Normalización para Métricas Automáticas: Traducción de Idioma Arábigo y Inglés

Sherri Condon*, Gregory A. Sanders†, Dan Parvaz*, Alan Rubenstein*, Christy Doran*, John Aberdeen*, y Beatrice Oshika*

*The MITRE Corporation
7525 Colshire Drive
McLean, Virginia 22102

†National Institute of Standards and Technology
100 Bureau Drive, Stop 8940
Gaithersburg, Maryland 20899–8940

{scondon, dparvaz, Rubenstein, cdoran, aberdeen, bea}@mitre.org / gregory.sanders@nist.gov

Abstract

El Agencia de Proyectos Avanzados Blanda (DARPA) programa de Comunicación y Traducción de Lenguaje para Uso Táctico (TRANSTAC) ha experimentado con el uso de métricas automáticas para la traducción de diálogos de la palabra. Para traducciones al inglés, las puntuaciones de BLEU, TER, y METEOR correlacionan bien con las opiniones humanas, pero las puntuaciones de traducciones a árabe correlacionan con las opiniones humanas menos fuertemente. Este texto proporciona evidencia para soportar la hipótesis de que las medidas automáticas de árabe se encuentran inferiores debido a la variación y desviación en el árabe, demostrando que operaciones de normalización mejoran la correlación entre puntuaciones de BLEU y la satisfacción humana con el contenido semántico — así como entre puntuaciones de BLEU y las opiniones humanas sobre el éxito de la transferencia del significado de palabras individuales del inglés al árabe.

1 Introducción

El objetivo de los programas TRANSTAC es demostrar la capacidad de desarrollo rápido y de implementación de sistemas de traducción de doble vía que permitan a los hablantes de diferentes lenguas comunicarse con uno en otro en situaciones tácticas reales. El caso de uso principal es el de conversaciones entre personal del ejército que habla sólo inglés y civiles locales que hablan otros idiomas.

La estrategia de evaluación adoptada para los programas TRANSTAC ha sido el desarrollo de dos tipos de evaluaciones: evaluaciones en vivo en las que los usuarios interactúan con los sistemas de traducción de acuerdo con varios protocolos y evaluaciones offline en las que los sistemas procesan grabaciones de audio y transcripciones de interacciones. Detalles de los métodos de evaluación TRANSTAC se describen en Weiss et al. (2008), Sanders et al. (2008) y Condon et al. (2008).

Las evaluaciones también ofrecen la oportunidad de explorar la aplicabilidad de medidas automatizadas de traducción de diálogo hablado y comparar estas medidas con las decisiones humanas de un panel de hablantes bilingües. Cuando se comparan los puntajes sistemáticos (puntajes de datos de un sistema) se ha obtenido correlaciones de puntuaciones de BLEU, TER, METEOR, y puntuaciones basadas en las decisiones humanas (Sanders et al., 2008). Cuando los datos son más finos en el nivel de sistema, sin embargo, las correlaciones de las decisiones humanas con las medidas automatizadas de traducción de maquinaria (MT) son mucho más bajas.

Las evaluaciones también ofrecen la oportunidad de estudiar los resultados de aplicar mediciones de MT automatizadas a lenguas diferentes al inglés. Estudios de los mediciones se han centrado principalmente en la traducción al inglés y otros idiomas europeos relacionados al inglés. El
TRANSTAC data present some significant differences between the automated measures of translation into English vs. Arabic. In particular, the results produced by five TRANSTAC systems in July 2007 and by the best-scoring three of those five 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.

Another difference concerns the relative values of the automated measures. There is evidence from the human judgments that the systems’ translations from English into Arabic are better than the translations from Arabic into English, and speech recognition error rates (word error rate) for English source-language utterances were much lower than for Arabic (Condon et al., 2008). Yet the scores from automated measures for translation from English to Arabic have consistently been significantly lower than for translation from Arabic to English.

We hypothesize that several features of Arabic are incompatible with assumptions that are fundamental to these 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: the estimates consistently correlate lower with 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 report describes experiments we have conducted to assess the extent to which scores are affected by the features and the extent to which those effects can be mitigated by normalization operations applied to Arabic texts before computing automated measures.

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 uses METEOR to score translations of the recorded scenarios with a measure that incorporates recall on the unigram level. METEOR and BLEU scores routinely have high correlation with each other. For those reasons, we will report only BLEU 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; Owczarzak et al., 2007; Turian and Melamed, 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.

