# On The Evaluation of Machine Translation Systems Trained With Back-Translation

**Sergey Edunov** **Myle Ott** **Marc'Aurelio Ranzato** **Michael Auli**
Facebook AI Research

## Abstract

Back-translation is a widely used data augmentation technique which leverages target monolingual data. However, its effectiveness has been challenged since automatic metrics such as BLEU only show significant improvements for test examples where the source itself is a translation, or *translationese*. This is believed to be due to translationese inputs better matching the back-translated training data. In this work, we show that this conjecture is not empirically supported and that back-translation improves translation quality of both naturally occurring text as well as translationese according to professional human translators. We provide empirical evidence to support the view that back-translation is preferred by humans because it produces *more fluent* outputs. BLEU cannot capture human preferences because references are translationese when source sentences are natural text. We recommend complementing BLEU with a language model score to measure fluency.

## 1 Introduction

Back-translation (BT; Bojar and Tamchyna 2011; Sennrich et al. 2016a; Poncelas et al. 2018a) is a data augmentation method that is a key ingredient for improving translation quality of neural machine translation systems (NMT; Sutskever et al. 2014; Bahdanau et al. 2015; Gehring et al. 2017; Vaswani et al. 2017). NMT systems using large-scale BT have been ranked top at recent WMT evaluation campaigns (Bojar et al., 2018; Edunov et al., 2018; Ng et al., 2019). The idea is to train a target-to-source model to generate additional synthetic parallel data from monolingual target data. The resulting sentence pairs have synthetic sources and natural targets which are then added to the original bitext in order to train the desired source-to-target model. BT improves generalization and can be used to adapt models to the test domain by adding appropriate monolingual data.

Parallel corpora are usually comprised of two types of sentence-pairs: sentences which originate in the source language and have been translated by humans into the target language, or sentences which originate from the target language and have been translated into the source language. We refer to the former as the *direct* portion and the latter as the *reverse* portion. The setup we are ultimately interested in is models that translate direct sentences.

Translations produced by human translators, or *translationese* tend to be simpler and more standardized compared to naturally occurring text (Baker, 1993; Zhang and Toral, 2019; Toury, 2012). Several recent studies found that such reverse test sentences are easier to translate than direct sentences (Toral et al., 2018; Graham et al., 2019), and human judges consistently assign higher ratings to translations of target original sentences than to source original sentences. These studies therefore recommend to restrict test sets to source original sentences, a methodology which has been adopted by the 2019 edition of the WMT news translation shared task.

Unfortunately, automatic evaluation with BLEU (Papineni et al., 2002) only weakly correlates with human judgements (Graham et al., 2019). Furthermore, recent WMT submissions relying heavily on back-translation mostly improved BLEU on the reverse direction with little gains on the direct portion (Toral et al. 2018; Barry Haddow's personal communication and see also Appendix A, Table 7; Freitag et al. 2019).

This finding is concerning for two reasons. First, back-translation may not be effective after all since gains are limited to the reverse portion. Improvements on reverse sentences may only be due to a better match with the back-translated training sentences in this case. Second, it may further reduce



*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

our confidence in automatic evaluation, if human judges disagree with BLEU for systems trained with back-translation. Indeed, human evaluations of top performing systems at WMT'18 (Bojar et al., 2018) and WMT'19 (Bojar et al., 2019) did not agree with BLEU to the extent that correlation is even negative for the top entries (Ma et al., 2019).

In this paper, we shed light on the following questions. First, do BT systems only work better in the reverse direction? Second, does BLEU reflect human assessment for BT models? And if that is not the case, why not and how can we alleviate the weaknesses of BLEU?

Our contribution is an extensive empirical evaluation of top-performing NMT systems to validate or disproof some of the above conjectures. First, we show that translationese sources are indeed easier to translate, but this is true for both NMT systems trained with and without back-translated data. Second, we confirm that human assessment of BT systems poorly correlates with BLEU. Third, BLEU cannot capture the higher quality of back-translation systems because the outputs of both back-translation and non back-translation models are equally close to the translationese references. Fourth, we show that BT system outputs are significanlty more fluent than the output of a system only trained on parallel data, and this may explain the human preference towards BT generations. Finally, we recommend to improve automatic evaluation by complementing BLEU with a language model score which can better assess fluency in the target language while avoiding the artifacts of translationese references.

