Evaluation of Machine Translation Errors in English and Iraqi Arabic

Sherri Condon, Dan Parvaz, John Aberdeen, Christy Doran, Andrew Freeman, and Marwan Awad
The MITRE Corporation
McLean, VA
{scondon, dparvaz, Aberdeen, cdoran, afreeman, mawad}@mitre.org

Abstract

Errors in machine translations of English-Iraqi Arabic dialogues were analyzed at two different points in the systems' development using HTER methods to identify errors and human annotations to refine TER annotations. The analyses were performed on approximately 100 translations into each language from 4 translation systems collected at two annual evaluations. Although the frequencies of errors in the more mature systems were lower, the proportions of error types exhibited little change. Results include high frequencies of pronoun errors in translations to English, high frequencies of subject person inflection in translations to Iraqi Arabic, similar frequencies of word order errors in both translation directions, and very low frequencies of polarity errors. The problems with many errors can be generalized as the need to insert lexemes not present in the source or vice versa, which includes errors in multi-word expressions. Discourse context will be required to resolve some problems with deictic elements like pronouns.

1. Introduction

A limitation of both human judgments and automated measures of translation quality is that they are not diagnostic. Analyses of the errors produced by machine translation (MT) systems have the potential to focus research aimed at improving translation performance. However, translation error annotation is problematic because there are many ways to translate a single expression from one language to another. It is difficult to annotate specific errors in a consistent way to yield quantifiable results because what constitutes an error depends on what is considered correct. A team of linguists at The MITRE Corporation had an opportunity to analyze translation errors from 4 English-Iraqi Arabic speech translation systems at two different stages of development, and this paper presents the methods adopted to identify errors along with some results of the analyses.

We developed a methodology that incorporates techniques from the Translation Error Rate (TER) family of measures (Snover et al., 2006) in order to facilitate identification and annotation of errors. In the methods used to compute the Human Translation Error Rate (HTER) score, human editors create a reference translation of a source input that is as close as possible to the system's machine translation of that input, while preserving the content of the source. We adopted this process so that the differences between system hypotheses and reference translations are more likely to reflect problems in the MT rather than variation among possible satisfactory translations.

2. Related Work

Methods for evaluating MT output have proliferated rapidly: 39 automated measures were submitted to the NIST 2008 Metrics for Machine Translation Challenge (Przybocki, Peterson & Bronsart, 2008). However, there have been few attempts to analyze MT errors so that researchers can identify the types of errors that are made. Llitjós, Carbonell & Lavie (2005) created a hierarchical taxonomy of errors for use in refining rules of transfer-based MT systems. They do not describe which types of errors were identified, but Vilar et al. (2006) extended the typology and used it to annotate errors in translations from Spanish to English, English to Spanish, and Chinese to English. Three types of inputs from audio data were compared: clean text transcripts, verbatim transcripts, and the output of automated speech recognition (ASR).

The error annotation reported in Vilar et al. was performed by human annotators using reference translations and a tool that highlights the differences between the MT and a reference translation. Aside from explaining that the use of reference translations "must be done with care" (Vilar et al., 2006, p. 697), there is no description of the annotation process, but the results provide a rich source of generalizations about the types of errors produced by the RWTH statistical MT system.

Popović & Ney (2007) combine part-of-speech tagging with Word Error Rate (WER) and Position Independent Word Error Rate (PER) measures (Tillman et al., 1997) to provide automated error analyses of the same translations that Vilar et al. (2006) analyzed. The results of the automated analyses are comparable to the results of the human analyses and provide error frequencies for a variety of syntactic and morphosyntactic classes without the time and expense required by human annotators.

Popović & Ney do not specify whether they used the clean text transcripts or the verbatim ones, but Vilar et al. note that the latter include some ungrammatical constructions which present an additional source of errors. It remains to be seen whether conventional parsers and morphological analyzers can provide accurate automated analyses of highly disfluent speech data such as the verbatim transcripts we analyzed.

