Automated Metrics for Speech Translation

Sherri Condon, Mark Arehart, Christy Doran, Dan Parvaz, John Aberdeen, Karine
Megerdoomian, Beatrice Oshika, and Greg Sanders[†]

The MITRE Corporation
7525 Colshire Drive
McLean, VA 22102
00-1-703-983-5522

[†]National Institute of Standards and Technology
100 Bureau Drive, Stop 8940
Gaithersburg, MD 20899-8940
00-1-301-975-4451

{scondon/marehart/cdoran/dparvaz/aberdeen/karine/bea}@mitre.org, [†]gsanders@nist.gov

ABSTRACT

In this paper, we describe automated measures used to evaluate machine translation quality in the Defense Advanced Research Projects Agency's Spoken Language Communication and Translation System for Tactical Use program, which is developing speech translation systems for dialogue between English and Iraqi Arabic speakers in military contexts. Limitations of the automated measures are illustrated along with variants of the measures that seek to overcome those limitations. Both the dialogue structure of the data and the Iraqi Arabic language challenge these measures, and the paper presents some solutions adopted by MITRE and NIST to improve confidence in the scores.

Categories and Subject Descriptors

D.2.8 [**Software Engineering**]: Metrics – *performance measures*
I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *machine translation*

General Terms

Measurement, Performance, Experimentation

Keywords

Speech translation evaluation, automated translation metrics, machine translation

1. INTRODUCTION

While human judgments are considered to be the gold standard for evaluating translation performance, it is the development of automated evaluation metrics that has facilitated significant advances in machine translation technology during the last decade. Unlike evaluation methods that involve human judgments, automated measures provide rapid, reliable feedback with relatively low cost. Both human judgments and automated metrics are limited in ways that are still not fully understood, and this report reveals some additional characteristics concerning the application of automated measures to speech translation between English and Iraqi Arabic.

The Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) program has experimented with several evaluation strategies and metrics. Since the inception of the Defense Advanced Research Projects Agency (DARPA) speech translation programs, a MITRE team has coordinated with system developers to design collection methods for training data and evaluation methods to measure progress. More recently, The National Institute of Standards and Technology (NIST) has directed these efforts, and the MITRE team has focused on automated metrics.

The evaluations have focused on the basic functionality of speech recognition and machine translation, and a major goal has been tests that incorporate users and domains which are representative of the military uses for which the systems are designed. Consequently, a significant challenge of developing useful evaluation methods for the TRANSTAC program has been the conflict between replicability and authenticity. Test conditions resembling real-world conditions 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. The inputs in the offline evaluation are the same for each system, and the automated measures used to evaluate system performance on those inputs produce scores in the same way each time they are computed. Therefore, the same tests can be repeated as the systems mature. Automated measures such as BiLingual Evaluation Understudy (BLEU) [10], Translation Edit Rate (TER) [13], and Metric for Evaluation of Translation with Explicit word Ordering (METEOR) [1] have been developed and widely used for translations of text and broadcast material, which have very different properties than dialogue. The TRANSTAC evaluations provide an opportunity to explore the applicability of automated metrics to translation of spoken dialogue and to compare these metrics to human judgments from a panel of bilingual judges.

The evaluations also offer a chance to study the results of applying automated MT metrics to languages other than English. Studies of the measures have primarily involved translation to English and other European languages related to English. The TRANSTAC data present some significant differences between the automated measures of translation into English and Arabic, and our research has provided some insights into the reasons for these differences.

2. AUTOMATIC TRANSLATION METRICS

2.1 The BLEU Measure

A fundamental problem of translation evaluation is that there are many possible translations from a source language input to a target language output. The IBM researchers who developed BLEU in 2001 provided a partial solution to this problem by creating test sets with more than one translation for each input. The machine translation output is then compared to these reference translations, and a score is computed based on the number of n-grams in the output that match the references. For example, Figure 1 provides a sample machine translation from Iraqi Arabic to English along with 4 reference translations.

In Figure 1, 11 of the 12 words in the system output can be matched to words in the reference translations, producing a score of 11/12 for unigram matches. There are 11 bigrams (sequences of 2 words) in the system output, and 5 of them correspond to bigrams in the reference translations: *he has, stomach pain, pain and, and always,* and *pain in*. Therefore, the bigram score is 5/11. The trigram score is 1/10: only *pain and always* can be matched to the references, and there are no matching 4-grams. The BLEU score is computed by micro-averaging [4] the n-gram scores of all the outputs in the test corpus, for n = 1, 2, 3, and 4. Then the geometric mean of the four n-gram averages is computed. Finally, the result is multiplied by a "brevity penalty."

