An Extensive Exploration of Back-Translation in 60 Languages

Paul McNamee and **Kevin Duh**
Human Language Technology Center of Excellence
Johns Hopkins University
mcnamee@jhu.edu kevinduh@cs.jhu.edu

Abstract

Back-translation is a data augmentation technique that has been shown to improve model quality through the creation of synthetic training bitext. Early studies showed the promise of the technique and follow on studies have produced additional refinements. We have undertaken a broad investigation using back-translation to train models from 60 languages into English; the majority of these languages are considered moderate- or low-resource languages. We observed consistent gains, though compared to prior work we saw conspicuous gains in quite a number of lower-resourced languages. We analyzed differences in translations between baseline and back-translation models, and observed many indications of improved translation quality. Translation of both rare and common terms is improved, and these improvements occur despite the less natural synthetic source-language text used in training.

1 Introduction

Back-translation was applied to statistical machine translation at least as far back as 2009 (Bertoldi and Federico, 2009) with modest gains being reported. Sennrich *et al.* (2016) applied back-translation in NMT and obtained gains of 2-3 BLEU for English/German and about 4 BLEU in Turkish to English. This renewed interest in back-translation and it became a popular technique used in WMT evaluations, particularly in high-resource settings.

Research continued in back-translation, with a paper by Hoang *et al.* (2018) who studied iterative back-translation, where the reverse model is itself improved through back-translation. In low-resource scenarios they observed gains of about 1.5 BLEU, however, the marginal gain of repeated iterations is small. Many studies conducted experiments where a high resource language pair was sampled to artificially create a "low" resource dataset, however, we are concerned that such simulations are not a good proxy due to dissimilar scripts, atypical subject matter, and noisy training data common in low-resource bitext. A few studies have looked look at *bona fide* low-resource language pairs. One example is Xia *et al.* (2019) who found 3+ BLEU point gains in several languages, and even an 8 point gain in Azerbaijani to English.

Other influential works in back-translation include: Edunov *et al.* (2018) who investigated the optimal amount of monolingual data to use in high-resource pairs; Imankulova *et al.* (2017) who examined filtering out lower-quality synthetic bitext pairs; Marie *et al.* (2020) who examined weighting synthetic exemplars differently than human-produced bitext; and Edunov *et al.* (2020) and Graça *et al.* (2019) who studied use of sampling.

Our goal in this study is to reexamine the use of back-translation through extensive experimentation in moderately and low-resourced languages. We believe that this is the largest study to date in terms of the number of languages for which back-translation effectiveness has been analyzed. We describe our experimental setup in Section 2. In Section 3 we compare back-translation to a baseline model for 60 source languages. An analysis of these results is provided in Section 4. In Section 5 we examine the amount of synthetic data to use in six languages. And in Section 6 we report on experiments using repeated back-translation in 13 languages.

2 Methods

In this section we describe model training and the evaluation datasets we use for evaluation.

2.1 Training

Neural machine translation models were trained with the Transformer (Vaswani et al., 2017) using Amazon's Sockeye (v2) toolkit (Apache-2.0) (Hieber et al., 2020). Data was obtained from public sources, in particular, bitext downloadable from the OPUS portal (Tiedemann, 2012). Preprocessing steps included: running the Moses tokenizer;

Findings of the Association for Computational Linguistics: ACL 2023, pages 8166–8183

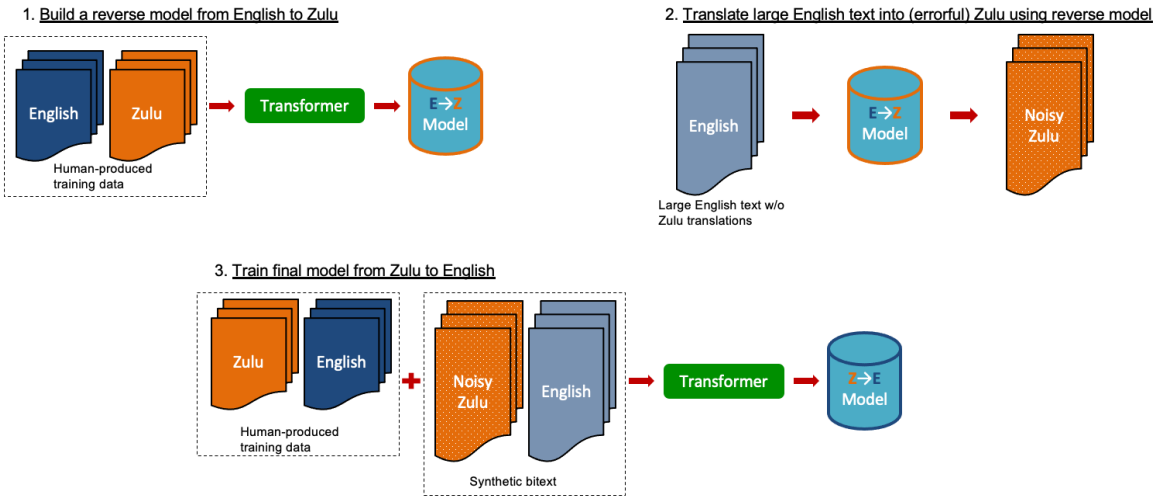

Figure 1: Steps to conduct back-translation, illustrated for a Zulu-to-English model.

removal of duplicate lines; and, learning of subword units using the *subword-nmt* toolkit. Case was retained.

Key hyperparameters include: use of 6 layers in both encoder and decoder; 1,024 dimensional embeddings; 16 attention heads; 4,096 hidden units per layer; 30,000 subword byte pair encoding (BPE) unit, separately in source and target languages; batch size of 1,024; the Adam optimizer with an initial learning rate of 2×10^{-4}. The models were thus trained with a straightforward implementation of the Transformer.

To perform back-translation we used monolingual English text from the web-crawled news portion of the Leipzig corpus[1]. This consisted of 7 million sentences of web-scraped news from 2014 to 2020.[2] There are 1 million sentences available from each year. The training process with back-translation is depicted in Figure 1. In Step 1 a reverse model is trained from the ultimate target language (here always English) *to* the ultimate source language. In Step 2 inference is performed using the reverse model on monolingual text. Finally, in Step 3 a forward model is trained, using a concatenation of the original human-produced bitext and the synthetic bitext from Step 2. This model is independently trained; the only difference compared to a baseline model (labelled 'Base' in Table 1 below) is that the training data has been supplemented.

2.2 Evaluation

The FLORES-101 (Goyal et al., 2022) dataset was created by Facebook Research using content from Wikipedia (*e.g.,* news, travel guides). Translations are from English into 100 other languages, with an emphasis on obtaining translations in lower-resourced languages. 3,001 sentences were split into *test*, *devtest*, and *dev* partitions; we report results on the 1012 sentence *test* set.

