Five Shades of Noise: Analyzing Machine Translation Errors in User-Generated Text

Marlies van der Wees **Arianna Bisazza** **Christof Monz**
Informatics Institute, University of Amsterdam
{m.e.vanderwees,a.bisazza,c.monz}@uva.nl

Abstract

It is widely accepted that translating user-generated (UG) text is a difficult task for modern statistical machine translation (SMT) systems. The translation quality metrics typically used in the SMT literature reflect the overall quality of the system output but provide little insight into what exactly makes UG text translation difficult. This paper analyzes in detail the behavior of a state-of-the-art SMT system on five different types of informal text. The results help to demystify the poor SMT performance experienced by researchers who use SMT as an intermediate step of their UG-NLP pipeline, and to identify translation modeling aspects that the SMT community should more urgently address to improve translation of UG data.

1 Introduction

User-generated (UG) text such as found on social media and web forums poses different challenges to statistical machine translation (SMT) than formal text. This is reflected by poor translation quality for informal genres (see for example Figure 1), which is typically measured with automatic quality metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), or TER (Snover et al., 2006). These scores alone, however, only reflect the overall translation quality, and do not provide any insight in what exactly makes translating UG text hard. While such knowledge is crucial for improving SMT of UG text, surprisingly little work on error analysis for SMT of user-generated text has been reported.

In (Arabic):	قالت عشان العيال متزعلش
Reference:	she said so the kids do not feel upset
MT output:	she said because of the sons
In (Chinese):	你 路上 慢 点
Reference:	take your time
MT output:	you are on the road to slow points

Figure 1: SMS examples with poor SMT output.

Moreover, the notion of user-generated content only partially specifies the exact nature of documents. What all documents that can be classified as being UG have in common is the fact that they have been written by a lay-person, as opposed to a journalist or professional author, and that they have not undergone any editorial control. UG text also tends to express the writer's opinion to a larger degree than news articles which generally strive for balance and nuance. Within UG text, we can distinguish several subclasses, including (i) message and dialog-oriented content such as short message service (SMS) texts, Internet chat messages, and transcripts of conversational speech, (ii) commentaries to news articles, often expressing an opinion about the corresponding articles and relating the content to the reader's situation, and (iii) weblogs, which can bear some resemblance to editorial pieces published by news organizations.

While UG text processing tasks are becoming more and more common, the research in SMT is still mostly driven by formal translation tasks[1], and existing error analysis approaches are only partially useful for UG. In this work, we conduct a series of analyses on five different UG benchmark sets for two language pairs, Arabic-English and Chinese-English, with the goals of (i) explaining the typically poor SMT performance observed for UG texts, and (ii) identifying translation modeling

[1] One of the very few exceptions is NIST OpenMT 2015, which focusses entirely on translating informal genres.

Proceedings of the ACL 2015 Workshop on Noisy User-generated Text, pages 28–37,
Beijing, China, July 31, 2015. ©2015 Association for Computational Linguistics

aspects that should be addressed to improve translation of UG data. We not only contrast our observations with two news data sets, but we also show that SMT quality can vary significantly across different types of UG content, and that different UG types exhibit dissimilar error distributions. Specifically, we summarize our main findings as follows:

- The SMS and chat benchmarks are the most distant from formal text at all the analyzed levels. Errors in other types of UG are often more similar to news errors than to those in SMS and chat messages.

- SMT model coverage dramatically deteriorates for phrases of length 3 or longer in most of the UG benchmarks.

- Errors due to out-of-vocabulary (OOV) words in the *source* text substantially increase in number for UG data sets, but are considerably less common than errors due to *source-target* OOVs, i.e., phrase pairs that are not covered by the SMT models.

2 Related Work

Identifying and analyzing different types of SMT errors is an essential step towards the development of translation approaches that can achieve more robust performance, and has been the focus of earlier work. Popović and Ney (2011), for example, combine word error rates with morpho-syntactic information to classify errors into five categories; inflectional errors, reordering errors, lexical errors, word deletions, and word insertions. Irvine et al. (2013) use word alignment links to quantify incorrect lexical choices, and determine how such errors change when shifting domains. Other work on SMT error analysis studies the effect of domain adaptation on SMT, for example by examining in which stage of the SMT pipeline the available in-domain data can best be used (Duh et al., 2010), or whether it is more promising to improve either phrase extraction or scoring (Bisazza et al., 2011; Haddow and Koehn, 2012).

