Applying Automated Metrics to Speech Translation Dialogs

Sherri Condon*, Jon Phillips*, Christy Doran*, John Aberdeen*, Dan Parvaz*, Beatrice Oshika*, Greg Sanders† and Craig Schlenoff†

*The MITRE Corporation
McLean, Virginia 22102

†National Institute of Standards and Technology
Gaithersburg, Maryland 20899

E-mail: {scondon, jphillips, cdoran, aberdeen, dparvaz, bea}@mitre.org, {gregory.sanders, craig.schlenoff}@nist.gov

Abstract

Over the past five years, the Defense Advanced Research Projects Agency (DARPA) has funded development of speech translation systems for tactical applications. A key component of the research program has been extensive system evaluation, with dual objectives of assessing progress overall and comparing among systems. This paper describes the methods used to obtain BLEU, TER, and METEOR scores for two-way English-Iraqi Arabic systems. We compare the scores with measures based on human judgments and demonstrate the effects of normalization operations on BLEU scores. Issues that are highlighted include the quality of test data and differential results of applying automated metrics to Arabic vs. English.

1. Introduction

The Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) program is a Defense Advanced Research Projects Agency (DARPA) research and development program. The goal of the TRANSTAC program is to demonstrate capabilities for rapid development and fielding of two-way translation systems that enable speakers of different languages to communicate with one another in real-world tactical situations. The primary use cases involve US military personnel in limited conversations with local foreign language speakers. Several prototype systems have been developed for military and medical screening domains in Iraqi Arabic, Mandarin, Farsi, Pashto, and Thai on both PDA and laptop-grade platforms.

Since the inception of the DARPA speech translation programs, a MITRE team has coordinated with system developers to collect training data and design evaluation methods. More recently, The National Institute of Standards and Technology (NIST) has directed the effort to assess the progress of system development and evaluate the systems' readiness for fielding. This report is one of several that describe the evaluation methods developed for the TRANSTAC program (Sanders et al., 2008; Weiss et al., 2008).

In the initial stages of development, the focus of evaluations has been on the basic functionality of speech recognition and machine translation, and a major goal has been tests incorporating users and domains that are representative of the military uses for which the systems are designed. Consequently, a major challenge of developing useful evaluation methods for the TRANSTAC program has been the conflict between replicability and authenticity. Test conditions that most closely resemble real-world conditions of use require spontaneous interaction between representative users with meaningful goals in realistic situations and environments. However, these conditions are not repeatable due to the inevitable variation in human behavior.

The strategy adopted for TRANSTAC evaluations has been to conduct two types of evaluations: live evaluations in which users interact with the translation systems according to several different protocols and offline evaluations in which the systems process audio recordings and transcripts of interactions. Because the inputs in the offline evaluation are the same for each system, we analyzed translations using automated metrics. Measures such as BiLingual Evaluation Understudy (BLEU) (Papineni et al., 2002), Translation Edit Rate (TER) (Snover et al., 2006), and Metric for Evaluation of Translation with Explicit word Ordering (METEOR) (Banerjee & Lavie, 2005) have been developed and widely used for translations of text and broadcast material, which have very different properties than dialog. The TRANSTAC evaluations provide an opportunity to explore the applicability of automated metrics to translation of spoken dialog. The evaluations also offer a chance to study the results of applying automated metrics to languages other than English, since studies of the measures have primarily involved translation to English and other European languages closely related to English.

The report begins with a brief discussion of measures of translation quality and issues that have been raised concerning automated metrics. Section 3 describes training data collected for the TRANSTAC program, and section 4 explains the procedures used to create offline test corpora for the 2007 evaluations of English-Iraqi Arabic speech translation systems. Section 5 presents results of the offline evaluations and compares them with measures involving human judgments. Section 6 explores the effects of normalization operations on BLEU scores and preliminary results concerning data quality issues.

2. Previous Work

Researchers have recognized that translation quality is multi-faceted and that human judgments of even more specific qualities such as fluency and fidelity are not always reliable (King, 1996; Turian, Shen & Melamed, 2003). Given the unevenness and cost of human judgments, researchers have welcomed automated measures such as BLEU and have proposed a plethora of alternative methods, all of which involve comparisons to one or more reference translations.

In contrast, evaluations of speech translation have relied on human judgments such as the binary or ternary classifications adopted by CMU (Gates et al., 1996) and Verbmobil (Nübel, 1997) researchers, which combine assessments of accuracy and fluency. Other methods use abstract semantic representations of the source utterances and require human judges to score structural elements of those representations separately. CMU researchers use the interlingua Interchange Format to represent utterance intent and content (Levin et al., 2000), and Belvin, Rieheman, & Precoda (2004) use predicate-argument structures. The TRANSTAC program has experimented with several types of human scoring, and these are described in Sanders et al. (2008).