TICO-19 (Anastasopoulos et al., 2020) was created as a domain-specific test set to support customization and evaluation of translation models that would be useful during the SARS-COV-2 pandemic. English content from PubMed and various Wikipedia.org projects was translated into 9 higher-resourced languages and 26 lower-resourced languages. The data is provided as *test* (2,100 sents) and *dev* (971 sents) partitions, though we use all 3,071 sentences for testing. Translations are available in 19 of the 60 languages that we studied.

Samples of *flores101* and *tico19* sentences can be found in the Appendix. Translations were scored using case-insensitive BLEU scores (Papineni et al., 2002) calculated with *sacrebleu* (Post, 2018).[3]

3 Results

Scores for the baseline models (Base) and for models trained using back-translation (BT) are shown in Table 1.

The top tier of languages experience gains of 1-2 BLEU points ($\sim$ 4% relative gain); the middle tier sees gains averaging about 5 BLEU (23% relative gain); and, the least resourced languages see average gains of about 8 BLEU (70% relative gain). Languages such as Burmese, Gujarati, Kannada, and Khmer attain *roughly double the score* of their

[1] https://wortschatz.uni-leipzig.de/en/download

[2] Experiments in Sections 5 & 6 use slightly different data.

[3] BLEU+case.lc+numrefs.1+smooth.exp+tok.13a+version.1.4.14

			flores101					*tico19*				
Code	Language	Bitext	M2M	Base	BT	Δ	%	M2M	Base	BT	Δ	%
heb	Hebrew	33.2M	37.9	44.0	45.4	+1.4	+3.2%					
srp	Serbian	32.3M	40.7	42.8	43.4	+0.6	+1.4%					
ind	Indonesian	26.4M	39.6	42.4	44.3	+0.9	+2.1%	43.9	45.1	46.5	+1.4	+3.1%
slv	Slovenian	25.2M	33.4	35.3	36.3	+1.0	+2.8%					
slk	Slovak	22.1M	37.6	38.3	39.7	+1.4	+3.7%					
est	Estonian	21.0M	35.8	37.7	38.5	+0.8	+2.1%					
kor	Korean	15.0M	25.6	29.3	31.0	+1.7	+5.8%					
lit	Lithuanian	14.9M	32.6	33.0	35.0	+2.0	+6.1%					
vie	Vietnamese	14.3M	33.2	35.5	36.7	+1.2	+3.4%					
lav	Latvian	14.2M	34.3	34.9	37.8	+2.9	+8.3%					
fas	Farsi	11.4M	29.9	35.1	37.6	+2.5	+7.1%	30.1	34.3	35.7	+1.4	+4.1%
bos	Bosnian	10.8M	37.6	39.0	41.2	+2.2	+5.6%					
swh	Swahili	9.9M	34.2	40.4	42.8	+2.4	+5.9%	33.0	38.5	40.8	+2.3	+6.0%
ukr	Ukrainian	9.0M	36.3	36.9	39.3	+2.4	+6.5%					
hin	Hindi	8.7M	34.8	35.2	40.9	+5.7	+16.2%	42.6	45.5	49.5	+4.0	+8.8%
tgl	Tagalog	6.3M	27.9	40.3	43.4	+3.1	+7.7%	40.9	49.6	54.8	+5.2	+10.5%
msa	Malay	6.1M	39.4	35.9	39.6	+3.7	+10.3%	45.6	41.2	45.3	+4.1	+10.0%
cat	Catalan	5.2M	43.4	40.3	43.5	+3.2	+7.9%					
isl	Icelandic	5.0M	29.5	31.3	33.7	+2.4	+7.7%					
mkd	Macedonian	4.8M	40.3	40.2	41.6	+1.4	+3.5%					
mlt	Maltese	4.2M	–	49.2	53.5	+4.3	+8.7%					
ben	Bengali	4.0M	28.6	27.3	33.3	+6.0	+22.0%	33.9	30.7	37.8	+7.1	+23.1%
afr	Afrikaans	3.0M	52.7	52.8	53.8	+1.0	+1.9%					
xho	Xhosa	3.0M	18.5	30.9	35.4	+4.5	+14.6%					
zul	Zulu	2.8M	17.9	30.5	35.2	+4.7	+15.4%	24.7	33.8	38.9	+5.1	+15.1%
sna	Shona	2.5M	–	20.5	23.4	+2.9	+14.1%					
gle	Irish	2.4M	1.2	34.4	37.6	+3.2	+9.3%					
hau	Hausa	2.2M	13.9	25.2	29.8	+4.6	+18.3%	16.9	25.9	31.3	+5.4	+ 20.8%
tam	Tamil	1.7M	10.8	20.1	28.6	+8.5	+42.3%	11.6	19.7	29.4	+9.7	+49.2%
urd	Urdu	1.7M	24.6	23.6	29.4	+5.8	+24.6%	26.0	26.5	31.1	+4.6	+17.4%
yor	Yoruba	1.4M	4.8	11.3	14.9	+3.6	+31.9%					
kat	Georgian	1.4M	16.1	17.5	22.3	+4.8	+27.4%					
mal	Malayalam	1.3M	22.9	19.0	31.7	+12.7	+66.8%					
azj	Azerbaijani	1.2M	8.7	13.5	18.5	+5.0	+37.0%					
jav	Javanese	1.2M	23.0	12.4	20.0	+7.6	+61.3%					
mar	Marathi	1.1M	23.5	17.9	29.1	+11.2	+62.6%	24.0	19.5	30.0	+10.5	+53.8%
nya	Nyanja	1.1M	–	15.6	20.4	+4.8	+30.8%					
bel	Belarusian	1.1M	15.2	14.1	16.8	+2.7	+19.1%					
hye	Armenian	983k	22.1	25.4	32.7	+7.3	+28.7%					
amh	Amharic	950k	14.3	19.8	29.5	+9.7	+49.0%					
tel	Telegu	908k	–	23.6	35.9	+12.3	+52.1%					
npi	Nepali	787k	14.0	16.5	29.9	+13.4	+81.2%	23.9	20.2	35.7	+15.5	+76.7%
som	Somali	786k	3.3	14.6	21.5	+6.9	+47.3%	3.0	8.8	12.0	+3.2	+36.4%
cym	Welsh	772k	26.7	40.5	50.8	+10.3	+25.4%					
lin	Lingala	768k	4.0	11.6	19.2	+7.6	+65.5%	6.5	9.2	15.9	+6.7	+72.8%
lug	Ganda	768k	4.0	6.3	11.2	+4.9	+77.8%	8.6	9.3	15.9	+6.6	+71.0%
mya	Burmese	734k	8.4	10.3	19.8	+9.5	+92.2%	12.6	11.1	19.9	+8.8	+79.3%
nso	Pedi	718k	4.0	21.8	31.8	+10.0	+45.9%					
glg	Galician	692k	38.2	33.5	37.0	+3.5	+10.4%					
ceb	Cebuano	691k	21.4	25.6	32.6	+7.0	+27.3%					
orm	Oromo	667k	–	4.5	7.3	+2.8	+62.2%	–	5.6	9.6	+4.0	+ 71.4%
kaz	Kazakh	635k	5.4	16.5	25.7	+9.2	+55.8%					
khm	Central Khmer	634k	14.3	10.5	19.8	+9.3	+88.6%	21.4	14.6	26.1	+11.5	+78.8%
ibo	Igbo	568k	12.5	14.3	19.7	+5.4	+37.8%					
mon	Mongolian	559k	15.8	10.8	18.9	+8.1	+75.0%					
guj	Gujarati	410k	1.6	14.6	29.4	+14.8	+101.3%					
kan	Kannada	390k	0.8	8.3	18.7	+10.4	+125.3%					
tgk	Tajik	386k	–	9.7	17.0	+7.3	+75.3%					
pan	Panjabi	326k	16.3	16.4	27.4	+11.0	+67.1%					
kir	Kirghiz	318k	–	7.6	13.4	+5.8	+76.3%					