## 2 Related Work

Back-translation has been originally introduced for phrase-based machine translation (Bojar and Tamchyna, 2011). For back-translation with neural machine translation, there is a large body of literature building upon the seminal work of Sennrich et al. (2016a), from large-scale extensions with sampling (Edunov et al., 2018; Ott et al., 2018) or tagging (Caswell et al., 2019) to its use for unsupervised machine translation (Lample et al., 2018) as well as analysis (Poncelas et al., 2018b) and iterative versions (Hoang et al., 2018).

More similar to our work, Toral et al. (2018) analyzed performance of trained state-of-the-art NMT systems in direct and reverse mode. They observe that translationese is simpler to translate and claimed that gains for such systems mostly come from improvements in the reverse direction.

Concurrent to our work, Graham et al. (2019) find that automatic evaluation with BLEU does not align with the hypothesis that reverse sentences are easier to translate instead. Unfortunately, their findings are not very conclusive because they do not control for the change of actual content, as sentences in one direction may be extracted from documents which are just harder to translate. In this work we correct for this effect by comparing translations of source original sentences with their double translations. Graham et al. (2019) also observe that BLEU does not reliably correlate with human judgements. While they consider a large variety of systems trained in various ways, we instead focus on the comparison between the same NMT system trained with and without back-translated data.

Earlier work on statistical machine translation models argued in favor of using source original data only to train translation models (Kurokawa et al., 2009), language models for translation (Lembersky et al., 2011), and to tune translation models (Stymne, 2017). All these studies base most of their conclusions on automatic evaluation with BLEU, which is problematic since BLEU is not reliable and this procedure may overly optimize towards translationese references.

Freitag et al. (2019) proposed a post-editing method to turn translationese system outputs into more natural text. As part of their evaluation, they also observed that human assessments poorly correlate with BLEU. While we confirm some of these observations, our goal is an in-depth analysis of the evaluation of NMT systems trained with back-translated data. We provide empirical evidence corroborating the hypothesis that the discrepancy between BLEU and human assessment is due to the use of translationese references, and we provide a constructive suggestion on how to better automatically evaluate models trained with BT.

## 3 Experimental Setup

In the next sections we first discuss the datasets and models used. Then, we report BLEU evaluations showing a big discrepancy between the gains obtained by a BT system in forward versus reverse direction compared to a baseline trained only on parallel data. This is followed by a series of hypotheses about the reasons for this discrepancy, and

empirical studies in support or to disprove these hypotheses. We conclude with a recommendation for how to better evaluate NMT systems trained with BT.

### 3.1 Training Datasets

We consider four language directions: English-German (En-De), German-English (De-En), English-Russian (En-Ru) and Russian-English (Ru-En).

For En-De, we train a model on the WMT'18 news translation shared task data. We used all available bitext excluding the ParaCrawl corpus. We removed sentences longer than 250 words as well as sentence-pairs with a source/target length ratio exceeding 1.5. This results in 5.18M sentence pairs. For back-translation, we use the same setup as the WMT'18 winning entry for this language pair which entails sampled back-translation of 226M German newscrawl sentences (Edunov et al., 2018).[1]

For De-En, En-Ru, Ru-En we use all parallel data provided by the WMT'19 news translation task, including Paracrawl. We remove sentences longer than 250 words as well as sentence-pairs with a source/target length ratio exceeding 1.5 and sentences which are not in the correct language (Lui and Baldwin, 2012). This resulted in 27.7M sentence-pairs for En-De and 26M for En-Ru.

For the back-translation models we use the top ranked Facebook-FAIR systems of the WMT'19 news shared translation task.[2] The parallel data and pre-processing of those systems is identical to our baselines which are trained only on parallel data (Ng et al., 2019). As monolingual data, the WMT'19 newscrawl data was filtered by langid, resulting in 424M English and 76M Russian monolingual sentences. For En-De and De-En models use a joined byte-pair encoding (BPE; Sennrich et al. 2016b) with 32K split operations, and for En-Ru and Ru-En separate BPE dictionaries for the source and target with 24K split operations.

