Mach Translat (2012) 26:159–176
DOI 10.1007/s10590-011-9105-x

Evaluation of 2-way Iraqi Arabic–English speech translation systems using automated metrics

Sherri Condon · Mark Arehart · Dan Parvaz · Gregory Sanders · Christy Doran · John Aberdeen

Received: 12 July 2010 / Accepted: 9 August 2011 / Published online: 22 September 2011

Abstract The Defense Advanced Research Projects Agency (DARPA) Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) program (http://1.usa.gov/transtac) faced many challenges in applying automated measures of translation quality to Iraqi Arabic–English speech translation dialogues. Features of speech data in general and of Iraqi Arabic data in particular undermine basic assumptions of automated measures that depend on matching system outputs to reference translations. These features are described along with the challenges they present for evaluating machine translation quality using automated metrics. We show that scores for translation into Iraqi Arabic exhibit higher correlations with human judgments when they are computed from normalized system outputs and reference translations. Orthographic normalization, lexical normalization, and operations involving light stemming resulted in higher correlations with human judgments.

Approved for Public Release: 11-0118. Distribution Unlimited. The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. Some of the material in this article was originally presented at the Language Resources and Evaluation Conference (LREC) 2008 in Marrakesh, Morocco and at the 2009 MT Summit XII in Ottawa, Canada.

S. Condon (✉) · M. Arehart
The MITRE Corporation, McLean, VA, USA
e-mail: scondon@mitre.org

D. Parvaz
The MITRE Corporation, Orlando, FL, USA

G. Sanders
National Institute of Standards and Technology, Gaithersburg, MD, USA

C. Doran · J. Aberdeen
The MITRE Corporation, Bedford, MA, USA

Keywords Arabic · Machine translation · Evaluation · Automated metrics · Speech translation

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 supporting development of two-way speech translation systems. The goal of the TRANSTAC program is to demonstrate capabilities to rapidly develop and field 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. For several years, the program focused on two-way translation between Iraqi Arabic and English, which provided an opportunity to explore the applicability of automated metrics to translation of spoken dialogue. The evaluations also offered a chance to study the results of applying automated metrics to languages other than English.

Since the inception of the DARPA speech translation programs for military domains, a MITRE team has coordinated with system developers and researchers to collect training data and design evaluation methods. More recently, The National Institute of Standards and Technology (NIST)[1] has directed efforts to assess the progress of system development and evaluate the systems' readiness for fielding. 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 (Weiss et al. 2008) and offline evaluations in which the systems process audio recordings and transcripts of interactions (Condon et al. 2009). The inputs in the offline evaluation are the same for each system, and 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) (Bannerjee and Lavie 2005) have been developed and widely used for translations of text and broadcast material, which have very different properties than dialogue.

The TRANSTAC evaluations have also provided an opportunity to compare automated metrics to human judgments from a panel of bilingual judges. When comparing system-level scores by pooling all data from a given system, high correlations (typically above 0.9) have been obtained among BLEU, TER, METEOR, and scores based on human judgments (Sanders et al. 2008). When the data are more fine-grained than system-level, however, the correlations of the human judgments to the automated metrics for machine translation (MT) are much lower. In addition, the results produced by five TRANSTAC systems in July 2007 and by the best-scoring three of those five

[1] Certain commercial equipment, instruments, software, or materials may be identified in this article in order to specify the experimental procedures adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the equipment, instruments, software, or materials are necessarily the best available for the purpose.

systems in June 2008 revealed that the correlations between the automated MT metrics and the human judgments are lower for translation into Arabic than for translation into English (Condon et al. 2009).

We hypothesize that several features of Arabic are incompatible with assumptions that are fundamental to many automated measures of MT quality. These features of Arabic contrast with properties of English and most of the other Indo-European languages to which automated metrics have been applied. The consequence of these differences is that automated MT measures give inaccurate estimates of the success of translation into Arabic compared to languages like English: for Arabic the automated measures consistently correlate lower with utterance-level human judgments of semantic adequacy and with human judgments of how successfully the meaning of content words is transferred from source to target language.

This article begins in Sect. 2 by highlighting some of the challenges we faced in applying automated measures of translation quality to TRANSTAC Iraqi Arabic-English speech translation data. Section 3 describes the methods used to prepare evaluation data and the corpora for which scores are reported. In Sect. 4, the normalization operations that were experimented with are presented, and Sect. 5 discusses the scores that were computed using normalization procedures. Section 6 offers conclusions and directions for further research.

2 Challenges for automated metrics

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 and Riehemann 2003; Lavie et al. 2004; Owczarzak et al. 2007). In addition to BLEU, the TRANSTAC program has used METEOR and TER to score translations of the recorded scenarios. METEOR incorporates recall as well as precision and scores co-occurrences on the unigram level. The relative weights of precision and recall are language-specific. METEOR also incorporates a language-specific fragmentation penalty that serves a role similar to the role of higher-order n-grams in BLEU. TER scores were computed in initial evaluations. TER, METEOR and BLEU scores for the TRANSTAC data routinely have high correlations with each other (Sanders et al. 2008). For that reason, we will report only BLEU[2] results here.

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 et al. 2007;

[2] We use a variant of BLEU (`bleu_babylon.pl`) provided by IBM that produces the same result as the original IBM version of BLEU when there is no value of n for which there are zero matching n-grams. 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. This modification is deemed an advantage when scoring individual sentences, because zero matches on longer n-grams are then fairly likely.

