Evaluating Speech Translation Systems: Applying SCORE to TRANSTAC Technologies

Craig Schlenoff
National Institute of Standards and Technology
100 Bureau Drive, Stop 8230
Gaithersburg, MD 20899
301-975-3456
craig.schlenoff@nist.gov

Brian Weiss
National Institute of Standards and Technology
100 Bureau Drive, Stop 8230
Gaithersburg, MD 20899
301-975-4373
brian.weiss@nist.gov

Michelle Potts Steves
National Institute of Standards and Technology
100 Bureau Drive, Stop 8940
Gaithersburg, MD 20899
301-975-3537
michelle.steves@nist.gov

Greg Sanders
National Institute of Standards and Technology
100 Bureau Drive, Stop 8940
Gaithersburg, MD 20899
301-975-4451
greg.sanders@nist.gov

Fred Proctor
National Institute of Standards and Technology
100 Bureau Drive, Stop 8230
Gaithersburg, MD 20899
301-975-3425
frederick.proctor@nist.gov

Ann Virts
National Institute of Standards and Technology
100 Bureau Drive, Stop 8230
Gaithersburg, MD 20899
301-975-5068
ann.virts@nist.gov

ABSTRACT

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 translation systems that enable speakers of different languages to communicate with one another in real-world tactical situations without an interpreter.

The National Institute of Standards and Technology (NIST), along with support from MITRE and Appen Pty Ltd., have been funded to serve as the Independent Evaluation Team (IET) for the TRANSTAC Program. The IET is responsible for analyzing the performance of the TRANSTAC systems by designing and executing multiple TRANSTAC evaluations and analyzing the results of the evaluation.

To accomplish this, NIST has applied the SCORE (System, Component, and Operationally Relevant Evaluations) Framework. SCORE is a unified set of criteria and software tools for defining a performance evaluation approach for complex intelligent systems. It provides a comprehensive evaluation blueprint that assesses the technical performance of a system and its components through isolating variables as well as capturing end-user utility of the system in realistic use-case environments.

This document describes the TRANSTAC program and explains how the SCORE framework was applied to assess the technical and utility performance of the TRANSTAC systems.

Categories and Subject Descriptors

I.2.7 [**Computing Methodologies**]: Natural Language Processing – *machine translation, speech recognition and synthesis*

General Terms

Algorithms, Measurement, Performance, Experimentation, Human Factors, Languages

Keywords

Performance evaluation, speech-to-speech translation system, SCORE, TRANSTAC

1. INTRODUCTION[1]

Performance evaluation of advanced technologies can often be very challenging. It is the authors' belief that the design of an effective evaluation is as much a research issue as is the

[1] 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.

PerMIS'09, September 21-23, 2009, Gaithersburg, MD, USA.
ACM 978-1-60558-747-9/09/09

technology development itself. One must be able to accurately answer questions such as:

- Does the overall system do what it claims to do?
- What are the factors that would cause the overall system to fail?
- Is the system useful to the end-user (whether it be military, law enforcement, first responders, industry, etc.)?
- What are the key situations that the technology would be most useful for?
- How well do the individual components of the system perform and what is their impact on the performance of the overall system?
- How can we isolate specific capabilities of the system and test their performance?

In order to address this, the SCORE Framework (System, Component, and Operationally Relevant Evaluations) Framework was developed. SCORE is a unified set of criteria and software tools for defining a performance evaluation approach for complex intelligent systems. It provides a comprehensive evaluation blueprint that assesses the technical performance of a system and its components through isolating and changing variables as well as capturing end-user utility of the system in realistic use-case environments. [1]

The paper is organized as follows: Section 2 gives an overview of the TRANSTAC effort; Section 3 describes the SCORE framework; Section 4 describes how the SCORE Framework was applied to assess the TRANSTAC systems, Section 5 describes the metrics used in the TRANSTAC program, and Section 6 concludes the paper.

2. OVERVIEW OF THE DARPA TRANSTAC PROGRAM[2]

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 translation systems that enable speakers of different languages to communicate with one another in real-world tactical situations without an interpreter.

Several prototype systems have been developed under this program for numerous military applications including force protection and medical screening. The technology has been demonstrated on PDA (personal digital assistant) and laptop platforms. NIST was asked to assess the usability of the overall translation system and to individually assess each component of the system (the speech recognition, the machine translation, and the text-to-speech).