Automated metrics were selected for the TRANSTAC offline evaluation because each system processes the same set of recorded inputs, and reference translations can be prepared for that set of utterances. Also, as Lita, Rogati & Lavie (2005) observe, BLEU and measures derived from BLEU have become de facto standards in the MT community. As automated measures are used more extensively, researchers learn more about their strengths and shortcomings, which allows the scores to be interpreted with greater understanding and confidence. Some of the limitations that have been identified for BLEU are very general, such as the fact that its precision-based scoring fails to measure recall, rendering it more like a document similarity measure (Culy & Riehemann, 2003; Lavie, Sagae, & Jayaraman, 2004; Owczarzak, van Genabith, & Way, 2007). In addition to BLEU, the TRANSTAC program uses METEOR to score translations of the recorded scenarios with a measure that incorporates recall on the unigram level.

METEOR also addresses another problem that has been associated with BLEU. The ability of BLEU to take into account many possible translations for a given segment of language depends on the number of reference translations that are available for comparison. METEOR uses WordNet synonyms to allow for lexical variation that is not present in reference translations. Also, METEOR uses stemming to remove inflectional affixes that may prevent translations from matching due to minor variation. For example, TRANSTAC data includes variation such as contractions (*I'm* vs. *I am*) because it is transcribed from spoken interactions. However, these enhancements are available only for English, and we investigate the effects of normalization operations on automated scores for both English and Iraqi Arabic in section 6.

A known limitation of the BLEU metric is that it only indirectly captures sentence-level features by counting n-grams for higher values of *n*, but syntactic variation can produce translation variants that may not be represented in reference translations, especially for languages that have relatively free word order (Chatterjee, Johnson & Krishna, 2007; Owczarzak, van Genabith, & Way, 2007; Turian, Shen, & Melamed, 2003). The TRANSTAC program also uses TER to measure translations of the recorded scenarios, which allows for some syntactic variation because any number of contiguous words can shift positions in a single edit.

Another issue that is relevant to TRANSTAC evaluations concerns the quantity of data required for reliable automated measures. Offline evaluations are conducted during the live evaluations so that the number of inputs that systems process is limited by available scheduling time. Also, training data is difficult to collect (see Section 3) so that it is important to hold as little as possible back for evaluation. Fortunately, some recent work suggests that samples as small as 300 sentences can be sufficient to correctly detect significant differences between systems, though bootstrap sampling is recommended (Koehn, 2004; Zhang & Vogel, 2004).

A related concern is the length of the inputs, which has particular importance for TRANSTAC data because spoken utterances tend to be shorter than written ones. For example Turian, Shen, & Melamed (2003) report that samples of reference translations from TIDES corpora averaged about 31 words per sentence, whereas 30 words is considered a maximum for inputs to the TRANSTAC speech translation systems. In the next section the data collected to train TRANSTAC systems is described, and additional features of those data that might affect the use of automated metrics are discussed.

3. TRANSTAC Training Corpora

Initially, TRANSTAC stakeholders agreed that domains and use cases should be narrowly defined in order to provide realistic goals for the speech translation systems. However, it quickly became clear that even the most routine interactions can easily veer out of domain when, for example, the driver at a checkpoint tries to explain why he has a sack of money in the trunk. Interviews with veterans of military operations in Iraq and Afghanistan initially resulted in about 50 scenarios that were used to elicit interactions in 6 domains, including checkpoints, searches, infrastructure surveys (sewer, water, electricity, trash, etc.), and training. Later, another 30 scenarios were developed with more diverse topics such as medical screening, inspection of facilities, and recruiting for emergency service professionals.

Scenarios provide each role-player with a description that sets the scene, identifies the role of the speaker, provides some background and motivation for the speaker, and may describe an outcome for the encounter. For example, the military speaker might be asked to imagine that he is at a checkpoint, that a car driven by a young man has approached, that a search of the car revealed a large bag of cash in the trunk, and that the man is detained for further questioning. Scenarios included an example interaction or suggested topics for discussion. Role-players were coached to prepare for their roles before recording.

A variety of protocols were used in order to take advantage of role-players available at different data collection events and to maximize the number of interactions that were recorded. Most of the 630 dialogs recorded in 2005 consisted of a male English speaking soldier or Marine interacting with a male Iraqi Arabic speaking civilian via a male bilingual interpreter, and another 570 were recorded in 2006 using the same protocol, except that about 25% of the participants were female. This protocol made it possible to obtain a maximum amount of speech from the very limited time that we had access to military personnel. Most of these dialogs were recorded using a protocol in which an inoperable telephone handset or similar prop was passed to each role-player before he or she could begin to talk, which minimized overlap among the speakers.