Owczarzak et al. 2007; Turian et al. 2003). It is possible to run the BLEU metric on only unigrams, and as will be explained later, that ability appears to be important for accurately evaluating the advantages of the work in the current study.

The fundamental assumption of BLEU which is problematic for our data is that several reference translations adequately allow for legitimate variation in the translations (Callison-Burch et al. 2006). This assumption is also true of most of the 39 automated measures submitted to the NIST 2008 Metrics for Machine Translation Challenge (Przybocki et al. 2009). Measures based on exact matching of system outputs to references, including the Word Error Rate (WER) measure used to score automatic speech recognition (ASR), are at a disadvantage when applied to data that contains much variation which is unrelated to translation quality. Some of the variability in our data is a feature of speech in general, while some results from features of Iraqi Arabic and other Semitic languages. Each of these sources of variation is discussed below.

2.1 Challenges from properties of Iraqi Arabic

Arabic, like other Semitic languages, has both a morphology and an orthography which are not immediately amenable to current approaches in automated MT scoring. All approaches to date make the following assumptions concerning the texts:

1. *Ease of tokenization*. Current scoring code assumes a relatively trivial means of tokenization, i.e., along white space and punctuation. Many languages, especially most Indo-European ones, orthographically separate articles and particles (prepositions, etc.) This means of tokenization isolates prepositions from noun phrases and object pronouns from verbs. In contrast, orthographic conventions in Arabic attach frequently used function words to the related content word. As an example, the Arabic *lilbrnAmj*[3] ('to the program') consists of three separate elements (*l-* 'to', *Al-* 'the', *brnAmj* 'program'). So a scoring program encountering *lbrnAmj* ('to [a] program') without further tokenization would score it as entirely wrong. However, once properly stemmed and tokenized, it becomes clear that the only element missing is the definite article, which means that it is 2/3 correct.
2. *Concatenative morphology*. In addition to morphological elements that are affixal in nature, Arabic has a morphology which interleaves roots, usually consisting of three or more consonants, with patterns (e.g., geminate the middle consonant and place a *t-* at the beginning) and characteristic vowels (*a* for perfect tense) to create new forms. So the root *kfr* (general semantic area: 'sacrilege', 'blasphemy') combined with the nominal pattern *taCCiyC* (generally, 'causing one to do X') results in the surface form *takfiyr* ('accusation of blasphemy'). Not only is this interleaving pattern used for coining words, it is the preferred method of forming masculine plurals.
3. *Non-defective script*. Languages written with Roman script have some orthographic representation (however imperfect) of both vowels and consonants, which aids both in the dictionary lookup process and in stemming or lemmatizing. Arabic

[3] Arabic strings here are written according to Buckwalter notation (2001): *llbrnAmj* = للبرنامج.

is written in a defective script in which most vowels (the so-called "short" vowels) are usually unwritten. Therefore, it is often difficult to find with any certainty whether two similarly written forms are actually the same word (e.g., the Arabic *ktAb* might either be *kitAb* 'book' or *kut~Ab* 'writers'. Determining which form is which on an automated basis, when possible, will depend on paying careful attention to usage, which the scoring programs generally do not do.

4. *Uniform orthography*. Although short vowels are typically not represented in Arabic script, they may be rendered using diacritic notations. The number of distinct forms in which a word may occur is multiplied by these diacritics, other diacritics that are variably included in Arabic spellings, and additional orthographic variation that is unique to specific characters and morphemes. Therefore, measures that depend on exact matching of word forms may fail to match forms that differ in superficial ways.
5. *Constrained word order*. Arabic word order is not as free as in some languages, but it is more variable than in languages like English. Automated measures depend on word order to provide indirect assessments of fluency and coherence, using *n*-gram matching (in BLEU) or other methods of tracking word order differences between system hypotheses and reference translations (METEOR, for example, looks at how many "chunks" of contiguous words match between hypothesis and reference translations). For languages with highly variable word order, reference translations may not (and often will not) capture all allowable orders, especially since translators may be influenced by the structure of the source text.

The normalization experiments reported here do not solve all of the problems of applying automated measures to languages like Arabic. However, they do provide some estimates of the degree to which these problems influence scores obtained by automated measures as well as promising directions for resolving some of the problems.

2.2 Challenges from speech data

Collection of training and evaluation data was a challenge for the TRANSTAC program. In spite of attempts to define narrow domains and use cases that would provide realistic goals for the speech translation systems, it quickly became clear that even the most routine interactions can easily veer out of domain: for example, when 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 six domains, including checkpoints, searches, infrastructure surveys (sewer, water, electricity, trash, etc.), and training. Later, additional 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. In monolingual dialogues, two speakers role-played scenarios in Iraqi Arabic. Bilingual dialogues consisted of an English speaking soldier or Marine interacting with an Iraqi Arabic speaker via a bilingual interpreter. Speakers were required to address each other directly without addressing the interpreter, as they would when using a translation device. Data were not collected using translation devices in order to obtain a maximum amount of speech from the very limited time that we had access to military personnel. Evaluation data were collected using the same protocols.

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).

Example 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. Translators were instructed to ignore filler expressions and partial words, but they tended to vary in how closely they represented repetitions or false starts.

