Identifying Fluently Inadequate Output in Neural and Statistical Machine Translation

Marianna J. Martindale† **Marine Carpuat**‡ **Kevin Duh**° **Paul McNamee**°
†iSchool, ‡Dept. of Computer Science, University of Maryland, College Park, USA
°HLTCOE, Johns Hopkins University, Baltimore, USA
mmartind@umiacs.umd.edu, marine@cs.umd.edu
kevinduh@cs.jhu.edu, mcnamee@jhu.edu

Abstract

With the impressive fluency of modern machine translation output, systems may produce output that is fluent but not adequate (*fluently inadequate*). We seek to identify these errors and quantify their frequency in MT output of varying quality. To that end, we introduce a method for automatically predicting whether translated segments are fluently inadequate by predicting fluency using grammaticality scores and predicting adequacy by augmenting sentence BLEU with a novel Bag-of-Vectors Sentence Similarity (BVSS). We then apply this technique to analyze the outputs of statistical and neural systems for six language pairs with different levels of translation quality. We find that neural models are consistently more prone to this type of error than traditional statistical models. However, improving the overall quality of the MT system such as through domain adaptation reduces these errors.

1 Introduction

Recent work has shown that well-trained, in-domain neural machine translation (NMT) systems can produce translations that, at the sentence level, are rated on par with human reference translations (Hassan Awadalla et al., 2018). Part of this success comes from the impressive improvements in fluency of NMT output compared to previous MT paradigms (Bentivogli et al., 2016; Toral and Sánchez-Cartagena, 2017; Koehn and Knowles, 2017). However, NMT has also been shown to sometimes produce output that is low adequacy and even unrelated to the input–particularly when not trained on sufficient in-domain data (Koehn and Knowles, 2017). Because of NMT's uncanny ability to produce fluent output, these translations may not just be inadequate but *fluently inadequate*. The fluency of *fluently inadequate* translations may mislead users into trusting the content based on fluency alone–particularly in the context of other fluent and adequate translations (Martindale and Carpuat, 2018).

Mitigating the effects of fluently inadequate translations first requires understanding the scale of the problem and what situations are likely to generate these errors. The general success and high system level quality of NMT suggests that fluently inadequate translations are rare, but we cannot say how rare without a means of automatically identifying *potentially* fluently inadequate translations in large collections of MT output.

In this work, we propose a method to automatically detect fluently inadequate translations based on the underlying characteristics of fluency and adequacy. We view fluently inadequate translations as translations that are fluent, well-formed sentences that could have been written by a human, and that do not preserve the meaning of the reference. In practice, given a reference translation r and MT hypothesis h, we consider h to be fluently inadequate if $fluency(h) > \tau_f$ and $adequacy(h, r) < \tau_a$, where τ_a and τ_f are minimum fluency and adequacy thresholds respectively. We define novel fluency and adequacy metrics for this purpose, building on prior work on grammaticality detection and comparisons of multisets applied to word embeddings (Section 2).

We conduct two sets of experiments. First, we

evaluate the fluency and adequacy metrics, establishing that they can be used for the task of detecting fluently inadequate translations, and set thresholds τ_a and τ_f empirically in Sections 3.1 and 3.2. We then conduct an automatic analysis to assess how frequent these errors are in neural and statistical machine translation (SMT) systems for a variety of languages and varying levels of model quality and train/test domain match. We find that fluently inadequate translations are more common in NMT overall, especially when there is less training data and when there is a mismatch between training and test data.

2 Approach

2.1 Predicting Fluency

We propose to score fluency using metrics introduced for the related task of detecting grammaticality, which scores the well-formedness of a sentence. Lau et al (2016) take an unsupervised, language modeling approach to predicting grammaticality. Based on the intuition that well-formedness errors will be caused by one or more incorrect or out of place words, they introduce scores based not only on sentence probability, but also on scores that focus on lowest word probabilities in a segment. Specifically, given a 5-gram language model, the following scores are computed:

$$Mean\ LP = \frac{\sum_{n=1:N} \log p_5(w_n|w_{n-1}, ...)}{N} \quad (1)$$

$$Norm\ LP = \frac{\sum_{n=1:N} \log p_5(w_n|w_{n-1}, ...)}{\sum_{n=1:N} \log p_1(w_n)} \quad (2)$$

$$Word\ LPmin_n = min_n \left\{ -\frac{\log p_5(w)}{\log p_1(w)} \right\} \quad (3)$$