The vast majority of SMT research, including the above described work on error analysis, is evaluated on data containing *formal* language. Work on SMT of *informal* text mostly targets reduction of OOV words in the source text, for example by correcting spelling errors (Bertoldi et al., 2010), normalizing noisy text to more formal text (Banerjee et al., 2012; Ling et al., 2013a), or enhancing the training data with bilingual segments extracted from Twitter (Jehl et al., 2012; Ling et al., 2013b). Other work improves SMT of UG text by combining statistical and rule-based MT (Carrera et al., 2009), or models trained on formal and informal data (Banerjee et al., 2011). Finally, Roturier and Bensadoun (2011) conduct a comparative study to determine the ability of several SMT systems to translate UG text, but they do not examine what errors the systems make. To our knowledge, our work is the first that looks inside an SMT system to systematically inspect its behavior across a diverse spectrum of UG text types.

3 Experimental setup

We perform our error analysis on two language pairs, Arabic-English and Chinese-English.

3.1 Evaluation sets

For both language pairs we use evaluation sets for five types of user-generated text: SMS messages, chat messages, manual transcripts of phone conversations (called Conversational Telephone

	Dev set		Test set		
Genre	Lines	Tokens	Lines	Tokens	Refs
SMS	2.7K	23.3K	7.6K	44.9K	1
Chat	3.5K	22.5K	7.1K	44.5K	1
CTS	2.4K	23.1K	3.6K	40.6K	1
Comments	1.1K	25.8K	1.7K	45.5K	1
Weblogs	0.8K	14.6K	1.3K	39.9K	4
News 1	1.0K	26.9K	1.6K	46.3K	1
News 2	1.0K	34.4K	1.4K	46.6K	4

Table 1: Statistics of the Arabic-English UG (top) and contrastive news (bottom) evaluation sets. Tokens are counted on the Arabic side.

	Dev set		Test set		
Genre	Lines	Tokens	Lines	Tokens	Refs
SMS	1.8K	15.3K	4.2K	36.3K	1
Chat	4.0K	25.6K	6.0K	45.7K	1
CTS	2.2K	25.1K	2.9K	44.8K	1
Comments	1.0K	26.5K	1.5K	41.0K	1
Weblogs	0.5K	8.8K	0.7K	14.4K	4
News 1	0.8K	24.5K	1.5K	41.9K	1
News 2	1.2K	29.4K	0.7K	17.7K	4

Table 2: Statistics of the Chinese-English UG (top) and contrastive news (bottom) evaluation sets. Tokens are counted on the Chinese side.

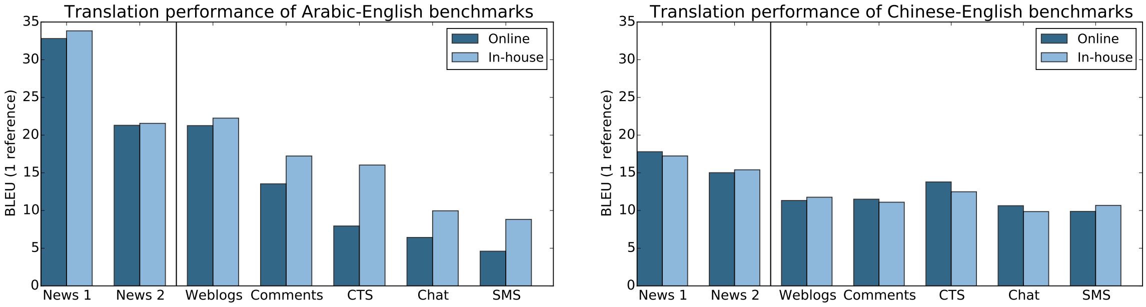

Figure 2: Translation performance of baseline experiments for various Arabic-English (left) and Chinese-English (right) data sets, measured in case-insensitive BLEU for one reference translation.

Speech (CTS)), weblogs, and readers' comments to news articles. The first four data sets originate from BOLT and NIST OpenMT, and are distributed by the Linguistic Data Consortium (LDC), while the last data set is crawled from the web. All UG experiments are contrasted with two news data sets; the news portions of NIST evaluation sets, and web-crawled news articles.

For Arabic-English, the web-crawled news articles and comments originate from the Gen&Topic data set (van der Wees et al., 2015), in which both genres cover the same distributions over various topics. Consequently, any observed differences between the news and UG portions of this data set can be entirely attributed to genre differences and not to potential topical variation.

We have created similar-sized benchmark sets as much as possible, however sometimes limited by availability. Tables 1 and 2 show the data specifications of the Arabic-English and Chinese-English evaluation sets, respectively.[2]

3.2 SMT systems

All experiments presented in this paper are performed with our in-house state-of-the-art system based on phrase-based SMT and similar to Moses (Koehn et al., 2007). Our Arabic-English system is built from 1.75M lines (52.9M source tokens) of parallel text, and our Chinese-English system from 3.13M lines (55.4M source tokens) of parallel text. We tokenize all Arabic data using MADA (Habash and Rambow, 2005), ATB scheme, and we segment the Chinese data following Tseng et al. (2005). Both systems use an adapted 5-gram English language model that linearly interpolates different English Gigaword subcorpora with the English side of our bitexts, containing both news and UG data.