Table 1: BLEU scores on the *flores101* and *tico19* benchmarks for our baseline bilingual models (Base), back-translation models trained with the addition of 7M back-translated English sentences (BT), and Facebook's M2M model. Δ shows the absolute BLEU improvement between BT and Base; % shows relative gain. The rows are sorted by training data size for Base (column: Bitext), where the top group has ten million or more lines of training bitext, middle group has between one to ten million lines, and bottom group has less than 1 million lines.

baseline models. While some languages just improve a poor model to a slightly less poor model (*e.g.,* Oromo, 4.5 to 7.3, +62%), several cases are languages that move from a score of 10 to 15 to a score between 20 and 30, an adjustment from poor to good.

Across the different languages the gains on the *tico19* benchmark track gains on the *flores101* test set. This indicates that we did not just get lucky in picking good monolingual data to use for back-translation, since the synthetic bitext works well on both the news/travel text (*flores101*) and the health domain benchmark (*tico19*).

On both tests sets, in every instance, back-translation conferred gains. There is a strong inverse relationship between the amount of training data used in the baseline model and the improvement in BLEU score with back-translation. This is clear in Figure 2 where the less-resourced language are plotted towards the left. It was not clear that models in the impoverished languages would improve given the questionable quality of their reverse models, yet large gains are indeed seen.

We ran a bootstrap resampling test (Koehn, 2004) comparing BT with Base: with the exception of Serbian, all BLEU improvements in BT are statistically significant ($p < 0.05$). This expands the observation of (Guzmán et al., 2019) which measured large BT gains for an earlier version of *flores101* consisting of Nepali and Sinhala.

To give context to our baseline models we also report performance using the 1.2 billion parameter M2M100 model released by Facebook (Fan et al., 2022), which was trained on 7.5 billion sentence pairs. Note that our bilingual models often outperform the multilingual M2M.

In Figure 3 we show examples of translations. Consider the first example, about Portuguese explorer Vasco da Gama. In the Kazakh training data, the explorer's name never occurs, neither in Kazakh nor English. But in the synthetic bitext, the name appears eight times in the monolingual English, and it is correctly back-translated in Kazakh once, along with a couple of partially correct translations and errors. This is apparently enough to learn how to decode the name properly.

4 Analysis of Results

We now provide various analyses to better understand the results in Section 3. Specifically, we are interested to learn why and how back-translation

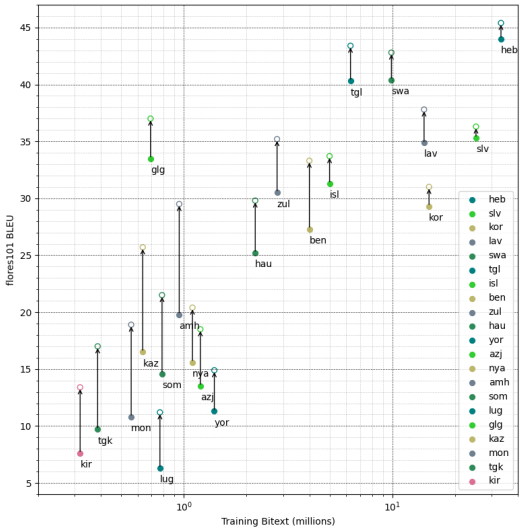

Figure 2: Gains in BLEU on *flores101* using back-translation. The horizontal axis (log scale) is the amount of training data in the Base model. To improve readability only one out of every three languages is plotted.

(BT) improves upon the baseline (Base).

Are the improvements in BT consistent across evaluation metrics? Yes. The histograms in Figure 4 summarize the translation quality in terms of BLEU, chrF (Popović, 2015), and TER (Snover et al., 2006). The BLEU plot corresponds to results in Table 1, and the rightward shift of the BT curve compared to the Base curve indicates the general improvements in BLEU. Both the chrF plot and TER plot shows similar trends of increasing chrF score and decreasing TER score for BT. The improvements are especially pronounced in the low chrF and high TER regions, consistent with our finding about BLEU improving most for low-resource languages.

What kinds of words are translated correctly? Figure 5 shows the precision/recall of out-of-vocabulary (OOV) and high-frequency words, calculated using the `compare-mt` tool (Neubig et al., 2019). We define OOV words as words in the test-set that do not occur in the training text of Base, while frequent words are those with over 1,000 occurrences. For this analysis, MT hypotheses and references in English were processed with the Moses tokenizer (Koehn et al., 2007). We observed improvements in both precision and recall on both classes of words. Figure 3 gave an example of improvement in OOV translation, but in Figure 5 we see that BT improves word precision and recall across the board. In fact, the high-frequency words

Kazakh	Еуропалық ықпал мен отаршылдық 15 ғасырда басталды, өйткені португалдық саяхатшы Васко да Гама Еуропадан Үндістанға дейінгі мұйіс жолын тапты.
Ref	European influence and colonialism began in the 15th century, as Portuguese explorer Vasco da Gama found the Cape Route from Europe to India.
Base	The European promoted and the colonial started in the 15th century, since the Portuguese traveler Vassko also found the way to Hama from Europe to India.
BT	The European influence and colonialism began in the 15th century, as the Portuguese traveler Vasco da Gama found his way up the horn from Europe to India.
Mongolian	Тухайлбал тэдэнд эрдэнэшиш, улаан лооль, төмс, кокоагын аль нь ч байгаагүй ба эртний ямар ч Ром хүн цацагт хяруул хэзээ ч амсаж байгаагүй.
Ref	For instance, they didn't have corn, nor tomatoes, nor potatoes, nor cocoa, and no ancient Roman ever tasted a turkey.
Base	For instance, they had no idea of treasure, tomatoes, potatoes, cocktail, and even ancient Rome had never experience a turkey.
BT	In particular, they had none of the corn, tomatoes, potatoes and coconuts, and in ancient times no Romans had ever tasted the turkeys.

Figure 3: Example translations. Shown are the source sentence, the reference, the baseline model translation, and the translation when back-translation is utilized.