The brevity penalty is assessed because without it, the score would not reflect portions of the reference translations that were completely missed. For example, suppose we added *and I don't know what to do* to each of the reference translations. Without the brevity penalty, the BLEU score would not be affected. The brevity penalty lowers the BLEU score in proportion to the difference between the number of words in the system outputs and the number of words in the reference translations whose lengths are closest to the lengths of the outputs (combined across the entire test corpus).

The example illustrates some of the limitations of the BLEU metric. Although the system output is not fluent English, the meaning expressed in the reference translations is easily inferred from the system output. The BLEU score cannot discriminate between a translation like the system output in Figure 1 and a translation like (1), which has the same number of matching n-grams.

(1) *he has some abdomen and always my and he says in his*

The n-gram matching treats all words equally, regardless of their significance for the meaning. In the extreme case, a semantically loaded word like *not* is treated no differently than an optional conjunction like *and*.

It has been observed that BLEU and measures derived from BLEU have become de facto standards in the machine translation community [7]. 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 observed earlier that the measure primarily reflects the accuracy of the words that the system produced with only a brevity penalty to assess what the system may have missed. This makes the measure more like a document similarity measure [9]. In fact, researchers often use information retrieval terms to describe this problem. BLEU scores measure *precision*: the proportion of words or documents that were correctly translated or retrieved compared to the total words or documents that were translated or retrieved. BLEU scores do not measure *recall*: the proportion of words or documents that were correctly translated or retrieved compared to the total words or documents that should have been translated or retrieved.

2.2 The METEOR Measure

Researchers have proposed dozens of alternative measures that seek to improve on BLEU, while retaining the basic insight of comparing system outputs to multiple reference translations. Many of these measures were compared in the NIST Metrics for Machine Translation 2008 Evaluation (MetricsMATR08) [8]. In addition to BLEU, the TRANSTAC program uses METEOR to score translations of the recorded dialogues. METEOR incorporates a unigram recall score that can yield higher correlations with human judgments than BLEU scores [1].

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 solely on the number of reference translations that are available for comparison. In contrast, METEOR accepts synonyms defined in a resource called WordNet [17], allowing additional options that are not present in reference translations. For example, METEOR would recognize the equivalence of *pain* and *ache* in Figure 1. METEOR also uses stemming to remove inflectional affixes that may prevent translations from matching due to minor variation. For example, after stemming, METEOR would match *cries* and *crying* in Figure 1 because they are both forms of the verb *cry*. However, these enhancements are available only for English: there is no equivalent of WordNet for Iraqi Arabic, and Arabic affixes are often ambiguous out of context, making it difficult to stem words accurately.

The METEOR score is computed by aligning the system output to the closest reference translation as in Figure 2. After stemming, *cries* and *crying* are considered a match, as are *saying* and *says*. In Figure 2, three words of the reference translation (in boldface) are not matched to the system output, and three words of the system output (not boldface) do not match the reference translation.

Ref 1: he has some pain in his stomach and always cries and complains about stomach pain

Ref 2: he has some pain in his stomach and he always cries and says I have a stomach pain

Ref 3: he has some stomach pain and always cries saying my stomach hurts

Ref 4: he has a stomach ache and he always cries and says my stomach hurts

System: he has stomach pain and always crying he says pain in stomach

Figure 1: Sample Reference Translations and System Output

| Ref 3: | he has **some** stomach pain and always cries | saying | **my** | stomach **hurts** |
| **System: he has** | **stomach pain and always crying** he **says** | pain | in **stomach** |

Figure 2: METEOR Alignment of System Output and Reference Translation

Therefore, recall is 9/12 and precision is 9/12. A weighted F-score (harmonic mean of recall and precision) is computed with a penalty if any of the words have been aligned out of order. Recall is weighted more heavily than precision, though this can be adjusted by the user.

Unlike the BLEU score, the METEOR score for the significantly poorer translation in (1) is lower than for the system output in Figures 1 and 2. For (1), recall is 8/12, precision is 7/12, and then the score is lowered by a penalty that applies to the match of *my* because *my* occurs in a different position in the two sequences.