Arabic, like other Semitic languages, has both a morphology and an orthography which are not immediately amenable to current approaches in automated MT scoring (and training, for that matter). 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 llbrnAmj (‘to the program’) consists of three separate elements (l_to, Al_the, brnAmj_program). So a scoring program encountering llbrnAmj (‘to a program’) without further tokenization would score it as

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.  

Arabic strings here are written according to Buckwalter orthographic conventions in Arabic attach frequently used function words to the related content word. As an example, the Arabic llbrnAmj (‘to the program’) consists of three separate elements (l_to, Al_the, brnAmj_program). So a scoring program encountering llbrnAmj (‘to a program’) without further tokenization would score it as

1. 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.

2. Arabic strings here are written according to Buckwalter orthographic conventions in Arabic attach frequently used function words to the related content word. As an example, the Arabic llbrnAmj (‘to the program’) consists of three separate elements (l_to, Al_the, brnAmj_program). So a scoring program encountering llbrnAmj (‘to a program’) without further tokenization would score it as

للبرواميج
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 which 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 *takfýr* (“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 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. Consequently, 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 definitely 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.

3 **English-Arabic Directional Asymmetry**

Evaluation data analyzed in this paper is recorded audio input from human speakers engaged in dialog scenarios. The TRANSTAC scenarios have included checkpoints, searches, infrastructure surveys (sewer, water, electricity, trash, etc.), training, medical screening, inspection of facilities, and recruiting for emergency service professionals.

The gold-standard metrics for translation adequacy are commonly deemed to be judgments from a panel of bilingual human judges. In TRANSTAC, we have a panel of five bilingual judges for each evaluation, and we obtained utterance-level judgments of semantic adequacy, initially on the four-value scale in Figure 1 and later on the seven-value scale in Figure 2. The seven-value scale has an explicit numeric interpretation as equally-spaced values, and the numeric interpretation was presented to the judges during their training (that is, the judges knew we would 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.

- Completely adequate
- Tending adequate
- Tending inadequate
- Inadequate

Figure 1: Four-value scale for semantic adequacy
We also had a highly literate native speaker of each source language mark the content words (nouns, verbs, adjectives, adverbs, important prepositions and quantifiers) in the source utterances and asked bilingual judges to say whether the meaning of each of these pre-identified content words was successfully transferred in the translation; we then calculated the probability of successful transfer of content words (Sanders et al., 2008).

Both methods of obtaining human judgments of semantic adequacy result in higher scores for translations from English into Arabic than for translations from Arabic into English. Data from the June 2008 evaluation, using the seven-point scale for semantic adequacy, averaged approximately one point higher for translations into Arabic. Earlier evaluations using the four-point scale showed the same pattern, as illustrated by Figure 3.

The same contrast holds for the human judgments that assess whether each content word in the English input was successfully translated, deleted, or substituted in the system output: the scores are higher for English to Arabic than for Arabic to English translations (Sanders et al., 2008). Moreover, the live evaluations provide a similar pattern of results: humans judged that system performance was better for translation from English to Arabic than from Arabic to English (Condon et al, 2008). Therefore, all of the evaluations involving human judges produce directional asymmetries that suggest translations into Arabic are better than translations into English.

One automated measure, the word error rate (WER) from the speech-recognition stage, suggests that translation from English to Arabic should be better than translation from Arabic to English: the average WER of the top 3 systems in the June 2008 evaluation was 13.5 for English and 31.1 for Arabic. This should account for some of the superior performance of the translations into Arabic because it is difficult for machine translation to overcome speech recognition errors.

Yet Figure 3 shows that the BLEU scores exhibit the opposite asymmetry: scores for translation from English to Arabic are considerably lower than scores from Arabic to English, and this pattern holds for METEOR and TER scores, too. Though it can be argued that the values of these scores should not be compared across languages, the concern is that these differences reflect serious flaws in the measures for languages like Arabic. In fact, the automated measures achieve higher correlations with human judgments for translation to English than for translation to Arabic.

4 Normalization for English and Arabic

We hypothesize that the primary responsibility for the contrasts in directional asymmetry between automated measures and human judgments lies in the features of Arabic described in section 2. Human judges are capable of ignoring minor variation in order to comprehend the meaning of language in context. The examples in (1) illustrate that even in the absence of context, errors in inflectional morphology do not prevent communication of the...
sender’s message.

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.

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 dialogs.