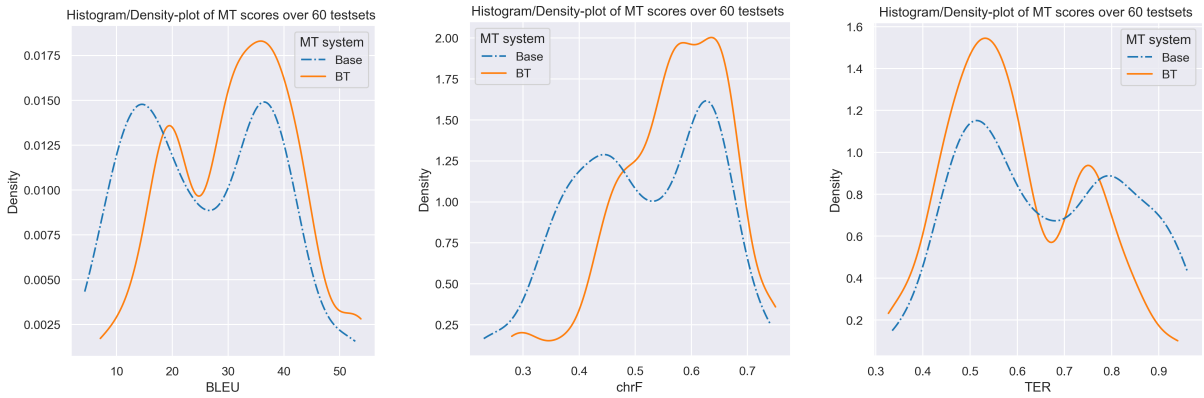

Figure 4: Visualization of Base and BT system scores over the 60 *flores101* testsets. For easier comparison, the BLEU/ChrF/TER scores are tallied into a histogram, then converted to a density plot by kernel density estimation. Note that the scores are generally bimodal. Our explanation of the bimodality is based on the fact that 20-25% of the languages we studied have between 1M and 3M lines of training data; below 1M lines the scores tend to be below 20 BLEU, and above 3M lines the scores tend to be over 30 BLEU. There simply are not as many languages with quantities of training data in the intermediate range. When comparing BT against Base, there is a rightward shift of improving BLEU and chrF scores and a leftward shift of improving (*i.e.,* decreasing) TER.

lead to the most BLEU gain.

We also conduct a word accuracy analysis that groups words by part-of-speech tags. The English reference and hypotheses are tagged with CoreNLP (Manning et al., 2014), then the respective precision and recall values are calculated. We average over the 60 testsets and report the resulting F_1 scores in Table 2. We note that the F_1 measure increases across all parts-of-speech for BT, with the largest gains in nouns, particles, and verbs.

How accurate are the reverse models? Does the reverse model need to be highly accurate for back-translation to perform well? This is a question that is especially pertinent to low-resource conditions. We measure the accuracy of the reverse model that synthesized the 7 million lines of BT data. Since the reference is in a non-English lan-

POS	Share	F_1	%
CC: coord. conjunction	3.3%	0.88	+3%
CD: card. number	1.8%	0.80	+6%
DT: determiner	9.5%	0.67	+7%
IN: preposition	12.3%	0.61	+8%
JJ∗: all adjectives	7.7%	0.56	+12%
MD: modal	1.1%	0.54	+8%
NN∗: all nouns	28.0%	0.61	+14%
PRP: personal pronoun	1.9%	0.61	+10%
RB∗:: all adverbs	4.3%	0.49	+9%
RP: particle	0.2%	0.29	+19%
TO: to	1.3%	0.72	+5%
VB∗: all verbs	14.5%	0.47	+14%
WDT: Wh-determiner	0.6%	0.46	+11%
WP: Wh-pronoun	0.2%	0.52	+13%
WRB: Wh-adverb	0.2%	0.56	+10%
All other tags	13.1%	0.78	+2%

Table 2: F_1 measure of BT word accuracy by POS tag. % indicates the percent improvement over F_1 of Base. Share is the proportion of the tag in the *flores101* testset.

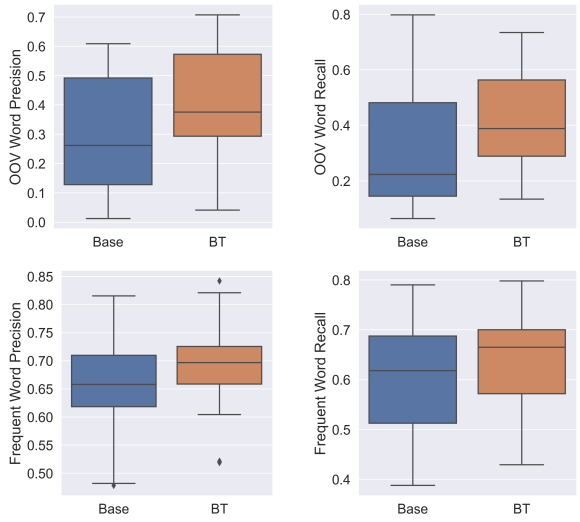

Figure 5: Precision and Recall of OOV (top) and Frequent (bottom) words, aggregated over the 60 *flores101* testsets.

guage, we compute BLEU using sentence-piece tokenization (spBLEU) for consistency of evaluation across languages. Figure 6 is a scatterplot where the x-axis is the BLEU score of a forward baseline model (*e.g.,* zul-eng) and the y-axis is the spBLEU of the reverse model (*e.g.,* eng-zul) trained on the same bitext. For most language-pairs, we see a strong correlation between the two BLEU scores, which is reasonable because both forward and reverse models are trained on the same bitext. For about a fifth of the language pairs, reverse model spBLEU is significantly low (*e.g.,* in the range 0-10) compared to forward model BLEU. These are mostly models for Indian languages (tam, kan, mal) or languages that may be challenging to segment (mya, khm); nevertheless, the BT gains are still rather impressive in these languages. These results suggest that the reverse model does not need to be highly accurate for back-translation to be effective.

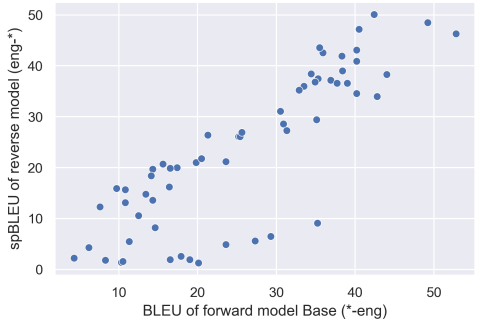

Figure 6: Scatterplot of forward and reverse model accuracies on *flores101*.

What does the BT bitext look like? We attempt to characterize the BT training data by comparing statistics on the foreign and English sides. Figure 7 (top) shows the out-of-vocabulary (OOV) rate (by word types) of the baseline bitext compared to the backtranslated bitext (which includes the baseline bitext). We observe that the English OOV rate on the *flores101* test set is on average 4.5% for Base, and this drops significantly to 1.5% for BT. This shows that the BT data improves coverage on the *flores101* vocabulary. Previous work has shown that one explanation for back-translation's success is the improved coverage in domain mismatch conditions (Dou et al., 2020). We believe there is certainly some of this effect, but the improvements in both *flores101* and *tico19* imply that domain coverage is not the only reason for improvement.