2.3 The TER, STER and HTER Measures

Another limitation of the BLEU metric is that it only indirectly captures sentence-level properties such as word order by counting n-grams for values of *n* that are greater than one. But syntactic variation can produce translation variants that may not be represented in reference translations, especially for languages that have relatively free word order [2,15]. For example, in the sample in Figures 1 and 2, the word *always* could appear in a variety of positions as illustrated in (2) for reference #4.

(2) a. he has a stomach ache and he always cries and says my stomach hurts (original reference)
    b. he has a stomach ache and he cries always and says my stomach hurts
    c. he has a stomach ache and always he cries and says my stomach hurts

Although (2b) and (2c) may seem to be slightly less natural, they are certainly acceptable English forms. In other languages, word order is much freer than in English so that 3 or 4 reference translations will provide only a fraction of the options. METEOR allows users to adjust the word order penalty, but each word that must be moved or shifted in order to align with the reference translation is penalized separately. Therefore, when entire phrases in a language can freely occur in several positions, the translation is penalized for each word in the phrase.

The TRANSTAC program has also experimented with the TER metric to measure translation quality. Unlike METEOR, TER allows any number of contiguous words to shift positions in a single move. Computation of the TER score is based on the Levenshtein edit distance measure for string matching [3], which counts the number of insertions, deletions, and substitutions required to transform one string into another. Figure 3 shows how the alignment in Figure 2 would be edited to transform the system output into the reference translation. The deletions and substitutions that transform *he says pain in* into *saying my* could have been aligned differently with no effect on the number of deletions and substitutions.

The edit distance score is usually normalized by dividing the number of edits by the length of one of the strings, which would produce a score of 7/12 in Figure 3. When more than one reference translation is available, the denominator is the average length of the reference translations.

Levenshtein edit distance does not allow for the possibility of aligning words that are out of order, as TER does. TER permits movement of words or contiguous sequences of words in order to align them, and the shifts are counted as edits along with insertions, deletions, and substitutions. With a slightly different reference translation, Figure 4 shows how allowing a shift produces a lower edit distance score. The TER score in figure 4 is 7/13, whereas the Levenshtein edit distance score treats one *he* as a deletion and the other as an insertion, yielding a score of 8/13 for the same pair. (Lower TER scores reflect better performance.)

TER does not recognize synonyms, though the Semantic Translation Error Rate (STER) does use WordNet to align synonyms [17]. Instead, the inventors of TER introduced a variant that requires human intervention: Human Translation Error Rate (HTER). TER and HTER were developed for another DARPA machine translation program, Global Autonomous Language Exploitation (GALE), for which the machine translation evaluation is also conducted by NIST. In order to compute HTER, a human "post editor" edits the system output to produce a new reference translation that is maximally similar to the system output, while preserving the meaning of the reference translation. For example, a maximally similar reference for the system output in Figures 1-3 is (3).

(3) he has stomach pain and always **cries** he says **I have** pain in **my** stomach

Computing TER using the reference translation in (3) results in a score of 4/15 (errors are in boldface), which is a significant improvement over the TER score computed using any of the reference translations in Figure 1. The lower error rate seems appropriate given our intuition that the meaning of the reference translations can easily be inferred from the system output. In contrast, consider the much poorer translation in (1), which is repeated as (4a).

(4) a. he has some abdomen and always my and he says in his
    b. he has some abdomen pain and always cries and he says my stomach hurts

Ref 3:	he has some	stomach pain and always cries		saying		my	stomach hurts
System: he has		**stomach pain and always crying**	**he**	**says**	**pain**	**in**	**stomach**
Edits:	insertion		substitution deletion	substitution	deletion	substitution	deletion

Figure 3: TER Alignment of System Output with Reference Translation and Edits

Ref 3:	he has some	stomach pain and	**he**	always cries		saying		my	stomach hurts
System:	he has	**stomach pain and**		always **crying**	**he**	**says**	**pain**	**in**	**stomach**
Lev Edits:	insertion	insertion		substitution	deletion	substitution	deletion	substitution	deletion
TER Edits:	insertion	shift$_1$		substitution	[1]	substitution	deletion	substitution	deletion

Figure 4: TER vs. Levenshtein Edit Distance

(4a) receives a TER score of 8/12 when compared to the closest reference translation in Figure 1, which is a higher error rate than the system output's score of 7/12. The HTER score results in an even greater difference between the two translations: compared to the maximally similar reference translation in (4b), the HTER score for (4a) is 5/14 compared to 4/15 for the system output.