Another source of variation in the data is the fact that orthographic conventions typically represent more formal varieties of language than are found in conversation. In addition to the inevitable typographic errors produced by human transcribers and translators, which are then incorporated in the MT systems via the training data, the following problems may prevent system outputs from matching references:

1. British lexical variants such as *bonnet* for hood of car, *boot* for trunk of car and the morphological variant *learnt* (data were processed in Australia)
2. Lexical variation among Iraqi Arabic, Modern Standard Arabic, and other regional varieties
3. In both languages, some words are pronounced very differently than they are conventionally spelled, such as *wanna*, *gonna*, *cuz*, *could of*, and *kinda* in English
4. A special case of (3) in English is contractions such as *I'm* vs. *I am* and *isn't* vs. *is not*
5. Abbreviations in both languages may be written in various ways: *IRS* vs. *I.R.S*, and American abbreviations may be pronounced in Arabic, e.g, /ay-ar-ϵs/, with

the consequence that they are no longer abbreviations because the sounds do not represent Arabic letters
6. Transliteration variants of names such as *Ahmad/Ahmed* in English
7. Compounds in English may be written three ways: with hyphens or spaces or as single words (*black-market, black market, blackmarket*)

The data were transcribed and translated according to guidelines compiled in consultation with TRANSTAC developers, but if guidelines are too extensive, they are difficult for transcribers and translators to follow, and variability seems inevitable. All of this variation makes it difficult to apply automated measures of translation quality that depend on matching system outputs to reference translations, even when several reference translations are available.

3 Evaluation data

The scores reported here are from two evaluations conducted in June 2008 and November 2008. In each evaluation, systems processed two test sets: one set, referred to as the Open set was held out of training data and included speakers who were also recorded in the training data. A different Open set was used in each evaluation and then released to the researchers. The Sequestered set was collected at the same time and in the same way as the training data, but a separate group of speakers was used so that the speakers were new to the systems. The Sequestered set was used in both evaluations so that improvements could be measured against identical test data. A subset of each test set was selected for human judgments, which were obtained by the evaluation team at NIST. For each set, Table 1 provides the quantities of utterance units,[4] which were the translation units compared in scoring, and the number of dialogues from which the utterance units were selected. Utterances in the human judged set were selected from both the Open set and the Sequestered set.

Because there was a limited amount of data available for evaluation, the dialogues and utterance units within those dialogues were selected by hand. Dialogues were selected for domain representation, fluency and rich content. Utterances in the dialogues were selected to represent the sort of utterances that the systems are expected to translate. For the most part, the utterances in the dialogues were left intact, but when they could be eliminated without disrupting the coherence of the conversation, utterances (or sequences of utterances) were removed if they were extremely long, disfluent, or incoherent. Also, utterances consisting of only a single acknowledging "okay" (or a similar phrase) were limited to one or two per dialogue.

Dialogues were also selected to match the proportion of male and female speakers in the training data. Most of the speakers in the test data were male. In the June Open set, all of the English speakers were male and about 25% of the Iraqi utterances were from female speakers. The November Open set included one dialogue with a female English speaker (about 10% of the English utterances), and 22% of the Iraqi utterances

[4] Each speaker's turn was transcribed as at least one utterance unit. Long turns were separated into more than one utterance unit, but transcribers tended to be inconsistent in separating the units, partly because they were instructed to identify points where there was maximal acoustic separation.

Table 1 Test set sizes in utterance units

Source language	June 2008 evaluation			November 2008 evaluation		
	English	Iraqi Arabic	Dialogues	English	Iraqi Arabic	Dialogues
Open set	656	579	14	618	689	10
Sequestered set	810	664	13	810	664	13
Human judged set	109	97	13	93	108	13

were from female speakers. In the Sequestered set, all the English speakers were male and 36% of the Iraqi utterances were from female speakers. In recruiting Iraqi speakers for data collection, efforts were made to identify speakers who spoke Baghdad dialect. The military English speakers were recruited based on their experience in Operation Iraqi Freedom.

Test data was processed by four DARPA-funded statistical MT systems, labeled here as systems A, B, C, and D (A–D). Each utterance was processed as a separate audio input, and system logs recorded the speech recognition results along with the text translation. Each test set was also processed using text transcriptions of the audio inputs in order to measure translation quality without speech recognition errors, but the scores presented here are from the audio inputs.

4 Data analysis and normalization

In addition to computing automated measures of translation, WER was computed for the speech recognition results, and human judgments were obtained for the human judged set. WER was measured using the NIST SCLite scoring software (SCLite 2009). In scoring English ASR, NIST first modifies the reference transcriptions and system outputs to reduce the kind of variation described in Sect. 2.2. 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. This approach to computing WER served as a model for the normalization operations that we investigated.

4.1 Human judgments of translation quality

The gold-standard metrics for translation adequacy are commonly deemed to be judgments from a panel of bilingual human judges. For each TRANSTAC evaluation, a panel of five bilingual judges provided utterance-level judgments of semantic adequacy on the seven-value scale in Fig. 1 The seven-value scale has an explicit numeric interpretation as equally-spaced values, and the judges were instructed to interpret the seven values as equally-spaced. The judges were instructed that when they were torn between two of the labeled choices, to choose the unlabeled choice between. NIST emphasized the numeric interpretation to avoid having to treat the judgments as categorical during data analysis.

