Performance Evaluation of Speech Translation Systems

Brian A. Weiss[1], Craig Schlenoff[1], Greg Sanders[1], Michelle P. Steves[1], Sherri Condon[2], Jon Phillips[2], and Dan Parvaz[2]

[1]National Institute of Standards and Technology
100 Bureau Drive, Gaithersburg, MD, USA 20899
Email: bweiss@nist.gov, schlenof@nist.gov, gsanders@nist.gov, msteves@nist.gov

[2]Mitre Corporation
7535 Colshire Drive, McLean, VA, USA 22102
Email: scondon@mitre.org, jphillips@mitre.org, dparvaz@mitre.org

Abstract

One of the most challenging tasks for uniformed service personnel serving in foreign countries is effective verbal communication with the local population. To remedy this problem, several companies and academic institutions have been funded to develop machine translation systems as part of the DARPA (Defense Advanced Research Projects Agency) TRANSTAC (Spoken Language Communication and Translation System for Tactical Use) program. The goal of this program is to demonstrate capabilities to rapidly develop and field free-form, two-way translation systems that would enable speakers of different languages to communicate with one another in real-world tactical situations. DARPA has mandated that each TRANSTAC technology be evaluated numerous times throughout the life of the program and has tasked the National Institute of Standards and Technology (NIST) to lead this effort. This paper describes the experimental design methodology and test procedures from the most recent evaluation, conducted in July 2007, which focused on English to/from Iraqi Arabic.

1. Overview

The Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) program is a Defense Advanced Research Projects Agency (DARPA) advanced technology research and development program. The goal of the TRANSTAC program is to demonstrate capabilities to rapidly develop and field free-form, two-way speech-to-speech translation systems that enable speakers of English and other languages to communicate with one another in real-world tactical situations when an interpreter is unavailable. To date, several prototype systems have been developed for force protection and medical screening domains for multiple languages. Systems have been demonstrated on both PDA (Personal Digital Assistant) and laptop-grade platforms with varying performance.

The primary use case involves US military personnel and foreign language speakers. The military personnel will be trained to use the systems with the assumption that the foreign language users will receive system-provided instruction at the beginning of an interaction.

The National Institute of Standards and Technology (NIST), along with support from MITRE, was funded to serve as the Independent Evaluation Team (IET) for Phase 2 of the TRANSTAC Program. In this role, NIST was responsible for evaluating the performance of five TRANSTAC systems in January and July of 2007 for communication between English and Iraqi Arabic speakers. This report presents the evaluation methodology used in the July 2007 TRANSTAC system evaluations. However, detailed results of the evaluations cannot be reported due to restrictions on releasing the data.

2. System Description

English and Iraqi Arabic speech translation systems developed by five technology teams were evaluated in the July 2007 event. The teams included BBN, Carnegie Mellon University (CMU), Fluential, IBM, and SRI. Each system's architecture consists of three primary components: Automated Speech Recognition (ASR) of the spoken input in the source language, (2) Machine Translation (MT) from the input source language to an output target language, and (3) Text-to-Speech (TTS) generation of spoken output in the target language. The systems translated in both directions (to and from English). In addition, the laptop-based systems that were evaluated include user interfaces of varying complexity, most of which display both English and Iraqi Arabic translations as they are processed. Although the systems have visual interfaces, each also has an eyes-free mode in which the user operates the system using only an external control device such as a mouse or buttons on a microphone. All of the systems employ external microphones that are either handheld or close-talking headsets. Each system also has several pre-programmed commands that the speakers may invoke, as necessary. They include *Please repeat*, *I don't understand*, etc.

3. Evaluation Design

The IET adopted an evaluation approach that is expected to scale well as the technologies develop, thus allowing for valid assessments of performance improvements over time. Evaluation tasks included developing a scalable testing approach (using a recently designed framework), securing participants for testing, and formulating scenarios for training and evaluation.

3.1 Developing a Scalable Testing Approach

The TRANSTAC Phase 2 evaluations were based on previous TRANSTAC evaluations developed by MITRE in Phase 1 of the program, but also incorporated some new procedures and evaluation types that were not previously employed. All testing approaches were designed to easily scale alongside the developing system capabilities.

Per the DARPA Broad Agency Announcement, the following two types of tests were the focus for the

TRANSTAC Phase 2 evaluation:

1) System usability testing - providing overall scores to the capabilities of the whole system.
2) Software component testing – evaluating individual components of a system to see how well they perform in isolation.