Because it was difficult to schedule speakers with military experience, some of the dialogs recorded in 2005 and another 1235 interactions recorded in 2006 followed a protocol in which both roles were played by Iraqi Arabic speakers and no interpreter was required. Other protocols among the 2005 recordings included military English speakers playing both roles with an Iraqi Arabic speaker interpreting so that Arabic versions were recorded, too. Additional data were collected by eliciting answers to prerecorded questions from native Iraqi Arabic speakers, and one of these collections was designed to elicit names of people, places, and organizations.

All of the interactions were transcribed orthographically, and the transcriptions were translated into the other language (English to Arabic or Arabic to English) by professional transcribers and translators. Transcription and translation conventions were developed with input from developers, NIST, the Linguistic Data Consortium (LDC), and MITRE. Portions of the Arabic data were transcribed phonetically, and some diacriticized lexica were created. Transcriptions included timestamps at the beginning and end of each segment. Some recordings, transcriptions, and translations from the 2006 data were not distributed to the developers so that they could be used for evaluation. These data are referred to as the *reserved* data (see section 4).

The data collection protocols resulted in speech that differs from the inputs that users produce when interacting with speech translation devices. Users communicating via a translation device quickly realize that they must speak clearly, avoid false starts and filler expressions such as ‘uh,’ and keep inputs 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. Some examples are provided in (1).

(1) a. then %AH how is the water in the area what's the -- what's the quality how does it taste %AH is there %AH %breath sufficient supply?

b. the -- the first thing when it comes to %AH comes to fractures is you always look for %breath %AH fractures of the skull or of the spinal column %breath because these need to be* these need to be treated differently than all other fractures.

c. would you show me what part of the -- %AH %AH roughly how far up and down the street this %breath %UM this water covers when it backs up ?

The examples in (1) illustrate the filler expressions such as ‘um’ and ‘uh,’ which are transcribed ‘%UM’ and ‘%AH,’ and false starts, which are represented by dashes, in the data. In the initial reserved scenarios for the January evaluation, 27% of the segments contained dashes, 42% contained annotations with ‘%’ and 28% contained annotations for unintelligible speech. 10% of the training data contains segments that are 26 or more words in length, including 5% over 30 words, many of which exceed 40 and even 50 words.

Another source of mismatches between training data and live evaluation inputs is in the transcription. Transcribers were instructed to divide sequences of speech from a single speaker into smaller units at reasonable logical break points. The guidelines indicate that there was ongoing clarification of this directive, and it is clear that divisions were inconsistently applied. For example, (1a) contains four separate questions, and (1b) was divided in the middle of a sentence where the asterisk appears in the text. There can be good reasons not to separate every distinct sentence-like unit in a steady stream of speech. If speakers do not pause between these units, then the speech cannot be divided cleanly due to coarticulation.

4. Selection of Data for Evaluation

Training data were collected, processed, and released as separate corpora based on the data collection events at which they were produced. In order to identify a representative reserved set from each corpus, the vocabulary in each dialog was analyzed to provide the following information:

1. Total word tokens and word types in the dialog
2. Number of tokens and types that are unique to the dialog
3. Percentage of tokens and types in the dialog that occur in other dialogs
4. Number of times a word in the dialog appears in the corpus: average for all words

From the dialogs that were in the mid-range for the percentage of word types that occurred in other dialogs, reserved dialogs were chosen so that each scenario topic was covered, a variety of speakers were represented, and the score in (4) above was maximized. Ideally, the speakers in the reserved data would be distinct from the speakers in the training data, but there were not enough recordings to achieve this goal for the smaller data collections that included English speakers. Approximately 10% of the recordings were reserved.

Before each evaluation event, the sets of reserved scenarios were analyzed, and a summary of information relevant to selecting the scenarios was produced. This information included the scenario topics, gender of the speakers, the number of English and Arabic utterances, and information about the lengths of utterances in the scenarios. Selection of specific audio inputs for the offline evaluation required several passes through the pool of scenarios available for the offline corpus. In the first pass, complete dialogs for the offline evaluation were selected. For the January 2007 evaluation, there were very few dialogs available to select from, and most of the available dialogs had to be included.

From the selected dialogs, individual utterances were identified as candidates for the offline audio inputs. Utterances were selected to satisfy the following goals:

1. Proportions of male and female speakers are similar to proportions in the training set
2. Utterance lengths do not exceed 30 words with preference for 5 - 15 words in length
3. Minimize the frequency of false starts, pauses and filled pauses
4. Avoid utterances that do not preserve structural and semantic coherence
5. Avoid utterances that appear to overlap with other utterances according to the timestamps
6. At least 400 utterances in each language

After an initial pass through the dialogs to select utterances for an initial count, a second pass finalized the choices by eliminating additional utterances that were less desirable according to the criteria, while still preserving the goal of 400 inputs per language. Because there was a minimal amount of data, only the worst offenders of criteria 2-4 were excluded. The final set of offline inputs in January 2007 consisted of 415 English inputs and 437 Arabic inputs. Timestamps were used to segment the audio recordings into a separate clip for each input.