[2] Due to DARPA restrictions, the results of the evaluation cannot be published. Instead, this paper will focus on the evaluation approach as opposed to the results.

All of the TRANSTAC systems work fundamentally the same. Either English speech or an audio file is fed into the system. Automatic Speech Recognition (ASR) processes the speech to recognize what was said and generates a text file of the speech. That text file is then translated to another language using Machine Translation (MT) technology. The resulting text file is then spoken to the foreign language speaker using Text-To-Speech (TTS) technology. This same process then happens in reverse when the foreign language speaker speaks. This is shown in Figure 1.

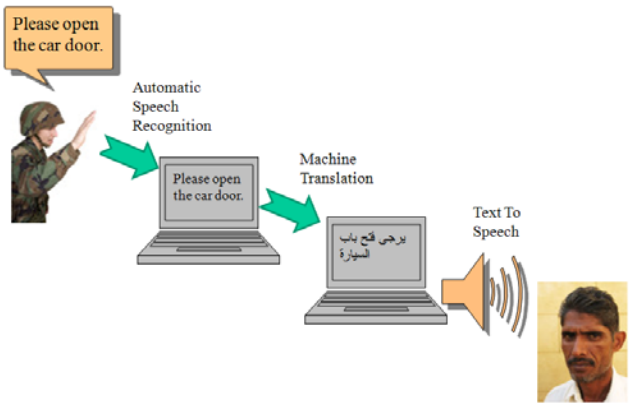

Figure 1: How Speech Translation Works

3. OVERVIEW OF THE SCORE FRAMEWORK

The SCORE Framework [2] [3] has been developed at the National Institute of Standards and Technology (NIST) over the past three years to provide formative evaluations of advanced technologies that are still under development. SCORE is built around the premise that, in order to get a true picture of how a system performs in the field, it must be evaluated at the component level, the system level, the capability level and within operationally-relevant environments.

SCORE is unique in that:

- It is applicable to a wide range of technologies, from manufacturing to defense systems
- Elements of SCORE can be decoupled and customized based upon evaluation goals
- It has the ability to evaluate a technology at various stages of development, from conceptual to fully mature
- It combines the results of targeted evaluations to produce an extensive picture of a systems' capabilities and utility

To date, SCORE has been used to evaluate a wide range of advanced technologies, including Soldier-worn sensor systems, technologies allowing real-time multimedia information sharing among Soldiers in the field, two-way speech translation systems, and autonomous robotic platforms. It has been the foundation for ten technology evaluations involving Soldiers and Marines from

around the country. SCORE has been used as the basis of two DARPA programs to evaluate advanced technologies.

SCORE defines five evaluation goal types, as shown in Figure 2:

- *Component Level Testing – Technical Performance* – involves decomposing a system into components to isolate those subsystems that are critical to system operation.
- *Capability Level Testing – Technical Performance* – involves decomposing a system into capabilities (where the complete system is made up of multiple capabilities). A capability can be thought of as an individual functionality, such as the ability for a sensor system to send and receive a picture or the ability for a translation system to identify and translate names (discussed below).
- *Capability Level Testing – Utility Assessments* –assesses the utility of an individual capability. The benefit of this evaluation type is that specific capability utility and usability to the end-user can still be addressed even when the system and user-interface are still under development.
- *System Level Testing – Technical Performance* –assesses the system as a whole, but in an ideal environment where test variables can be isolated and controlled. The benefit is that tests can be performed using a combination of test variables and parameters, where relationships can be determined between system behavior and these variables and parameters based upon the technical performance analysis.
- *System Level Testing – Utility Assessments* –assesses a system's utility, where utility is defined as the value the application provides to the system's end-user. In addition, usability is assessed. which includes effectiveness, learnability, flexibility, and user attitude towards the system.

Considering each of these evaluation elements, SCORE takes a tiered approach to measuring the performance of intelligent systems. At the lowest level, SCORE uses elemental tests to isolate specific components and then systematically modifies variables that could affect the performance of those components to determine those variables' impact. Typically, this is performed for each relevant component with the system. At the next level, the overall system is tested in a highly structured environment to understand the performance of individual variables on the system as a whole. Then, individual capabilities of the system are isolated and tested for both their technical performance and their utility using task tests. Lastly, the technology is immersed in a longer scenario that evokes typical situations and surroundings in which the end-user is asked to perform an overall mission or procedure in a highly-relevant environment which stresses the overall system's capabilities. Formal surveys and semi-structured interviews are used to assess the usefulness of the technology to the end-user.