The IET adapted the System, Component, and Operationally-Relevant Evaluation (SCORE) framework to achieve the two TRANSTAC evaluation goals. SCORE is a framework built around the premise that, in order to get a holistic picture of how a system performs in the field, it must be evaluated at the component level, the system level, and in operationally-relevant environments (Schlenoff et al., 2007). Each of these evaluation types provides insight into different aspects of the performance and value of the test systems. It is only by looking at the results of all of the evaluations that one can gain a comprehensive (but not necessarily complete) picture of the overall system performance when used in the field. There is no substitute for testing a technology in the actual use-case environment, but it can be informative to test systems in controlled and/or simulated field conditions until they are ready for operational environments.

The SCORE framework is adapted for use in the TRANSTAC evaluations by conducting system level evaluations with live, operationally-relevant dialogues in which both technical performance and usability are assessed. Software components were evaluated through the use of pre-recorded inputs processed by systems during the July evaluation. The component evaluation is referred to as the *offline* evaluation, which contrasts with the live system evaluations.

3.2 Evaluation Approaches

For the live evaluations, military subject matter experts (SMEs) speaking English and foreign language experts (FLEs) speaking Iraqi Arabic communicated using the TRANSTAC systems. They were asked to role-play structured scenarios that designate items of information which the SME must convey to the FtHE LE or elicit from the FLE by asking questions. The FLE was provided with specific (but not scripted) responses so that the content of the dialogues would be the same for all systems. Each dialogue was stopped after 10 minutes, which made it possible to compare the number of items successfully conveyed by each speaker in the test period. In addition, questionnaires were provided to the SMEs and FLEs at the conclusion of their dialogues to gauge their perception of the TRANSTAC systems. The results of the live evaluation address the goal of testing system usability (item 1 above).

For the offline component level testing, recorded inputs were selected from a sample of the training data collected by the IET. A small percentage of the training data was held back from the developers to be used for the evaluation set. The systems processed inputs in audio format, logging both the recognition output to test the systems' ASR capabilities and the MT output. Transcriptions of the audio were also processed to test the systems' MT capabilities independent of speech recognition.

Both of these evaluation approaches are designed to measure the progressive development of the TRANSTAC systems' technical capabilities and to predict the impact these technologies will have on user performance within a range of scenarios. The scenarios' content was crafted to provide a reasonable level of difficulty for the TRANSTAC systems at their current state of development, while creating the opportunity to evaluate the TRANSTAC systems in the future at their expected rate of growth and improvement.

Live evaluations of the TRANSTAC systems were conducted in two different venues, which are referred to as the *lab* evaluation and the *field* evaluation. Testing in these venues is described in the subsequent sections.

3.2.1 Lab Evaluation

The lab evaluations are designed to test the TRANSTAC systems in an ideal environment with no background noise and stationary participants. This environment provides the IET and the developers with an estimate of the best that the systems can do at their current stage of development. Because similar lab evaluations were performed earlier in the TRANSTAC program, it is useful to continue performing lab evaluations for a long term comparison of the systems' progress.

Ten structured scenarios were performed in individual ten-minute time windows for each system. The system laptops were placed on a table, and the speakers were seated at the table. Each scenario was enacted by a different pair of speakers, but the speakers assigned to a scenario remained the same for each system.

After each response from the FLE, the SME relayed the response to an IET member who recorded the SME's reported information. This procedure was intended to provide the IET with a view of the SME's comprehension of system outputs, and it also served as a check for the rare occasions when a SME reported a response that was different than the system's output because he knew what the response should be after performing the scenario several times. The same procedure was followed in the field evaluations.

3.2.2 Field Evaluations

The purpose of the field evaluations was to test the TRANSTAC systems in a more realistic environment. Specifically, the field evaluations introduced controlled background noise, and because the sessions were conducted outside, there was some uncontrolled noise. SMEs carried the TRANSTAC systems in backpacks, and the speakers were mobile during the evaluation. Ten structured scenarios were performed in this setting in separate ten-minute time windows with speaker and scenario assignments consistent across all systems.

One field scenario was performed twice: once with background noise at about 80 decibels and once without background noise. Another field scenario was constructed to be 'out of domain' in order to test systems in a common

civilian interaction for which no training data had been provided to the systems and the developers.
The field evaluation focused on the TRANSTAC systems' performance in an environment that was more representative of operational conditions. Although the environment was not completely realistic, it introduced a variety of factors that were not present in the lab environment. For example, a checkpoint scenario was enacted using a real vehicle, which provided an opportunity for nonverbal responses (in addition to verbal ones) such as opening doors and also required speakers to move around the vehicle.