While parallel data is scarce in general, the situation is much worse for UG data, where there are hardly any sizable parallel corpora for any language pair. As a consequence, the training data of both systems comprises 70-75% news data, mostly LDC-distributed, and 25-30% data in various other genres (weblogs, comments, editorials, speech transcripts, and small amounts of chat data), mostly harvested from the web. Per language pair, all experiments use the same SMT models, but we tune parameters separately for each benchmark set using pairwise ranking optimization (PRO) (Hopkins and May, 2011).

To put the results of our system into perspective, we also run a first series of experiments on a well-known and established online SMT system.

4 Error analysis and results

We perform four series of experiments, each with the goal of answering different questions about SMT for UG text:

1. How large is the gap in translation quality between news and different types of UG data? (§4.1). To answer this question, we measure the BLEU score of two state-of-the-art SMT system outputs on all our data sets.

2. What kind of translation choices does the SMT system make for UG data? To answer this question, we measure phrase lengths used during the translation (or decoding) process (§4.2).

3. What translation choices could have been made by the SMT system? To answer this question, we compute mono- and bilingual coverage of the SMT models (§4.3).

[2] Note that two evaluation sets contain four reference translations instead of one. To allow for fair comparison, we average the scores of the four references in all our analyses.

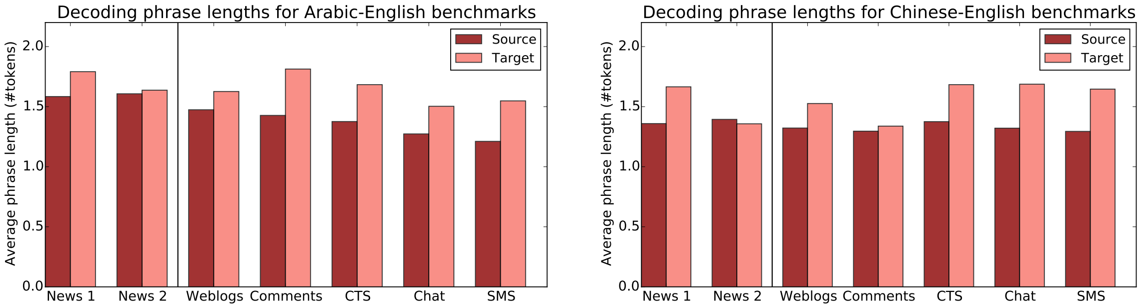

Figure 3: Average source-side and target-side phrase lengths used during decoding.

4. Why did the SMT system make the translation choices that it made? What errors are observed for each benchmark, and how often? To answer these questions, we reimplement the word-alignment driven error analysis approach by Irvine et al. (2013) and perform a qualitative analysis on the results (§4.4).

4.1 Overall translation quality

A first important indication of SMT quality across different genres can be given by translation quality measures that are based on the similarity between the SMT output and a reference human translation. To estimate the gap in translation quality between news and UG text, but also among various types of UG text, we measure the BLEU scores (1 reference) of our in-house SMT system and that of the online system on all our evaluation sets.

The results in Figure 2 (left) show that translation quality differs greatly between the Arabic-English data sets. In particular, the News 1 data set (from NIST) yields considerably higher BLEU scores than all other evaluation sets, including the News 2 (web-crawled) set, which represents the same genre but is visibly more difficult to translate. On the other end of the spectrum, we see that translation quality of the SMS and chat data sets is very poor. Note that our in-house system is optimized per genre, whereas the online system is optimized for general language and speed.

For Chinese-English (Figure 2, right) the differences in BLEU are less pronounced, both across the different data sets and between the two SMT systems. Still, translation quality is worse for the UG data sets than for news, indicating that also for this language pair translating UG text is more challenging than translating news.

As all subsequent analyses require system-internal information, we carry out the experiments with our in-house system only.

4.2 Translation phrase length analysis

Most state-of-the-art SMT systems, including our in-house system, are phrase-based, with translations being generated phrase by phrase rather than word by word (Koehn et al., 2003). An abundant use of small phrases during decoding indicates that the system is not taking advantage of the model's ability to memorize large contextual and possibly non-compositional translation blocks. It is therefore interesting to measure the average phrase length (i.e., number of tokens) used by the system, for the source as well as the target language (Figure 3). For Arabic-English we see that source-side phrases are noticeably longer for both news benchmarks than for the UG data sets. The average target-side phrase length, on the other hand, shows less correlation with the genres of the data sets. Similar trends are observed for Chinese-English, however differences are less extreme.