Approved for Public Release: 10-1174. Distribution Unlimited. The views, opinions, and/or findings contained in this article are those of the author/presenter and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense.

3. Methods

Our approach to error annotation has been to adopt the methods that are used to compute the HTER score for the Defense Advanced Research Projects Agency's (DARPA) Global Autonomous Language Exploitation (GALE) program. For HTER, The National Institute of Standards for Technology (NIST) post-editing tool was used to facilitate this process (NIST, 2008) and the guidelines for post-editing were similar to the GALE guidelines (NIST, 2007). The tool allows the editor to view the TER score for the machine translation compared to the editor's reference translation so that the editor can experiment with alternative wordings and see which is scored closest to the machine translation. In addition, the tool displays reference translations produced for each source transcription by independent human translators. For the TRANSTAC data, each human post-editor was able to view 4 reference translations of each source.

Three native English speakers annotated the translations from Arabic to English, and three Arabic speakers annotated the translations from English to Arabic. For the latter, two non-native Arabic linguists each annotated half of the data, and a third native Arabic speaker reviewed the annotations. Differences were reconciled by the relevant annotator pairs. For half of the Arabic to English translations, each translation was annotated by two annotators, and the differences were reconciled in meetings of all three annotators.

After producing the customized reference translations, annotators used the TER output to classify each error. The TER score computes the number of deletions, insertions, substitutions, and shifts of words that are required to modify the machine translation so that it matches the reference translation. The tool that computes TER aligns the machine translation and the reference translation and annotates each error by associating an abbreviation *I*, *D*, or *S* with the word that is inserted, deleted, or substituted respectively. For shifts, the "@" symbol appears in the position that the shifted word has moved from. The errors identified by these TER notations were used to classify the errors into the following common grammatical classes: *pronoun*, *noun*, *verb*, and *other* (prepositions, adjectives, conjunctions, and adverbs). The result is an annotation that combines the TER classifications with the word class annotations. For example, *ivp* was used to annotate an inserted verb or predicate element such as a modal or auxiliary.

Two additional syntactic categories were adopted in an attempt to limit the annotations associated with each error to a single category. One category, labeled *Pro-Verb* in this paper, was used to annotate errors involving a single element consisting of a pronoun and verb. In translations to English, these were insertions or substitutions of contractions such as *I'm*, and in translations to Arabic, the annotation was used if the subject was expressed only by inflection on the verb and both the verb and the subject inflection were wrong. The second category, labeled *Subject Person*, was used exclusively to annotate Arabic verbs when the subject person inflection did not agree with the subject.

In addition to the annotations described above, errors that changed the polarity or speech act of an input were noted, as were words that were wrongly transliterated (or replaced by question marks) and errors caused by word sense ambiguity. A set of annotation guidelines was compiled so that annotators would complete the annotation process as consistently as possible. In addition to producing reference translations, annotators occasionally changed the TER alignment when it did not capture the relations between the MT and the reference in a satisfactory way. Usually these changes were made to align words in the same syntactic categories.

Not all differences between machine and human translation noted by TER were annotated as errors. For example, inflectional differences such as the singular *guard* vs. the plural *guards* were annotated as *null* to indicate that the error was insignificant. Other differences annotated as null were deleted or inserted articles *(the, a)*, deleted *and* (if the absence did not affect the meaning), repetitions, and synonyms. Subject inflection on Arabic verbs that did not agree with the subject was usually annotated as null if the error was in number or gender, unless the annotator believed that the result would be ambiguous or confusing for the listener. In contrast, subject person inflection was annotated using the special categories described above.

