the exemplars as being correctly judged — that is, to be no more forgiving and no harsher. They will be asked to prefer one of the four choices that have anchor values, and to choose a value in-between only when they are "on the fence" between two of the anchor values.

4.5. Utterance Review

Post evaluation, each utterance will be reviewed and a determination will be made if the utterance is properly categorized for the type of error that was intended, or to include additional error types that may have occurred.

5. System Requirements

5.1. Hardware

System hardware requirements were agreed upon by all teams during the March 7th 2012 bi-weekly telephone conference. All teams agreed to use the BBN proposed laptop and purchase the exact specification (see: www.nist.gov/itl/iad/mig/boltp1.cfm) as distributed on Monday February 27th by BBN.

Each system will be using the same microphone, the SuperMic as developed and built by BBN. Each team will receive at least one SuperMic and the system is to use this microphone for the evaluations. Spare SuperMics will be on hand during the evaluation.

5.2. Input

An evaluation trial will begin by a test 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 system output file format which will record the necessary information for evaluation (see: www.nist.gov/itl/iad/mig/boltp1.cfm) in a system log file. There is to be one log file for each starting utterance.

Systems much use audio to communicate clarifications requests to the system operator. The system operator will not have visual access of the laptop's screen. If a system has a replay feature to replay the system clarification request, such interaction will only occur if the system operator requests that the clarification be replayed. A test moderator will operate the system to initiate the playback.

5.4. System Capabilities

In this section we define a few functions that Activity B systems must accommodate for evaluation purposes.

1. System must be able to accept as text input a trial ID. The trial ID will be an alphanumeric string that uniquely identifies the TOPIC and UTTERANCE (HA001-01). This trial ID is to be incorporated into the name of the corresponding log file.
2. The system must be able to abort the current trial and initialize to the starting state.

a. In the case of unintended user error, the system should be re-initialized to the state that accepts the trial ID (again).
b. In the case of the system taking more than 60 seconds to respond, the operator must be able to abort the trial and move to the next trial ID.

3. Systems must display on the screen, the information being written to the log file.
    a. To include recognized text
    b. To include text of clarification requests

6. Evaluation Protocols

6.1. System Installation

Activity B evaluation will occur at a hotel[1] in Gaithersburg Maryland. Team members will be required to be present to initialize their system start-up and shutdown their system on each evaluation day.

Team members will be asked to document (sign a written statement) that their systems is in proper working order prior to each day's testing.

On day 1 of the evaluation week, team representatives may be present in their own testing room to review and assist in the training of system operators. Teams will have up to 30 minutes to demonstrate how to operate their system to groups of 3-4 system operators, and pairs of NIST evaluation team members.

NIST will maintain possession of the systems throughout the testing week, taking them off-site and recharging them if necessary, each evening. On successive evaluation days systems will be rotated amongst the available rooms.

At the conclusion of the evaluation NIST will retain possession of the systems (not the SuperMics) for a period of 6-months. These systems may be utilized to assist in planning future phases of BOLT.

6.2. Evaluation Subjects (System Operators)

NIST will recruit 10-15 system operators to utilize the systems during the evaluation. These subjects will meet basic (and minimal) user requirements. Their primary and first language will be English. They will be required to have some post High School education and military experience. A qualified subject will be free of heavy regional accents as determined by NIST testers. Both male and female subjects will be utilized.

Prior to the evaluation week, system operators will be provided written guidelines (see: www.nist.gov/itl/iad/mig/boltp1.cfm) that explain the evaluation procedures and goals. They will be instructed to be cooperative but not enabling.

6.3. Evaluation Rooms

[1] Hilton; 620 Perry Parkway; Gaithersburg; MD; 20877

Each system will be set-up in quasi identical hotel meeting rooms. Systems will not be moved from room-to-room within an evaluation day, but will be evaluated in a different room each day of the test.

6.4. Testing

System operators will be assigned a set of evaluation trials (starting utterance). Over the course of the evaluation each operator will implement their entire set of evaluation trials using each system.

All systems will be tested simultaneously.

There will be two evaluation moderators and two system operators present in each room during a testing session. Each moderator will have specific responsibilities:

MODERATOR-A:

- Manages the System. Enters trial ID. Confirms the ready state.
- Records oddities for initial utterance (identifies additional error type, most commonly ASR errors)
- Records if intended error type is produced by the initial utterance.

MODERATOR-B:

- Manages the system operators. Reviews scripted utterance, decides if restart is necessary.
- Supplies additional information about the scenario – true intent.
- Guides operator through unexpected situations.

System operators will alternate after each completed scenario (estimated 15 minutes or less).

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

7. Schedule

BOLT Activity B Evaluation Schedule	
Date(s)	Event
April 17	**Training Topics available**
May 16	**Due date to share data resources (target side)**
June 6	**Private English Vocabulary (50 words) due (provided to NIST)**
June 11-15	**Dry Run (@ NIST)**
9-11:30 am June 14	*SRI*
1-3:00 pm June 14	*SAIC*
9-11:30 am June 15	*IBM*
1-3:00 pm June 15	*BBN*
July 23-27	**Evaluation Week**
July 23	*System Installation (AM)* *System operator Training (PM)*
July 24	*Evaluation Day 1*
July 25	*Evaluation Day 2*
July 26	*(if necessary) Evaluation Day 3*
July 27	*(if necessary) Evaluation Day 4*
August 13-17	**Human Assessment**
August 24	**Results to DARPA**
August 31	**Results to Teams**

7.1. Training Topics and Starting Utterances

NIST will distribute 10-20 sample topics and starting utterance that can be used for system development. These topics will be created by the same team that creates the evaluation topics.

7.2. Dry Run

A dry run will be required for each team. NIST will schedule an appointment for each site to come to NIST and have a NIST staff member interact with the system in the presence of team representatives, potential test moderators, and program representatives. The system must be in an evaluation-ready-state and accept the trial ID. The log file for 3-5 interactions will be retrieved and analyzed by NIST to ensure proper formatting, and will be used to exercise the scoring pipeline.

7.3. Evaluation Week

Activity B will reserve one week for evaluation. This evaluation period is expected to consume one day for training, 2-3 days for evaluation and allows for one day reserved accommodating unforeseen problems.

7.4. Results and scoring

NIST will recruit human assessors to evaluate the system output using software developed by NIST (see Section 4). All system results will be provided first to the BOLT program manager, and shortly thereafter to the teams.

8. Appendix A – System Description

The system description must provide the following information:

8.1. Modifications to the Hardware (if any)

8.2. Data used in system training

8.2.1. Where obtained, how available to others

8.3. System Start-up instructions

8.4. Log file extraction instructions

8.5. System Shut-down instructions

9. Appendix B – Definitions

9.1. 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.

9.2. Scenario

A scenario is a description of background information that the system operator must know in order to respond to system clarification requests.

9.3. Starting Utterance

Grouped by topic, a starting utterance is a word-for-word script that the system operator will speak into the system to start an evaluation trial.

If the evaluation moderator determines that the starting utterance was improperly spoken, the evaluation trial will begin anew.

9.4. Evaluation Trial

An evaluation trial is defined by the entire interaction between the system operator and the system beginning with the initial utterance and ending with the production of a translation by the system.