$$Word\ LP_{n\%} = \frac{\sum_{w \in LP_{n\%}} -\frac{\log p_5(w)}{\log p_1(w)}}{|LP_{n\%}|} \quad (4)$$

$$Word\ LP_{mean} = \frac{\sum_{n=1:N} -\frac{\log p_5(w_n)}{\log p_1(w_n)}}{N} \quad (5)$$

where $logp_5$ is the 5-gram log probability, w_n is the n^{th} word and N is the number of words in the sentence. $Mean\ LP$ is the sentence n-gram log probability, normalized by length. $NormLP$ is the sentence n-gram log probability, normalized by sentence unigram log probability. The other metrics are focused on probability of individual words given the preceding words. Each word's 5-gram log probability is normalized by its unigram probability ($logp_1$), $Word\ LPmin_n$ is the n^{th} lowest normalized word probability, $LP_{n\%}$ is the lowest $n\%$ normalized word probabilities. Because there may be outliers that score artificially low, we introduce an additional variant, $WordLP_{mid}$, which uses LP_{mid}, the middle 50% of the normalized word probabilities:

$$Word\ LP_{mid} = \frac{\sum_{w \in LP_{mid}} -\frac{\log p_5(w)}{\log p_1(w)}}{|LP_{mid}|} \quad (6)$$

We expect that fluently inadequate output is being influenced by the training data more than the input text, so we build our language model based on the target side of the system training data rather than a large generic language model.

2.2 Predicting Adequacy

BLEU (Papineni et al., 2002) is a widely accepted baseline measure of MT quality at the system level and, as such, is an obvious choice for a baseline adequacy metric. However, it may not be well suited for this task. Segments with high BLEU scores more closely match the reference, indicating high adequacy, but translations that receive a lower BLEU score may be inadequate or they may be adequate with different word choice. For the purpose of detecting fluently inadequate translations, we can be confident that a segment with a high BLEU score is adequate, but low BLEU scores do not necessarily imply low adequacy.

To account for cases where a translation may be adequate but receive a low BLEU score, we need an adequacy metric that will be less affected by word choice. This suggests the need for comparing semantic representations rather than matching strings. For our baseline vector-based metric, we use the common, simple approach of comparing sentence embeddings generated by averaging the word embeddings for each word in the sentence. However, this approach does not directly compare any of the word vectors, only their sum, and there are many unrelated sentences that could produce the same sentence vector. We introduce an alternative word embedding based measure of sentence similarity that overcomes this flaw

to produce more a reliable adequacy metric, bag-of-vectors sentence similarity (BVSS).

BVSS Metric BVSS is an application of Saga (Similarity AGregation Application) introduced by Knox (2015). The Saga approach frames a task as an information similarity problem. Given a measure of information I for a multiset, the similarity between two multisets, X and Y, is the proportion of information from the union of X and Y that is found in both X and Y:

$$S(X,Y) = \frac{I(X) + I(Y) - I(X \cup Y)}{I(X \cup Y)} \quad (7)$$

The information measure in Saga uses single-linkage agglomerative clustering (Florek et al., 1951). If the items in a multiset are clustered according to similarity, more clusters indicate more disparate items and, therefore, more information. When we compare two multisets of items, X and Y, we first cluster each multiset separately to get $I(X)$ and $I(Y)$. We then pool all of the items and cluster again to get $I(X \cup Y)$. If the items in X are similar to the items in Y, those items will cluster together yielding fewer clusters than if they were different.

A nice feature of this approach is that in addition to the undirected similarity, we can modify Equation 7 to a directed form. The directed similarity to X of Y would be given by the proportion of information in Y that also appears in X:

$$S_X(Y) = \frac{I(X) + I(Y) - I(X \cup Y)}{I(Y)} \quad (8)$$

To compare sentences with this approach, we treat a sentence as a multiset of words and determine the similarity of words using the cosine similarity of their embeddings. Replacing X and Y in equation 7 with S for MT system output and R for reference gives us the *BVSS* metric:

$$BVSS(S,R) = \frac{I(S) + I(R) - I(S \cup R)}{I(S \cup R)} \quad (9)$$

The directed form provides a way to measure when information is lost (i.e., the reference has more information than the MT output) or hallucinated (the MT output has more information than the reference). We will use *BVSS-reference* and *BVSS-system* to refer to these directed similarities. *BVSS-reference* is the proportion of the information in the reference that is also in the MT output and *BVSS-system* is the proportion of the information in the MT output that is also in the reference:

$$BVSS_{reference} = \frac{I(S) + I(R) - I(S \cup R)}{I(R)} \quad (10)$$

$$BVSS_{system} = \frac{I(R) + I(S) - I(R \cup S)}{I(S)} \quad (11)$$

3 Detection Method Evaluation

Since there is no existing dataset with manual annotation of fluently inadequate translations, we first evaluate our fluency and adequacy prediction approaches comparing against direct assessment scores from WMT16 (Bojar et al., 2016) as 2016 was the only year in which human fluency judgments were collected. We then use our automated fluency scores on reference translations and automated adequacy scores on synthetic low adequacy "translations" to determine thresholds for high fluency and dubious adequacy.

3.1 Fluency Experiments

Task For WMT16, fluency judgments were collected for Czech-English (CS-EN), German-English (DE-EN), Finnish-English (FI-EN), Romanian-English (RO-EN), Russian-English (RU-EN), and Turkish-English (TR-EN) in the news shared task. Annotations were collected with the goal of system-level reliability, so many segments only have one judgment. To improve reliability, we use only segments where there are two or more judgments.

Model setup Fluency scores are based on a 5-gram language model. We built a 5-gram KenLM (Heafield, 2011; Heafield et al., 2013) language model using the monolingual news training data from WMT16.

Results For each of the metrics described in section 2.1, we calculated the Pearson correlation with the direct assessment scores for each of the language pair data sets and for all the data combined. Results are shown in Table 1. Although these correlations are lower than we would like, we find that for all language pairs and for the combined data, $WordLP_{mid}$ yields the highest correlation, so we will use this formula for our fluency prediction metric.

Fluency Metric	CS-EN	DE-EN	FI-EN	RO-EN	RU-EN	TR-EN	All
$MeanLP$	0.32619	0.21290	0.27686	0.25831	0.22792	0.32402	0.26974
$NormLP$	0.41271	0.26721	0.25297	0.22797	0.27404	0.2496	0.28037
$WordLPmin_1$	0.04490	0.01192	0.05817	0.02359	0.05289	0.04036	0.03745
$WordLPmin_2$	0.28831	0.23004	0.21382	0.21216	0.24384	0.20712	0.23121
$WordLP_{25\%}$	0.40021	0.25993	0.23916	0.20506	0.28920	0.21564	0.26748
$WordLP_{50\%}$	0.32168	0.26854	0.22640	0.19729	0.25738	0.20799	0.24382
$WordLP_{mean}$	0.38227	0.29371	0.26660	0.22748	0.30028	0.25658	0.28609
$WordLP_{mid}$	**0.42543**	**0.34306**	**0.34907**	**0.31295**	**0.34471**	**0.38615**	**0.35872**

Table 1: Pearson correlation between each of the fluency prediction metrics and the human fluency direct assessment scores for each language and across all languages.

	CS-EN	DE-EN	FI-EN	RO-EN	RU-EN	TR-EN	All
Percent fluent	59.22%	59.70%	56.79%	58.04%	60.80%	48.81%	57.21%
Precision	65.35	63.36	59.62	62.06	66.22	52.37	61.42
Recall	90.97	87.29	91.56	92.20	87.67	87.77	89.38
F1	76.06	73.42	72.21	74.18	75.45	65.60	72.81

Table 2: Precision, recall, and F1 on fluent translations for $WordLP_{mid}$ on system outputs for each language pair and on all system outputs. The percentage of outputs that were labeled fluent based on the human fluency judgments is also provided for reference.

Setting the fluency threshold Because our goal is to correctly label sentences as fluently inadequate rather than to provide an exact score, we must select a fluency threshold τ_f to label a translation as "fluent". To determine this threshold, we computed the $WordLP_{mid}$ scores for the reference translation sentences in the WMT16 news training data. To cover most examples while allowing for variance in human judgments, the threshold is set at the point where 90% of reference segments would be labeled as fluent. Precision, recall, and F1 scores for $WordLP_{mid}$, are shown in Table 2. Across all data sets we see high recall but the precision is not as high. Although this suggests that this metric might overestimate the fluency of translations, we are more concerned with comparing between systems than with the raw scores.