4. Data

The translations selected for the study were from an evaluation of 4 speech translation systems that were developed for the DARPA Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) program (Weiss et al., 2008). The systems used statistical MT engines, except that one system's English to Arabic translation was rule-based with a statistical fall-back. In addition to live evaluations, TRANSTAC evaluations include evaluations based on recorded dialogues, which allows all systems to be tested on identical inputs (Condon et al., 2008). For the latter, systems are also tested using transcriptions of the audio inputs in order to evaluate the translation capability without speech recognition errors. We used outputs from those text translations for the error analyses, and we chose the same subset of translations for which NIST obtained human judgments (Sanders et al., 2008).

Translations were analyzed from two evaluations conducted in July, 2007 and June, 2008. In the 2007 corpus, there were 95 source utterances translated from English to Iraqi Arabic, and 101 from Iraqi Arabic to English. In the 2008 corpus, we analyzed 109 source

Translation Direction	July, 2007	June, 2008
English to Iraqi Arabic	380	436
Iraqi Arabic to English	404	428

Table 1: Total Number of Translations Analyzed

	Pronouns		Verbs		Nouns		Other		Total	
Translation to →	English	Arabic	English	Arabic	English	Arabic	English	Arabic	English	Arabic
Deletion	0.109	0.065	0.121	0.026	0.038	0.036	0.056	0.086	0.323	0.214
Insertion	0.057	0.034	0.043	0.024	0.017	0.026	0.043	0.047	0.160	0.132
Substitution	0.095	0.059	0.090	0.077	0.048	0.119	0.116	0.132	0.348	0.387
Total	0.261	0.158	0.253	0.127	0.103	0.181	0.214	0.266	0.831	0.732

Table 2: Deletion, Insertion, and Substitution Errors in Word Classes: Proportion of 2007 Total Errors

utterances that were translated from English to Iraqi Arabic and 107 that were translated from Iraqi Arabic to English. The translations from four machine translation systems were processed and annotated so that several hundred were analyzed for each direction from each evaluation. Table 1 presents the totals.

5. Results

Tables 2 and 3 present the proportions of non-null errors annotated by type and word class in the 2007 corpus. (Together the tables sum to 1.) *Deletions* are words in the reference translation that do not occur in the machine translation, and *insertions* are words in the machine translation that do not occur in the reference translation. The TER *shift* annotation is labeled *Word Order* in this report. For translations to both languages, the most frequent type of error is substitution. In fact, because both the *Pro-V* and *Subject Person* categories in Table 2 are also substitutions in Arabic, nearly half of the total errors for translations into Iraqi Arabic are substitutions, and this proportion holds in the 2008 data, too (Tables 6 and 7).

For translations into English, substitutions are about one third of the total errors, and deletions account for another third of the total errors. Pronouns and verbs are the word classes with the highest frequencies of errors in translations to English: together they represent about half of the annotated errors. No single word class has a comparable frequency of errors in translations to Iraqi Arabic. However, if errors involving subject person inflection on the verb (*Pro-V* and *Subject Person*) are added to the verb errors, then about 22% of the errors in translations to Iraqi Arabic involve verbs.

The frequency of polarity reversals in which a negative becomes positive or vice versa is very low at less than 1% for translations to both languages. Errors that affect the speech act of the source are also very rare. These were only noted for the declarative/interrogative/ imperative contrast: other speech acts expressed by verbs or modals were classified as verb errors. All of the speech act errors occurred when a translation from Arabic to English omitted or inserted an inverted auxiliary, resulting in a statement instead of a yes/no question or vice versa. Also very rare were source words that were not translated but were transliterated or replaced with question marks, which are labeled *Untranslated* in Table 3.

Identifying errors caused by word sense ambiguity requires analysis of source inputs, but the HTER-based procedures we adopted do not require knowledge of the source utterances. In order to identify the frequency of word sense errors, a bilingual judge viewed the source for each annotated error in order to determine whether it should also be annotated as a word sense error. The proportion of errors attributed to word sense ambiguity in Table 3 is much smaller than reported in studies like Vilar et al. (2006). This result is likely due to the relatively narrow domains of the TRANSTAC test data.