[1]WMT'18 models are available at https://github.com/pytorch/fairseq/tree/master/examples/backtranslation and we used a single model.

[2]WMT'19 models are available at https://github.com/pytorch/fairseq/tree/master/examples/wmt19

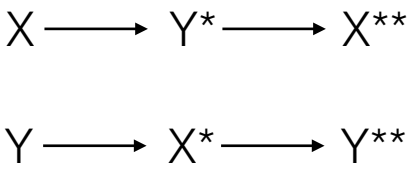


Figure 1: Illustration of the translations used in this work. $X$ represent sentences originating in the source language. $Y$ are sentences originating in the target language. A single $^*$ symbol represents a translation of an original sentence, while $^{**}$ represents a double translation, i.e. a translation of a translationese sentence. The original dataset consists of the union of ($X$, $Y^*$) pairs (direct mode) and ($X^*$, $Y$) (reverse mode). According to BLEU, a system trained with BT improves only in reverse mode. As part of this study we have collected double translations, which are useful to assess whether translationese inputs are easier to translate (by comparing performance when the input is $X^{**}$ versus $X$ and the reference is $Y^*$) and easier to predict (by comparing performance when the reference is $Y^{**}$ versus $Y$ and the input is $X^*$).

### 3.2 Sequence to Sequence Models

We train models using the big Transformer implementation of fairseq (Vaswani et al., 2017; Ott et al., 2019). All our models are trained on 128 Volta GPUs, following the setup described in Ott et al. (2018). For En-De we used single Transformer Big models without checkpoint averaging. For De-En and En-Ru we increased model capacity by using larger FFN size (8192) and we also used an ensemble of models trained with three different seeds.

In the remainder of this paper, we will refer to baseline NMT models trained **o**nly on **p**arallel data as **OP**, and to models trained on both parallel data and **b**ack-**t**ranslated data as **BT**.

### 3.3 Test sets and Reference Collection

In order to assess differences in model performance when inputting translationese vs. natural language (§4.2), we collected additional references which will be made publicly and freely available soon.3 These are sentence-level (as opposed to document level) translations which matches the training setup of our models. In Appendix B we confirm that our findings also apply to the original WMT document-level references.

Figure 1 illustrates the composition of the test set for each language direction which is divided into two partitions: First, the *direct* portion consists of sentences $X$ originally written in the source language which were translated into the target lan-

guage as $Y^*$. Additionally, we translated $Y^*$ back into the source language to yield $X^{**}$, a translationese version of $X$. Second, for the *reverse* portion, we have naturally occurring sentences in the target language $Y$ that were translated into the source as $X^*$. We also translated these into the target as $Y^{**}$ to obtain a translationese version of the original target. For each language pair we use the following data:

**English ↔ German.** We used newstest2014 that we separated into English-original and German-original sets. We then sampled 500 English-original and 500 German-original sentences from each subset and asked professional human translators to translate them into German and English respectively. In addition, we ask professional human translators to provide $X^{**}$ and $Y^{**}$ which are translations of $Y^*$ and $X^*$, respectively.

**English ↔ Russian.** For this setup we sampled 500 English-original sentences from the En-Ru version of newstest2019 and asked professional human translators to translate them into Russian at the sentence-level. Similarly, we sampled 500 Russian-original sentences from the Ru-En version of newstest2019 and obtained English references. We also collected double translations $X^{**}$, $Y^{**}$ of $Y^*$ and $X^*$, respectively. 3 The additional references are available at https://github.com/facebookresearch/evaluation-of-nmt-bt.

### 3.4 Human and Automatic Evaluation

Human evaluations and translations were conducted by certified professional translators who are native speakers of the target language and fluent in the source language. We rate system outputs using both source and target based direct assessment. In the former case, raters evaluate correctness and completeness on a scale of 1-100 for each translation given a source sentence. This method is the most thorough assessment of translation quality. It also has the additional benefit to be independent of the provided human references which may affect the evaluation. For target based direct assessment, raters evaluate closeness to the provided reference on a scale of 1-100 for each translation. This is easier since it only requires people fluent in one language, and it is the evaluation performed by recent WMT campaigns (Graham et al., 2017; Bojar et al., 2018).