In general, SMT systems incur higher model costs when utilizing many small phrases rather than few large phrases. If, in spite of that, a system selects many short phrases, which is the case for most of our UG benchmarks, this can be due to (i) unreliable translation probabilities or (ii) to the mere lack of correct translation options in the models. We investigate both issues in the following analyses.

4.3 Model coverage analysis

Next, we examine the translation model coverage for each data set, which tells us what phrases the system *could have* used for decoding. For each of our test sets, we create automatic word alignments using GIZA++ (Och and Ney, 2003), and extract from these the set of all reference phrase pairs using Moses' phrase extraction algorithm (Koehn et al., 2007). By comparing this set of phrase pairs to the available phrases in the SMT models, which

			Source phrase recall				Target phrase recall				Phrase pair recall			
Genre	BLEU	LM PP	1	2	3	4	1	2	3	4	1	2	3	4
News 1	33.8	65	99.7	88.9	56.3	26.1	99.7	91.1	61.5	29.6	84.9	54.4	23.6	8.1
News 2	21.5	86	99.6	88.1	53.7	21.8	99.5	88.1	53.4	23.6	77.4	46.9	18.8	5.9
Weblogs	22.3	152	99.2	80.5	40.6	13.5	99.5	86.3	48.9	17.8	78.4	41.5	12.9	2.9
Comments	17.2	117	97.7	80.2	43.0	15.3	99.7	89.8	55.3	21.9	59.1	33.2	11.1	2.8
CTS	16.0	103	97.4	66.3	25.1	6.4	99.8	90.8	54.3	21.5	66.7	25.7	6.1	1.0
Chat	10.0	179	94.1	56.0	19.4	4.7	98.6	86.1	47.3	16.7	60.8	21.3	4.5	0.8
SMS	8.8	196	93.7	57.8	17.5	3.3	99.1	86.3	47.0	14.6	62.0	21.1	3.7	0.4

Table 3: Target language model perplexity and translation model coverage of Arabic-English benchmarks. Phrase pair recall values are broken down by source phrase length. Intensities of the cell colors indicate relative recall values with respect to the best scoring benchmark (measured in BLEU).

			Source phrase recall				Target phrase recall				Phrase pair recall			
Genre	BLEU	LM PP	1	2	3	4	1	2	3	4	1	2	3	4
News 1	17.2	121	99.0	80.2	40.8	16.2	99.5	84.9	48.0	19.5	69.1	34.8	10.8	3.3
News 2	15.4	118	98.8	84.2	44.3	16.0	99.4	83.8	44.2	14.7	63.1	32.4	10.7	3.3
Weblogs	11.8	153	98.6	76.6	33.8	11.1	99.3	81.6	40.8	12.4	59.0	27.0	7.3	1.7
Comments	11.1	195	98.7	78.3	35.2	8.7	97.9	77.9	35.1	10.2	53.5	21.6	5.0	1.0
CTS	12.5	135	98.7	80.7	40.1	10.5	99.8	86.3	47.4	16.4	70.0	33.5	9.3	1.7
Chat	9.9	221	98.0	71.9	27.5	6.1	99.4	82.6	43.2	13.0	62.3	24.8	5.4	0.6
SMS	10.7	234	97.3	68.5	24.9	4.8	99.0	80.4	40.5	12.5	62.6	24.6	5.1	0.5

Table 4: Target language model perplexity and translation model coverage of Chinese-English benchmarks. See Table 3 for explanation on colors and categories.

have been extracted using the same procedure, we can compute the following statistics:

1. *Source phrase recall*, defined as the fraction of reference phrase pairs whose *source* side is found in the SMT models.
2. *Target phrase recall*, defined as the fraction of reference phrase pairs whose *target* side is found in the SMT models.
3. *Phrase pair recall*, defined as the fraction of reference phrase pairs whose source and target side are jointly found in the SMT models.

Low recall values indicate that the models lack phrases or phrase pairs that match the test data, which can be addressed by adding additional relevant training data or by generating new phrases. In addition, we measure language model perplexity as an indication of how predictable each benchmark is for the language model. Note that high perplexity corresponds to lower coverage.

The model coverage results for Arabic-English and Chinese-English are shown in Tables 3 and 4, respectively. All recall scores are broken down by phrase length, up to phrases of four tokens.[3] We use cell color intensity to represent relative recall values with respect to the best scoring benchmark according to BLEU, i.e., News 1. The results show that source phrase recall is substantially lower for the UG benchmarks than for news, particularly for longer phrases. Regarding target phrase recall, differences between various data sets and genres are much smaller. This suggests that many of the reference phrases could potentially be generated by the system, even for the UG data. However, to be able to output the available target phrases, the system needs a match with the input source phrases, which is exactly what is being measured with phrase pair recall. Here, we see that for the majority of single-word source phrases, the expected target phrase is accessible by the system. For longer phrases, though, there is again a drastic decline in recall, with almost no phrases of length 4 or longer having the expected target covered by the models. Similar to source phrase recall, this decline is notably bigger for UG than for news.