3.2.4 Eyes-Free/Hands-Free

Anticipating the use of TRANSTAC systems under conditions that require military personnel to keep their attention on their surroundings, rather than on the speech translation system, Phase 2 evaluations required systems to be eyes-free and to be operated with only minimal manual controls. The January 2007 evaluation was the first time that this constraint was placed on the developers. During the July testing, neither speaker was able to see the TRANSTAC screen or interact with the keyboard. The only feedback speakers received from the TRANSTAC system was aural, and most systems provided audio TTS playback of the English speech recognition (the systems did not demonstrate Iraqi Arabic audio confirmation). Physical interaction with the systems was limited to external devices designed by the developers that were plugged into the systems' A/V (Audio/Video) and/or USB (Universal Serial Bus) ports. Typically, these were mouse-sized control devices with several buttons, though one developer mounted buttons on a handheld microphone.

3.2.5 Noise-Masking

The recruitment of suitable FLEs presented a dilemma for recording both training data and live evaluation dialogues. It is desirable that FLEs understand Iraqi Arabic, but not English. The intended use of a translation device, reflected in typical evaluation procedures, is that the SME says something in English, the device translates the utterance into Iraqi Arabic, and the FLE responds to the system's speech output. If the FLE is bilingual, he or she can respond to the SME's English input, even if this action is unintentional. The same issue arises in recording sessions that collect training data when a human interpreter replaces the translation device.
There are very few Iraqi Arabic speakers in the United States who do not speak some English, and it was discovered in Phase 1 of the TRANSTAC Program that, as a general rule, people living in the United States who speak Iraqi Arabic but not English tend to be non-ideal for the TRANSTAC work due to various demographic issues. For example, they are often elderly individuals who find it difficult to engage in the role playing that scenarios require.
A solution to this dilemma has been to develop and apply a method that selectively masks English utterances so that the bilingual speaker cannot hear them. The bilingual speakers wear headphones that allow them to hear any Arabic speech, but when English is spoken, they hear a recording of white noise played loudly enough to inhibit understanding of the English speech. The solution also works well in the data collection sessions where translations are produced by a human interpreter.
For the live lab evaluation, both the SME utterances and the TRANSTAC system English confirmation outputs were masked from the FLE. This setup worked well since the lab evaluation had the two speakers engage one another while stationary throughout their entire dialogue. The masking was not employed in the field evaluation due to the amount of hardware and wiring needed to make it functional. This would not have worked well in the field conditions, where much more physical motion was necessary. A wireless method is being explored for future evaluations so that masking may be incorporated in subsequent field evaluations.

3.3 Evaluation Participants

The main participants who interacted with the TRANSTAC systems were the English speaking SMEs and the Iraqi Arabic FLEs. Ten Marines were present at the evaluation with five assigned to the field exercises and five to the lab evaluation. Ten Iraqi Arabic speakers participated with five assigned to the lab evaluation and five to the field evaluation. The Marines were all reservists and the Iraqis were all US citizens. Some of the Iraqis had served as translators supporting the US military in Iraq.
In addition to the SMEs and FLEs, there were a number of IET members who served various roles during the evaluation. These are listed below:

- Test Point of Contact (POC) – There was a test POC for the field evaluation, the lab evaluation, and offline evaluation (one each). Their main responsibility was to ensure that their test proceeded smoothly. They also ensured that the scenario stopped after 10 minutes (lab and field, only).
- Transcribers – There were two transcribers in the field and two in the lab evaluations; one for English speech and one for Arabic speech. The English transcriber transcribed verbatim all spoken English from both the SME and the TRANSTAC system. The Iraqi Arabic transcriber translated and then reproduced all Arabic speech from both the FLE and the TRANSTAC system in English. Translating the Arabic into English provided "transcripts" that could be analyzed by monolingual English speakers.
- Quality Assuror (QA) – There were two QA personnel in each of the live evaluations. One was an Iraqi Arabic speaker and one was an English speaker. The QAs ensured that the dialogue proceeded as intended and noted all of the responses that the SMEs reported retrieving from the FLEs.
- Data Specialist – The data specialist collected and managed all of the information from the TRANSTAC systems, along with the transcriptions and QA documents created by the evaluation team.