To rate a translation, we collected three judgements per sentence. We repeated the evaluation for sentences where all three raters provided judgements that differed by more than 30 points. Evaluation was blind and randomized: human raters did not know the identity of the systems and all outputs were shuffled to ensure that each rater provides a similar number of judgements for each system.

Following the WMT shared task evaluation (Bojar et al., 2018), we normalize the scores of each rater by the mean and standard deviation of all ratings provided by the rater. Next, we average the normalized ratings for each sentence and average all per-sentence scores to produce an aggregate per-system z-score. As automatic metric, we report case-sensitive BLEU using SacreBLEU (Post, 2018).[3] We also consider other metrics in Appendix C, but conclusions remain the same.

| src | ref | sys | en-de | de-en | en-ru | ru-en |
|---|---|---|---|---|---|---|
| $X$ | $Y^*$ | OP | 33.7 | 40.3 | 31.3 | 43.8 |
| | | BT | 32.3 | 38.6 | 31.9 | 41.2 |
| $X^*$ | $Y$ | OP | 31.3 | 43.0 | 40.5 | 31.8 |
| | | BT | 38.9 | 48.7 | 50.6 | 40.3 |

Table 1: BLEU for four language directions measured on source original sentences ($X \rightarrow Y^*$) as well as target original sentences ($X^* \rightarrow Y$) for a model trained on parallel data only (OP) as well as a back-translation model (BT). BT performs much better than OP on the reverse portion of the test set but BLEU shows no difference on the direct portion.

| src | ref | sys | en-de | de-en | en-ru | ru-en |
|---|---|---|---|---|---|---|
| $X$ | $Y^*$ | OP | 33.7 | 40.3 | 31.3 | 43.8 |
| | | BT | 32.3 | 38.6 | 31.9 | 41.2 |
| $X^{**}$ | $Y^*$ | OP | 39.7 | 46.9 | 42.8 | 49.9 |
| | | BT | 39.2 | 45.6 | 44.0 | 47.6 |

Table 2: BLEU for source original sentences ($X \rightarrow Y^*$) compared to the same sentence pairs with a translationese source ($X^{**} \rightarrow Y^*$). Translationese inputs are simpler to translate but BT and OP systems benefit equally from translationes inputs.

## 4 Results

### 4.1 Evaluating BT with Automatic Metrics

We first reproduce the known discrepancy between BT and OP in the reverse direction (target original

[3] SacreBLEU signature: BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.3.1

| src | ref | sys | en-de | | de-en | | en-ru | | ru-en | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | BLEU | human | BLEU | human | BLEU | human | BLEU | human |
| $X$ | $Y^*$ | OP | 33.7 | -0.18 | 40.3 | -0.07 | 31.3 | -0.66 | 43.8 | -0.37 |
| | | BT | 32.3 | **-0.05** | 38.6 | **0.03** | 31.9 | **-0.35** | 41.2 | **-0.12** |
| $X^*$ | $Y$ | OP | 31.3 | -0.01 | 43.0 | 0.06 | 40.5 | 0.06 | 31.8 | -0.02 |
| | | BT | 38.9 | **0.10** | 48.7 | **0.13** | 50.6 | **0.16** | 40.3 | **0.07** |
| $X^{**}$ | $Y^*$ | OP | 39.7 | -0.05 | 46.9 | 0.07 | 42.8 | -0.17 | 49.9 | -0.05 |
| | | BT | 39.2 | **0.03** | 45.6 | **0.16** | 44.0 | **-0.01** | 47.6 | **0.12** |
| $X^*$ | $Y^{**}$ | OP | 39.5 | -0.01 | 63.6 | 0.06 | 49.5 | 0.06 | 44.4 | -0.02 |
| | | BT | 41.8 | **0.10** | 61.2 | **0.13** | 50.4 | **0.16** | 38.7 | **0.07** |

Table 3: BLEU and human preference judgements on four language directions with a bitext-only model as well as a back-translation model (BT). BLEU shows no strong preference when the source is natural text ($X$) but professional human translators prefer BT regardless of whether the source is $X$ or translationese ($X^*$). Back-translation also does not overproportionally benefit from inputting translationese since both OP and BT show similar improvements when switching from $X$ to $X^{**}$ inputs. BT human scores are statistically significantly better at p=0.05 than the respective OP as per paired bootstrap resampling (Koehn, 2004).

sentences; $X^* \rightarrow Y$) and the forward direction (source original sentences; $X \rightarrow Y^*$).