The January 2007 process was so labor intensive that in the July 2007 procedure, we experimented with a purely random selection method. After selecting 20 appropriate scenarios from the reserved data, half of the utterances in each language were identified by selecting every *n* utterances, where *n* was chosen so that 200 utterances would be selected from all the segments of the language in the 20 scenarios. In order to maintain some dialog continuity, the remaining 200 utterances for each language were hand selected from 10 of the 20 scenarios following the procedures described above for the January evaluation, yielding 419 English utterances and 429 Iraqi Arabic utterances.

An additional set of 138 English and 141 Arabic utterances was prepared by editing transcripts of 5 dialogs and rerecording them without disfluencies. These *rerecorded* dialogs were produced in order to compare results on dialogs that were similar to the original offline inputs in structure and content, but without the repetitions, false starts, and fillers that characterize those inputs.

In both January and July 2007, text inputs were produced from the transcriptions of the selected segments, in order to provide measures of translation quality that were independent of the speech recognition. Consequently, each offline evaluation produced a set of results that included speech recognition WER for each language and BLEU, TER, and METEOR translation scores for spoken inputs as well as BLEU, TER, and METEOR scores for textual inputs.

5. Automated Measures and Human Judgments

WER was measured using the NIST SCLite scoring software. In scoring English ASR for the TRANSTAC evaluation, NIST first modified the reference transcriptions, replacing each occurrence of an English contraction with the most likely expansion for that occurrence in its context. Further, words such as *gonna*, *wanna*, *'em* and *'cause* that represent phonological reduction are replaced by the unreduced equivalent. Words that are usually written as a single word are replaced by that form. Hyphenated words are rewritten as multiple words (replacing hyphen by space). Similar re-writes are done to the system output, except that contractions are replaced by an alternation, so that either version can match the reference. The net result of normalizing the system output and reference transcription files is to increase the number of matches (lowering the WER), make fairer comparisons among systems, and increase repeatability.

The evaluation used a variant of BLEU (bleu_babylon.pl) provided by IBM that produces the same result as NIST's variant of BLEU when there is no zero match, except that the score is normalized between 0 and 1. For situations where zero matches occur, this implementation uses a penalty of log(0.99/# of n-grams in the hypothesis) to compute the final score. METEOR was modified to exclude some default normalizations that assume a Western orthography when it was run on the Arabic texts. Four reference translations were created for each input.

A sample of 95 English to Arabic and 101 Arabic to English translations was also scored using two methods that involved human judgments. In one method, which will be referred to as *Likert judgments*, bilinguals classified the translations as *completely adequate*, *tending adequate*, *tending inadequate* and *inadequate*. The speaker turns that were judged were drawn from 5 of the dialogs used for the automated scoring in Figure 1 and from the 5 rerecorded dialogs. The same translations were scored using another method, developed by NIST, in which each open class content word (c-word) in the source utterance was identified, and bilingual judges determined whether the word had been successfully translated, deleted, or substituted in the target utterance. The measure, which NIST refers to as *low-level concept transfer,* is computed as an odds score by dividing the number of c-words successfully translated by 1 minus the number of insertions, substitutions or deletions in the target. Details about these measures are provided in Sanders et al. (2008).

In the live evaluation, military English speakers and Iraqi Arabic speakers were asked to role play scenarios using the translation systems. In order to maintain consistency in the content of the unscripted interactions as they were repeated for each system, the same speakers were required to obtain and provide the same specific information using each system. Scores were based on a binary human judgment of translation adequacy for inputs produced in 20 ten-minute dialogs. The live evaluations are described in greater detail in Weiss et al., 2008.

Figure 1 presents the results of the automated scores for the rerecorded data pooled with the reserved data from the July 2007 evaluation. The five systems that were evaluated are consistently labeled with the letters A through E in all of the figures. It is important to note that the graphs in Figure 1 map the WER as 1 - WER and TER as (100 – TER)/100 to make them more comparable. For comparison, Figure 2 presents the results of the three other evaluation measures for each system. The scores labeled *Live* in Figure 2 are from the live evaluation and reflect the average proportion of 30 utterances, usually questions, successfully communicated by the English-speaking user and the average proportion of 30 responses successfully communicated by the Arabic speaker in 10 minutes. The scores labeled *Likert* are the human Likert judgments, and Figure 2 presents the proportion of translations that was judged to be either *adequate* or *tending adequate*. Finally, the scores labeled *Concept* were produced using the low-level concept transfer scoring described above. However, instead of odds ratios, the values in Figure 2 are computed by dividing the number of c-words judged to be correctly translated by the sum of the number judged to be correctly translated plus the sum of the errors.