Translation to →	English	Arabic
Word Order	0.139	0.169
Pro-Verb	0.013	0.007
Subject person	n/a	0.090
Polarity	0.009	0.002
Speech act	0.006	0
Untranslated	0.001	0
Word sense*	0.021	0.032
Total	0.168	0.268

*not included in total

Table 3: Other Annotated Errors: Proportion of 2007 Total Errors

We were fortunate to be able to annotate translation errors for the same 4 systems at two different stages of development, which allows us to observe any changes in error types and to compare error frequencies with automated measures of translation quality. Table 4 provides the frequencies of the TER scorer's automatic annotation of differences between the post-edited reference translations and the machine translations from English to Iraqi Arabic. It also provides the frequencies of TER annotations to which annotators assigned a non-null error type. Because there were more input utterances in the 2008 evaluation than in the 2007 evaluation, the non-null error frequencies are also normalized per input. Finally, BLEU scores computed in the offline evaluations are presented. The latter were computed on a larger sample of test data (500-600 inputs in each direction) from which the annotated corpora were selected. Details about computation of the BLEU scores and NIST's selection of the subset of inputs we annotated are available in Condon et al. (2008) and Sanders et al. (2008) respectively. Table 5 presents the same information for translation from Iraqi Arabic to English.

Tables 4 and 5 show that only about half of the differences between the post-edited reference translations and the machine translations to Arabic were viewed as significant errors by the annotators in translations to Arabic,

System	2007 TER Errors	2008 TER Errors	2007 Non-null Errors*	2008 Non-null Errors*	2007 BLEU Scores	2008 BLEU Scores
A	353	225	179 /1.84	134 /1.22	.321	.341
B	408	222	203 /2.09	132 /1.21	.195	.305
C	246	144	116 /1.19	87 /0.80	.276	.339
D	233	221	115 /1.19	104 /0.95	.233	.325

*raw frequency/normalized per input

Table 4: English to Arabic Error Frequencies and BLEU Scores

System	2007 TER Errors	2008 TER Errors	2007 Non-null Errors*	2008 Non-null Errors*	2007 BLEU Scores	2008 BLEU Scores
A	292	269	176 /1.74	180 /1.68	.479	.469
B	355	354	240 /2.38	223 /2.08	.364	.446
C	287	279	166 /1.64	175 /1.64	.468	.484
D	291	229	189 /1.87	146 /1.36	.468	.475

*raw frequency/normalized per input

Table 5: Arabic to English Error Frequencies and BLEU Scores

compared to about two thirds for translations to English. As mentioned in section 3, minor inflectional differences such as number, gender and even tense were not annotated as errors if the annotator judged that the speaker's intent could be inferred adequately from the translation and its context. In this respect, the annotations resemble the METEOR measure (Banerjee & Lavie, 2005), which employs stemming before matching reference and machine translations. Similarly, other differences were annotated as null if the annotator judged that these errors would be ignored by interlocutors who were tolerant of machine translation and were able to make sense of the machine outputs in the dialogue context. Also, some null annotations involved disfluencies and repetitions that did not affect the primary import of the input.

Tables 6 and 7 present the proportions of non-null errors annotated by type and word class in the 2008 corpus. By comparing with Tables 2 and 3, we can observe whether the significant improvements that systems achieved in translation quality produce any changes in the types of errors that we observe. For translations to Iraqi Arabic, most of the proportions differ by less than .04, but a few are somewhat greater. The proportion of pronoun errors is about .065 lower in the 2008 corpus and there is an increase of the same amount in the number of *Subject-person* errors, which can also be viewed as an increase in the proportion of verb errors.

5.1 Iraqi Arabic Errors

At 15.5%, errors in subject person inflection on verbs comprise one of the larger categories of errors in the 2008 translations to Iraqi Arabic. (1) provides an example in which the person inflection on the verb, which is the first person singular form that would be used when "I" is the subject, occurs with a third person plural subject *my marines*. "Ref" indicates the human post-edited translation, and "MT" indicates the system output.