- Noise-Masking Expert – The masking monitor ensured that the noise-masking setup was working properly (an expert setup the systems prior to the evaluation, but the systems were managed by the POCs during the test event).
- Questionnaire Administrator – Questionnaires were administered to the speakers after they completed their dialogues in both the field and the lab environments.
- Videographer - All of the live evaluation dialogues were videotaped with separate audio recordings for higher quality sound.

4. Demographics

Demographic information was self-reported by each participant via survey instruments and was collected during the evaluation event. Participants were asked to provide basic demographic information such as age and gender along with information on their speech and language influences, including languages they speak, places where they have lived, and language(s) spoken at home as children. They were also asked how often they use computers and how comfortable they are with using computers. Additionally, the English speakers were asked to provide information related to their military experience, such as rank, length of service, Military Occupation Specialty (MOS), and Operation Iraqi Freedom (OIF) deployment duration(s) and location(s). A summary of this demographic information includes:

- All Marines were male for the July 2007 evaluation
- Average participant age of the Marines was 32 with a range of 22 to 43 years of age
- All Marines use computers at both home and work
- All Marines had been deployed in Iraq for peacekeeping, peace enforcement, stability operations, and/or combat duties related to OIF

5. Participant Preparation

The day before the evaluation week was set aside for training the SMEs and FLEs. The major goals of this day were:

1. Familiarize SMEs and FLEs with the TRANSTAC program.
2. Prepare them for the evaluation workflow.
3. Allow them to practice their assigned scenarios to maximize their familiarity with their roles.

The third item above was emphasized to address problems that have occurred in previous evaluations. There appeared to be a distinct disadvantage for the system that was tested first because the SMEs and FLES were least familiar with the scenarios during the first trials with a system. Once the evaluation began and the SMEs and FLES became more familiar with the scenarios, they were quicker to respond and had an easier time formulating their contributions. To minimize this learning curve, the SMEs and FLEs were given time to become as familiar as possible with the scenarios prior to the evaluation.

Although these efforts improved the quality of the initial dialogues, there were still differences between results at the beginning and end of the evaluation. The system that was tested first was tested again in the lab, and performance measures improved for the repeated scenarios. Other factors that contribute to the learning effect are likely to result from participants' experience with speech translation systems as they identify successful ways to formulate their inputs and recover from errors in recognition or translation. They probably also become more adept at understanding the synthesized speech outputs. The next evaluation will be conducted with all systems tested in parallel, which should allow all systems to benefit equally from learning effects.

Another goal of the participant training day was to determine the best assignment of scenarios to the military SMEs. SMEs were questioned about their background and experience with the kinds of situations described in the evaluation scenarios. In many cases, it was possible to match SMEs with scenarios that were familiar to them.

The participant training day was devoted exclusively to familiarizing participants with the evaluation process and their roles in the scenarios. SMEs were trained to use the TRANSTAC systems during 90 minute sessions at the beginning of each day of testing. Developer teams conducted the training for their systems. For FLEs, the systems played a maximum 2 minutes of instructions at the start of each scenario.

6. Scenarios and Data Collection

In order to effectively assess the systems' performance on in-domain data, conversational audio data sets between English speaking military personnel and foreign language speaking civilians were recorded, transcribed, translated, and distributed to the technology teams prior to the evaluation so that they could use the data to train their systems. These conversations were motivated by operationally-relevant scenarios provided to each data collection participant. The scenarios were created by the IET with feedback from military and foreign language subject matter experts. A small portion of this audio data (known as the *representative set*) was not provided to the developers so that it could be used by the evaluation team to develop the structured scenarios employed in the live evaluations. In addition, the representative data was used to produce inputs for the offline evaluation.

6.1 Scenario Development for Data Collection

Developing scenarios for data collection followed a series of steps, each necessary to ensure the creation of representative, operationally-relevant scenarios. The first step was to review the existing scenarios (those used in prior evaluations) and to decide which scenarios should be reused and/or reworked for the July 2007 evaluation. The next step was creating new scenarios based upon information gathered from media articles and from experienced military personnel who participated in focus group interviews and role-playing exercises. Scenarios

were finalized after a review process that included personnel with various backgrounds examining each scenario for specific content.