The major difference between scores from the automated translation measures and those involving human judgments is the difference between scores for English to Arabic translations and those for Arabic to English translations. According to all the automated translation metrics, scores for the former are much lower than scores for the latter. This difference will be referred to as a *directional asymmetry*, and it occurs in spite of the fact that the speech recognition error rate ranges from 15% to 35% higher for Iraqi Arabic. In contrast, the scores involving human judgments exhibit the opposite asymmetry with English to Arabic translations scoring consistently, though sometimes only slightly, higher than Arabic to English translations.

For the most part, the BLEU, TER, and METEOR scores produce similar patterns of results, which closely match the patterns of human judgments for the Arabic to English translations. In the January 2007 evaluation, there were clear differences between higher scoring systems A, B, and C vs. lower scoring systems D and E, except in the concept transfer scoring, where differences between the two groups were smaller for Arabic to English translations. The July results exhibit a similar pattern for Arabic to English translations, and systems A, B, and C also performed better than systems D and E in the live evaluations. However, the automated English to Arabic scores are less consistent and less similar to the human judgments.

While systems A and B consistently score higher than the other systems, scores for systems C and D are closer for English to Arabic than for Arabic to English translations. In part, this reflects the more accurate English speech

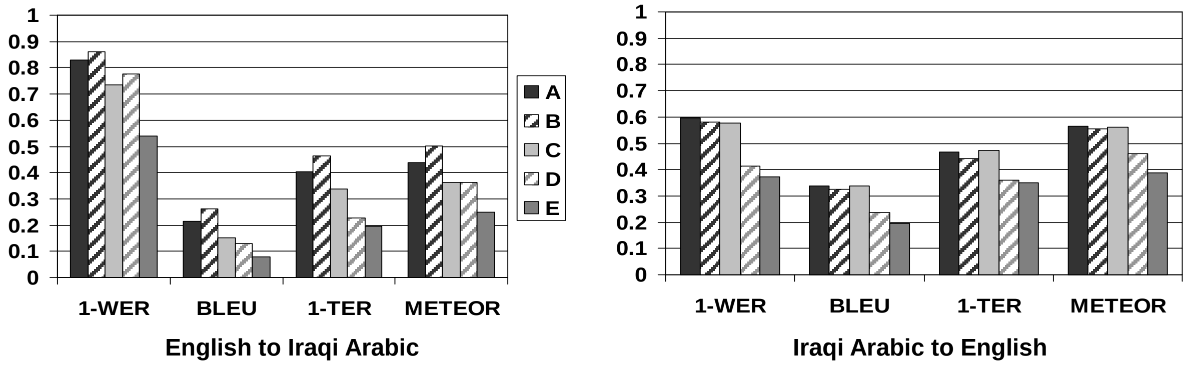

Figure 1: Automated Measures for Translations and Speech Recognition for Systems A - E

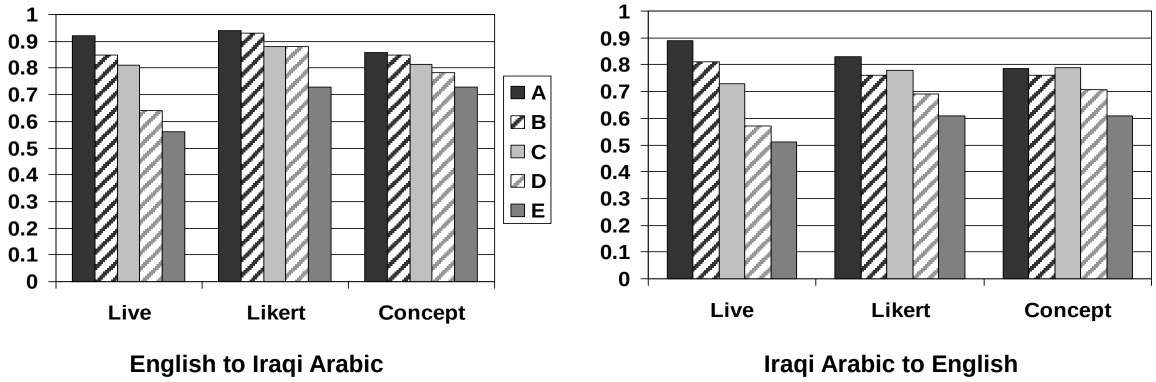

Figure 2: Translation Quality Measures Involving Human Judgments for Systems A - E

recognition achieved by system D. Like the BLEU and METEOR scores, the Likert and concept transfer measures also assign similar scores to systems C and D for English to Arabic translations. Sanders et al. (2008) provide more detail about the correspondence between the automated measures and the human judgments.