Table 1 shows that BT does not improve over OP on direct sentences ($X \rightarrow Y^*$) in aggregate. However, on the reverse portion BT does improve, and it does so by very large margins of between 5.7-10.1 BLEU. Appendix C shows that TER (Snover et al., 2006), BEER (Stanojevic and Sima'an, 2014), METEOR (Banerjee and Lavie, 2005) and BERTScore (Zhang et al., 2019) also do not distinguish very strongly between OP and BT for direct sentences.

A possible explanation for this result is that BT can better translate target-original test sentences because those sentences mimic the training data of BT. The BT training data (§3) consists largely of target original sentences-pairs with back-translated sources which could explain the discrepancy between performance of the BT system on the direct and reverse portions.

### 4.2 Translationese Benefits Both BT & OP

Translationese is known to be a different dialect with lower complexity than naturally occurring text (Toral et al., 2018). This is corroborated by the fact that this data is straightforward to identify by simple automatic classifiers (Koppel and Ordan, 2011). One possible explanation for why back-translation could be more effective for target original sentences is that the input to the system is translated language. This may give the BT system two advantages: i) the input is simpler than naturally occurring text and ii) this setup may be easier for the back-translation system which was trained on additional target original data that was automatically translated.

To test this hypothesis we feed source original sentences and translationese into our systems and compare their performance. We created a test setup where we have *both* a source original sentence ($X$) and a translationese version of it ($X^{**}$) which share a reference ($Y$), see §3.3. This enables us to precisely test the effect of translationese vs natural language.

Table 2 shows that BLEU is substantially higher when the input is translationese ($X^{**}$) compared to natural language ($X$), however, both BT and OP obtain comparable improvements. Therefore, the BLEU discrepancy between BT and OP in direct vs. reverse cannot be explained by BT gaining an advantage over OP through translationese inputs.

### 4.3 Human Evaluation Contradicts BLEU

The aforementioned experiments were evaluated in terms of BLEU, an automatic metric. To get a more complete picture, we ask professional human translators to judge translations using source-based direct assessment (unless otherwise specified, this is our default type of human evaluation; see §3.4).

Table 3 (first two sets of rows) shows that human judges prefer BT over OP regardless of whether sentences are source original ($X \rightarrow Y^*$) or target original ($X^* \rightarrow Y$). This is in stark contrast to the corresponding BLEU results.

Similar observations have been made in the two most recent WMT evaluation campaigns: at WMT'18 (Bojar et al., 2018), the large-scale sampled BT system of Facebook-FAIR (Edunov et al., 2018) ranked 6th in terms of BLEU while being ranked first in the human evaluation. The results of WMT'19 show a similar picture where a system relying on large scale back-translation ranked first in the human evaluation but only 8th in terms of BLEU (Bojar et al., 2019).

*We conclude that professional human translators prefer BT over OP - regardless of whether test sentences are source or target original.*

### 4.4 Human Evaluation is Robust

Our current observations could be explained by some idiosyncrasy in the human evaluation. To reject this hypothesis we performed both source-based and target-based assessment for all English-German systems of Table 3 using professional translators (§3.4) and computed the correlation between the two types of assessments. The correlation coefficient between source and target based assessment is 0.90 (95% confidence interval 0.55 - 0.98), which indicates that human evaluation is robust to the assessment type. This finding is consistent with other work comparing the two types of human evaluations (Bojar et al., 2018).

### 4.5 Why BLEU Fails in Direct Mode

Next, we investigate why BLEU does not agree with human judgements in direct mode. BLEU measures n-gram overlap between a model output and a human reference translation. In the case of direct sentences, the references are translationese.

We found earlier that BLEU does not distinguish between BT and OP even though professional human translators prefer BT. Given references are translationese, one possible explanation is that both systems produce translations which equally resemble translationese and thus BLEU fails to distinguish between them.