For each judge separately, and for each direction (to or from English) separately, the values were converted to standard normal deviates (mean 0.0 and standard

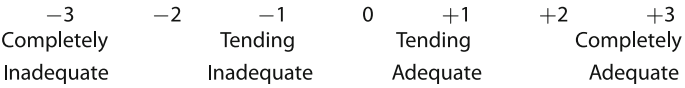

Fig. 1 Seven-value scale for semantic adequacy

Table 2 Orthographic normalization operations used in norm1 for Iraqi Arabic

Type of variation	Example	Normalization operation
Short vowel/shadda inclusions	جُمهُورِيَّة vs. جمهوريّة	Delete vowel and shadda diacritics
Explicit nunation inclusions	أَحيَانًا vs. أحيانا	Delete nunation diacritics
Omission of the hamza	شي vs. شيء	Delete hamza
Misplacement of the seat of the hamza	الطوارئ vs. الطواريء	Delete hamza
Variations where taa marbuta should be used	بالجمجمة vs. بالجمجمه	Replace taa marbuta with hah
Confusion between yaa and alif maksura	شى vs. شي	Replace alif maksura with yaa
Initial alif with or without hamza/madda/wasla	إسم vs. اسم	Replace with bare alif

deviation 1.0). The mean across the judges for each utterance is then taken, and the result is a normalized utterance-level Likert score (Likert 1932).

4.2 Normalization

Human judges are capable of ignoring minor variation in order to comprehend the meaning of language in context. The examples in (2) illustrate that even in the absence of context, errors in inflectional morphology do not prevent communication of the sender's message.

Example 2 a. two book (two books)
b. Him are my brother. (He is my brother)

In contrast, scores from automated MT metrics computed with reference to the correct versions in parentheses would be low because the inflected forms do not match. We hypothesize that scores will correlate more closely with human judgments when extraneous variation and some inflection are removed. For many Arabic strings, a complete morphological analysis is not possible without taking context into account because the surface forms are ambiguous, but a complete morphological analysis is not required to provide forms that can be matched by automated measures. We began by applying two types of normalization to both the English and Iraqi Arabic dialogues. System outputs and references are normalized by the same procedures.

Rule-based normalization, referred to as *Norm1*, focuses on orthographic variation. For Iraqi Arabic, a Perl script reduces seven types of variation by deleting or replacing variants of characters with a single form. Table 2 lists the seven types, provides examples of each, and describes the normalization operation that is applied in Norm1. These seven are attested orthographic variants in written Arabic and may therefore

occur in the reference translations. For English, the rules include operations that transform letters to lowercase,[5] replace periods with underscores in initialisms,[6] replace hyphens with spaces, and expand contractions. The latter include forms such as *this'll*, *what'll*, *must've*, *who're*, and *shouldn't*.

The second type of normalization, referred to as *Norm2*, is inspired by the normalization operations that NIST uses to compute WER for evaluation of automatic speech recognition. In addition to rule-based normalization operations such as replacing hyphens with spaces, NIST uses a global lexical mapping (GLM) that allows contractions and reduced forms such as *wanna* to match the corresponding un-contracted and unreduced forms. For Iraqi Arabic, the contractor that processes TRANSTAC training data produced a list of variant spellings of Arabic words from the transcription files, and TRANSTAC developers and researchers added to the list. MITRE vetted these lists to make sure that they reflected genuine free variation or spelling corrections. Many of the variants were caused by orthographic variation that is addressed in Norm1 and were excluded as redundant, but there were 300 mappings, including a few misspellings and typographical errors that are not corrected in Norm1 (e.g., شقد vs. شكد , يبها vs. بهاء). For English, the GLM is customized for the vocabulary of each evaluation to eliminate the kinds of spelling variation described in Sect. 2.2. The November GLM included 150 mappings.

The Norm2 operation uses a GLM mapping (x, y) like a conditional: if a word in the system translations or reference translations matches x, then it is rewritten as y. For Arabic, the x and y forms are normalized orthographically using the Norm1 rules because they are matched after Norm1 has applied. For English, we ensure that the x forms have Norm1 representations when the mapping is constructed.[7] Although each system may tokenize Arabic differently for processing, the Arabic (and English) translations that are logged for scoring are produced in conventional orthographic form.

For additional normalizations of Iraqi Arabic, we referred to the work of Larkey et al. (2007), who experimented with a variety of stemmers and morphological analyzers for Arabic to improve information retrieval scores. We produced a modified version of their Light10 stemmer. Word-initially, Light10 removes the conjunction *wa* (و), the definite article *al* (ال), prepositions *bi* (ب), *li* (ل), *fi* (ف), and the form كـ , which is used like English *like* or *as*. The prepositions and كـ are removed only if followed by the definite article *al*, which is removed only if the remainder of the word is at least two characters long. The conjunction may be removed without a following *al*, but only if the remainder of the word is at least three characters long. These constraints minimize

[5] Because the contrast between upper and lower case is irrelevant for speech output, systems were not expected to use conventional capitalization. The training data employed upper case only for names of persons, place, and organizations (which were also annotated).

[6] Initialisms such as *I.R.S.* are pronounced differently than acronyms such as *AIDS* and were distinguished in the training data by using underscores for the former (*I_R_S*), but no punctuation for the latter (*AIDS*). However, the data were not always consistent, and systems tended to produce a variety of representations. The underscore was chosen to eliminate confusion with other uses of the period/full stop.

[7] Except for case normalization, the English Norm1 operations do not typically impact items in the GLM, such as the mapping of *kinda* to *kind of* or of *blackmarket* to *black market*.