[3] The source-target phrase pair recall (last four columns) is split by source phrase length rather than target phrase length since source phrases are the actual input to the SMT system.

Looking at the differences between the various types of UG data, we see that the SMS and chat benchmarks are most severely affected by overall poor model coverage. As for weblogs, the target phrase recall is similar to SMS and chat, whereas both source phrase and phrase pair recall are much higher. For CTS and web comments, there are notable differences between model coverage for the two language pairs, despite similar BLEU scores. While comments have better coverage in the Arabic-English models, CTS has higher recall values for Chinese-English.

Finally, we see that language model perplexity is on average lower for Arabic-English than for the Chinese-English benchmarks. This is somewhat surprising given that perplexity is measured on the English side, but it can partially explain the low BLEU scores on, for example, the Chinese-English News 1 benchmark. All news benchmarks have relatively low perplexities, which is expected since the language model covers more news than UG data. Of the UG benchmarks, CTS has a remarkably low perplexity value, suggesting that for this genre the language model can potentially compensate for low translation model coverage.

4.4 WADE: Word Alignment Driven Evaluation

Next, to gain a more fine-grained insight in *why* our SMT system makes its translation choices, we reimplement an evaluation approach proposed by Irvine et al. (2013), which analyzes SMT error types at the word alignment level. The analysis exploits automatic word alignments between (i) a given source sentence and its reference translation, and (ii) the same source sentence and its automatic translation. Each aligned source-reference word pair is examined for whether the alignment link is matched by the decoder. Formally, f_i is a foreign word, e_j is a reference word aligned to f_i, $a_{i,j}$ is the alignment link between f_i and e_j, and H_i is the set of output words that are aligned to f_i by the decoder. If $e_j \in H_i$, the alignment link $a_{i,j}$ is marked as correct. Otherwise, $a_{i,j}$ is categorized with one of the following error types:

1. A SEEN error indicates an unseen source word, i.e., out-of-vocabulary (OOV) item. This error is assigned to $a_{i,j}$ if f_i does not appear in the phrase table used for translation. This type of error inversely correlates with length-1 source phrase recall (§4.3).
2. A SENSE error indicates an unseen target word. This error is assigned to $a_{i,j}$ if f_i does appear in the phrase table but never with translation candidate e_j.
3. A SCORE error indicates suboptimal scoring of translation options. This error is assigned to $a_{i,j}$ if f_i exists in the phrase table with translation candidate e_j, but another translation candidate is preferred by the decoder.

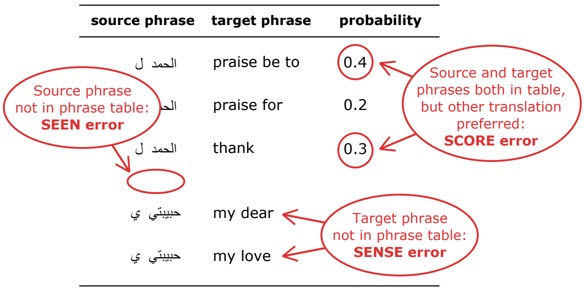
source phrase
target phrase
probability
الحمد ل
praise be to
0.4
الحمد
praise for
0.2
الحمد ل
thank
0.3
حبيبتي ي
my dear
حبيبتي ي
my love
Source phrase not in phrase table: SEEN error
Source and target phrases both in table, but other translation preferred: SCORE error
Target phrase not in phrase table: SENSE error

Figure 4: Graphical overview of SEEN, SENSE and SCORE errors in a toy phrase table.

Figure 4 shows a graphical representation of these error types and their 'location' in the phrase table. In addition to the listed error types, Irvine et al. define SEARCH errors as errors due to pruning in beam search, and refer to the complete set of errors as the S^4 *taxonomy*. For this analysis, however, SEARCH errors are indistinguishable from SCORE errors, and are therefore never assigned.

A final category that can be considered are *freebies*: OOVs that are copied over verbatim to the output sentence and accidentally match the reference translation (e.g., urls, proper nouns, etc.). For the language pairs that we study, they are very rare; at most 0.35% for Arabic-English (in CTS) and 0.63% for Chinese-English (in SMS). Manual inspection reveals that nearly all freebies are English words in the foreign source text. Since they are so rare, we omit freebies from our results.