6.2 Data Collection

Scenario recordings were collected by IET personnel at locations with adequate populations of Iraqi Arabic speakers. The representative data used for the July 2007 evaluation consisted of two dialogue types: one in which an English speaking military SME communicated with an Iraqi Arabic FLE using an interpreter and another in which two Iraqi Arabic FLEs conversed with one another in Arabic without an interpreter. Condon et al. (2008) provide details about the data collection protocols and the quantity of training data collected for development and testing of the TRANSTAC systems.

Phase 2 dialogues were recorded in studios that were able to support the noise-masking setup. Participants were recruited based on their background, experience, gender, and, in the case of Arabic speakers, on linguistic factors such as their dialects and time spent in Iraq. Prior to each recording, the SME, FLE, and interpreter studied the scenario that they had been assigned and practiced role playing with a rehearsal coach. Each dialogue was monitored by an Iraqi Arabic speaker who ensured that participants followed the protocol.

6.3 Representative Set Selection

The representative set is a compilation of dialogues that were withheld from the developers in order to produce evaluation scenarios and offline data. The training data were incrementally released to the developers because dialogues were collected across multiple recording sessions occurring over months of time. The representative set was created incrementally, too. A sequential process was developed to select dialogues from the data:

1. Sort the dialogues based upon the percentage of unique words each dialogue has in common with the other dialogues in the same collection.
2. Consider the middle 1/3 (approximately) of the sorted dialogues.
3. Sort this set of dialogs by scenario number.
4. Count the number of times each word in a dialogue appears in the other dialogues of the same collection and compute the average for the dialog.

The final step was somewhat subjective: at least one instance of each distinct scenario was selected while aiming for a variety of English and Iraqi Arabic speakers and a maximum value for the average in (4).

6.4 Scenario Adaptation for Evaluation

For the live evaluations, selected dialogues from the representative set were adapted as structured scenarios. The content of the selected dialogues was used to specify the information that role players conveyed when they enacted the evaluation scenarios. Both the SME and FLE were given structured scenario instructions that outlined the scenario background, set the scene, and presented the scenario's outcome. The SME's instructions listed specific information that the SME was required to convey to the FLE or elicit from the FLE by asking questions. These prompts were presented as simple phrases as opposed to complete sentences to prevent the SMEs from reading the lists verbatim.

Instead of prompts, the FLEs were provided with several paragraphs in English outlining the information they were supposed to convey to the SME when appropriately prompted. The pertinent information was **bolded** within the paragraphs, and the FLE was instructed to formulate the specified information in their own words in Iraqi Arabic. Each scenario provided over 35 prompts for the SME along with the responses that the FLE was required to produce for each prompt.

7. Scenario Selection for Evaluation

Because all live evaluation metrics depend on dialogues produced using structured scenarios, it was important to select and construct the scenarios carefully, while balancing a variety of goals. For the offline evaluation, not only the content, but also the quality of the speech had to be considered.

7.1 Field Structured Scenarios

Field scenarios were the first to be selected from the representative set because the field evaluation presented more constraints on appropriate scenarios. This step involved sifting through the representative set to determine which scenarios could be realized in our limited field environment. Dialogues were considered appropriate if they could support the use of a stationary vehicle, if they could be performed in the entrance of an interior or exterior doorway, or if they called for the role players to walk around in a limited area.

In addition to dialogues from the representative set, appropriate scenarios from the January 2007 evaluation were also identified. The IET wanted to include several previously evaluated scenarios in the field (and in the lab) to compare performance on the same scenarios in January and July. Several factors were considered when determining if a previous scenario would be viable again. They included:

- The number of prompts the SMEs were able to complete when using the scenario in the previous evaluation.
- New data that had been distributed to developers since January.
- Uniqueness. A January scenario is more likely to be selected if there are no new July scenarios that are similar in content.

Final selections from the "field appropriate" set were made to achieve a representative proportion of scenario content based upon the percentage of dialogues collected for each sub-domain in the training data.

7.2 Lab Structured Scenarios

Scenarios for the lab evaluations were selected from dialogues in the representative set that had not been

chosen for the field evaluation and from scenarios used in the January 2007 evaluation. In addition, two scenarios selected for the field evaluation were also performed in the lab, as a means of drawing comparisons between the field and lab conditions. One of these repeated scenarios was the one that was performed twice in the field (once with background noise and once without background noise) as an added comparison.