Table 3 Light10 suffixes separated in norm3a or removed in norm3b

Arabic suffix	Morphological features when attached to verbs (V) and nouns (N)
ها	V: 3rd person singular feminine object; N: possessive pronominal
ان	N: Dual number
ات	N: Feminine plural
ون	V: subject agreement
ين	N: Oblique masculine plural
يه	V: 3rd person singular masculine object; N: possessive pronominal
ية	N: Feminine nisba adjective, attributive
ه	V: 3rd person singular masculine object; N: possessive pronominal
ة	N: feminine singular (or singular of mass/collective noun)
ي	N: 1st person singular possessive pronoun, nisba adjective marker

the possibility of removing characters which are actually part of the word. We did not experiment with other tokenizers such as MADA + TOKAN (Habash and Rambow 2005) or AMIRA (Diab 2009), which are capable of distinguishing the prefixes from stems.

The suffixes that are removed are listed in Table 3. Norm1 renders some of these forms indistinct before the normalizations based on Light10 are applied. The Light10 stemmer is "light" because there is no attempt to remove other morphemes such as the prefixes that express aspect and subject agreement on verbs or the infixes that indicate plural nouns. Our primary concern is the affixes that often correspond to separate words in languages like English where highly probable sequences such as *and*+*the*+word or *in*+*the*+word can increase bigram and trigram scores. Separating the affixes can have the same effect when the affixes are produced correctly. Separating or deleting the affixes prevents incorrect affixes from impacting correct stems to which they are attached.

5 Results of scoring with normalization

5.1 Normalization effects on Iraqi Arabic to English BLEU scores

Figure 2 demonstrates the effects of the normalization operations by presenting BLEU scores for Iraqi Arabic to English translations from the November data. Results are presented for each test set, Open and Sequestered, from each system, A–D, for each type of normalization. Norm0 scores are computed using the system outputs and reference translations with no normalization. Norm1 scores are computed after applying rule-based operations, such as normalization of contractions, to reference translations and system outputs. Norm2 scores are computed after applying the GLM to the output of Norm1. The figures show that the effects of the normalization operations vary by system and data set. Although each normalization operation increases the BLEU score, the difference is often very small. The difference between Norm0 and Norm2 scores

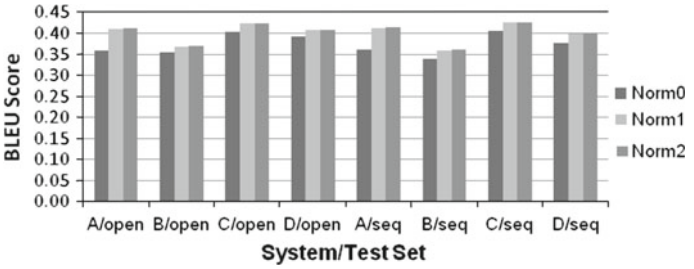

Fig. 2 November BLEU scores for Iraqi Arabic to English translations from systems A–D on open and sequestered test sets with normalization operations

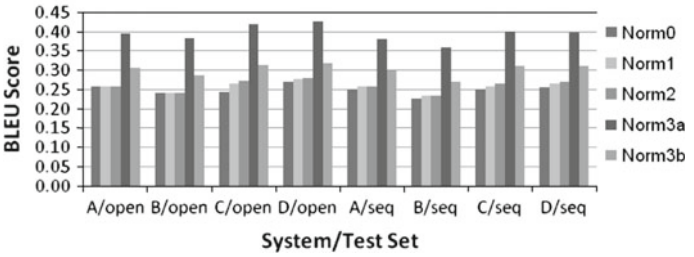

Fig. 3 November BLEU scores for English to Iraqi Arabic translations from systems A–D on open and sequestered test sets with normalization operations

for the June data ranged from .016 to .046 and averaged .022. For the November data the difference ranged from .016 to .053 and averaged .028.

5.2 Normalization effects on English to Iraqi Arabic BLEU scores

Figure 3 presents the BLEU scores for English to Iraqi Arabic translations from the November data. Results are presented for each test set, Open and Sequestered, from each system, A–D, for each type of normalization. Norm1 scores are computed after applying orthographic normalization to the un-normalized Norm0 reference translations and system outputs. Norm2 scores are computed after applying the GLM to the output of Norm1. Operating on the output of Norm2, we produced two additional normalized sets based on Light10: Norm3a separates the prefixes and the suffixes in Table 3, but does not remove them, while Norm3b removes the affixes. Norm3a allows comparisons to reference translations using all of the forms that are present in the texts. Some of those forms express grammatical categories such as definiteness, number, and gender that human judges might consider insignificant, but the affixes that represent pronominal objects and possessors are likely to be important for interpreting the meaning of the utterance.

Like the scores for Iraqi Arabic to English translations, most scores for English to Iraqi Arabic translations show modest increases from Norm1 and Norm2 operations compared to Norm0, averaging .014 in the June test set and .01 in the November set. Norm3a has the effect of increasing the number of words that are scored, introducing

a large number of unigrams that are likely to be scored as correct translations. This alone will increase scores from automated metrics. Scores from metrics such as BLEU that are based on *n*-gram co-occurrence statistics will also increase because Norm3a ensures that the order of prefix sequences such as *wa* + *al* + noun or *bi* + *al* + noun will match, thus increasing bigram and trigram matches. Figure 3 illustrates this effect with an average .15 increase in the scores from Norm0 to Norm3a. A similar increase was observed for the June test set. Increases from Norm0 to Norm3b averaged .05 on both test sets.