(1) Ref: المارينز مالتي رح يفتشون البيت
MT: المارينز مالتي رح أفتش البيت

Ref: AlmArynz mAlty rH yft$wn Albyt
Ref: the-Marines my will 3m-search-pl the-house

MT: AlmArynz mAlty rH >ft$ Albyt
MT: the-Marines my will 1sg-search the-house

Source: my marines are going to search the house

The system translation would be understood like "my marines I am going to search the house," which could be interpreted several ways.

Figure 1 presents the proportions of 5 classes of errors in the 2007 English to Iraqi Arabic translations. It demonstrates the variation in frequencies that occurs among the systems. System D is the translation engine which employed rule-based MT for English to Iraqi Arabic translations, and the much lower proportion of *Subject-person* errors for System D in Figure 1 suggests that rule-based approaches perform better when agreement inflection is required.

Although word order tends to be stricter in English than in Arabic, the frequency of word order errors is about the same in translations to Iraqi Arabic and translations to English. Many word order errors in both directions reverse the order of head nouns and their modifiers. This is another case in which a clash in the structures of the two languages results in MT errors: in English adjectival modifiers usually precede the noun, whereas in Arabic they usually follow. An example is in (2).

(2) Ref: عندهم التجهيزات إضافية
MT: عند إضافي التجهيزات

Ref: Endhm AltjhyzAt <DAfyp
Ref: at+them det+supplies additional+fem

MT: End <DAfy AltjhyzAt
MT: with additional det+supplies

Source: they have additional supplies

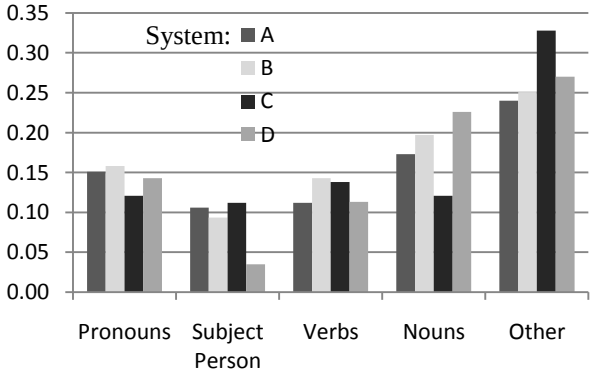

Figure 1: Proportions of Error Types in 2007 English to Iraqi Arabic Translations

	Pronouns		Verbs		Nouns		Other		Total	
Translation to →	English	Arabic	English	Arabic	English	Arabic	English	Arabic	English	Arabic
Deletion	0.112	0.039	0.102	0.022	0.046	0.031	0.075	0.053	0.334	0.144
Insertion	0.047	0.028	0.039	0.035	0.007	0.039	0.039	0.079	0.131	0.182
Substitution	0.102	0.024	0.087	0.072	0.052	0.103	0.081	0.153	0.323	0.352
Total	0.261	0.092	0.228	0.129	0.105	0.173	0.195	0.284	0.789	0.678

Table 6: Deletion, Insertion, and Substitution Errors in Word Classes: Proportion of 2008 Total Errors

An analysis of the word order errors in the 2008 English to Arabic translations shows that 27% of the word order errors reverse the order of a head, usually a noun, and its modifiers. This proportion does not include the order of modifiers that are expressed in the Arabic *idafa* or *construct* structure. Machine translations that reverse the order of nouns in an Iraqi Arabic *idafa* structure constitute a significant 40% of the word order errors. (3) provides an example.

(3) Ref: إجيت علمود أشوف محطة الكهرباء مالتكم
MT: إجيت علمود أشوف الكهرباء المحطة مالتكم

Ref: <jyt Elmwd >$wf mHTp AlkhrbA' mAltkm
Ref: came+1s in-order-to see+1s station-of det+electricity poss-2p

MT: <jyt Elmwd >$wf AlkhrbA' AlmHTp mAltkm
MT: came+1sin-order-to see+1s det+electricity det+station poss-2p

Source: I came out to take a look at your power station

Analysis of the 28% of errors annotated as *Other* in the 2008 English to Arabic translations revealed that a significant portion of the errors involved dependencies among multiple words. Some errors involved prepositional particles selected by verbs, as in (4).