To test this hypothesis and measure the closeness of system outputs with respect to translationese, we train two large transformer-based language models (Baevski and Auli, 2018). The first is trained on outputs produced by the En-De BT system, the second one on the outputs produced by the En-De OP system. The outputs are the translation of English Newscrawl 2018 comprising 76M sentences. We then evaluate the language models on source original sentences ($Y^*$) of newstest2015-2018.

| data | OP | BT |
|---|---|---|
| $Y^*$ | 37.2 | 36.8 |
| $Y$ | 82.2 | 57.4 |

Table 4: Perplexity on the source-original/translationese portion ($Y^*$) and the target-original portion of newstest2014-2018 ($Y$). We translate the English newscrawl training data with either OP and BT and train two language models on the outputs. Both BT and OP are equally close to translationese (first row), but BT is closer than OP to naturally occurring text (second row).

The first row of Table 4 shows that both language models achieve similar perplexity on $Y^*$ (37.2 VS 36.8), suggesting that the translations of BT and OP are equally close to translationese. Interestingly, both system outputs are closer to translationese than natural text since PPL on $Y^*$ is significantly lower than the PPL on $Y$ (second row of Table 4). This is also supported by BLEU being higher when using $Y^{**}$ as a reference compared to $Y$ for the same input $X^*$ (second and last row of Table 3).

*Our results support the hypothesis that the outputs of BT and OP are equally close to translationese. This in turn may explain why BLEU cannot distinguish between OP and BT in direct mode where the reference is translationese.*

### 4.6 BT Generates More Natural Text

Back-translation augments the training corpus with automatic translations from target original data. Training models on large amounts of target original data may bias BT systems to produce outputs that are closer to naturally occurring text. In contrast, OP systems have been trained on the original parallel data, a mix of direct and reverse data which contains a much smaller amount of target original sentences. This may explain why BLEU evaluation with translationese references (direct portion) does not capture the human preference for BT.

To understand this better, we conduct two experiments. The first experiment is based on the language models we trained previously (§4.5) to assess how close our systems are to translationese and naturally occurring text. The second experiment is based on a human study where native speakers assess the fluency of each system output.

For the first experiment we reuse the two language models from §4.5 to measure how close the system outputs are to natural text ($Y$). The second

| | BT | OP | draw |
|---|---|---|---|
| De-En | 28 | 16 | 63 |
| En-De | 50 | 33 | 18 |
| En-Ru | 37 | 21 | 42 |

Table 5: Human preference in terms of fluency for system outputs of BT and OP. Judgements are based on a pair-wise comparison between the two systems without the source sentence and conducted by native speakers. All results are based on 100 judgements and the preference of BT over OP is statistically significant at p=0.05.

| | BT PPL | OP PPL |
|---|---|---|
| De-En | 74.8 | 78.7 |
| En-De | 48.6 | 52.6 |
| Ru-En | 57.6 | 68.6 |
| En-Ru | 61.7 | 72.4 |

Table 6: Automatic fluency analysis with language models trained on the Common Crawl corpus in the respetive target language. BT receives lower perplexity (PPL) throughout, despite attaining the same BLEU score of OP, see Table 1.

row of Table 4 shows that the BT language model assigns much higher probability to naturally occurring text, $Y$, compared to the OP language model (82.2 VS 57.4 perplexity), suggesting that *BT does indeed produce outputs that are much closer to natural text than OP*. We surmise that this difference, which is captured by a language model trained on system outputs and evaluated on $Y$, could be at least partially responsible for the marked human preference towards BT translations.

In the second experiment, native speakers of English, German and Russian rate whether the output of OP is more fluent than the output of BT for 100 translations of the De-En, En-De and En-Ru systems. Human raters perform a pair-wise ranking and raters can only see two translations but not the source; the system identity is unknown to raters.

Table 5 shows that BT is judged to be significantly more fluent by native speakers than OP in three languages.

## 5 Improving BT Evaluation

In the previous sections, we gathered mounting evidence that BLEU fails at capturing the improved fluency of BT in direct mode. Next, we propose to use a language model to assess fluency as an additional measure to complement BLEU. Different to the setup above (§4.5, 4.6), where we used a separate LM for each system, we propose to use *a single* LM for all systems in order to simplify the evaluation.