As WADE errors are assigned at the fine-grained level of individual words, this analysis allows for (i) sentence-level visualization of errors, and (ii) collecting aggregate statistics of each error type for an entire evaluation set. By assembling the latter for various benchmarks, we can quantify global differences between genres or data sets. At the same time, by examining (i) we can gain insight in the nature of the different 'errors', which might be real mistakes, or, for instance, different lexical choices.

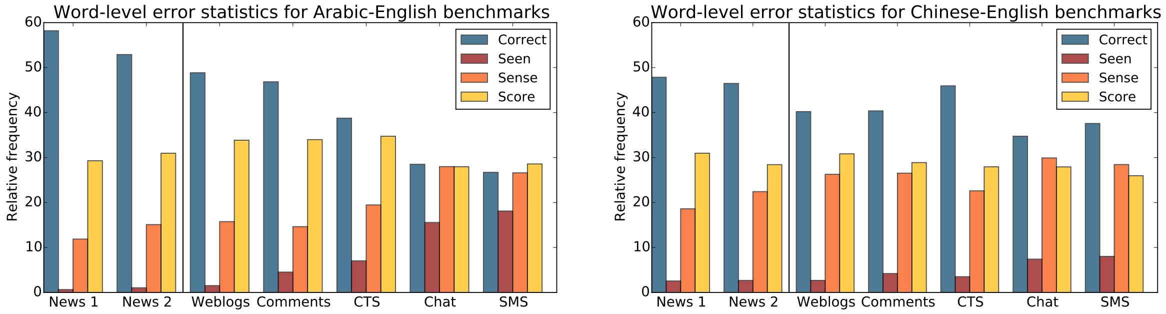

Figure 5: Aggregate error statistics for Arabic-English (left) and Chinese-English (right) benchmark sets.

Quantitative results. The aggregate error statistics for each data set are shown in Figure 5. To put our results into perspective, we recall the findings of Irvine et al. (2013). They find that for *formal* domains using a French-English system, 50–60% of the alignment links are correct, and SCORE errors are more common than SENSE errors, which in turn are more common than SEEN errors. While we observe a similar distribution for our Arabic-English news benchmarks, these numbers do not generalize to the Arabic-English UG benchmarks nor to any of the Chinese-English data sets.

First, the portion of SEEN errors increases dramatically for the Arabic-English UG translation tasks. For Chinese-English this trend is less pronounced yet also clearly observable. Next, SENSE errors also increase substantially for most of the UG data, making up the majority of the errors for Chinese-English SMS and chat. This indicates that a promising strategy for adapting SMT systems to translating UG data involves generating new target-side translation candidates that match the source phrases in the input sentences. Finally, we evaluate the fraction of SCORE errors. While this is the most commonly observed error type in most of the data sets, there seems to be very little correspondance with the genre or BLEU scores of the benchmarks. This is an interesting finding since most work in system adaptation for SMT focuses on better scoring of existing translation candidates (Matsoukas et al., 2009; Foster et al., 2010; Axelrod et al., 2011; Chen et al., 2013, among others). However, for UG translation tasks this does not appear as the most profitable approach.

Qualitative results. The generated sentence-level error annotations allow us to examine the various error types in detail. The first phenomenon that we repeatedly observe in the UG data are SEEN errors due to misspellings or, in the case of Arabic, dialectal forms. Two such examples are shown in Figures 6A and 6B: In the first, the SMT system does not recognize the dialectal form of verb negation 'mtzEl$', which is a morphologically complex word containing both a prefix and a suffix. In the second, the input word 'Almwb**b**Ayl' ('mobile') is wrongly spelled 'Almwy**y**Ayl'. It is interesting to note that 'b' and 'y' are very similar in the Arabic script. This type of errors is particularly frequent in chat and SMS, which can partly explain the different distribution of errors across the Arabic-English data sets (Figure 5).

Also frequently observed in the UG data are SMT lexical choices that are more formal than the reference translations. This is not surprising given the large amount of formal data in the SMT models, but it does illustrate the need for adaptation to UG data. Often, the optimal lexical choice is simply absent from the SMT models, resulting in SENSE errors. This can be observed in Figure 6A, where 'sons' is output instead 'kids', and in Figure 6C, where 'i understand' is output instead of the colloquial 'i got it'. In other situations, the annotated SCORE errors indicate that the correct choice was available to the SMT system without being selected for translation. For example in Figure 6D, the output 'my parents' is preferred to the more colloquial 'mom and dad' in the reference.