(4) Ref: رح تريد منه يفتح كل البيبان
MT: تريد له يفتح كل البيبان

Ref: rH tryd mnh yftH kl AlbybAn
Ref: future 2ms+want from+him 3ms+open all-of det+doors

MT: tryd lh yftH kl AlbybAn
MT: 2ms+want to+him 3ms+open all-of det+doors

Source: you would want him to open up all the doors

In contrast, there were only 3 instances of wrong prepositions in verb modifiers. The largest group of errors annotated as *Other* involves multi-word expressions such as "to where" in (5), which must be translated as a single connector meaning "in order that/in order to."

(5) Ref: نقدر ننطيك الفلوس علمود تطلع وتشتري المواد
MT: أقدر الفلوس وين تطلع وتشتري المواد

Ref: nqdr nnTyk Alflws Elmwd tTlE wt$try AlmwAd
Ref: can+1p 1p+give+2ms det+money in-order-to 2ms+go-up and+2ms+buy det+material

MT: >qdr Alflws wyn tTlE wt$try AlmwAd
MT: can+1s det+money where 2ms+go-up and+2ms+buy det+material

Source: we can give you funds to where you can go out and buy the materials

Translation to ⟶	English	Arabic
Word Order	0.171	0.166
Pro-Verb	0.003	0.000
Subject person	n/a	0.155
Polarity	0.017	0
Speech act	0.019	0
Untranslated	0.001	0
Total	0.211	0.321

Table 7: Other Annotated Errors: Proportion of 2008 Total Errors

23% of errors annotated as *Other* involved multiple word dependencies like those in (4) and (5). Most of these errors can be viewed as word sense ambiguities, and annotators found that careful analysis of the errors annotated as *Other* led them to increase their counts of word sense errors. Excluding the multi-word errors, another 17% of errors annotated as *Other* could be attributed to word sense ambiguities. Consequently, the total proportion of errors attributable to word sense ambiguities in the 2008 English to Arabic translations may be as high as 10%, which is similar to the proportions reported in Vilar et al. (2006).

5.2 English Errors

The high frequency of pronoun errors in translations to English is striking. In the 2007 translations to English, the frequencies of nouns and pronouns in the translations are nearly equal (about 18%), yet the frequency of pronoun errors is 2 or more times higher than the frequency of noun errors for every translation system. These errors are significant because many of the errors involve basic personal pronouns *(I, you, he, she, it,* etc.*)*. Consequently, the errors do not occur because the translation system has encountered rare or out-of-domain language. Instead, the errors are caused by significant linguistic differences between the two languages and by the fact that use of pronouns depends on their context. For example, in (6), the subject of the Arabic verb "understand" can be either "I" or "you."

(6) iftahamit
understand+past+1st or 2nd person singular subject
"I/you understood"

In the case of utterances like (6), the only way to determine the referent of the verb's subject is to use the context. Human interlocutors are very good at resolving this kind of ambiguity from the discourse context, but the

task has proven to be quite challenging for language processing systems. The example in (7) demonstrates a pronoun substitution in machine translation that is likely to have been caused by the ambiguity illustrated in (6).

(7) Ref: I saw his symptoms
MT: you see his symptoms

Another difference between English and Arabic can cause incorrect pronouns to occur in machine translations. Like Romance languages, Arabic has the property of gender associated with nouns. Every noun is either masculine or feminine, and there are no pronouns like *it*. Consequently, in order to select the correct English pronoun form, it is necessary to know the referent of the pronoun. Some examples of this problem are provided in (8).