3.2 Adequacy Experiments

Task and Data We assess adequacy metrics using the direct assessment adequacy scores and system outputs for all language pairs from WMT16 (Bojar et al., 2016). Adequacy judgments were collected for all submitted systems in all language pairs in the news shared task. These annotations were used to determine the system rankings in the news task and as gold standard quality judgments for the metrics shared task. For the metrics task, enough annotations were collected for each system-produced segment to establish segment-level reliability, while only enough judgments for system-level reliability were collected for the remainder of the segments for the news task. Because we need segment-level reliability, we use only the metrics subset of the data as gold standard human judgments, and we use the reference translations from the news subset in generating synthetic inadequate examples.

We use the standardized human direct assessment adequacy scores from WMT16 (Bojar et al., 2016) as gold standard in determining how well each adequacy metric correlates with human judgments. However, for binary questionable/acceptable adequacy judgments, we must be sure that the inadequate examples are clearly inadequate regardless of fluency and other MT quirks. The high correlation between human judgments of fluency and adequacy in Callison-Burch et al (2007) and Graham et al (2017) may indicate that human adequacy judgments are influenced by fluency, lowering the adequacy scores of disfluent translations. To ensure that our inadequate examples are truly inadequate, we rely on synthetic examples. We generate synthetic low adequacy translations by randomly selecting pairs of reference translations from the WMT16 news task

Adequacy Metric	CS-EN	DE-EN	FI-EN	RO-EN	RU-EN	TR-EN	All
BLEU	0.54275	0.41975	0.41460	0.48410	0.45093	0.50346	0.46242
Averaged Embeddings	0.43905	0.18998	0.31218	0.36303	0.23545	0.30257	0.29584
BVSS	0.61286	0.47068	**0.51856**	**0.56164**	**0.55478**	**0.58858**	**0.54306**
BVSS-Reference	**0.62178**	**0.47877**	0.49006	0.55619	0.51949	0.56264	0.53643
BVSS-System	0.53773	0.38887	0.45177	0.47698	0.50288	0.53687	0.46925

Table 3: Pearson correlation between each of the adequacy prediction metrics and the human adequacy direct assessment scores for each language and across all languages.

	Prec.	Recall	F1
BLEU	94.33	99.08	96.65
Averaged Embeddings	84.56	**99.15**	91.28
BVSS	**99.39**	99.04	**99.22**
BVSS-Reference	99.00	99.03	99.01
BVSS-System	99.17	99.03	99.10
BLEU+BVSS	**99.61**	**99.81**	**99.71**

Table 4: Precision, recall, and F1 on BLEU, BVSS, BVSS-Reference, BVSS-System, and BLEU with BVSS and BVSS-System on the questionable adequacy test set with thresholds calculated based on predicted adequacy scores for the synthetic low adequacy dev data.

and treating one as synthetic MT output and the other as reference. We split these synthetic examples into dev and test sets. The dev synthetic examples are used in choosing the binary acceptable/questionable adequacy threshold τ_a as described below. The test synthetic examples are used as the questionable adequacy items in our adequacy precision/recall test set, with acceptable adequacy items chosen from actual WMT16 submissions. Because we are looking for extreme inadequacy and the systems in WMT16 were of competitively high quality, we use segments with direct assessment scores in the top 90% as acceptable adequacy in the test set.

Model setup Our vector-based metrics are based on word embeddings. We use the pre-trained aligned Wikipedia fastText word vectors (Joulin et al., 2018; Bojanowski et al., 2017).

Results For each metric defined in Section 2.2, we calculated the Pearson correlation with the direct assessment scores for each of the WMT16 language pair data sets and for all the data sets combined (Table 3). The averaged sentence embeddings had the lowest correlation across all language pairs. BVSS-System performed similarly well compared to BLEU, but BVSS and BVSS-Reference both outperformed BLEU.

Setting the adequacy threshold As with fluency, our goal for the adequacy metric is to correctly label a sentence as questionable adequacy rather than to provide an exact score. We used each candidate adequacy metric described in section 2.2 to score the segments in the synthetic low adequacy dev set, and set adequacy threshold τ_a for each metric such that 99% of dev set examples would be labeled inadequate. The precision, recall, and F1 on the synthetic test set using this threshold for each metric is shown in Table 4. We see that as with correlation scores, the Averaged Embeddings have much lower precision than BLEU or any of the BVSS metrics, and the BVSS metric have higher precision than BLEU.

Because of the potentially complementary differences in BLEU and BVSS, we also tested combinations of BLEU and the highest-performing vector-based metric, BVSS. We combine the metrics by marking a translation as questionable adequacy only if both metrics would label it as questionable. We see a slight improvement in F1 with the combination, and we adopt this metric for labeling segments as questionable adequacy.