The HTER measure does not need to use WordNet or stemming because the human post editor can incorporate synonyms and adjust inflections. Also, unlike human judgments of translation quality, HTER does not require bilingual judges. Monolingual post editors can produce the customized reference translations from a single reference translation. Although the HTER measure appears to be more sensitive than the TER measure, it requires human intervention. Therefore, the significant advantages that automated measures obtain by eliminating the time, expense, and variability of human evaluations are lost. Consequently, the TRANSTAC program has not used the HTER metric.

2.4 More Issues for Automated Metrics

One shortcoming of automated measures of translation quality is shared by human judgments, which are typically obtained by asking bilinguals to rate system outputs on a scale that ranges from poor to perfect. Neither automated measures nor human judgments provide feedback that is diagnostic or that specifies the problems in less-than-perfect translations. In fact, BLEU is designed to be computed on an entire test corpus, using micro-averaging and calculating the brevity penalty based on all of the references and system outputs in the test set. NIST micro-averages the HTER scores when reporting evaluation results for GALE [Le, personal communication]. The claim is often made that automated measures cannot be expected to correlate well with human judgments at a sentence or utterance level: the high correlations that are reported compare corpus-level scores among translation systems so that the statistic is typically based on only a dozen or fewer data points. The MetricsMATR08 evaluation computed both utterance and corpus level correlations, and the former were much lower [9].

Another issue that is relevant to TRANSTAC evaluations concerns the quantity of data required for reliable automated measures. TRANSTAC 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 to assess the significance of differences in scores [15].

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 report that samples of reference translations from TIDES corpora averaged about 31 words per sentence [15], whereas 30 words is considered a maximum for inputs to the TRANSTAC speech translation systems. All of the automated measures of translation quality have been developed and tested using text data, whereas TRANSTAC data is speech data, which is structured very differently. In the next section the data collected for TRANSTAC systems is described, and additional features of those data that might affect automated metrics are discussed.

3. TRANSTAC TEST DATA
3.1 Data Collection

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. Eventually scenarios were consolidated into six broad categories: checkpoints, civil affairs, facility inspections, medical, training, and joint operations.

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. Large amounts of Iraqi Arabic data can be collected if Arabic speakers interact in Arabic. For authentic military English, dialogues were recorded in which an American soldier or Marine interacted with an Iraqi Arabic speaking civilian via a bilingual interpreter. This protocol made it possible to obtain a maximum amount of speech from the very limited time that we had access to military personnel. In earlier data collection events, 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. Later, lights were used to signal when

participants could begin to speak. 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 diacriticized lexica were created. Transcriptions included timestamps at the beginning and end of each segment. Some recordings, transcriptions, and translations 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 3.2).

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 speech 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 resembles ordinary conversation with high frequencies of filler expressions, pauses, breaths, and unclear speech as well as lengthy utterances. Some examples are provided in (5).

(5) 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 (5) 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.

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 has been ongoing clarification of this directive, and it is clear that divisions were inconsistently applied. For example, the single segment in (5a) contains four separate questions, and (5b) 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 co-articulation.

3.2 Selection of Evaluation Data
The TRANSTAC offline evaluations have primarily used two types of recorded dialogues. Reserved test data are subsets of the training data that are held back for evaluation instead of delivered to researchers for system development. Although reserved sets can be maximally representative of the training data, they are not ideal test sets because systems have been exposed to the voices and speech patterns of the speakers during training. Therefore, a special data collection using speakers who do not appear in any training data was conducted in order to create a test set that is sequestered for re-use.

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 collection, the vocabulary in each dialogue was analyzed to provide the following information:

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

From the dialogues that were in the mid-range for the percentage of word types that occurred in other dialogues, reserved dialogues were chosen so that each scenario topic was covered, a variety of speakers were represented, and the score in (4) above was maximized. Approximately 10% of the recordings were reserved.