Another phenomenon, particularly common for Chinese-English UG translations, is that idioms are translated in small chunks, thereby losing their meaning as a phrase. In Figure 6D, the characters '说', '一', and '声' mean 'to say', 'one', and 'sound', respectively. The phrase '说一声' as a whole means 'talk a bit about something' but is not covered by the SMT models. Similarly, '你路上慢点' in Figure 6E literally means 'you on the road slow a bit', which, if covered by the models, could have been translated into 'be careful on

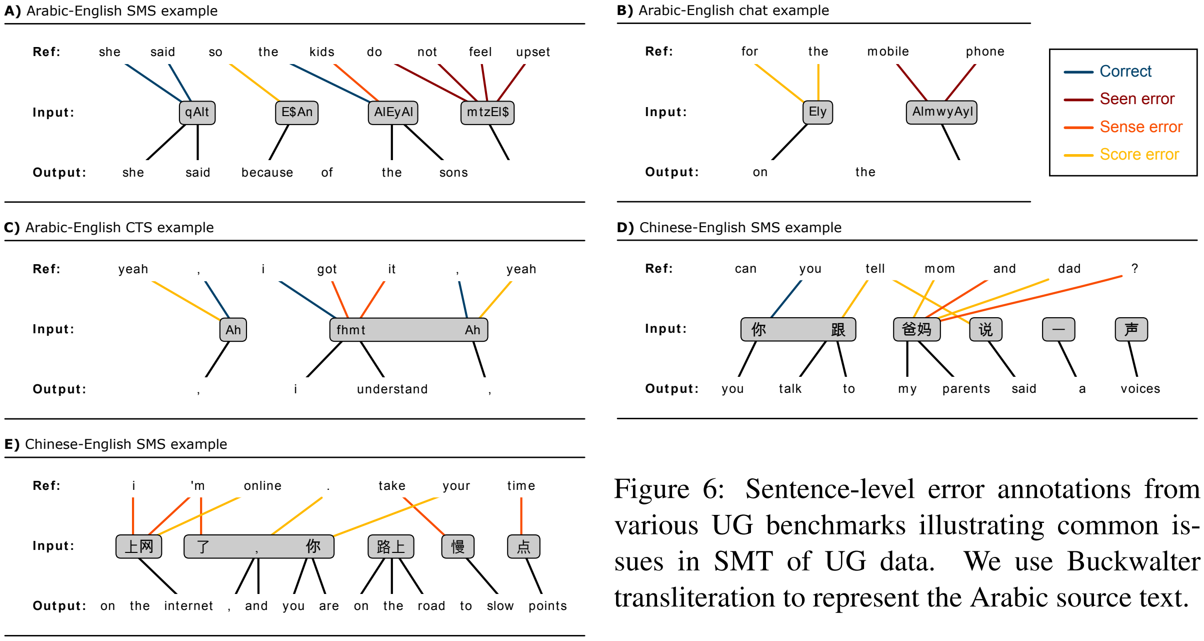
A) Arabic-English SMS example
Ref: she said so the kids do not feel upset
Input: qAlt E$An AlEyAl mtzEl$
Output: she said because of the sons
B) Arabic-English chat example
Ref: for the mobile phone
Input: Ely AlmwyAl
Output: on the
Correct
Seen error
Sense error
Score error
C) Arabic-English CTS example
Ref: yeah , i got it , yeah
Input: Ah fhmt Ah
Output: , i understand ,
D) Chinese-English SMS example
Ref: can you tell mom and dad ?
Input: 你 跟 爸妈 说 一 声
Output: you talk to my parents said a voices
E) Chinese-English SMS example
Ref: i 'm online . take your time
Input: 上网 了 , 你 路上 慢 点
Output: on the internet , and you are on the road to slow points

Figure 6: Sentence-level error annotations from various UG benchmarks illustrating common issues in SMT of UG data. We use Buckwalter transliteration to represent the Arabic source text.

your way' or 'take your time'. These examples illustrate that the low phrase pair recall for longer phrases severely complicates SMT of UG data.

A final recurring issue in SMS and chat messages is the omission of first person pronouns, see for example Figure 6E. The Chinese source phrase '上网了' literally means 'get online' (+ auxiliary word marking past tense). A native speaker understands that this concerns the sender, which is reflected by a first person pronoun in the reference. The SMT system, on the other hand, cannot infer the subject of this phrase and instead generates a translation without pronouns.

Other, less common, types of errors occurring in the UG data are due to inconsistent segmentation or tokenization of input text, which mostly affects rare words, emoticons, and repeating punctation. Finally, SEEN errors for named entities are overall rare but occur in both news and UG benchmarks.

5 Conclusions and future directions

Translating user-generated (UG) text is a difficult task for SMT. To explain the poor translation quality observed for UG data, we have performed a detailed error analysis on two language pairs (Arabic-English and Chinese-English) and five different types of UG data (SMS, chat, CTS, weblogs, and comments). Our quantitative results show among others that (i) UG data is translated with shorter source phrases than news, (ii) UG translation model coverage deteriorates substantially for longer phrases, and (iii) phrase-pair OOVs pose a bigger challenge to UG translation tasks than source OOVs. In our qualitative analysis we found that common issues in UG data include (i) OOVs due to misspellings or Arabic dialectal forms, (ii) lexical choices that do not reflect colloquial formulations, (iii) phrasal idioms being translated word by word, and (iv) omitted first person pronouns in SMS and chat.