7.3 Offline Dialogues

Utterances for the offline evaluation were taken from dialogues in the representative set. There was a concern that the dialogues collected with human interpreters had features that were quite different than the input users produce when they actually use the systems. Users communicating via speech translation devices quickly realize that they must speak clearly, avoid false starts and filler expressions such as 'uh,' and keep their input short and simple. In contrast, the training data resembled ordinary conversation with high frequencies of filler expressions, pauses, breaths, and unclear speech as well as lengthy utterances.

Because inputs for the offline evaluation were taken directly from the recordings of dialogues in the representative set, the dialogues were selected to minimize disfluent and ill-formed utterances. They were also selected to be representative of scenarios and speaker genders in the training data. Specific utterances within the dialogues were selected in two ways. Half of the utterances were selected randomly from 20 dialogues and half were selected by hand from 10 of those dialogues. The randomly selected utterances were identified by concatenating the dialogues and selecting every n^{th} utterance, where n was the number that would yield about 200 utterances from the total number of utterances in the set for each language. The 200 additional utterances selected by hand for each language minimized disfluencies while preserving the coherence of dialogue exchanges.

Comparing performance on the random vs. hand selected utterances provides an estimate of the effect of disfluencies on offline system scores. Another estimate was provided by rerecording 5 dialogues without disfluencies, which added about 140 inputs for each language to the data that systems processed during the offline evaluation. Condon et al. (2008) provide details about the procedures adopted for the offline evaluation.

8. Metrics

The IET intends the metrics to reflect the end goal of the TRANSTAC program: the deployed use of speech-to-speech MT technology that enables consistently successful communication between American military users and foreign personnel. The TRANSTAC community is in agreement that measures should focus on (1) the semantic adequacy of the translations, leading to justified user confidence in the system's translations, and (2) the ability of an English speaker and foreign language speaker to successfully carry out a task-oriented dialogue in a narrowly focused domain of known operational need under conditions that reasonably simulate use in the field.

Human judgments of the semantic adequacy of the translations and of successful concept transfer are important metrics for the TRANSTAC program. These human judgments of translation quality are generally regarded as a gold standard measure of translation quality and are usually considered valid even when comparing systems that take widely varying approaches to machine translation.

One measure adopted for TRANSTAC evaluations reflects operational task success, bringing into play the usability of the system, the ability of its users to correct misunderstandings, and any abilities the system may have to exploit models of the operational tasks/scenarios. This measure, which has come to be known as *high level concept transfer*, assesses the speakers' success in conveying the information specified in the structured field and lab scenarios.

Another measure adopted to address the semantic adequacy of the translations assesses the end-to-end pipelined performance of the three core technologies (ASR, MT, and TTS) for a sample of inputs from the offline data. A panel of bilingual judges rated the semantic adequacy of the translations by assigning a Likert-type score to each utterance, choosing from a four-point scale:

- *Completely_adequate*
- *Tending_adequate*
- *Tending_inadequate*
- *Inadequate*

In addition, an analyst who is a native speaker of each source language identified the low level elements of meaning (low level concepts) in the sample and then asked the panel of bilingual judges to identify which low level concepts were successfully transferred into the target-language output (where failures are deletions, substitutions, or insertions of concepts). Progress from one evaluation to the next may be presented as an odds ratio. Odds of successful concept transfer is a more quantitative measure of translation adequacy than the Likert-type judgments of semantic adequacy, whereas the Likert-type judgments give the bilingual judges the opportunity to take into account the relative importance of the various concepts, which the low level concept transfer measure does not. Combining the Likert-type judgments of semantic adequacy with the odds of successful transfer of a low level concept gives a fuller picture of semantic adequacy.

Also, SMEs and FLEs in the live evaluations were surveyed in utility assessments to provide formative feedback to the TRANSTAC system developers.

For the full set of offline data, a suite of automated metrics was calculated to enable the developers to better understand the contributions of individual components to the end-to-end success of their systems. The hope is to identify automated metrics that can be run quickly and easily, yet will correlate strongly with judgments of semantic adequacy provided by bilingual judges.