Before each evaluation event, the sets of reserved dialogues were analyzed, and a summary of information relevant to selecting the test dialogues was produced. This information included the scenario topics, gender of the speakers (most were male), 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 requires several passes through the pool of dialogues available for the offline corpus. In the first pass, complete dialogues for the offline evaluation are selected based on the authenticity of the content, the range of scenarios, and the variety of speakers.

From the selected dialogues, 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 dialogues 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 at least 400 inputs per language. In order to preserve the content and coherence of the dialogues, only the worst offenders of criteria 2-4 were excluded. As more data was collected, the number of utterances was raised to 600. The sequestered test set was selected in a similar manner. It includes 810 English utterances and 664 Arabic utterances.

Timestamps were used to segment the audio recordings into a separate clip for each input. In addition, text inputs were

produced from the transcriptions of the selected segments in order to provide measures of translation quality that were independent of speech recognition. Consequently, offline evaluations produced a set of results that included speech recognition word error rate (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.

4. CORRELATIONS AMONG MEASURES

Speech recognition performance is important because recognition errors usually result in translation errors. The speech recognition word error rate was measured using the NIST SCLite scoring software, which computes a score derived from Levenshtein edit distance by comparing system recognition outputs to transcriptions of the speech [12]. To address the variation that occurs in speech, NIST modifies 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. Compound 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.

For each evaluation, a sample of approximately 100 English-to-Arabic and 100 Arabic-to-English translations from the offline test data 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*. More recently these judgments have been modified to the 7-point scale in Figure 5. 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

+3 Completely adequate
+2
+1 Tending adequate
 0
−1 Tending inadequate
−2
−3 Inadequate

Figure 5: Seven-value scale for semantic adequacy

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 [11].

Tables 1 and 2 show how the system scores from automated measures correlate with each other, with the human Likert judgments, and with the low level concept scores. Because TER and WER scores are error rates, they are subtracted from 1 to allow a positive correlation. "Concept Odds" refers to the low level concept measure described above, while "%Adequate" is the percent of utterances that were judged completely adequate in the Likert judgments. The correlations are typical of the correlations that developers of automated metrics of translation quality report. They are very high, but are based on only 5 systems and only on the samples of approximately 100 translations for each direction.

Figures 6 and 7 present the scores obtained for each automated measure and each human-judged measure, including the live dialogues. In the latter, military English speakers and Iraqi Arabic speakers were asked to role play scenarios using the translation systems. 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 [16]. The figures show similar patterns for all of the automated measures and for the human judged measures based on the offline data. The pattern for the live data is somewhat different, but for the most part, systems A-C score higher than D and E.

	BLEU	METEOR	1 - TER	Concept Odds	%Adequate	1 - WER
BLEU	1					
METEOR	0.994	1				
1 - TER	0.994	0.993	1			
Concept Odds	0.955	0.919	0.928	1		
%Adequate	0.969	0.937	0.951	0.994	1	
1 - WER	0.958	0.968	0.974	0.872	0.888	1

Table 1: English to Arabic Pearson Correlations among Measures for January 2007 Systems

	BLEU	METEOR	1 - TER	Concept Odds	%Adequate	1 - WER
BLEU	1					
METEOR	0.974	1				
1 – TER	0.982	0.945	1			
Concept Odds	0.978	0.990	0.972	1		
%Adequate	0.979	0.988	0.930	0.971	1	
1 - WER	0.813	0.906	0.756	0.847	0.880	1