3.3 Selected Scoring Method

Based on the fluency and adequacy evaluations in Sections 3.1 and 3.2, we select $WordLP_{mid}$ and the BLEU+BVSS combination to label segment translations as fluently inadequate.

The results on segment level fluency and adequacy prediction tests show that neither metric is perfect at the segment level. However, the impact of segment-level errors is lessened when segment level scores are aggregated to compare across systems.

Data source	Arabic	Chinese	Farsi	German	Korean	Russian
Subtitles	30M	11M	6.2M	22M	1.4M	26M
UN v1	18M	-	-	-	-	-
WMT17	-	25M	-	5.8M	-	25M
LDC	1.3M	-	-	-	-	-
All General	49M	36M	6.2M	28M	1.4M	51M
TED	174K	169K	114K	152K	164K	180K
TED Test	1982	1982	1982	1982	1982	1982

Table 5: Number of segments in General Domain and TED training and test data for all languages

4 System-Level Analysis of Fluently Inadequate Translations

Koehn and Knowles (2017) showed that in out-of-domain and low-resource settings NMT produces lower quality output than SMT and they include examples where the NMT produced translations that were fluent but unrelated to the input. We seek to quantify this observation by estimating how often such fluently inadequate translations occur in SMT and NMT systems in different domain mismatch and training data settings. We score the output of 36 MT systems according to the percentage of fluently inadequate translations using the method described above.

4.1 MT Systems

We use a set of neural and phrase-based statistical MT models built from the same general domain data and adapted to translate a more specific domain, namely, transcripts of TED talks. We selected six languages to cover a range of resource availability scenarios and language families: Arabic, Chinese, Farsi, German, Korean and Russian.

4.1.1 Data

The number of segments of training and test data for each language is summarized in Table 5. The same tokenization was performed for all systems for a given language, and the tokenized data was split into subwords for NMT training using byte pair encoding (BPE) (Sennrich et al., 2016). The BPE models were trained separately on the source and target language with 30K BPE symbols.

All languages used data from the OpenSubtitles[1] corpus (Tiedemann, 2009) in the General domain training and dev data sets. The Chinese, German, and Russian models used additional parallel corpora from WMT17[2] (Bojar et al., 2017). For the Arabic models, we added data from the Linguistic Data Consortium (LDC)[3] and the UN v1 corpus[4] (Ziemski et al., 2016).

The domain for the In-Domain and Domain-Adapted models was TED talks. Training, dev, and test sets for the domain were from the Multi-target TED Talks Task (MTTT) corpus (Duh, 2018). All systems, regardless of training setting, were tested on the TED domain test set.

Fluency scores for each system were generated based on a language model built on the English side of its primary training data. As noted in Section 2.1, it is important that the language model match the training data, and we expect this to be particularly true when the test set is out-of-domain. We therefore use only the General domain data for both the General models and the adapted models, while the In-Domain models use the in-domain training data. Thresholds were calculated in a similar manner to the thresholds on the WMT16 data: thresholds for $WordLP_{mid}$ were calculated based on the General domain training data and thresholds for sentence BLEU and BVSS based on synthetic data built from the TED training data.

4.1.2 Statistical MT Systems

The statistical systems were built using the Apache Joshua toolkit[5] (Post et al., 2015). We tested three SMT models for each language: Joshua General, Joshua In-Domain, and Joshua Domain-Adapted, which were trained respectively on the General domain data, on the TED training data and on both. Language models for all systems were built from the English side of the training data. The Domain-Adapted model was tuned

[1] http://www.opensubtitles.org/

[2] http://www.statmt.org/wmt17/translation-task.html

[3] LDC2004T18, LDC2007T08, and LDC2012T09

[4] UN v1 is included in the Russian and Chinese WMT17 data

[5] http://cwiki.apache.org/confluence/display/JOSHUA/

	Arabic	German	Farsi	Korean	Russian	Chinese
Joshua General	23.50	30.65	13.41	6.34	24.49	14.79
Joshua TED Only	24.49	28.72	16.56	9.81	21.85	13.32
Joshua Adapted	27.11	31.35	17.71	10.24	25.23	15.70
Sockeye General	29.6	34.59	22.22	11.56	**28.6**	15.92
Sockeye TED Only	27.42	32.25	21.31	14.4	22.9	16.18
Sockeye Adapted	**35.37**	**39.9**	**27.92**	**17.22**	**28.6**	**20.37**

Table 6: BLEU scores for all systems

on TED dev data.