5.3 Normalization effects on relative rankings

The normalization operations have the potential to alter the rankings of the systems, and this possibility is realized in Figs. 2 and 3. In Fig. 2, system A scores range from .017 to .033 lower than system D scores at Norm0 for both the Open and Sequestered evaluation sets, but they are slightly higher than the system D scores at Norm2. System C scores remained higher than A and D by about .01 or .02. We can compare these rankings to rankings from the human judgments described in Sect. 4.1, which are based on subsets of the Open and Sequestered sets used to compute the scores in Fig. 2 and were obtained only for systems A, C, and D. The November human judgments of systems C and D were nearly identical, but system A achieved about 7% higher proportions of utterances judged as +2 and about 5% higher proportions judged as +3. Consequently, the Norm2 scores in Fig. 2, which rank system A closer to systems C and D, are more similar to the human judgments than the Norm0 results, but the relative rankings are not the same.

For translations from English to Iraqi Arabic, in Fig. 3, most of the Norm3 scores reverse the rankings of systems A and C compared to the Norm0, scores. In the human judgments, the total proportion of utterances judged as +1, +2, or +3 was nearly identical for systems C and D and about 5% more for system A. However, system D had about 10% more utterances judged as +3 than systems A and C. Consequently, it is not clear which normalization level best reflects those judgments.

5.4 Utterance and dialogue correlations to human judgments

The difficulty of comparing system-level scores to human judgments motivates a finer-grained analysis. Because we hypothesized that the normalized BLEU scores would more accurately reflect the performance of the translation systems, we computed correlations between the normalized scores and the human judgments using the subsets of the evaluation data for which we had human judgments (see Table 1 in Sect. 3). Correlations to human judgments typically compare system-level scores, but we had only four systems to compare, and we have seen that system level comparisons are inconclusive.

BLEU scores were not designed to be computed on single utterances, whereas all the human judgments were obtained for single utterances. Therefore, in addition to comparing scores of individual utterances, we also took advantage of another unit in the human-judged data: the utterances were selected as contiguous excerpts from 13

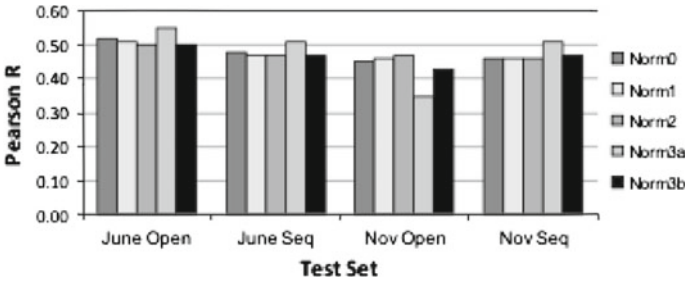

Fig. 4 Utterance level correlations of human judgments and BLEU scores computed with normalization operations for June and November open and sequestered subsets of English to Iraqi Arabic translations

dialogues (six from the Open set and seven from the Sequestered set). BLEU scores were computed for each dialogue unit of the three systems for which human judgments were obtained, which provided 18 scores from the Open set and 21 scores from the Sequestered set to compute the Pearson R correlations

For the dialogue level correlations, it was necessary to combine the human judgments for the translations in each dialogue, which consisted of 7–10 English to Iraqi Arabic translations. As described in Sect. 4.1, the judgments from each rater were normalized and the mean of the five normalized ratings was used for an utterance level score. For the dialogue level score, we computed the mean of the utterance level scores.

Figure 4 presents the utterance level correlations of human judgments and BLEU scores for English to Iraqi Arabic translations. Utterances from each June and November, Open and Sequestered evaluation set are correlated separately. In three of the four sets, the Norm3a correlation is clearly the highest, but otherwise there is no clear pattern in the correlations. The range of correlations, from .35 to .55 is similar to the range of correlations obtained from pairwise correlations of the human judges, which is .32 to .58.

In addition to applying BLEU in the standard way (computing the geometric average of matches on unigrams, bigrams, trigrams, and 4-grams), we also computed BLEU using just unigram matches (an option in the BLEU scoring software). Because unigram-only matching effectively gives no weight to fluency or word order, the unigram-only values for BLEU are more a measure of semantic adequacy of the words in the machine translation output. We believe this is an advantage for languages like Arabic with freer word order. Looking only at unigrams gives the translations no extra credit for additional n-grams, especially the extra n-grams that are present in Norm3a, so that these unigram-only values put the Arabic scoring on a more theoretically equal footing with English.

Figure 5 presents the same utterance level correlations of human judgments and BLEU scores for English to Iraqi Arabic translations as in Fig. 4, except that BLEU scores are computed with unigram-only matching. In most of the normalization conditions, correlations are the same or higher when BLEU scores are computed with unigram matching. One exception to this trend is the Norm3a correlations for the June test sets. When the BLEU scores were not affected by the higher numbers of 2-, 3- and

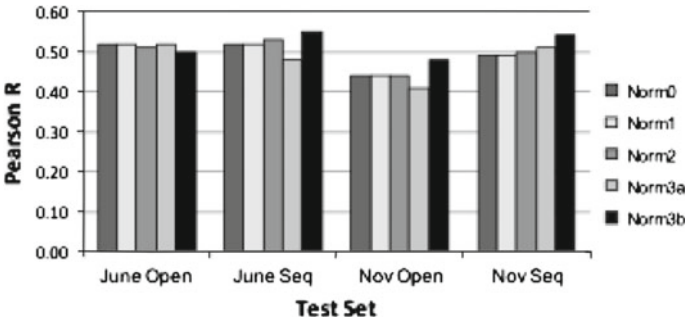

Fig. 5 Utterance level correlations of human judgments and unigram BLEU scores computed with normalizations for June and November open and sequestered subsets of English to Iraqi Arabic translations