Table 2: Arabic to English Pearson Correlations Among Measures for January 2007 Systems
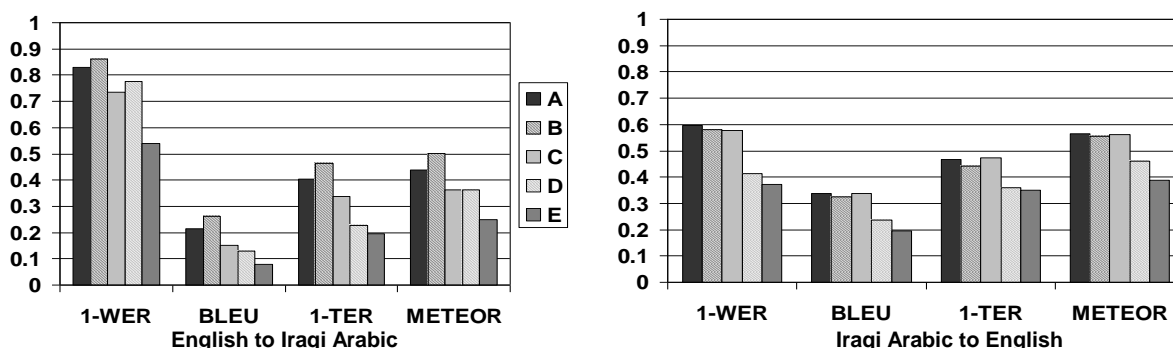
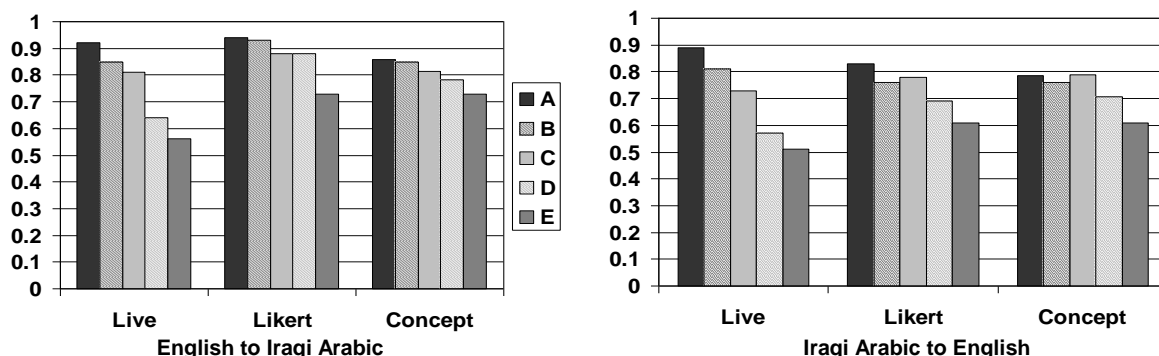**Figure 6: Automated Measures for Translations and Speech Recognition for January 2007 Systems A - E**

Figure 7: Translation Quality Measures Involving Human Judgments for January 2007 Systems A – E

5. CHALLENGES FROM ARABIC

One fact about the patterns of scores has persisted in subsequent evaluations. Although the human judgments consistently suggest that translation from English to Arabic is more successful than translation from Arabic to English, the automated measures consistently suggest the opposite. Moreover, the WER for English is much lower than for Arabic, which should also make translation more accurate, as suggested by the human judgments, but not the automated measures. It cannot be expected that scores from automated metrics will be comparable across languages, but the concern is that the measures may be less indicative of translation performance for languages like Arabic.

Several properties of Arabic challenge assumptions of automated measures. For example, it is assumed that words can be separated by spaces or punctuation, but six high-frequency words in Arabic, including the equivalents of „the' and „and' are attached to the word that follows them in Arabic orthography. The orthography of Arabic is extremely variable, with diacritic elements frequently omitted, so that string matching may fail due to a minor difference that would not obstruct understanding. Also, word order is freer in Arabic than in English. Furthermore, Arabic is more highly inflected than languages like English, though these differences have little effect on meaning. The examples in (6) illustrate that even in the absence of context, errors in inflectional morphology do not prevent communication of the sender's message.

(6) a. two book (two books)
     b. Him are my brother. (He is my brother)

BLEU scores computed with reference to the correct versions in parentheses would be very low because the inflected forms do not

match. METEOR provides a stemming operation that addresses this problem for English, but for many Arabic strings, complete stemming is not possible because the forms are ambiguous. Instead we experimented with light stemming, which has proven to be helpful in information retrieval tasks [6].

While NIST's normalization of references and outputs for computing WER has been uncontroversial, similar processes had not been proposed for automated measures of translation quality. However, normalization appears to be a simple way of handling superficial variation that would adversely affect accurate scoring of translations, just as it does for scoring WER. The TRANSTAC program has introduced normalization procedures for both English and Arabic to reduce variability before scoring with automated metrics. Norm1 performs rule-based normalization such as replacing contractions with full forms in English and removing all diacritics in Arabic. Norm2 performs word-based normalization such as the spellings of Arabic names in English. We experimented with two consequences of light stemming in Arabic: Norm2a separates the affixes, but does not delete them, while Norm2b deletes the affixes.