Another result from the automated measures concerns a system which employs rule-based translation for English to Arabic. There is increasing evidence that BLEU scores tend to be higher and better correlated with human judgments for statistical machine translation systems, especially those that have been optimized using BLEU, compared to systems developed using approaches that are not n-gram based (Callison-Burch, Osborne, & Koehn, 2006; Coughlin, 2003; Doddington, 2002). The system that uses rule-based translation for English to Arabic also employs a statistical translation component as a fallback, which made it possible to compute scores using both types of translation within the same system. Using the statistical method, the system's scores increase .012 for BLEU and .013 for both 1-TER and METEOR.

6. Normalization for Automated Metrics

Because automated scoring approaches depend on the ability to match word forms in system outputs to the reference texts, spelling variation, punctuation, encoding issues and morphological variation can result in mismatches between what are otherwise semantically equivalent words. These issues motivate the stemming procedures that are employed for METEOR measures of translation to English. For morphologically rich languages such as Arabic, there is an even greater likelihood that inflectional differences with little effect on meaning will result in lower scores, which could contribute to the directional asymmetries observed in the automated metrics in Figure 1.

MITRE is investigating methods to normalize the inflection variation while trying to leave only the distinctions that affect meaning. Meanwhile, our focus has been on orthographic and lexical variation. For each language, we used two types of normalization. *Rule based* normalization involves a series of language-specific rules to deal with systematic issues such as removing certain types of punctuation or conflating two Unicode forms for an equivalent Arabic character to a single representation. The *GLM* normalization employs a list of lexical substitutions and is used to conflate spelling variants, following NIST's use of global lexical mappings for computing WER. Normalization is applied to both system outputs and reference texts.

For English, rule-based normalization involved the application of three simple rules to remove end of sentence punctuation, orthographic case, and underscores from acronyms (e.g., *I_D* => *ID*). The case-normalization rule is redundant for WER since the NIST WER-scorer ignores case by default. GLM-normalization involved 194 lexical substitutions for contractions (e.g., *you're* => *you are*) and spelling variants (*kilometre*=> *kilometer*, *Muhammad*=>*Mohamed*). Figure 3 presents BLEU scores for Arabic to English translations computed without normalization (NORM0), with rule based normalization only (NORM1), and with both rule based and GLM normalization (NORM2).

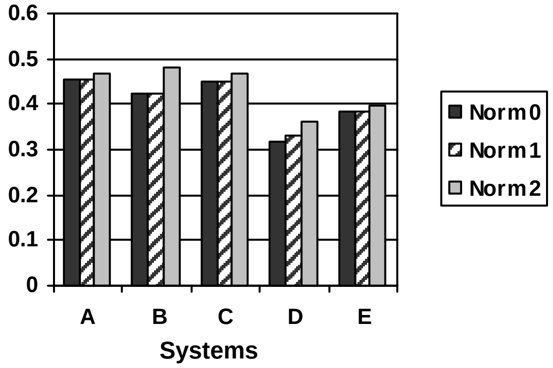

Figure 3: Iraqi Arabic to English BLEU Scores with 2 Types of Normalization

Scores were largely unaffected by NORM1, except for system D, which experienced a large improvement due to the removal of end-of-sentence punctuation in the system outputs. NORM2 more consistently resulted in improvements across all the systems, and the average increase in BLEU scores from NORM0 to NORM2 was 0.28. WER for English speech recognition was also computed using the normalization procedures, and the average WER after NORM2 decreased from 28.06% to 25.28%.

For the rule-based normalization of Iraqi Arabic, an extensive list of character conversions was applied to remove short-vowels and diacritics, and some additional rules normalized *teh marbuta* and various *alif-hamza* variants. NORM2 used a series of 1495 lexical substitutions. These substitutions were generated by examining the reference evaluation texts for possible spelling variants. However, many of the lexical substitutions were redundant with normalizations applied by the rules in NORM1.

Figure 4 presents BLEU scores for English to Arabic translation computed without normalization (NORM0), with rule based normalization only (NORM1), and with both rule based and GLM normalization (NORM2). NORM1 resulted in improvements for all systems, while the NORM2 substitutions proved to be mostly redundant. The average increase in BLEU scores from NORM0 to NORM2 was 0.14. The average Iraqi Arabic WER after NORM2 decreased from 48.54% to 46.56%.