Rule-based normalization, referred to as Norm1, focuses on orthographic variation. For Arabic, a Perl script reduces seven types of variation by deleting or replacing variants of characters with a single form. Table 1 lists the seven types, provides examples of each, and describes the normalization operation that is applied in Norm1. These seven are acceptable orthographic variants in written Arabic and may therefore occur in the reference translations. For English, the rules include operations that transform letters to lowercase, 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 word error rate (WER) for evaluation of automatic speech recognition. It is standard practice for NIST to normalize system outputs and reference transcriptions when computing WER, though it is not standard to apply similar operations when computing automated measures of translation quality. 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. Most of the variants were caused by orthographic variation that is addressed in Norm1 so that Norm2 tends to be redundant with Norm1 for Arabic. But there are a few misspellings and typographical errors that are not corrected in Norm1 (e.g., بَلْلَهِ إِبْنِ يَزَادَهُ تَمِيمَةٍ, بَلْلَهِ إِبْنِ يَزَادَهُ تَمِيمَةٍ).

For English, regular contractions are expanded by the Norm1 rules, but the GLM includes forms without apostrophes such as aren’t. Reduced forms include gotta, gonna, ‘til, ‘cause, and ‘em. Because the contractor, Appen Ltd., is an Australian firm, some British spellings such as centre and vandalise are also included. Other mappings link various forms of abbreviations (e.g., C.P.R.), spelling errors in the reference texts, and spellings of Arabic names. Finally, the mapping separates nouns from the form ‘s when these occur in reference transcriptions and translations. Where appropriate, these were later hand-normalized as contractions. Ambiguous contractions with ‘d were also normalized by hand into the appropriate forms with would or had.

For the additional normalizations of Arabic that are the focus of this paper, we referred to the work of Larkey et al. (2007). They experimented with a variety of stemmers and morphological analyzers for Arabic to improve information retrieval scores. We produced a modified version of light10 for our use. At the beginnings of words, light10 removes the conjunction wa ( ﯾ ), the definite article al (ال), prepositions bi ( ب), li ( ل), fi ( ف ), and the form ك, which is used like English like or as, but is grammatically like a noun. The forms are removed only

<table>
<thead>
<tr>
<th>Type of Variation</th>
<th>Example</th>
<th>Normalization Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short vowel / shadda inclusions</td>
<td>ندرّ يّ ف. ندرّ ق.</td>
<td>Delete vowel and shadda diacritics</td>
</tr>
<tr>
<td>Explicit nunation inclusions</td>
<td>أَجَرْ وَأَحْجَرْ</td>
<td>Delete nunation diacritics</td>
</tr>
<tr>
<td>Omission of the hamza</td>
<td>و. ون.</td>
<td>Delete hamza</td>
</tr>
<tr>
<td>Misplacement of the seat of the hamza</td>
<td>أَلْءَطَأْرُ</td>
<td>Delete hamza</td>
</tr>
<tr>
<td>Variations where taa marbuta should be used</td>
<td>بَيْنَ الْعَابِدِين</td>
<td>Replace taa marbuta with haa</td>
</tr>
<tr>
<td>Confusion between yaa and alif maksura</td>
<td>يّ ف. يّ ش.</td>
<td>Replace alif maksura with yaa</td>
</tr>
<tr>
<td>Initial alif with or without hamza/madda/waslal</td>
<td>يّ أَسْمَ</td>
<td>Replace with bare alif</td>
</tr>
</tbody>
</table>

Table 1: Orthographic Normalization Operations Used in Norm1 for Iraqi Arabic
if followed by the definite article *al*, which is removed only if the remainder of the word is at least 2 characters long. These constraints minimize the possibility of removing characters which are actually part of the word. The conjunction may be removed without a following *al*, but only if the remainder of the word is at least 3 characters long.

These forms are not prefixes in the sense of bound morphemes attached at the beginnings of words. They are independent words that are conventionally spelled as part of the following word. In contrast, all the suffixes removed by light10 are bound morphemes. The suffixes that are removed are listed in Table 2. 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 inflexes that indicate plural nouns. Our primary concern is the prefixes because they are free morphemes with rigid word order, and separating them produces sequences that more closely resemble similar parts of speech in languages like English.

We produced two versions of normalizations based on light10: Norm2a separates the forms, but does not remove them, while Norm2b removes the separated forms.