The language model is trained on a large monolingual dataset *disjoint* from the monolingual dataset used for generating back-translated data for BT training. This restriction is critical, otherwise the language model is likely to assign higher probably to BT generations simply because training and evaluation sets overlap. To train these language models we sample 315M, 284M and 120M monocrawl sentences for each of the three target languages, namely English, German and Russian, respectively.

The language model is used to score the outputs of BT and OP on the direct portion of the test set. If two systems have similar BLEU scores, then a lower perplexity with the LM indicates higher fluency in the target natural language. This fluency assessment is complementary to BLEU which in turn is more sensitive to adequacy.

Table 6 shows that the language model assigns lower perplexity to BT in all four setups. This shows that a language model can help to assess the fluency of system output when a human evaluation is not possible.

In future work, we intend to further investigate how to best combine BLEU and language model scoring in order to maximize correlation with human judgements, particularly when evaluating BT in direct mode. Meantime, practitioners can use this additional metric in their evaluation to break ties in BLEU scoring.

## 6 Conclusions

According to our findings, back-translation improves translation accuracy, for both source and target original sentences. However, automatic metrics like BLEU fail to capture human preference for source original sentences (direct mode).

We find that BT produces outputs that are closer to natural text than the output of OP, which may explain human preference for BT. We recommend distinguishing between direct and reverse translations for automatic evaluation, and to make final judgements based on human evaluation. If human evaluation is not feasible, complementing standard metrics like BLEU with a language model (§5) may help assessing the overall translation quality.

In the future, we plan to investigate more thor-

oughly the use of language models for evaluating fluency, the effect of domain mismatch in the choice of monolingual data, and ways to generalize this study to other applications beyond MT.

## Acknowledgements

We thank Barry Haddow for initially pointing out the BLEU discrepancy between the forward and reverse portions of the WMT 2018 test set.

## References

Alexei Baevski and Michael Auli. 2018. Adaptive input representations for neural language modeling. *arXiv*, abs/1809.10853.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proc. of ICLR*.

Mona Baker. 1993. Corpus linguistics and translation studies: Implications and applications. *Text and technology: In honour of John Sinclair*, 233:250.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *In Proc. of ACL Workshop on Intrinsic and Extrinsic Evaluation Measure for Machine Translation*.

Ondrej Bojar and Ales Tamchyna. 2011. Improving translation model by monolingual data. In *Proc. of WMT*.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proc. of WMT*.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proc. of WMT*.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proc. of WMT*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proc. of EMNLP*.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. Text repair model for neural machine translation. *arXiv*, abs/1904.04790.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional Sequence to Sequence Learning. In *Proc. of ICML*.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):330.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. Translationese in machine translation evaluation. *arXiv*, abs/1906.09833.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proc. of 2nd Workshop on Neural Machine Translation and Generation*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proc. of EMNLP*.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proc. of ACL*.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In *Proc. of MT Summit*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based and neural unsupervised machine translation. In *EMNLP*.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2011. Language models for machine translation: Original vs. translated texts. In *Proc. of EMNLP*.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proc. of ACL: Demonstrations*.

Qingsong Ma, Johnny Wei, Ondrej Bojar, and Yvette Graham. 2019. Results of the wmt19 metrics shared task: Segment-level and strong mt systems pose big challenges. In *Proc. of WMT*.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's wmt19 news translation task submission. In *Proc. of WMT*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proc. of NAACL: Demonstrations*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proc. of WMT*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL*.

Alberto Poncelas, Dimitar Sht. Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018a. Investigating backtranslation in neural machine translation. *arXiv*, 1804.06189.

Alberto Poncelas, Dimitar Sht. Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018b. Investigating backtranslation in neural machine translation. *arXiv*, 1804.06189.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proc. of WMT*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proc. of ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proc. of ACL*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proc. of AMTA*.

Milos Stanojevic and Khalil Sima'an. 2014. Beer: Better evaluation as ranking. In *Proc. of WMT*.

Sara Stymne. 2017. The effect of translationese on tuning for statistical machine translation. In *Proc. of NoDaLiDa*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proc. of NIPS*.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proc. of WMT*.

Gideon Toury. 2012. *Descriptive translation studies and beyond: Revised edition*, volume 100. John Benjamins Publishing.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proc. of NIPS*.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. *arXiv*, abs/1906.08069.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv*, abs/1904.09675.