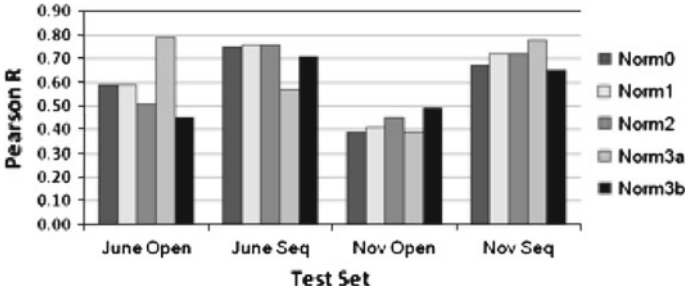

Fig. 6 Dialogue level correlations of human judgments and 4-gram BLEU scores computed with normalization operations for June and November open and sequestered subsets of English to Iraqi Arabic translations

4-gram matches produced by separating the affixes, the correlations were lower for those sets. Yet the Norm3a correlation is considerably higher (by .06) for the November Open test set, which highlights the sensitivity of these approaches to the content of these small test sets. Whereas the Norm3a condition results in the highest correlations for three of the four test sets when BLEU is computed in the usual way, the Norm3b condition produces the highest correlations for three of the four sets when BLEU is computed with unigram-only matching. In each of the latter, the Norm3b correlation resulting from unigram-only matching is higher than the highest correlation produced when BLEU is computed with 2-, 3-, and 4-gram matching.

Figure 6 presents the dialogue level correlations of human judgments and BLEU scores obtained by combining the English to Iraqi Arabic translations from each dialogue as if they comprised a single document. This method provides longer sequences to compute BLEU scores, as the measure was designed to be used, rather than single utterances. BLEU scores were otherwise computed in the usual way. A significant consequence of computing the scores in this manner is that correlations in Fig. 6 are mostly higher, and in some cases much higher, than those obtained by computing utterance level BLEU scores in Fig. 4. For example, the June Open set Norm3a correlation increases from .55 to .79, and the November Sequestered set Norm3a correlation increases from .51 to .78. While Norm3a provides the highest correlations

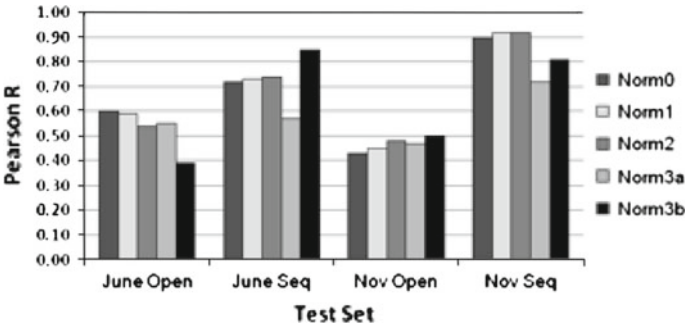

Fig. 7 Dialogue level correlations of human judgments and unigram BLEU scores computed with normalizations for June and November open and sequestered subsets of English to Iraqi Arabic translations

for dialogues from two test sets, the Norm1 and Norm2 correlations are the highest in the June Sequestered set, and the Norm3b correlation is the highest in the November Open set. The latter set differs from all of the other sets in that the Norm0, Norm1, and Norm2 correlations obtained by computing dialogue level BLEU scores are lower than those obtained with utterance level BLEU scores in Fig. 4, yet the Norm3a and 3b correlations are higher.

Figure 7 presents the results of the correlations to human judgments for English to Iraqi Arabic translations when BLEU scores are computed at the dialogue level with unigram-only matching. Some of the correlations in Fig. 7 are the highest obtained from the three methods of computing BLEU scores that have been presented, such as values of .92 for the Norm1 and Norm2 correlations in the November Sequestered set and the value of .85 for the Norm3b correlation in the June Sequestered set. In fact, these correlations and the .5 value obtained for the Norm3b correlation in the November Open set are the highest correlations for those data sets among all of the correlations in Figs. 4, 5, 6, and 7. Only the June Open set does not achieve the highest correlation value from dialogue level scores with unigram-only matching.

Although the results are not as clear as they could be, the experiments support the hypothesis that BLEU scores computed from normalized data can more accurately reflect the performance of translation systems. For each dataset, the highest correlation among all the methods of computing BLEU scores is achieved using a score based on normalized translations. This consequence applies to evaluation of all types of machine translation, but it is especially relevant to speech translation systems, for which variation in linguistic forms is a pervasive problem. However, the results do not provide clear evidence that one of the four types of normalization we experimented with is the most appropriate for the translations to Iraqi Arabic. When BLEU scores are computed in the usual manner, the Norm3a conditions tend to produce the highest correlations, but when BLEU scores are computed with unigram-only matching, the Norm3b conditions tend to be highest. The fact the some of the highest correlations were obtained using Norm3b and unigram-only matching supports the claim that human judges tend to focus on the content words and overlook grammatical details such as word order when interpreting language in communicative context.

We note here that the analyses in this section have focused on whether the normalizations increase the correlation of the BLEU scores with the human judgments, and not on whether the changes in BLEU scores are statistically significant or whether the changes in BLEU scores result in statistically significant differences among the systems.