4. APPLYING SCORE TO TRANSTAC

Technical performance of the individual components of the TRANSTAC system was performed using offline tests (represented by the red arrow in Figure 3). Both technical and utility performance of the entire system was performed using lab-based evaluations of a laptop-based system (represented by the gray arrows in Figure 3) and more field-friendly utility systems (represented by the green arrows in Figure 3). Utility evaluations were also performed out in the field with the field-friendly systems (represented by the blue arrow in Figure 3). Lastly, the specific capabilities of the TRANSTAC systems (such as their ability to recognize proper names) were tested both for their technical capability and their utility (represented by the purple arrows in Figure 3). Each of these tests is discussed in detail below.

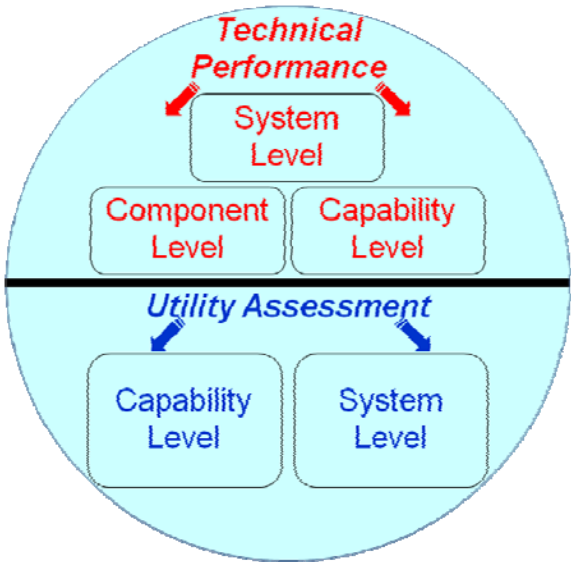

Figure 2: SCORE Architecture

4.1 Offline Evaluations

The offline evaluation was performed to assess the technical performance of the TRANSTAC systems at the component level. There were three primary components that were being tested: the Automated Speech Recognition, the Machine Translation, and the Text-to-Speech. The offline evaluation was performed so that the component evaluation would be conducted on identical inputs for all systems. In advance of the evaluation, research teams were provided with the required log formats for storing the results of the offline processing. They were also provided with sample offline data that could be used to develop logging scripts and produce sample outputs. A verification script was provided to check the output for log format errors.

During the offline evaluation, research teams provided the same versions of their systems that were used for the live evaluation. Research teams were provided with audio files for speech recognition and subsequent translation. Separately, they were provided with transcription files for text translation.

Each system processed approximately 1000 audio files of utterances in each language and stored the results in system logs. In the context of this paper, an utterance is the words spoken by a

human from the time s/he starts speaking to the time that the TRANSTAC system begins to translate. An utterance can contain one or many concepts (individual pieces of information), but efforts have been made to have a comparable average number of concepts among all offline utterances from one evaluation to the next.

For each audio file, the system stored the results of ASR, the translation based on ASR output, and time stamps marking the beginning and end of each process (recognition, translation, and TTS, if used). Outputs from the transcription inputs were the same except that results related to speech recognition were left blank. When processing was complete, a verification script was run on the logs to ensure that the output conformed to the required format. Logs were also checked for the correct number of outputs.

In addition to the above, thirty well-formed foreign language text strings were fed into the TTS engine of the TRANSTAC systems. These engines read in the text strings and output audio files which contained the spoken version of the text.

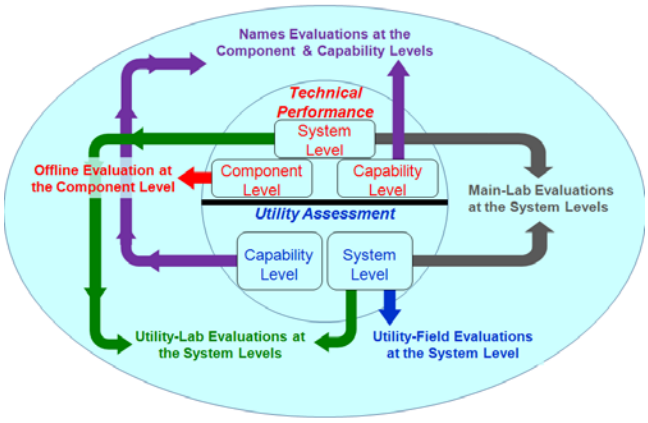

Figure 3: SCORE Applied to TRANSTAC

Analysis on the offline evaluation focused on component level performance of the TRANSTAC systems using automated metrics and human judgments. The following metrics were used to analyze the offline data:

- Human Judgment
  - Low-level concept transfer, performed by bilingual human judges
  - Likert judgment [4] at utterance level, performed by bilingual human judges
  - Likert judgment performed by bilingual human judges, to assess TTS
- Automated Metrics
  - Word Error Rate (WER) to assess ASR and TTS
  - METEOR, BLEU – to assess ASR and MT together

More details about these metrics can be found in Section 5.

4.2 Lab-Based Evaluations

The main difference between the offline evaluation described in Section 4.1and the live lab and field evaluations described in Sections 4.2 and 4.3 is that the live evaluations allow speakers to generate their own utterances of inquiries and responses while the offline evaluations uses scripted, recorded utterances by both speakers to provide an apples-to-apples comparison.

Figure 4 shows an example of one of the teams' TRANSTAC system. The main processing unit is a standard laptop, in which a head-mounted microphone (top right) and a speaker (bottom right) are plugged in. A hand-held control (bottom left) is as plugged into the laptop which allows the Soldier/Marine to let the system know when each speaker is about to talk. Each 'START' button corresponds to a different speaker and the button on the bottom allows the Soldier/Marine to replay the last audio that was output from the TRANSTAC system.

Lab-based evaluations were used to assess the technical capability and utility of the TRANSTAC systems at the systems level. Approximately twenty scenarios are used to assess the performance of the TRANSTAC systems in a lab setting. These scenarios have either been structured scenarios or spontaneous scenarios. Structured scenarios provide a set of questions to the English speaker that they needed to find answers to. The foreign language speaker was given the answers to those questions in paragraph format. A dialogue occurred between the two speakers and the number of answers that the English speaker was able to obtain was noted.

Figure 4: Example of Laptop-Based TRANSTAC System

For spontaneous scenarios, a brief paragraph was provided to the English and foreign language speaker to give them the proper background to carry on a meaningful conversation. The background could state that they were performing a census survey and were going house to house gathering information about peoples' living conditions. The direction that the Soldier/Marine

takes the conversation was up to them, as long as it is within the bounds of the scenario description. There are advantages and disadvantages to both types of scenario, which is outside the scope of this paper. However, in both cases, the goal was to measure the number of meaningful interactions that the Soldier/Marine and the foreign language speaker has in a finite amount of time.

In addition, after the interaction, questionnaires were provided to the English and foreign language speakers to gauge their perception of the TRANSTAC systems.

All scenarios were performed in an indoor environment, usually in a conference room of a hotel. The Soldier/Marine and the foreign language speakers were stationary, with the TRANSTAC system on the table between them. All lab scenario runs were performed in this environment, with each scenario occurring within a ten minutes period. Noise masking technology was deployed to stop the speakers from hearing each other. They could only respond to what came out the TRANSTAC system. The goal of this type of evaluation is to place the systems in what many would consider an ideal environment (no background noise, minimal movement, etc.) to get an upper bound on how well they could perform.

Because there were two physical systems (a laptop version and a more field-friendly) we used the same lab-based evaluation procedures for both systems.

For the lab-based evaluations, the following metrics were used to analyze the data:

- A count of high-level concepts found out by the Soldier/Marine in response to the questions he asked.
- Analysis of the questionnaire performed by Soldiers/Marines and foreign language speakers after each scenario in which they participated.

More details about these metrics can be found in Section 5.

4.3 Field-Based Evaluations

The field-based evaluations were used to assess the utility of the TRANSTAC systems at the system level. The field scenarios were performed outdoors with Soldiers/Marines wearing combat gear (body armor, helmet, gloves, etc.). They carried a "utility version" of the TRANSTAC systems while performing the scenarios. Following the scenarios, the Soldiers/Marines filled out questionnaires and participated in interview sessions with the evaluation team.