## A Forward/reverse BLEU for WMT'18 English-German systems

| system | fwd | rev | delta |
|---|---|---|---|
| online-Y | 47.1 | 30.3 | -16.8 |
| MMT-production-system | 51.8 | 36.7 | -15.1 |
| online-B.0 | 52.9 | 39.1 | -13.8 |
| NTT | 50.7 | 39.7 | -11.0 |
| Microsoft-Marian | 52.5 | 41.6 | -10.9 |
| KIT | 50.3 | 39.5 | -10.8 |
| LMU-nmt | 43.5 | 33.4 | -10.1 |
| uedin | 47.8 | 37.8 | -10.0 |
| online-A | 37.8 | 28.6 | -9.2 |
| JHU | 46.0 | 38.2 | -7.8 |
| online-F | 23.5 | 16.4 | -7.1 |
| UCAM | 48.9 | 42.1 | -6.8 |
| RWTH-UNSUPER | 16.7 | 12.0 | -4.7 |
| online-G | 25.9 | 22.5 | -3.4 |
| LMU-unsup | 15.2 | 14.3 | -0.9 |
| **Facebook-FAIR** | **45.8** | **46.1** | **0.4** |

Table 7: Forward/reverse BLEU for WMT'18 English-German systems.

Table 7 shows that a large-scale back-translation system, Facebook-FAIR, mostly improves BLEU on the reverse portion whereas it is outperformed by many other entrants in the forward portion.

## B Results with WMT references

| src | ref | sys | en-de | | de-en | | en-ru | | ru-en | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | BLEU | human | BLEU | human | BLEU | human | BLEU | human |
| $X$ | $Y^*$ | OP | 33.7 | -0.18 | 40.3 | -0.07 | 31.3 | -0.66 | 43.8 | -0.37 |
| | | BT | 32.3 | -0.05 | 38.6 | 0.03 | 31.9 | -0.35 | 41.2 | -0.12 |
| $X$ | $Y^*_{WMT}$ | OP | 28.7 | -0.18 | 35.4 | -0.07 | 31.8 | -0.66 | 39.7 | -0.37 |
| | | BT | 29.9 | -0.05 | 34.2 | 0.03 | 31.9 | -0.35 | 38.5 | -0.12 |

Table 8: BLEU results with respect to the original WMT references (document-level) and the sentence-level references used throughout this study. Sentence-level references result in higher BLEU but OP and BT still achieve very similar BLEU.

Table 8 shows that BLEU does not strongly distinguish between BT and OP, regardless of whether the reference was obtained at the document-level ($Y^*_{WMT}$) or at the sentence-level ($Y^*$).

## C Other metrics than BLEU

| src | ref | sys | en-de | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | human | BLEU | TER | BEER | METEOR | BERTScore |
| $X$ | $Y^*$ | OP | -0.18 | 33.7 | 0.466 | 0.635 | 0.531 | 0.849 |
| | | BT | -0.05 | 32.3 | 0.473 | 0.619 | 0.512 | 0.843 |
| $X^*$ | $Y$ | OP | -0.01 | 31.3 | 0.504 | 0.609 | 0.530 | 0.841 |
| | | BT | 0.10 | 38.9 | 0.431 | 0.652 | 0.580 | 0.866 |
| $X^{**}$ | $Y^*$ | OP | -0.05 | 39.7 | 0.403 | 0.677 | 0.590 | 0.878 |
| | | BT | 0.03 | 39.2 | 0.409 | 0.669 | 0.578 | 0.876 |
| $X^*$ | $Y^{**}$ | OP | -0.01 | 39.5 | 0.410 | 0.670 | 0.599 | 0.876 |
| | | BT | 0.10 | 41.8 | 0.383 | 0.683 | 0.610 | 0.884 |

Table 9: BLEU and other metrics as well as human preference judgements for English-German translations.

Table 9 shows results for automatic metrics other than BLEU (Papineni et al., 2002). The metrics TER (Snover et al., 2006), BEER (Stanojevic and Sima'an, 2014), METEOR (Banerjee and Lavie, 2005) and BERTScore (Zhang et al., 2019) show similar trends as BLEU, i.e., they do not indicate human preference of BT over bitext for the direct portion of the test set ($X \rightarrow Y^*$).