Finally, different types of UG exhibit dissimilar error distributions, demanding diverse strategies to improve SMT quality. For example, SMS and chat data might benefit from text normalization (Bertoldi et al., 2010; Yvon, 2010; Ling et al., 2013a) or otherwise resolving source OOVs, which also has been the main focus of previous work on SMT for UG. On the other hand, while research in domain adaptation for SMT often aims at better scoring of existing translation candidates, we have shown that for many UG tasks the most promising direction involves increasing phrase pair recall of the SMT models (i.e., reducing phrase pair OOVs), for example by paraphrasing (Callison-Burch et al., 2006) or translation synthesis (Irvine and Callison-Burch, 2014).

Acknowledgments

This research was funded in part by the Netherlands Organization for Scientific Research (NWO) under project number 639.022.213. We thank Rachel Cotterill, Nigel Dewdney, and the anonymous reviewers for their valuable comments.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation*, pages 65–72.

Pratyush Banerjee, Sudip Kumar Naskar, Johann Roturier, Andy Way, and Josef van Genabith. 2011. Domain adaptation in statistical machine translation of user-forum data using component level mixture modelling. In *Proceedings of the XIII Machine Translation Summit*, pages 285–292.

Pratyush Banerjee, Sudip Kumar Naskar, Johann Roturier, Andy Way, and Josef van Genabith. 2012. Domain adaptation in SMT of user-generated forum content guided by OOV word reduction: Normalization and/or supplementary data. In *Proceedings of the 16th Conference of the European Association for Machine Translation*, pages 169–176.

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2010. Statistical machine translation of texts with misspelled words. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 412–419.

Arianna Bisazza, Nick Ruiz, and Marcello Federico. 2011. Fill-up versus interpolation methods for phrase-based SMT adaptation. In *Proceedings of the 8th International Workshop on Spoken Language Translation*, pages 136–143.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24.

Jordi Carrera, Olga Beregovaya, and Alex Yanishevsky. 2009. Machine translation for cross-language social media.

Boxing Chen, Roland Kuhn, and George Foster. 2013. Vector space model for adaptation in statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1285–1293.

Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Analysis of translation model adaptation in statistical machine translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation (IWSLT 2010)*, pages 243–250.

George Foster, Cyril Goutte, and Roland Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 451–459.

Nizar Habash and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 573–580.

Barry Haddow and Philipp Koehn. 2012. Analysing the effect of out-of-domain data on SMT systems. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 422–432.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1352–1362. Association for Computational Linguistics.

Ann Irvine and Chris Callison-Burch. 2014. Hallucinating phrase translations for low resource MT. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 160–170.

Ann Irvine, John Morgan, Marine Carpuat, Hal Daumé III, and Dragos Stefan Munteanu. 2013. Measuring machine translation errors in new domains. *Transactions of the Association for Computational Linguistics*, 1:429–440.

Laura Jehl, Felix Hieber, and Stefan Riezler. 2012. Twitter translation using translation-based cross-lingual retrieval. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 410–421.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180.

Wang Ling, Chris Dyer, Alan W Black, and Isabel Trancoso. 2013a. Paraphrasing 4 microblog normalization. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 73–84.

Wang Ling, Guang Xiang, Chris Dyer, Alan Black, and Isabel Trancoso. 2013b. Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 176–186.

Spyros Matsoukas, Antti-Veikko I. Rosti, and Bing Zhang. 2009. Discriminative corpus weight estimation for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 708–717.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović and Hermann Ney. 2011. Towards automatic error analysis of machine translation output. *Computational Linguistics*, 37(4):657–688.

Johann Roturier and Anthony Bensadoun. 2011. Evaluation of MT systems to translate user generated content. In *Proceedings of the XIII Machine Translation Summit*, pages 244–251.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the Seventh Conference of the Association for Machine Translation in the Americas*, pages 223–231.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*, volume 171, pages 168–171.

Marlies van der Wees, Arianna Bisazza, Wouter Weerkamp, and Christof Monz. 2015. What's in a domain? Analyzing genre and topic differences in statistical machine translation. In *Proceedings of the Joint Conference of the 53th Annual Meeting of the ACL and the 7th International Joint Conference on Natural Language Processing of the AFNLP*.

François Yvon. 2010. Rewriting the orthography of SMS messages. *Natural Language Engineering*, 16(2):133–159.