The OOV rate on the foreign side presents an additional explanation. We use the Moses tokenizer and other language-specific tokenizers for this analysis. While the OOV rate on the foreign side is higher (10%), there is still considerable reduction by BT (7.5%). The only way for OOV rate to reduce on the foreign side is for the reverse model to generate via subword unit combinations new words that were previously not seen in the original bitext.

Finally, we train language models (4-gram *kenlm* (Heafield, 2011)) on both sides of the bitext for Base and BT, and measure the perplexity on *flores101* validation set. Here we use subwords as tokens to ameliorate the presence of OOV words, which complicates perplexity calculations. Figure 7 (bottom) shows that perplexity of a 4-gram trained Base English text is approximately 110 on average, and it drops to 100 for a 4-gram trained on BT English text. Surprisingly, the perplexity increases on the foreign side, growing from 75 to 85.

For perplexity, these are minor differences, but we make some conjectures: (1) The small change in perplexity is likely due to the BT data being relatively broad domain; if the BT data were selected to be very similar to the test set, the perplexities would drop much more significantly. (2) The upward trend in perplexity for BT on the foreign side suggests that the synthesized foreign text might not be wholly natural (see Appendix D). These texts do not improve monolingual perplexity, yet when paired as bitext they do improve MT accuracy.

Summary: BT improvements over Base are measured on multiple metrics, and translation improves across the board on all word types. The reverse

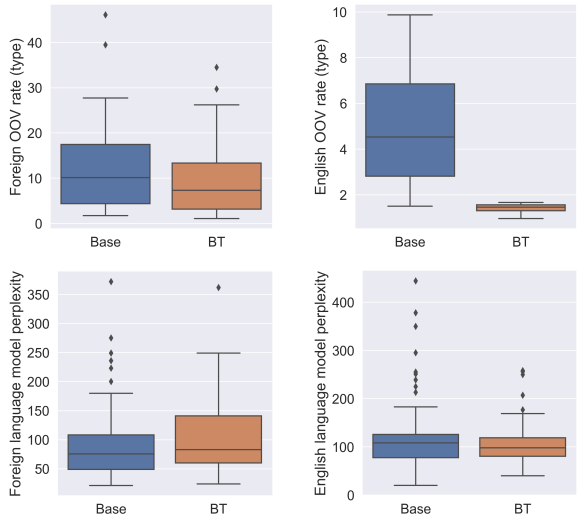

Figure 7: OOV rate (top row) and perplexity (bottom) of BT data on *flores101*, aggregated over 60 languages

model does not need to be highly accurate and the BT bitext (if broad-domain) does not need to be specifically matched to the test domain for BT to work effectively.

5 Monolingual Data Size

Some have studied the effect of the amount of monolingual text used in creating synthetic bitext. A common heuristic is to use a small multiple of the human-produced training bitext, for example two or three times the amount. We wanted to assess this ourselves, and we did this in six languages that varied in the amount of Base training data, from 300k lines of bitext up to 11 million.

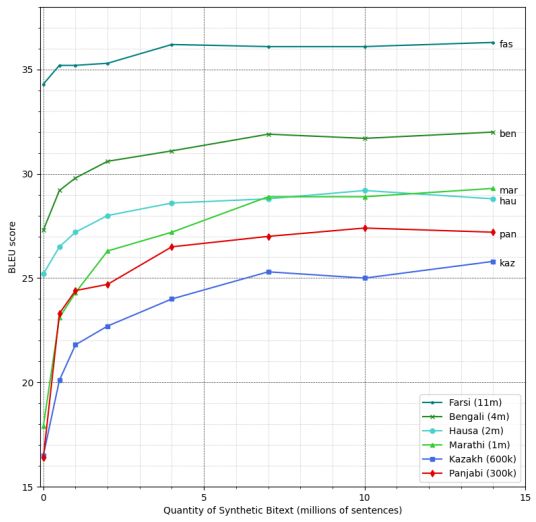

Figure 8: Comparing the amount of monolingual text used in back-translation.

In Section 3 we used 7 million sentences per language from the web-crawled English news portion of the Leipzig corpus. For these new experiments we expand to 14 million sentences from the years 2005 to 2020, training six additional models per language, each using differing amounts of monolingual text. When back-translation is used, we choose the most recent data up to our desired limit.

Figure 8 plots BLEU scores for six languages: Bengali, Farsi, Hausa, Kazakh, Marathi, and Panjabi. They vary in the amount of Base bitext from about 300k lines (Panjabi) up to 11 million lines (Farsi). At the left is the no back-translation condition, and proceeding left to right, larger amounts of synthetic bitext are used.

We make several observations from the plot. First, consistent with Table 1, the three least resourced languages show the greatest gains. Second, even the smallest amount of synthetic data considered, 500k sentences, produced tangible benefit. And third, the four rightmost conditions (*i.e.,* 4, 7, 10, and 14 million) are best, though there is little difference among them. Our earlier choice of 7 million sentences was felicitous.

We conclude that using even relatively small amounts of data can be effective, and that the risk of using too much data is low. For example, with the use of 14 million lines of synthetic bitext, the Panjabi model is using 40x more synthetic data than original human-produced bitext, and this still conveys large gains compared to using less data, and is nearly optimal compared to other choices.

6 Repeated Back-Translation

Earlier work in iterative back-translation (Hoang et al., 2018) showed small gains when first improving the reverse model, and then using that improved model to generate the final synthetic bitext. It makes sense that an improved synthetic bitext should have fewer errors and lead to an ultimately better model. We decided to investigate this method in thirteen languages, using just one attempt to improve the reverse model. This requires monolingual text in the source language to create synthetic data for the reverse model. For non-English text we used data from the OSCAR 22.01 corpus (Abadji et al., 2022), which was filtered to remove possibly problematic text[4] and then performed sentence splitting using *ersatz* (Wicks and Post, 2021).

[4]Anything marked as *adult*, *footer*, *header*, *noisy*, *short_sentences*, or *tiny*.

Lang	Bi/Monotext		*flores101* RBT	*flores101* Δ	*tico19* RBT	*tico19* Δ
tam	1.7M	5.4M	28.6	+0.3	29.4	-0.3
urd	1.7M	3.4M	29.4	0.0	31.1	-0.2
kat	1.4M	4.3M	22.3	+1.3		
azj	1.2M	4.0M	18.5	+0.4		
amh	950k	143k	29.5	+0.8		
tel	908k	1.7M	35.9	+0.8		
mya	734k	339k	19.8	+0.7	19.9	+1.0
kaz	635k	3.0M	25.7	+2.0		
khm	634k	171k	19.8	-0.9	26.1	+0.9
mon	559k	1.2M	18.9	+1.8		
guj	410k	1.1M	29.4	+2.8		
kan	390k	946k	18.7	+4.4		
tgk	386k	1.7M	17.0	+3.3		

Table 3: Results for repeated back-translation (RBT). Resultant BLEU scores are shown for the *flores101* and *tico19* benchmarks, along with the change in BLEU compared to the BT model from Table 1.