(8) MT: are taking care of it god willing and hopefully it will get better a little bit more

Ref: we are taking care of him god willing and hopefully he will get better soon

MT: of course I mean it is in good condition

Ref: of course I mean she is in good condition

The translations in (8) occur in hospital scenarios, and the third person pronouns refer to patients. We noted that it was much less likely for the systems to translate a pronoun that should have been neutral *it* as *he* or *she*, as if use of *it* had been over-generalized (or over-weighted).

The first example in (8) illustrates another difference between English and Arabic that produces problems for machine translation. Arabic is a language in which subject pronouns are not usually expressed. Like Spanish and Italian, Arabic has rich inflection on the verb that agrees with the subject in person, number, and gender. Consequently, the inflection on the verb can be adequate to indicate which pronominal subject the speaker intends, and in all three languages, the subject pronoun is usually produced only for emphasis. In contrast, English requires the subject pronoun to be expressed even when the verb form would permit no other subject[1]. The example in (9) demonstrates a pronoun error in which the machine translation has failed to include a subject pronoun because the Arabic did not.

(9) MT: was bitten by a scorpion
Ref: he was bitten by a scorpion

Pronoun deletions ranged from 28% to 62% of the pronoun errors produced by the four 2007 machine translation systems. The kind of error illustrated in (9) can be serious if the intended subject is not clear from the context, but at least the English-speaking listener is aware that there is a potential ambiguity in the reference due to a missing word. In contrast, substituting an incorrect pronoun, as in (7) and (8), can be a serious problem that leads to misunderstandings. 25% to 47% of the pronoun errors generated by the 2007 machine translation systems were pronoun substitutions, but not all of them were substitutions of one pronoun by another. Some were 'substitutions' of pronouns by other word classes, which occurs when TER aligns the machine translation and the reference translation. Most of those might be better classified as pronoun deletions.

Inserted pronouns are not as frequent as deleted or substituted pronouns. There are two likely sources of inserted pronouns. First, because it is often necessary for the translation to insert a pronoun subject when the Arabic source does not have one, there is a possibility that systems will mistakenly insert the pronoun subject when the source already has another subject expressed. An example is presented in (10).

(10) MT: at the same time those people they store them in this complex

Ref: at the same time those people store them in this complex

Second, Arabic and many other languages often require pronouns where English permits a gap. The most common examples are resumptive pronouns in relative clauses, as in (11).

(11) MT: it is about three kilometers from the point the checkpoint that he ran away from it

Ref: it is about three kilometers from the point the checkpoint that he ran away from

Pronoun errors like (10) and (11) detract from the fluency of the machine translation, though they are unlikely to lead to serious miscommunication. These errors were often annotated as *null* because the annotator judged that they did not interfere with comprehension of the speaker's intended meaning.

Not all of the errors annotated as pronoun errors involved personal pronouns. Errors involving any pronominals received this annotation, including indefinite pronouns *(someone, anything),* pleonastic pronouns *(it, there),* relative and interrogative pronouns *(which, when),* demonstrative pronouns *(this, that),* and pronominals corresponding to bare noun adverbials *(here, there).*

Another clash of linguistic properties results in a large proportion of the verb errors in translations to English. Over 40% of these errors involve the copula, the verb *be* that expresses identity and attribution. The copula is unique because it does not contribute meaning beyond a generic equivalence or attribution plus whatever tense or other verbal inflection it might carry. As a consequence, in Arabic and other languages such as Russian, the copula is omitted in the present tense. In (12), there is no Arabic word in the source utterance corresponding to *is* in the reference translation.

(12) MT: no sir all the family in the house
Ref: no sir all the family is in the house

Like pronoun errors, errors involving forms of the copula *be* cannot be attributed to structures or vocabulary that

[1] The one case in which the form of an English verb unambiguously determines its subject is *am*.

occur infrequently. Instead, they reflect linguistic differences that are challenging for translation systems. In the 2007 systems, errors involving a form of *be* either in the reference translation or in the machine translation ranged from 38% to 52% of the verb errors. Together, the pronoun errors and copula errors averaged over 40% of the total errors that the 2007 machine translation systems produced when translating from Iraqi Arabic to English.