The field environments were not intended to be completely representative of what the Soldiers/Marines would experience overseas. To replicate this type of environment would be a very difficult undertaking and it would not tell us much more than a more simplistic environment would. The reason for performing field evaluations was to subject the systems to the type of environmental variable that they would realistically be exposed to, such as wind, background noise, and the motion caused by the Soldier/Marine carrying the systems around with them. It also allowed the user to see how easy the system was to use while carrying around other gear such as bullet-proof vests and weapons.

Figure 5: Example TRANSTAC Utility System

An example of a utility version of the TRANSTAC system is shown in Figure 5. The "YOU" button on the microphone was meant to be push when the Soldier/Marnie was speaking (since they are the controller of the systems) and the "HIM" button was meant to be pushed when the foreign language speaker was speaking. A sample field environment that was used for testing is shown in Figure 6.

For the field-based evaluations, the following metrics were used to analyze the data:

- Analysis of the questionnaire performed by Soldiers/Marines and foreign language speakers after each scenario in which they participated.
- Semi-structured interviews with the Marine/Soldiers and foreign language speakers.

More details about these metrics can be found in Section 5.

Figure 6: Sample Field Environment

4.4 Proper Names Evaluations

The proper names evaluation was an example of a capability evaluation used to assess specific functionalities of the TRANSTAC system. The goal of the capability evaluation was to isolate specific functionalities of a system and test its performance with scenarios that are tailored to stress that functionality. The evaluation team focused on the ability for the TRANSTAC system to identify and convey proper names in a dialogue. In this context, proper names were people names, street names, and city names that were being conveyed from the foreign language speaker to the English speaker. Three unique, names-laden scenarios were created as scripted dialogues and recorded by unique speakers. Each scenario was very rich in proper names; they typically contained approximately 50 to 55 proper names within the 30 to 40 foreign language utterances. This recorded data was used to create the offline names evaluation.

The offline names evaluation was run similar to that of the other offline evaluations. Specific recorded utterances were selected and fed directly into the TRANSTAC systems. However, the metrics from this test focus on how the systems specifically handle the translations of the proper names, as discussed below.

The live names evaluation was run in a different manner than that of the live lab evaluation. The speakers were provided with the scripted names scenarios and instructed to read them verbatim into the TRANSTAC system. After hearing TRANSTAC translation of the English utterance, the foreign language speaker responded with their corresponding scripted utterance which again was spoken into the TRANSTAC system. That foreign language utterance was then translated into English. If the English speaker was able to understand the name that was translated/conveyed by the TRANSTAC system, they noted that and moved on to the next utterance. If the English speaker was unable to ascertain a name from the TRANSTAC output, then they were able to rephrase their original English utterance in any manner they saw fit. Likewise, the foreign language speaker, upon hearing the TRANSTAC output once the English speaker rephrased their utterance, could rephrase theirs accordingly to convey the desired name. The output of this evaluation produced both technical performance and utility assessment data. This took the form of measuring the number of names successfully transferred per unit time and collecting survey responses from the end-users regarding their specific names interactions. There were three names scenarios that were performed during the evaluation.

To evaluate the live and offline names evaluation, each TRANSTAC output was analyzed to see how well the proper name was translated from the foreign language to English. This was performed by a panel of human judges. A score was provided to each output which classified each name translation as either:

- Right name, right pronunciation
- Right name, wrong pronunciation
- Name translated as word (these were the cases where a proper name can also have a separate meaning… Black could be a person's last name or a color)
- Wrong name translation
- Name not recognized

5. METRICS APPLIED TO TRANSTAC

In order to get a comprehensive picture of the performance of the TRANSTAC system, a large number of performance metrics were used when evaluating the systems. Many of these metrics are described below. The TRANSTAC community is in agreement that the two aspects that best characterize the performance of the systems are: (1) the semantic adequacy of the translations, leading to justified user confidence in the system's translations, and (2) the ability of Marines/Soldiers and foreign language speakers 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. The metrics that were use to assess these capabilities are:

1. **High-Level Concept Transfer:** Semantic adequacy of the translations was assessed by bilingual judges telling us *whether* the meaning of each utterance came across. The high-level concept metric is the number of utterances that are judged to have succeeded. Thus, failed utterances are not directly scored (other than taking up time). The high-level concept metric is an efficiency metric which shows the number of successful utterances per unit of time, as well as accuracy. This metric is roughly quantitative.

2. **Likert Judgment:** A judgment of the semantic adequacy of the translations was performed by having a panel of bilingual judges rate the semantic adequacy of the translations, an utterance at a time. We asked our panel of five bilingual judges to assign a Likert-type score to each utterance, choosing from a seven-point scale.