4.1.3 Neural MT Systems

The neural systems were built using Sockeye[6] (Hieber et al., 2017). The systems used two LSTM layers in both encoder and decoder with hidden size 512 and word embeddings dimension 512. We used a batch size of 4096 and created a checkpoint every 4000 mini-batches. Our systems employed the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 0.0003. As with the SMT, we built three models for each language: Sockeye General, Sockeye In-Domain, and Sockeye Domain-Adapted. The Sockeye General and In-Domain models were trained with the same data as the corresponding SMT models. The Sockeye Domain-Adapted models were trained using continued training on TED data starting from the Sockeye General model as in Luong at al (2015) and Freitag and Al-Onaizan (2016).

4.2 System Analyses

We compute the percentage of fluently inadequate translations in the system output of all MT and SMT systems to determine the effect of MT paradigm and training data on the occurrence of fluently inadequate translations.

Although the language pair and system varies, we can directly compare the output of the systems because the test data for all systems is from the *Multi-target* TED corpus. Note that in the corpus, the source is English and the other languages are translations while our task is translating into English. This means that if there are human translation errors or non-literal translations, the source will be inconsistent across languages but the reference will be the same. Table 7 shows English references for two different segments in Farsi and Chinese that yielded fluently inadequate MT output, along with their corresponding source and system outputs. For some segments the human translation (our source) may have slightly different meaning from the original (our reference), but the fluently inadequate examples we seek to identify are much further in meaning from both the source and reference. For instance, in the Chinese-English example the Chinese adds information that must be inferred from context in the original English. The Chinese literally translates to "Crow parents also teach their children these kinds of skills." The Sockeye TED and Joshua outputs reflect this additional information, but the Sockeye General output is fluent but completely unrelated to the reference.

Figure 1 shows the percent of segments labeled as fluently inadequate for each system. Even the highest percentage (Chinese-English Sockeye General) is less than 2%. Based on the high recall and low precision scores for the fluency metric in Section 3.1, we expect that we are overpredicting fluently inadequate translations so the actual percentage may be even lower. This confirms that these errors are indeed rare.

We also see from Figure 1 that the NMT models for Korean and Chinese, the languages most typologically different from English, have the highest levels of fluently inadequate translations on out-of-domain models. Although they have similarly high percentages of fluently misleading and similar amounts of in-domain training data, the Chinese domain-adapted model improves much more than the Korean domain-adapted model.

We compare the percent fluently inadequate segments to system BLEU scores in Figure 2. Based on the definition of our metric for fluently inadequate translations, translations with high sentence BLEU cannot be labeled fluently inadequate, so we expect a strong negative correlation between system BLEU and the percent fluently inadequate. We do see this negative correlation, but we can also see a clear difference in the percent fluently inadequate for the SMT vs NMT systems.

[6]http://github.com/awslabs/sockeye

System	FA-EN Example	ZH-EN Example
Source	انگیزه های زیرکانه تری داشته باشید	乌鸦 父母 还 教会 自己 的 孩子 这样 的 技巧 呢 。
Reference	get smarter incentives .	parents seem to be teaching their young .
Joshua General	terry زیرکانه have motives .	parents also teach their children the skills like this ?
Joshua TED	you have the انگیزه زیرکانه needed	parents crow can also teach our kids that the skills that .
Joshua Adapted	terry motives inspired .	the parents teach their children such skills .
Sockeye General	have a more subtle motivator .	**i 'm afraid i 'm not going to have to go to bed .**
Sockeye TED	**there 's a lot of gamers .**	and the crow parents taught their kids like this .
Sockeye Adapted	have smarter motivations .	and their parents also taught their children how to do it .

Table 7: Reference translation and example translations from the Farsi-English and Chinese-English systems. Fluently inadequate examples in bold.

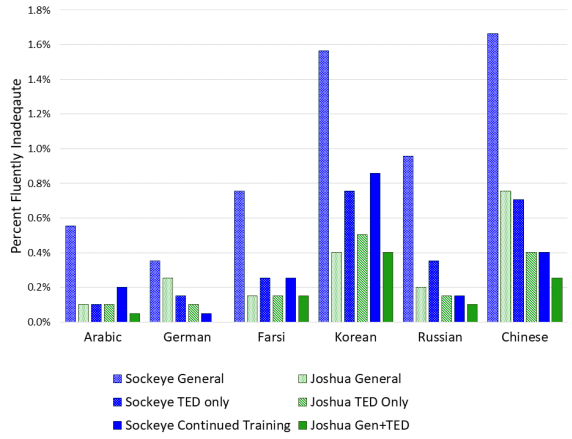

Figure 1: Segments labeled as fluently inadequate for General, TED, and Domain-Adapted Sockeye and Joshua models for all languages.