6. Conclusions

For the most part, the error types that have been quantified in this research are predictable from the structural differences between English and Iraqi Arabic. These differences include (1) head and modifier order, particularly for nouns, (2) subject inflection on verbs in Iraqi Arabic, (3) unexpressed pronoun subjects in Iraqi Arabic, (4) gender inflection on pronouns, and (5) unexpressed copula in Iraqi Arabic. Of the many structural differences between English and Iraqi Arabic, the ones above are undoubtedly salient in the TRANSTAC corpora because the structures occur frequently in speech. Consequently, the problems occur not for lack of examples, but because of linguistic differences that are difficult to resolve with statistical approaches to machine translation.

Many of these errors share another property which is known to challenge MT: they require inserting linguistic material into the translation that was not present in the source. Or they require the reverse: refraining from inserting linguistic material into the translation to correspond to items in the source. This failure to achieve one-to-one correspondence between the translation and the source also tends to be a feature of multi-word expressions, which are known to be sources of difficulty for MT and other language processing operations.

Some errors involve more than structural differences between the two languages. The pronoun errors in particular often require knowledge of the discourse context to resolve. Iraqi Arabic speakers communicate effectively even though the past forms of verbs do not distinguish between first and second person subjects because the context of the interaction makes it clear whether the speaker is talking about "I" or "you." Similarly, knowing whether to translate an Arabic "he" or "she" as "it' requires knowledge of the referent of the pronoun. These problems will continue to challenge machine translation systems until they are resolved.

Finally, this research has provided some estimates of the relative proportions in which the error types impact machine translation. These estimates can be used to guide efforts to improve translation quality.

7. 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.

8. References

Banerjee, S. and 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.

Condon, S., Sanders, G., Parvaz, D., Rubenstein, A., Doran, C., Aberdeen, J., and Oshika, B. (2009). Normalization for Automated Metrics: English and Arabic Speech Translation. *Proceedings of MT Summit XII.*

Llitjós, A., Carbonell, J., and Lavie, A. (2005). A framework for interactive and automatic refinement of transfer-based machine translation. In *Proc. Of the 10th Annual Conf. of the European Association for Machine Translation (EAMT)*, Budapest, Hungary,May.

NIST. (2007). Post Editing Guidelines for GALE Machine Translation Evaluation Version 3.0.2. http://projects.ldc.upenn.edu/gale/Translation/Editors/GALEpostedit_guidelines-3.0.2.pdf

NIST. (2008). MTPostEditor_V1.2.0.jar. V1.2.2 available at Linguistic Data Consortium (LDC). GALE: Machine Translation Post-Editing Resource page for current post-editors. http://projects.ldc.upenn.edu/gale/Translation/Editors/

Popović, M. and Ney, H. (2007). Word Error Rates: Decomposition over POS Classes and Applications for Error Analysis. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 48–55.

Przybocki, M., Peterson, K., and Bronsart, S. (2008). *Official results of the NIST 2008 "Metrics for MAchine TRanslation" Challenge (MetricsMATR08)*, http://nist.gov/speech/tests/metricsmatr/2008/results/

Sanders, G., Bronsart, S., Condon, S., and 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. *Proceedings of LREC 2008.*

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

Tillmann , C., Vogel, S., Ney, H., Zubiaga, A., and Sawaf, H. (1997). Accelerated DP Based Search for Statistical Translation. In European Conference on Speech Communication and Technology, pp. 2667-2670.

Vilar, D., Xu, J., D'Haro, L., and Ney, H. (2006). Error Analysis of Statistical Machine Translation Output. In *Proc. of the Fifth Int. Conf. on Language Resources and Evaluation (LREC)*, pp. 697–702.

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