   +3 Completely_adequate
   +2
   +1 Tending_adequate
   0
   –1 Tending_inadequate
   –2
   –3 Inadequate.

   The judges were provided with a substantial set of exemplars showing utterances which were deemed to correspond to the four values (completely adequate, tending adequate, tending inadequate, inadequate) and were asked to choose the in-between values only if on the fence between two of those values.

3. **Low-Level Concept Transfer:** A directly quantitative measure of the transfer of the low-level elements of meaning in each utterance. In this context, a low-level concept is a specific content word (or words) in an utterance. For example, the phrase "The house is down the street from the mosque." is one high-level concept, but is made up of three low-level concepts (house, down the street, mosque).

   We had an analyst who is a native speaker of each source language identify the low-level elements of meaning (low-level concepts) in representative sets of input utterances from

the offline datasets and then asked a panel of five bilingual judges to tell us 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 — the Likert-type judgments give the bilingual judges the opportunity to take into account the relative importance of the various concepts while the low-level concept transfer does not. [4]

4. **Automated Metrics:** A suite of automated metrics, intended to enable the research team to better understand what aspects of performance account for the end-to-end success of their systems. We hope to identify automated metrics that can be run quickly and easily yet will correlate strongly with judgments of semantic adequacy provided by bilingual judges. For speech recognition, we calculated Word-Error-Rate (WER) — using SCTK version 2.2.2 and automated procedures for normalizing the hypothesis and reference texts. For machine translation, we calculated BLEU [5] using four reference translations. We also measured MT performance by calculating a metric called METEOR defined by Alon Lavie of CMU. For both English and Dari, METEOR was run in the mode where it scores only exact matches (no stemming or synonymy). [6]
5. **TTS Evaluation:** To assess the performance of a TTS component, human judges listened to the audio outputs of the TTS evaluation and compared them to the text string of what was fed into the TTS engine. They then gave a Likert score from 1-5 (five being the best) to indicate how understandable the audio file was in comparison to what was fed into it. In addition, these human judges transcribed what they heard in the audio file in the foreign language and then these transcriptions were compared to the input text files using Word Error Rate.
6. **Surveys/Semi-Structured Interviews:** After each live scenario, the Soldiers/Marines and the foreign language speakers filled out a detailed survey asking them about their experiences with the TRANSTAC systems. The surveys explored how easy the system was to use, how well they perceived it worked, and errors that the users encountered when interacting with the system. In addition, after the field scenarios, semi-structured interviews were performed with all of the participants in which questions such as "What did you like?, What didn't you like? and What would you change?" were explored.

6. METRICS COMPARISON

Although I cannot discuss detailed results in this paper due to DARPA restrictions, I can discuss, at a meta-level the level of consistency that was found by applying these metrics to the teams' TRANSTAC output. For the purpose of this comparison, I will show the rank ordering of the teams' performance by applying the follow metrics described in Section 5: high-level concept transfer, low-level concept transfer, Likert judgment, BLEU, and METEOR. This is shown in Table 1.

Table 1: Metrics Comparison

LANGUAGE DIRECTION	Metric	Team 1	Team 2	Team 3
Dari to English	High Level Concept Transfer	1	2	3
Dari to English	Low-level Concept Transfer	1	2	2
Dari to English	Likert Judgment	1	2	2
Dari to English	BLEU	1	2	2
Dari to English	METEOR	1	2	2
English to Dari	High Level Concept Transfer	1	2	3
English to Dari	Low-level Concept Transfer	1	2	2
English to Dari	Likert Judgment	1	1	1
English to Dari	BLEU	1	2	2
English to Dari	METEOR	1	1	1

As shown in Table 1, the numbers under Team 1, Team 2, and Team 3 show their relative score compared to each other teams when applying the metrics in the second column. For example, Team 1 had the highest relative score applying the high-level concept transfer metric looking at the translation from Dari to English. Team 2 had the second highest score and Team 3 had the third highest score. When two teams have the same number in the same row, it means that the scores were not statistically significant enough to be able to say that one score was better than the other. For example, Team 2 and Team 3 have very comparable scores when applying the low-level concept transfer metric in the Dari to English direction; hence they are both listed as the second ranked team.