8.1 High level Concept Transfer in Lab and Field Evaluations

For each system, the 20 scenarios from the lab and field evaluations were scored for high level concept transfer. The analysis was performed by reviewing the QA notes, transcriptions of the interactions, system log files, and the audio recordings of each scenario. Each FLE response in the structured scenarios was treated as a high level concept, and the speakers' success in eliciting and conveying the concepts was scored by two judges. The following analyses were recorded for each dialogue:

- The number of concepts that the SME addressed (some prompts were skipped or never reached).
- The amount of time the SME spent trying to retrieve each concept (independent of their success).
- A strict score for each concept, which is either a score of 1 or 0. If the concept was transferred completely, independent of how many attempts it took, the score is 1. If any part of the concept is missing, then the score is 0. For example, if the concept was "the house down the road from the mosque" and the response was "the house down the road", the strict scoring would be 0.
- A loose score for each concept, which is a score of 1, 0.5, or 0. If the concept was transferred completely correctly, independent of the number of attempts, the score would be 1. If part of the concept was transferred, the score would be 0.5. Using the same example concept above, if the response transferred was "the house down the road", then the loose scoring would be 0.5. If none of the concept was transferred correctly, the score would be 0.
- A proper question score for each SME utterance, which is a score of either 1 or 0. A score of 1 indicates that the English speech was adequately translated into Iraqi Arabic. The score is independent of the number of attempts.
- A proper answer score for each FLE response, which is either 1 or 0. A score of 1 indicates that the Iraqi Arabic speech was adequately translated into English. The score is independent of how many attempts were made. If the English utterance was not adequately translated, the Arabic speaker should not have responded. If the Arabic speaker did respond, the utterance is not counted as a proper answer or as a retrieved concept.
- The total number of attempts required to retrieve the answer. An attempt is defined by the number of times that the SME needed to phrase or rephrase the question in an effort to elicit the answer.

8.2 Low level Concept Transfer for Offline Evaluations

The low level concept transfer analysis was developed to assess the semantic adequacy of translations using a method that is more fine-grained than global adequacy scores, but simpler than more complex analyses based on predicate-argument representations (Belvin, Rieheman, & Precoda, 2004) or Interchange Format (Levin et al., 2000). The idea is that the low level elements of meaning are the open-class words: the nouns, verbs, adjectives, and adverbs. Also included are the quantifiers and prepositions that the analyst deemed important to the meaning of the utterance, along with most pronouns. A native-speaker analyst with knowledge of linguistics identified the low level elements of meaning in a sample of about 100 translations from English and 100 to English, for each non-English language from the offline dataset.

For each low level concept, a panel of bilingual judges recorded whether the concept was successfully translated, substituted, or deleted. They also noted insertions of concepts. An interface allowed the judges to view a transcription of the input, the system translation, and the low level concepts that had been identified in the transcription. Judges received training on the task and were able to refer to codified guidelines.

Scores are reported as an odds ratio by dividing the number of concepts successfully translated by 1 minus the number of insertions, substitutions or deletions in the target. Details about the measure are provided in Sanders et al. (2008).

8.3 Likert Scores for Offline Evaluations

Immediately after the bilingual judges completed the analysis of low level concept transfer for the concepts in an utterance, the judges would then assign that same utterance to one of the four adequacy levels. The judges were instructed to first decide whether the translation was more adequate than inadequate, or vice-versa. After making that binary decision, the judges were then to decide the degree of adequacy or inadequacy, making their choice from a four-point Likert-type scale.

Although all five bilingual judges on the panel tended to rank order the systems and/or scenarios the same, they differed as to their judgments of what level of performance constituted the four Likert-type levels. Some judges were consistently harsher and some consistently easier, with the harshest and easiest judge averaging about one Likert-level apart. The judges did receive instruction for the decisions, including a set of examples for each of the four Likert levels. This is discussed in more detail by Sanders et al. (2008).

8.4 Automated Metrics

The automated metrics focus on the core technologies using system outputs from the offline evaluation. System logs preserved the results of ASR, and translations were produced both from the speech inputs and from transcriptions of the speech inputs so that MT could be measured with and without ASR errors.

For ASR, we calculated Word-Error-Rate (WER) — using the Speech Recognition Scoring Toolkit (SCTK) version 2.2.2 and the standard NIST procedures for normalizing the hypothesis and reference texts, thus giving English WER values that should be directly comparable to previous large-scale NIST evaluations of automatic speech recognition (Information Access Division, 2007). BLEU scores (Papineni et al, 2002) were calculated for MT. MT

performance was also measured by calculating METEOR and Translation Edit Rate (TER) scores (Bannerjee & Lavie, 2005; Snover et al., 2005). TER was calculated using TerCom version 6b. METEOR normalization was modified to handle Arabic text. For all three languages, METEOR was run in the mode where it scores only exact matches (no stemming or synonymy).