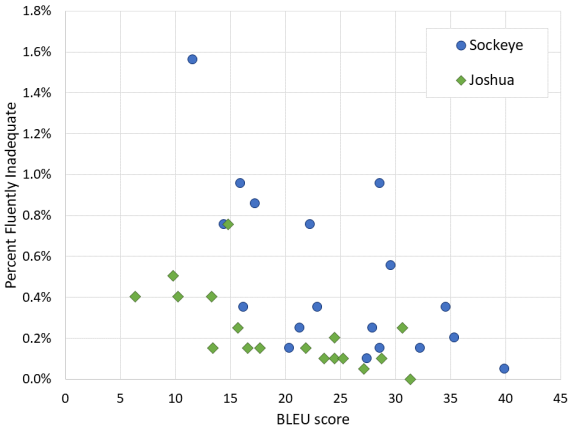

Figure 2: Segments labeled as fluently inadequate vs BLEU score for all Sockeye and Joshua models for all languages.

The NMT systems with low BLEU scores have much higher percentage of fluently inadequate translations than the similarly low-scoring SMT systems. This follows the suggestion in Koehn and Knowles (2017) that NMT is more prone to producing output that is disconnected from the source text when trained with insufficient or out-of-domain data. Indeed, we can see in Figure 1 that the NMT consistently has a higher percentage of fluently inadequate translations than the SMT.

Because our fluency metric relies on language models very similar to the language models used in the SMT systems, we might suspect that the fluency metric is biased towards the SMT models, potentially making SMT output more likely to be labeled as fluently inadequate. However, Figure 3 shows that the NMT systems still consistently have more segments labeled as fluent compared to SMT systems with similar BLEU score. This agrees with prior work showing that NMT output is more fluent than SMT and suggests that while the fluency metric likely leads to overprediction of fluently inadequate translations, it does not do so in a way that favors one paradigm over the other.

We also measured the percentage of fluently inadequate translations on the development set during training. Figure 4 shows that the percent fluently inadequate levels off very quickly, flattening after a few checkpoints on the in-domain model.

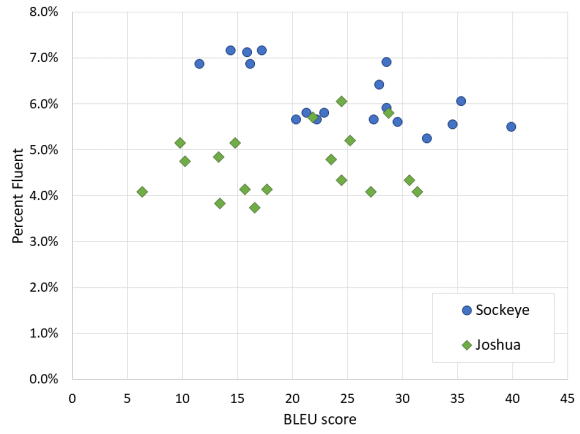

Figure 3: Segments labeled as fluent vs BLEU score for all Sockeye and Joshua models for all languages.

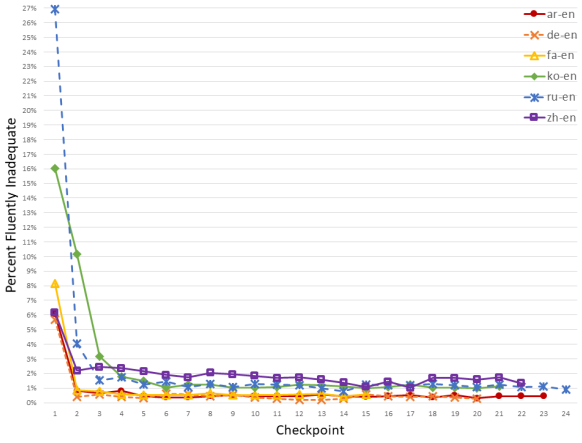

Figure 4: Percent fluently inadequate at each checkpoint during in-domain training

5 Related Work

MT quality metrics are judged based on their correlation with human judgments, and recently that has meant human adequacy judgments (Bojar et al., 2017). This indicates that any of the common MT metrics such as BLEU (Papineni et al., 2002) or METEOR (Banerjee and Lavie, 2005) may also serve as baseline adequacy scores. However, they incorporate elements of fluency while we wish to separate fluency and adequacy.