The table shows that there is significant comparability in the overall results when applying different metrics. In the Dari to English direction, Team 1 consistently was ranked #1 in all of the metrics applied and Team #2 was consistently ranked #2. The only difference was that there was a statistical difference between Teams 2 and Team 3 when applying the low-level concept transfer metric, where there was not a statistical difference when applying the other metrics.

When looking at the English to Dari direction in Table 1, Team 1 came out with the highest relative rank in all five metrics again. However, Teams 2 and Team 3's scores varied depending which metrics was applied. Looking at Team #3, it was ranked third when applying the high-level concept transfer metric but was tied for first when applying the Likert judgment and METEOR metrics. In situations like this, one usually defaults to the metrics which involves humans, which is sometimes referred to as ground truth or the gold standard. The first three metrics (high-level concept transfer, low-level concept transfer, and Likert) all involved human judges. Unfortunately, this still doesn't provide much insight as Team 3 is ranked #3, #2, and #1, respectively. As

such the only conclusion we can draw from this is that Team #1 appears to be superior overall, while Team #2 and Team #3 are roughly tied for second.

7. CONCLUSION

In this paper, we have discussed the SCORE Framework and shown how it was applied to the DARPA TRANSTAC program. Using SCORE, we were able to evaluate the performance of speech translation systems by looking at the performance of:

- the systems at the component level using offline evaluations,
- the performance of the overall system in ideal environments using lab evaluations,
- the performance of the system in operationally-relevant environments using field test, and
- the specific capabilities of the systems to evaluate proper names.

By putting together the results of all of these evaluations, we are able to gain a much more comprehensive evaluation of an overall system performance.

SCORE has proven to be an invaluable evaluation design tool for the NIST Evaluation Team and was the backbone of eleven DARPA evaluations: six for the DARPA ASSIST program (not discussed in this paper) and five for TRANSTAC program. It is expected to play a critical role in the remaining ASSIST and TRANSTAC evaluations.

The SCORE framework is applicable to domains beyond emerging military technologies and those solely dealing with intelligent systems. Personnel at NIST are applying the SCORE framework to the virtual manufacturing automation competition (VMAC) [7] and the virtual RoboRescue competition [8] (within the domain of urban search and rescue). Their intent is to develop elemental tests and vignette scenarios to test complex system capabilities and their component functions. The framework has proven to be highly adaptable and capable of meeting most any evaluation requirement.

ACKNOWLEDGMENTS

The authors would like to acknowledge the DARPA TRANSTAC program manager, Dr. Mari Maeda, and the members of the NIST IET for their continued support.

REFERENCES

[1] B. Weiss, C. Schlenoff, G. Sanders, M. Steves, S. Condon, J. Phillips, and D. Parvaz, "Performance Evaluation of Speech Translation Systems," in *Proceedings of the LREC 2008 Conference* Morocco: 2008.

[2] C. Schlenoff, M. Steves, B. Weiss, M. Shneier, and A. Virts, "Applying SCORE to Field-Based Performance Evaluations of Soldier Worn Sensor Technologies," *Journal of Field Robotics*, vol. 24, no. 8/9, pp. 671-698, Sept.2007.

[3] B. Weiss and C. Schlenoff, "Evolution of the SCORE Framework to Enhance Field-Based Performance Evaluations of Emerging Technologies," in *Proceedings of the 2008 Performance Metrics for Intelligent Systems (PerMIS) Conference* Gaithersburg, MD: 2008.

[4] G. Sanders, S. Bronsart, S. Condon, and C. Schlenoff, "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 the LREC 2008 Conference* Morocco: 2008.

[5] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)* Philadelphia, PA: 2002, pp. 311-318.

[6] S. Condon, J. Phillips, C. Doran, J. Aberdeen, D. Parvaz, B. Oshika, G. Sanders, and C. Schlenoff, "Applying Automated Metrics to Speech Translation Dialogs," in *Proceedings of the LREC 2008 Conference* Morocco: 2008.

[7] S. Balakirsky and R. Madhavan, "Advancing Manufacturing Research Through Competitions," in *Proceedings of the SPIE Defense Security and Sensing* Orlando, FL: 2009.

[8] S. Balakirsky, C. Scrapper, and S. Carpin, "The Evolution of Performance Metrics in the RoboCup Rescue Virtual Robot Competition," in *Proceedings of the Performance Metrics for Intelligent Systems Workshop* Gaithersburg, MD: 2007.