6 Conclusions and related work

Measures of translation quality based on exact matching of word forms are challenged by speech translation, especially when the translation involves languages like Iraqi Arabic. Much of the problem can be attributed to the inherent variability of speech and to orthographic variation, which is especially severe in Arabic languages. The inflectional morphology and relatively free word order in Arabic increase the likelihood that system outputs will not match reference translations. We demonstrate that the effects of these linguistic differences can be mitigated by normalization processes that reduce orthographic variation and delete or separate affixes and function words. Evidence was presented that normalization operations can improve the correlation of BLEU scores to human judgments of translation quality.

We hypothesized that features with minimal effect on meaning such as inflection and word order have little impact on judgments that focus on semantic quality. This may be especially true for dialogues, where disfluencies and inference from context are the norm. However, the correlation results did not provide clear support for this hypothesis. The normalization that most consistently produced the highest correlations with human judgments was one in which affixes were included rather than deleted. Of course, some of those affixes express significant content, especially the suffixes that represent pronominal objects. A better test of the hypothesis would be a normalization that deleted affixes expressing grammatical nuance such as agreement, while preserving those that represent significant content.

In demonstrating the potential advantages of using light stemming to improve the validity of automated measures of translation, we have drawn from the work of Larkey et al. (2007), which demonstrated the advantages of using light stemming for information retrieval. It also appears that light stemming can provide benefits for training speech translation systems. Shen et al. (2007) obtained increases in BLEU scores of Arabic to English translation by processing training data with a series of normalization operations similar to the ones we investigated. Working with an early TRANSTAC training corpus, Riesa et al. (2006) report significant increases in BLEU scores when morpheme segmentation and orthographic normalization on both English and Iraqi Arabic are employed as vocabulary optimization steps before training. Habash and Sadat (2006) experimented with several types of tokenization and several training set sizes for translation from Arabic to English. They demonstrated increases in BLEU scores over a baseline for all types of stemming at all training set sizes. They also reported improvements when the test data was a different genre than the training data.

Acknowledgements We are grateful to everyone who contributed to the TRANSTAC program, especially researchers at SRI International, Carnegie Mellon University, IBM, and BBN/Raytheon, our colleagues at NIST, and TRANSTAC Program Manager Dr. Mari Maeda, whose vision made this research possible.

We would also like to express our appreciation to the reviewers of the article, including the special issue guest co-editors, whose attentive reading and thoughtful comments led to significant improvements.

References

Bannerjee S, Lavie A (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

Buckwalter T (2001) Arabic transliteration. http://www.qamus.org/transliteration.htm

Callison-Burch C, Osborne M, Koehn P (2006) Re-evaluating the role of BLEU in machine translation research. Proc EACL 2006:249–256

Chatterjee N, Johnson A, Krishna M (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 485–490

Condon S, Sanders G, Parvaz D, Rubenstein A, Doran C, Aberdeen J, Oshika B (2009) Normalization for automated metrics: English and Arabic speech translation. In: Proceedings of MT summit XII, Ottawa, Ontario, Canada, pp 33–40

Culy C, Riehemann S (2003) The limits of *n*-gram translation evaluation metrics. In: Proceedings of the MT summit IX, New Orleans, LA, USA, pp 71–78

Diab M (2009) Second generation tools (amira 2.0): Fast and robust tokenization, pos tagging, and base phrase chunking. In: MEDAR 2nd international conference on arabic language resources and tools, Cairo, Egypt

Habash N, Rambow O (2005) Arabic tokenization, morphological analysis, and part-of-speech tagging in one fell swoop. In: Proceedings of ACL, Ann Arbor

Habash N, Sadat F (2006) Arabic preprocessing schemes for statistical machine translation. In: Proceedings of the North American chapter of NAACL, New York

Larkey LS, Ballesteros L, Connell ME (2007) Light stemming for Arabic information retrieval. In: Soudi A, van den Bosch A, Neumann G (eds) Arabic computational morphology: knowledge-based and empirical methods. Springer, New York, pp 221–243

Lavie A, Sagae S, Jayaraman S (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

Likert R (1932) A technique for the measurement of attitudes. Arch Psychol 140:1–55

Owczarzak K, van Genabith J, Way A (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, Zhou WJ (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of ACL 2002, pp 311–318

Przybocki M, Peterson K, Bronsart S, Sanders G (2009) The nist 2008 metrics for machine translation challenge–overview, methodology, metrics, and results. Mach Trans 23:71–103

Riesa J, Mohit B, Knight K, Marcu D (2006) Building an English-Iraqi Arabic machine translation system for spoken utterances with limited resources. In: Proceedings of interspeech 2006: ICSLP ninth international conference on spoken language processing, p 2012

Sanders G, Bronsart S, Condon S, Schlenoff C (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, Marrakesh, Morocco

SCLite (2009) SCLite–NIST multi-modal information group. http://www.itl.nist.gov/iad/mig/tools/

Shen W, Delaney B, Anderson T, Slyh R (2007) The MIT-LL/AFRL IWSLT-2007 MT system. In: IWSLT 2007: international workshop on spoken language translation, Trento, Italy

Snover M, Dorr B, Schwartz R, Makhoul J, Micciula L (2006) A study of translation error rate with targeted human annotation. In: Proceedings of AMTA 2006, pp 223–231

Turian JP, Shen L, Melamed ID (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, Parvaz D (2008) Performance evaluation of speech translation systems. In: Proceedings of LREC 2008, Marrakesh, Morocco