This is a somewhat less controlled experiment, as the amount of monolingual text in OSCAR varies by the language. After the filtering mentioned above we used all of the remaining text.

Our results are shown in Table 3. On *flores101*, positive gains were seen in 11 of 13 cases (a tie for Urdu; a loss in Khmer). Changes tended to be small, except in the lesser resourced languages, where gains of between 2.0 and 4.4 points were achieved. On *tico19*, the changes were relatively small, with two minor losses (Tamil and Urdu), and two gains of about a point (Burmese and Khmer).

Back-translation requires training two separate models, one after the other. However, with the extra step of improving the reverse model, we must train a third model. Based on these results, the added expense of improving the reverse model is likely only worthwhile for languages with less than one million lines of human-produced bitext.

7 Related Work

BT for low-resource languages: Most papers on this topic examines some aspect of BT with experiments on specific low-resource languages, e.g.: Telegu (Dandapat and Federmann, 2018); Gujarati (Bawden et al., 2019); Lithuanian, Gujarati (Xu et al., 2019); Tagalog, Swahili, Somali, Turkish (Niu et al., 2019); Swahili (Sánchez-Martínez et al., 2020); Bribri (Feldman and Coto-Solano, 2020); Vietnamese (Li et al., 2020); Tamil, Inuktitut (Chen et al., 2020). Our contribution is orthogonal in that we have an expansive exploration over 60 moderate and low-resource languages.

Two recent survey papers on low-resource translation (Ranathunga et al., 2021; Haddow et al., 2022) mention the importance of data augmentation and back-translation in particular, though neither highlights the outsized impact of back-translation compared to higher resourced settings.

BT variants: Although we use only the most simple BT technique, there are many advanced variants that may be interesting as future work. In addition to the papers on sampling, filtering, and weighting mentioned in the introduction, BT can be improved with meta-learning (Pham et al., 2021), transliteration (Karakanta et al., 2018), data selection (Soto et al., 2020), tagging (Caswell et al., 2019), lexical/syntactic diversity (Burchell et al., 2022).

BT for multilingual models: We focus on bilingual models, but BT for multilingual models is an area of growing interest. Fan *et al.* (2022) observed consistent, yet small gains in multilingual models (seemingly less than 2 BLEU, cf. their Figs. 4 & 6). Our experiments were exclusively bilingual and to-English, with larger gains in low-resource conditions, though direct comparison is not possible.

In a follow-on study (NLLB Team et al., 2022), Meta develop a larger version of the FLORES data in 200 languages, and built a massively multilingual many-to-many model. As part of that wide-ranging work, they conducted experiments with back-translation (their Sec. 6.4.1). Their best results used statistical MT to generate the synthetic bitext. Consistent with our results in translation to English, they found gains largest in "very low" resource languages (50.9 vs. 46.1 chrF++), but using multilingual mixture-of-experts models.

8 Conclusions

By revisiting back-translation for an expansive list of 60 mid- and low-resource languages we have come to a better understanding of the landscape. We found that:

- Back-translation improves performance in moderately resourced languages, but is significantly more effective in improving translation quality in low-resource languages with less than 1 million lines of training bitext.
- Translation of rare terms is improved due to increased lexical coverage in the synthetically generated bitext; however, translation of frequently occurring terms is also improved.
- Even when initial models are of low quality,

and the synthetic bitext contains noise, significant gains still occur.

- The risk of using too much synthetic data is low.
- Repeated back-translation imparts only minor gains, except in some of the least resourced cases we studied.

Limitations

Aside from the reverse models used in back-translation (which we did analyze in Section 4), we only studied translation of language pairs into English. Using data augmentation techniques like back-translation where English is not the target language, or is neither the source or target language is certainly worthy of study, but was out of scope in the present work. We did however, include many source languages that are typologically different from English (see Table 8 in the Appendix).

In order to study the effectiveness of BT in a large number of languages we relied on extant multilingual datasets, namely *flores101* and *tico19*. The direction of human translation when building these datasets was from English into another language.

We did not run repeated trials on our experiments. Many models required training for a couple of GPU-weeks on V100s, and additional trials would have added significant computational expense. We believe the trends we have identified are sufficiently clear and supported by the statistical analysis in Section 4.

Ethics Statement

Our goal in this work is to contribute to an understanding of how and when back-translation can be successfully employed when translating out of moderate- and low-resource languages. We believe that improving translation where English is the target language has utility both for its 1.5 billion L1 and L2 speakers globally, as well as for those non-English speakers whose content can be made accessible to additional communities.

State-of-the-art systems will make errors, including failing to resolve ambiguity, mistranslating proper names, hallucinations, subject-verb disagreement, among others. These errors could lead to harms if automated translations are used injudiciously by end users. Translation in low-resource conditions is inherently error-prone, however, based on our results, we believe that using back-translation will often lead to more robust translations.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. *arXiv e-prints*, page arXiv:2201.06642.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Rachel Bawden, Nikolay Bogoychev, Ulrich Germann, Roman Grundkiewicz, Faheem Kirefu, Antonio Valerio Miceli Barone, and Alexandra Birch. 2019. The University of Edinburgh's submissions to the WMT19 news translation task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 103–115, Florence, Italy. Association for Computational Linguistics.

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 182–189, Athens, Greece. Association for Computational Linguistics.

Laurie Burchell, Alexandra Birch, and Kenneth Heafield. 2022. Exploring diversity in back translation for low-resource machine translation. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 67–79, Hybrid. Association for Computational Linguistics.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

Peng-Jen Chen, Ann Lee, Changhan Wang, Naman Goyal, Angela Fan, Mary Williamson, and Jiatao Gu. 2020. Facebook AI's WMT20 news translation task submission. In *Proceedings of the Fifth Conference on Machine Translation*, pages 113–125, Online. Association for Computational Linguistics.

Sandipan Dandapat and Christian Federmann. 2018. Iterative data augmentation for neural machine translation: a low resource case study for english–telugu. In *Proceedings of the Conference of the European Association for Machine Translation*.

Zi-Yi Dou, Antonios Anastasopoulos, and Graham Neubig. 2020. Dynamic data selection and weighting for iterative back-translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5894–5904, Online. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2021. *Ethnologue: Languages of the World.* SIL International, Dallas, Texas.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846, Online. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2022. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Isaac Feldman and Rolando Coto-Solano. 2020. Neural machine translation models with back-translation for the extremely low-resource indigenous language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Miguel Graça, Yunsu Kim, Julian Schamper, Shahram Khadivi, and Hermann Ney. 2019. Generalizing back-translation in neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 45–52, Florence, Italy. Association for Computational Linguistics.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

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

Felix Hieber, Tobias Domhan, Michael Denkowski, and David Vilar. 2020. Sockeye 2: A toolkit for neural machine translation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 457–458, Lisboa, Portugal. European Association for Machine Translation.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Aizhan Imankulova, Takayuki Sato, and Mamoru Komachi. 2017. Improving low-resource neural machine translation with filtered pseudo-parallel corpus. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 70–78, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Alina Karakanta, Jon Dehdari, and Josef van Genabith. 2018. Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, 32:167–189.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej 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, Prague, Czech Republic. Association for Computational Linguistics.