Adequacy is, essentially, semantic equivalence and the goal of SemEval's Semantic Textual Similarity (STS) task is to measure the degree of semantic equivalence between two sentences (Cer et al., 2017). The cross-lingual version of the task is similar enough to quality estimation that one of the data sets for 2017 actually came from the WMT quality estimation task. However, the STS systems performed much worse on the MT data than when tested on the Stanford Natural Language Inference (SNLI) Corpus data for the same language pair, with the top system achieving a correlation of only 34 compared to 83. These models are also complex and for use in combination with fluency, we prefer a simpler approach for this study.

Although grammaticality focuses on well-formedness while fluency includes all aspects of "sounding natural," the metrics used to predict grammaticality may still prove to be good measures of fluency. Lau et al (2016) take an unsupervised, language modeling approach to the task of predicting grammaticality as described in Section 2.1. They used two types of test data. One was generated by round-tripping sentences through Google Translate and the other was generated by extracting example sentences from a syntax textbook. The MT-generated English data is most similar to our problem, and the most effective models for that data were the word-based scores from the language model.

6 Conclusion

We have introduced an approach to automatically detect fluently inadequate translations in machine translation output based on automatic fluency and adequacy metrics. Applying this technique to a diverse set of statistical and neural MT systems, we found that although fluently inadequate translations are rare, NMT does appear to be consistently more prone to this type of error compared to SMT. Improving the match between training and test with continued training on in-domain data reduces these errors. These findings raise several questions for future work: How often are fluently inadequate translations actually misleading to human users? How can we detect fluently inadequate translations without reference translations?

Acknowledgments

This work is supported in part by an Amazon Web Services Machine Learning Research Award. The views contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the sponsors. Part of this work was done at the JHU SCALE 2018 workshop and we would like to thank all our team members for helpful discussions, particularly John Farina and David Yuen.

References

Banerjee, Satanjeev 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 and/or summarization*, volume 29, pages 65–72.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas. Association for Computational Linguistics.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany, Aug. Association for Computational Linguistics.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, Jun. Association for Computational Linguistics.

Cer, Daniel, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. In *Proceedings of SemEval 2017*.

Duh, Kevin. 2018. The multitarget ted talks task. `http://www.cs.jhu.edu/~kevinduh/a/multitarget-tedtalks/`.

Florek, Kazimierz, Jan Łukaszewicz, Julian Perkal, Hugo Steinhaus, and Stefan Zubrzycki. 1951. Sur la liaison et la division des points d'un ensemble fini. In *Colloquium Mathematicae*, volume 2, pages 282–285.

Freitag, Markus and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*.

Graham, Yvette, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Hassan Awadalla, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, Will Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving Human Parity on Automatic Chinese to English News Translation, Mar.

Heafield, Kenneth, Ivan Pouzyrevsky, Jonathan H Clark, and Philipp Koehn. 2013. Scalable modified kneser-ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 690–696.

Heafield, Kenneth. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

Hieber, Felix, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *arXiv preprint arXiv:1712.05690*.

Joulin, Armand, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Kingma, Diederik P. and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Knox, Steven W. 2015. Extending pairwise element similarity to set similarity efficiently. Presentation at MAA MathFest, Washington, DC, 8.

Koehn, Philipp and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In *Workshop on Neural Machine Translation*, Vancouver, BC. arXiv: 1706.03872.

Lau, Jey Han, Alexander Clark, and Shalom Lappin. 2016. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

Luong, Minh-Thang, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Martindale, Marianna J. and Marine Carpuat. 2018. Fluency Over Adequacy: A Pilot Study in Measuring User Trust in Imperfect MT. In *Proceedings of the 13th Conference of The Association for Machine Translation in the Americas*, pages 13–25, Boston, MA, USA, March.

Papineni, Kishore, 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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Post, Matt, Yuan Cao, and Gaurav Kumar. 2015. Joshua 6: A phrase-based and hierarchical statistical machine translation system. *The Prague Bulletin of Mathematical Linguistics*.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics.

Tiedemann, Jörg. 2009. News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*, volume 5, pages 237–248.

Toral, Antonio and Víctor M. Sánchez-Cartagena. 2017. A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1, pages 1063–1073, Valencia, Spain. Association for Computational Linguistics.

Ziemski, Michal, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1. 0. In *LREC*.