Finally, preliminary results of our investigations of data quality are available. The motivation for this effort was to estimate the extent to which scores are affected by the disfluencies in the training data. Speech directed to a human interpreter differs from the more careful speech that users employ when they use the translation systems. We compared system scores for the inputs that were hand-selected to minimize disfluencies to scores for the

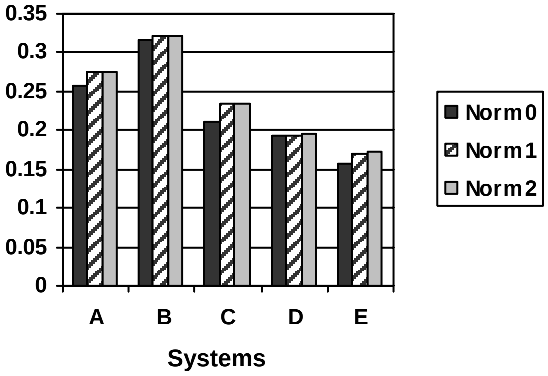

Figure 4: Iraqi Arabic to English BLEU Scores with 2 Types of Normalization

inputs that were randomly selected. Raw comparisons show that systems performed better on the hand-selected inputs compared to the randomly selected inputs, especially for scores that depend on processing Iraqi Arabic speech. Table 1 presents the average WER and BLEU scores that the systems obtained on the randomly selected data and the hand selected data. We are performing additional analyses to assess the statistical significance of these results.

In contrast to the results in Table 1, there were no differences in the BLEU scores for translations from transcripts of the speech, which suggests that the differences in the BLEU scores are entirely due to the speech recognition problems introduced by the more disfluent speech in the random sample. The fact that there were no differences between translations from the text versions of the two sets of input has another consequence. Randomly selected inputs did not maintain the conversational coherence of the inputs, whereas we tried to preserve English-Arabic exchanges such as questions and answers in the hand-selected inputs. The inputs were presented for processing in order, respecting the adjacency and sequencing of the exchanges. The lack of differences in the translation scores for random vs. hand-selected textual inputs suggests that systems were not affected by the presence or absence of context for the utterances.

Another set of results also provides some evidence that speech disfluencies in the test data cause decreases in scores from automated measures that would not occur with the more careful speech that users adopt when they interact with the translation systems. Scripts for the rerecorded dialogs were produced from the same dialogs that hand-selected (and randomly selected) inputs were drawn from. Consequently, there are 86 English and 80 Arabic inputs that have the same content, except that disfluencies are eliminated in the rerecorded dialogs. Unfortunately, the speakers in the rerecorded dialogs are different so that the results are confounded by speaker differences, including the fact that speakers in the original test data also occurred in the training data.

Average of 5 systems	Random	Hand
English WER	26.26	24.48
Iraqi Arabic WER	48.86	45.08
Arabic to English BLEU	0.289	0.316
English to Arabic BLEU	0.163	0.176

Table 1: Average WER and BLEU Scores Based on Random vs. Hand Selection of Speech Inputs

Table 2 presents the average WER and BLEU scores that the systems obtained on the original inputs and the rerecorded inputs. The improvement in English WER is entirely due to a decrease of 25% in system E: there was essentially no difference in the English WER for 3 systems, and system B's WER increased almost 8% on the rerecorded inputs. Reflecting the WER differences, System B's BLEU score was about .05 lower for the rerecorded inputs, whereas systems D and C each obtained a BLEU score about .04 higher for the rerecorded inputs. In contrast, the results for Iraqi Arabic were consistent: all systems performed much better on the rerecorded dialogs with large decreases in WER and smaller increases in BLEU scores. Of course, because the text versions of the inputs are nearly identical, there are no differences in the scores for those versions.

Average of 5 systems	Original	Rerecorded
English WER	26.36	23.70
Iraqi Arabic WER	50.76	35.54
Arabic to English BLEU	0.260	0.334
English to Arabic BLEU	0.178	0.187

Table 2: Average WER and BLEU Scores Based on Original vs. Rerecorded Speech Inputs

7. Conclusions and Further Research

This report presents results of applying familiar automated metrics to data that has not typically been evaluated using those measures. BLEU, TER, and METEOR scores have primarily been used to evaluate textual data or transcripts of broadcast data, whereas TRANSTAC speech is dialogic. Moreover, those automated measures have been more extensively used to measure translation to English and other European languages than translation to languages like Arabic. The results presented here suggest that the different automated measures produce very similar patterns of scores for the five systems that were evaluated, and in the case of translations from Arabic to English, these patterns resemble those from the human judgments of subsets of the same data (Likert and concept transfer scores).