Hongzheng Li, Jiu Sha, and Can Shi. 2020. Revisiting back-translation for low-resource machine translation between chinese and vietnamese. *IEEE Access*, 8:119931–119939.

Yin Lou, Rich Caruana, and Johannes Gehrke. 2012. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, page 150–158, New York, NY, USA. Association for Computing Machinery.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky.

2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Benjamin Marie, Raphael Rubino, and Atsushi Fujita. 2020. Tagged back-translation revisited: Why does it really work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5990–5997, Online. Association for Computational Linguistics.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A tool for holistic comparison of language generation systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 35–41, Minneapolis, Minnesota. Association for Computational Linguistics.

Xing Niu, Weijia Xu, and Marine Carpuat. 2019. Bidirectional differentiable input reconstruction for low-resource neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 442–448, Minneapolis, Minnesota. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Hieu Pham, Xinyi Wang, Yiming Yang, and Graham Neubig. 2021. Meta back-translation. In *International Conference on Learning Representations (ICLR)*, Online.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2021. Neural machine translation for low-resource languages: A survey.

Felipe Sánchez-Martínez, Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, Mikel L. Forcada, Miquel Esplà-Gomis, Andrew Secker, Susie Coleman, and Julie Wall. 2020. An English-Swahili parallel corpus and its use for neural machine translation in the news domain. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 299–308, Lisboa, Portugal. European Association for Machine Translation.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Xabier Soto, Dimitar Shterionov, Alberto Poncelas, and Andy Way. 2020. Selecting backtranslated data from multiple sources for improved neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3898–3908, Online. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Rachel Wicks and Matt Post. 2021. A unified approach to sentence segmentation of punctuated text in many languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3995–4007, Online. Association for Computational Linguistics.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796, Florence, Italy. Association for Computational Linguistics.

Nuo Xu, Yinqiao Li, Chen Xu, Yanyang Li, Bei Li, Tong Xiao, and Jingbo Zhu. 2019. Analysis of back-translation methods for low-resource neural machine translation. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II*, page 466–475, Berlin, Heidelberg. Springer-Verlag.

A Testset Examples

Examples from the *flores101* test partition are shown in Table 4. The text is rich in named-entities and multiword expressions. Examples from *tico19* are shown in Table 5. The language contains terminology specific to the medical and public health communities, and some texts are written in a scientific style.

B Correlation between *flores101* and *tico19*

In Section 3 we mentioned that the relative gain on the public-health related *tico19* dataset tracked the improvement seen on *flores101*. Figure 9 is a scatterplot of the relative gains of both datasets. We calculated Pearson's correlation coefficient to be 0.979.

C Can we find features that quantitatively explain BT improvements?

We attempt to define features x for each language-pair and build a glassbox regression model to predict y, defined as the percentage improvement when comparing BT BLEU with Base BLEU (*e.g.,* the column % in Table 1). The goal is to find explainable features that predict when BT improvement will be large or small. As a glassbox model, we use the Explainable Boosting Machine (EBM),

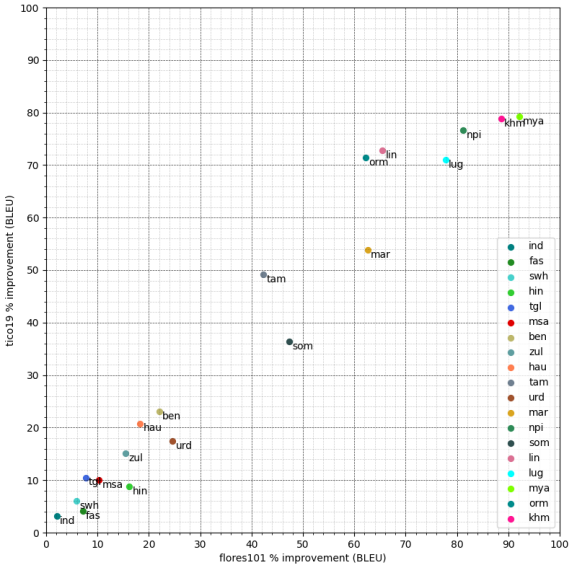

Figure 9: Scatterplot of relative gain on *tico19* vs. *flores101* for BT model.

which is introduced in (Lou et al., 2012) and implemented in Nori et al. (2019):

$$g(y) = \beta_0 + \sum_j f_j(x_j) \quad (1)$$

Here, g is a link function (identity for regression), x_j is a feature we manually define, and f_j the shape function for feature x_j that is learnt through bagging and gradient boosting. The advantage of EBM over conventional linear regression is that the f_j can be of arbitrary shape (leading to low mean-squared error) and yet can be easily interpretable (similar to decision trees).

We define the following features:

- `train_token`: number of tokens for training, in millions
- `oov_type`: the OOV rate, by type
- `tt_ratio`: type-to-token ratio, number of distinct word types divided by number of tokens (computed on the testset)
- `perplexity`: perplexity of the aforementioned 4-gram language model

Each feature is prefixed with (`en`,`fr`) to indicate that it is computed on the English or foreign side, respectively. Additionally, each feature is suffixed with (`1`, `2`) to indicate that it is computed on Base (`1`) or BT (`2`).

We have available only 60 "samples" for EBM: a random 85% is used for fitting the EBM and

The JAS 39C Gripen crashed onto a runway at around 9:30 am local time (0230 UTC) and exploded, closing the airport to commercial flights.
Around 11:29, the protest moved up Whitehall, past Trafalgar Square, along the Strand, passing by Aldwych and up Kingsway towards Holborn where the Conservative Party were holding their Spring Forum in the Grand Connaught Rooms hotel.
Nadal's head to head record against the Canadian is 7–2.

Table 4: Several examples from *flores101*. Only the English text is shown.

In ca 14% cases, COVID-19 develops into a more severe disease requiring hospitalisation while the remaining 6% cases experience critical illness requiring intensive care.
On 11 March 2020, the Director General of the World Health Organization (WHO) declared COVID-19 a pandemic.
Patients with severe respiratory symptoms have to be supported by extracorporeal membrane oxygenation (ECMO), a modified cardiopulmonary bypass technique used for the treatment of life-threatening cardiac or respiratory failure.

Table 5: Several examples from *tico19*. Only the English text is shown.

15% for test. While the sample size is small, the model is simple and the coefficient of determination (R^2) on the test set is a reasonable 0.7. We show the EBM interpretation results in Figure 10. According to this model, the `en_train_token_2` and `fr_oov_type_1` are the top two features for predicting the improvement in BLEU (y). A visualization of the the shape functions show that low values of `en_train_token_2` lead to high score (high y); this coincides with the previous observation that lower-resourced languages saw more improvements in BLEU. The shape function for `fr_oov_type_1` shows an interesting step function at around 10, meaning that systems with a foreign word OOV rate greater than 10% had a large amount to gain in BLEU.