8.5 Post Scenario/Session Questionnaires for Lab and Field Evaluations

Both SMEs and FLEs were asked to complete questionnaire survey instruments following each scenario in which they participated. They were administered an additional post-session questionnaire for each system on completion of the 10 scenarios for the lab or field evaluation in which they participated. Some questions required free-form responses, and others employed a Likert-scale response format from 1 to 5, with 1 being the lowest and 5 being the best score possible.

The Likert-scale statements that were evaluated after each scenario include:

- *What the system said made sense to me*
- *Based upon my experience in this interaction, I would use this system for future similar interactions*

Some of the post-session statements and questions that were evaluated are:

- *The <my language> words were put together in a way that was coherent and comprehensible*
- *How confident were you in the system's ability to help you communicate effectively?*

In addition to individual system results, survey responses provided some contrasts between the lab and field evaluations. In general, SMEs and FLEs who participated in the lab scenarios assigned lower ratings to the systems than SMEs and FLEs who participated in the field scenarios. In addition, FLEs assigned lower ratings than SMEs.

9. Conclusions

Detailed results of the evaluations cannot be reported due to restrictions on releasing the data. However, some anonymous results are presented in Sanders et al. (2008) and Condon et al. (2008), including comparisons among the different measures that were employed in the TRANSTAC evaluations.

The NIST IET learned numerous lessons throughout the TRANSTAC Phase 2 evaluation process and expects to use this knowledge in the design and implementation of the Phase 3 evaluations. Some measures have methodological implications. For example, the high level concept analysis indicates that systems generally performed better in the field than in the lab. SMEs and FLEs appeared to be more engaged in the field scenarios, though other factors may have contributed to better performances in the field: English utterances were not masked in the field, and speakers may have produced more easily processed speech in reaction to the background noise.

The order in which systems were evaluated seems to impact their performance, and a measurable difference was obtained between the performance of the first system evaluated and the performance of the same system several days later. Some improvements currently being explored include scheduling all systems' scenarios in parallel, conducting all evaluations in field conditions, and introducing noise-masking in the field environment.

10. NIST disclaimer

Certain commercial products and software are identified in this paper in order to explain our research. Such identification does not imply recommendation or endorsement by NIST, nor does it imply that the products and software identified are necessarily the best available for the purpose.

11. Acknowledgements

This work is funded through the DARPA TRANSTAC program and the authors greatly acknowledge the support of the TRANSTAC program manager, Dr. Mari Maeda.

12. References

Bannerjee, S. and A. Lavie (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005), pp. 65-73.

Belvin, R., Riehemann, S., and K. Precoda. (2004). A Fine-Grained Evaluation Method for Speech-to-Speech Machine Translation Using Concept Annotations. In *Proceedings of LREC 2004*, pp. 1427-1430.

Condon, S., Phillips, J., Doran, C., Aberdeen, J., Parvaz, D., Oshika, B., Sanders' G. and C. Schlenoff. (2008) Applying automated metrics to speech translation dialogs. In *Proceedings of LREC 2008.*

Information Access Division (2007). Tools – Evaluation Tools. http://www.nist.gov/speech/tools/.

Levin, L. Gates, D., Lavie, A., Pianesi, F., Wallace, D., Watanabe, T., and M. Woszczyna. (2000). Evaluation of a practical interlingua for task-oriented dialogue. In *Proceedings of the NAACL-ANLP 2000 Workshop on Applied interlinguas: practical applications of interlingual approaches to NLP, Volume 2*, pp. 18-23.

Papineni, K., Roukos, S., Ward, T., and W-J. Zhu. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pp. 311-318.

Sanders, G., Bronsart, S., Condon, S., and C. Schlenoff. (2008). Odds of successful transfer of low-level concepts: A key metric for bidirectional speech-to-speech machine translation in DARPA's TRANSTAC program. In *Proceedings of LREC 2008.*

Schlenoff, C., Steves, M., Weiss, B., Shneier, M., and A. Virts. (2007). Applying SCORE to field-based performance evaluations of soldier worn sensor technologies. *Journal of Field Robotics*, Volume 24, (8-9), pp. 671-698.

Snover, M., Dorr, B., Schwartz, R., Makhoul, J., Micciulla, L., Weischedel, R. (2005). A Study of Translation Error Rate with Targeted Human Annotation. LAMP-TR-126, CS-TR-4755, UMIACS-TR-2008-58, University of Maryland, College Park.