The patterns of automated scores for translations from English to Iraqi Arabic are less similar to the Likert and concept transfer judgments data with the latter exhibiting smaller differences among the systems. An unexpected result that we are continuing to investigate is the strong directional asymmetry in the translation measures. English to Arabic translations receive lower BLEU, TER, and METEOR scores than Arabic to English translations in spite of the fact that error rates for English speech recognition are much lower than for Arabic. And this pattern contrasts with the human judgments, which assign higher scores to the English to Arabic translations. We are testing the hypothesis that the freer word order and more extensive inflectional morphology of Arabic result in lower scores from automated metrics.

A significant change in the offline evaluations for the next phase of the TRANSTAC program will be production of test sets with speakers that do not appear in the training data. In addition, we are recommending that some training data be collected using the speech translation devices so that systems can be trained on data that reflects features of speech that users adopt when they use the devices. We look forward to analyzing the differences between the human and machine mediated dialogs.

The TRANSTAC evaluations continue to provide fertile ground for the development of measures to assess speech translation. They have provided opportunities to experiment with a variety of evaluation methods and metrics. They are also providing a richer understanding of automated metrics applied to dialog and to non-Western languages, which will afford greater confidence in the results of those measures.

8. Acknowledgments

We take this opportunity to appreciate TRANSTAC's DARPA program manager, Dr. Mari Maeda, whose leadership has sustained the program's successes. Many colleagues contributed to the enormous effort required to conduct the TRANSTAC evaluations and analyze the results, including Luma Ateyah, Sebastién Bronsart, Tony Downs, John Garofolo, Amir Hekmati, Greg Krill, Yale Marc, Emile Morse, Frank Nasrawi, Larry Sager, Michelle Potts Steves, Ann Virts, and Brian Weiss.

9. References

Banerjee, S. and A. Lavie. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, 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.

Callison-Burch, C., Osborne, M., and P.,Koehn. (2006). Re-evaluating the role of BLEU in Machine Translation Research. In *Proceedings of EACL 2006*, pp. 249-256.

Chatterjee, N., Johnson, A., and M. Krishna. (2007). Some Improvements over the BLEU Metric for Measuring Translation Quality for Hindi. In *Proceedings of the International Conference on Computing: Theory and Applications 2007*, pp. 485-90.

Coughlin, D. (2003). Correlating automated and human assessments of machine translation quality. In *Proceedings of MT Summit IX*, pp. 63-70.

Culy, C. and S. Riehemann. (2003). The Limits of N-gram Translation Evaluation Metrics. In *Proceedings of the MT Summit IX, AMTA*, pp. 71-18.

Doddington, G.. (2002). Automatic Evaluation of Machine Translation Quality Using N-Gram Co-Occurrence Statistics. In *Proceeding of the Second International Conference on Human Language Technology*, pp. 138-145.

Gates, D., Lavie, A., Levin, L., Waibel, A., Gavalda, M. Mayfield, L., Woszczyna, M., and P. Zhan. (1996). End-to-End Evaluation in JANUS: a Speech-to-speech Translation System. In *Proceedings of the ECAI Workshop on Dialogue Processing in Spoken Language Systems*, pp. 195-206.

King, M. (1996). Evaluating natural language processing systems. *Communications of the ACM*, 39, pp. 73-79.

Koehn, P. (2004). Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of EMNLP 2004.*

Lavie, A., Sagae, S., and S. Jayaraman. (2004). The Significance of Recall in Automatic Metrics for MT Evaluation. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004)*, pp. 134–143.

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.

Lita, L.V., Rogati, M., and A. Lavie. (2005). BLANC: Learning Evaluation Metrics for MT. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pp. 740–747.

Nübel, R. (1997). "End-to-End Evaluation in Verbmobil I," In *Proceedings of MT Summit VI*, pp. 232-239.

Owczarzak, K., van Genabith, J., and A. Way. (2007). Dependency-Based Automatic Evaluation for Machine Translation. In *Proceedings of HLT-NAACL 2007 AMTA Workshop on Syntax and Structure in Statistical Translation*, pp. 80-87.

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.*

Snover, M., Dorr, B., Schwartz, R., Makhoul, J., and L. Micciula. (2006). A Study of Translation Error Rate with Targeted Human Annotation. *In Proceedings of AMTA 2006*, pp. 223-231.

Turian, J.P., Shen, L. and I. D. Melamed. (2003). Evaluation of Machine Translation and Its Evaluation. In *Proceedings of MT Summit 2003*, pp. 386-393.

Weiss, B., Schlenoff, C., Sanders, G., Steves, M., Condon, S., Phillips, J., and Parvaz, D. (2008). Performance Evaluation of Speech Translation Systems. In *Proceedings of LREC 2008*.

Zhang, Y. and S. Vogel. (2004). Measuring confidence intervals for the machine translation evaluation metrics. In *Proceedings of TMI 2004*, pp. 85-94.