We also experimented with an option in the BLEU metric that uses only the unigram scores to allow for the freer word order in Arabic. We used the human judged subset of the June 2008 evaluation consisting of 109 English utterances (1431 words) and 96 Iraqi Arabic utterances (1085 words) in excerpts from 13 dialogs, each including about 7 exchanges. Table 1 provides Pearson's correlations among all the measures we have discussed for the English to Iraqi Arabic translations. Each correlation is computed over 39 data points (scores from 3 systems on excerpts from 13 dialogs). Correlations to the word-error-rate (WER) from

	English input WER Norm2	BLEU 1 Norm2	BLEU 4 Norm2	BLEU 1 Norm2a	BLEU 4 Norm2a	BLEU 1 Norm2b	BLEU 4 Norm2b	Likert Semantic Adequacy	Content Word AdjProbCor
WER Norm2	1								
BLEU_1 Norm2	-0.23	1							
BLEU_4 Norm2	-0.03	0.81	1						
BLEU_1 Norm2a	-0.33	0.77	0.63	1					
BLEU_4 Norm2a	-0.18	0.81	0.89	0.79	1				
BLEU_1 Norm2b	-0.43	0.82	0.51	0.80	0.61	1			
BLEU_4 Norm2b	-0.38	0.76	0.63	0.64	0.66	0.84	1		
Likert Sem Adeq	**-0.63**	**0.50**	0.19	**0.60**	0.41	**0.75**	0.63	1	
Adj Prob Correct	**-0.67**	**0.35**	0.07	**0.59**	0.30	**0.67**	0.48	**0.86**	1

Table 1: Pearson's *R* Correlations among the Metrics and Normalizations: June 2008 English to Iraqi Arabic

automated recognition of the English speech input are included in the first column. Next are correlations of Norm2, Norm2a, and Norm2b computed with BLEU_1 (BLEU with unigrams only) and with BLEU_4 (the more usual version with unigrams through 4-grams). Correlations with the two human-judgment metrics are highlighted with grey background: "AdjProbCorrect" is based on the low-level concept transfer score described in section 4.

The highest correlation in Table 1 is between the two types of human judgments. Also, it appears that WER is a good predictor of translation quality for the TRANSTAC systems. There is a steady increase in correlation from Norm2 to Norm2a to Norm2b. Norm2b scores correlate with the human judgments considerably more strongly than is the case for the Norm 2 and Norm2a scores. We believe this shows that human judges are more sensitive to errors on content words than to errors on the functional elements that are removed from Norm2b, but are only separated in Norm2a.

6. CONCLUSIONS

This report describes automated measures of translation quality, their limitations, and the issues encountered when applying the measures to speech translation data and to Arabic data. The report contributes to the research community's understanding of these measures, which have significantly advanced the development of machine translation systems.

7. REFERENCES

[1] 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.

[2] 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.

[3] Cohen, W., Ravikumar, P, and Fienberg, S. 2003. A Comparison of String Distance Metrics for Name-Matching Tasks. *IIWeb 2003: 73-78*.

[4] Internet Information. Course Wiki. Accessed August, 2009. http://ilps.science.uva.nl/Teaching/II0607/twiki/bin/view/Main/MeetingFeb22#1_1_1_Averaging.

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

[6] Larkey, L. and Connell, M. 2007. Light stemming for Arabic information retrieval. In Arabic Computational Morphology: Knowledge-based and empirical method, A. Soudi, A. van den Bosch, and G. Neumann, Eds,, Springer Verlag.

[7] Lita, L.V., Rogati, M., and A. Lavie. (2005). BLANC: Learning Evaluation Metrics for MT. In Proceedings of HLT/EMNLP, pp. 740–747.

[8] Metrics for Machine Translation Evaluation (Metrics MaTr08). Accessed August, 2009. http://www.itl.nist.gov/iad/mig/tests/metricsmatr/2008.

[9] 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.

[10] 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.

[11] 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.

[12] SCLite. NIST Multi-Modal Information Group. Accessed August, 2009. http://www.itl.nist.gov/iad/mig/tools/

[13] 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.

[14] Subramanian, K., Stallard, D., Prasad, R., Saleem, S., and Natarajan, P.   Semantic translation error rate for evaluating translation systems. IEEE Workshop on Automatic Speech Recognition & Understanding, 2007, pp. 390-395.

[15] 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.

[16] 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.

[17] WordNet. Accessed August, 2009. http://wordnet.princeton.edu.