<table>
<thead>
<tr>
<th>Arabic Suffix</th>
<th>Morphological Features When Attached to Verbs (V) and Nouns (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ْاء ان الس ن</td>
<td>V: 3rd person singular feminine object; N: possessive pronominal</td>
</tr>
<tr>
<td>ْان</td>
<td>N: Dual number</td>
</tr>
<tr>
<td>ْة</td>
<td>N: Feminine plural</td>
</tr>
<tr>
<td>ْة الن</td>
<td>N: Nominative masculine plural; V: subject agreement</td>
</tr>
<tr>
<td>ْة</td>
<td>N: Oblique masculine plural</td>
</tr>
<tr>
<td>ْي</td>
<td>V: 3rd person singular masculine object; N: possessive pronominal</td>
</tr>
<tr>
<td>ْي ن</td>
<td>N: Feminine nisba adjective, attributive</td>
</tr>
<tr>
<td>ْي</td>
<td>V: 3rd person singular masculine object; N: possessive pronominal</td>
</tr>
<tr>
<td>ْة</td>
<td>N: feminine singular (or singular of mass/collective noun)</td>
</tr>
<tr>
<td>ْة</td>
<td>N: 1st person singular possessive pronoun, nisba adjective marker</td>
</tr>
</tbody>
</table>

Table 2: Suffixes Removed in Light10 Stemming

Norm2a allows comparisons to reference translations using all of the forms that are present in the texts and handles the free morphemes like independent words, as they would be in a language like English. Norm2a has the effect of increasing the number of words that are scored, introducing a large number of unigrams (single words) that are likely to be scored as correct translations. This alone can increase scores from automated metrics. Scores from metrics such as BLEU that are based on n-gram co-occurrence statistics will also increase because Norm2a 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 4 presents the BLEU scores for English to Iraqi Arabic translation from 579 speech inputs before and after normalization. The normalization operations are cumulative: Norm2 is applied to the output of Norm1, while Norm2a and Norm2b are applied to the output of Norm2. The orthographic normalization in Norm1 led to slight increases in the BLEU scores for all systems (average .009). Norm2b increased BLEU scores an average of almost .04 above the Norm1 and Norm2 scores. Norm2a resulted in a large increase that averaged .148, boosted by the additional n-gram matches.

These additional n-gram matches could be an important confounding factor in our assessments of the advantages of these normalizations. One way to evaluate this confounding factor is to compare the results of analyses for Norm2b, which removes the affixes. The other approach we took to examine the effects of the extra n-grams is that 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. Because looking at only unigrams gives the translations no extra credit for additional n-grams, especially the extra n-grams that are present in Norm2a, these unigram-only values put the Arabic scoring on a more theoretically equal footing with English.

5 Normalization Results

We restrict our report to the BLEU metric in order to compare unigram scores, but results from other automated metrics such as METEOR are similar.

The correlations are based on a subset of the recorded offline evaluation data consisting of 109 English utterances (1431 words) and 96 Iraqi Arabic utterances (1085 words) in excerpts from 13 dialogs, each including about 7 exchanges. A fragment of typical input follows (with translation of the Arabic in square brackets).

E: well we can do certain things for you at this time
E: but you still have to go through the M.O.I. or the Mili--Ministry of Interior for some of your requisitions

A: ﻞﺯﺍﺭﺓ َ ﺃﺱﺃﻝٍﻡ ﻉﻝّ ﺍ. ﻙﺭٴﻡ

[%AH I’ll try to go to the Ministry and ask them about--about ((unintelligible)) %BREATH and we’ll see (literally: God is generous)]

E: well we can do in that process we will assist you with that process and maybe speed up their end of the %AH of the dealings with this

Table 3 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 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 in the right-hand two columns and bottom two rows: –Adj Prob Correct” refers to the adjusted probability correct score for transfer of content words described in section 3.

The highest correlation in Table 3 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.

Although Norm2b BLEU scores are more highly correlated to scores from human judgments than BLEU scores based on other normalizations, the highest correlation is achieved using BLEU_1 instead of BLEU_4. Correlations with the human-judgment metrics are always much lower for BLEU_4 than for BLEU_1. This result suggests our human judges were more tolerant of word order differences than the BLEU_4 metric expects.
Both types of human judgments focus on semantic quality, which may reduce the effect of word order.

Conclusion

Puzzling asymmetries between automated measures of English to Iraqi Arabic and Iraqi Arabic to English translations can be attributed to orthographic variation, inflectional morphology, and relatively free word order in Arabic. We demonstrate that the asymmetric effects of these linguistic differences can be mitigated by normalization processes that reduce orthographic variation and delete or separate affixes and function words.

Correlations to human judgments suggest 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 dialogs, where disfluencies and inference from context are the norm.

In demonstrating the advantages of using light stemming to improve the validity of automated measures of translation, we have drawn from Larkey et al.’s (2007) research, 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 BLEU score increases by processing training data with a series of normalization operations similar to the ones we investigated.

In future work, we will explore different combinations of deleting vs. separating affixes along with enhancements to take into account pronominal endings for the 2nd person (more common in spoken discourse) and other forms unique to Iraqi Arabic. Nevertheless, the first approximation presented here has been productive.

References