We should note that this EBM analysis only shows correlation, not causation.

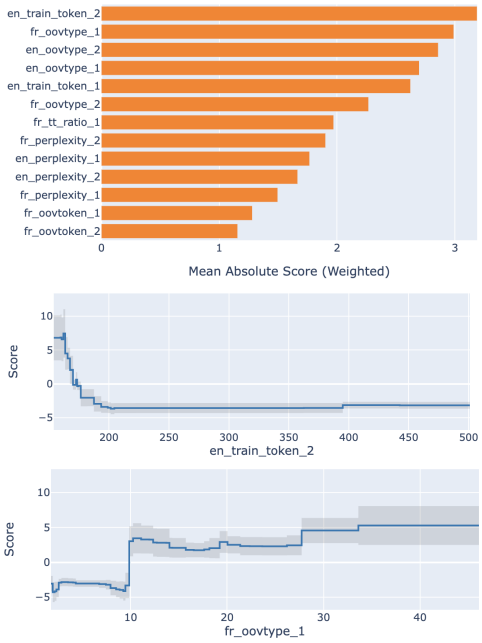

Figure 10: EBM: explaining when BT improves the most. The top figure shows which features are most important in the EMB. The bottom two figures are example shape functions.

D Quality in Reverse Models

In Section 4 we mentioned that back-translation can still be effective despite significant noise in the reverse models. In fact, in some languages, significant numbers of exact match hallucinations are produced. Some frequently repeated lines from the Javanese synthetic bitext are listed in Table 6.

Detecting and filtering out implausible sentence pairs is one approach to mitigate this problem Imankulova et al. (2017), however, in our work we simply removed any duplicates, so that at most one spurious example remained instead of possibly thousands. Despite the residual noise, back-translation is remarkably effective in these low-resource languages. In Table 7 we list the number of unique lines of back-translated text (*i.e.,* on the non-English side) for certain languages in which we observed this problem.

E Computational Expense

Our computing infrastructure consisted of a mix of NVIDIA V100 32GB and A100 40GB machines. We estimate that model training and decoding required 41,000 GPU-hours for the experiments reported in this paper. We are not able to estimate the actual carbon footprint incurred due to many factors involved, but we can estimate it for a given scenario as follows. If we take 250 watts (the rating for a V100), that is 10.25 MWh. If we assume a CO_2e emission of 432 kg/MWh, we end up with:

$$\frac{10.25 \text{ MWh}}{1} \times \frac{432 \text{ kg}}{\text{MWh}} \times \frac{1 \text{ ton}}{907.19 \text{ kg}} = 4.9 \text{ tons} \quad (2)$$

Count	Sentence
6,784	]]]] iku kecamatan ing Kabupaten Sumba Tengah Propinsi Nusa Tenggara Wétan. **GT:**]]]] is a district in Central Sumba Regency, East Nusa Tenggara Province.
6,684	Motorola C115 ya iku tilpun sélulèr kang diprodhuksi déning pabrikan Motorola. **GT:** Motorola C115 is a mobile phone produced by Motorola.
1,528	Kutha iki dumunung ing sisih kidul. **GT:** The city is located in the south.
1,463	Nokia N80 ya iku tilpun sélulèr kang diprodhuksi déning pabrikan Nokia. **GT:** Nokia N80 is a mobile phone produced by the manufacturer Nokia.
1,269	Kuwi sing paling penting banget. **GT:** That is the most important thing.
1,246	Kemangga iki racaké akèh tinemu ing Amérika Sarékat. **GT:** Many of these mangoes are found in the United States.

Table 6: Some commonly repeated lines in the Javanese synthetic bitext, with a English translation below obtained from Google Translate (GT). These lines are hallucinations due to the impoverished English-to-Javanese reverse model.

Lang	Uniq	Base	BT	%
Indonesian	6,901,700	42.4	44.3	+1.4%
Oromo	6,746,709	4.5	7.3	+62.2%
Kannada	6,716,601	8.3	18.7	+125.3%
Javanese	6,446,456	12.4	20.0	+61.3%
Tajik	6,018,847	9.7	17.0	+75.3%

Table 7: Number of unique lines (foreign side) of synthetic bitext. In total, 6,920,211 lines were back-translated. Little duplication is present in the Indonesian data, but the problem is significant in Oromo, Kannada, Javanese, and and Tajik. Base and BT BLEU scores and relative improvement are from Table 1.

Further, if we assume the data center power usage effectiveness (PUE) is 1.5 and there are no additional offsets for renewable energy, the CO_2e emission might be $4.9 \times 1.5 = 7.35$ tons.

Our Transformer models average about 275 million parameters.

F Language Properties

Table 8 lists some of the properties of the languages investigated in this work.

Code	Language	Family	Script	Speaker	Example Region	Type: MorphoSyntax, Phonology, etc.
heb	Hebrew	Afro-Asiatic, Semitic	Hebrew	9.4m	Israel	SVO, 22c/5v/4d
srp	Serbian	Indo-European, Balto-Slavic	Cyrillic	10.3m	Serbia	SVO, 7 cases, 25c/5v
ind	Indonesian	Austronesian, Malayo-Polynesian	Latin	199.0m	Indonesia	SVO, 19c/6v/3d
slv	Slovenian	Indo-European, Balto-Slavic	Latin	2.2m	Slovenia	SVO, 6 cases, 21c/8v/2d
slk	Slovak	Indo-European, Balto-Slavic	Latin	7.2m	Slovakia	SVO, 6 case, 27c/10v/4d
est/ekk	Estonian	Uralic, Finnic	Latin	1.2m	Estonia	SVO, 14 cases
kor	Korean	Koreanic	Hangul	81.5m	South Korea	SOV, 6 cases, 21c/8v/12d
lit	Lithuanian	Indo-European, Balto-Slavic	Latin	2.9m	Lithuania	SVO, 6 cases, 37c/10v, tonal
vie	Vietnamese	Austro-Asiatic, Mon-Khmer	Latin	76.8m	Vietnam	SVO, 25c/11v/20d, 6 tones
lav/lvs	Latvian	Indo-European: Balto-Slavic	Latin	2.0m	Latvia	SVO, 5 case, 25c/11v/5d
fas/pes	Farsi	Indo-European, Indo-Iranian	Arabic	74.2m	Iran	SOV, 23c/6v
bos	Bosnian	Indo-European, Balto-Slavic, Slavic	Latin	2.7m	Bosnia&Herzegovina	SVO, 7 cases, 25c/5v
swh	Swahili	Niger-Congo, Atlantic-Congo	Latin	69.2m	Tanzania	SVO, 18 noun classes
ukr	Ukrainian	Indo-European, Balto-Slavic	Cyrillic	33.2m	Ukraine	SVO, 7 cases, 30c/6v
hin	Hindi	Indo-European, Indo-Iranian	Devanagari	600.5m	India	SOV, 30c/10v/2d
tgl	Tagalog	Austronesian, Malayo-Polynesian	Latin	25.7m	Philippines	VSO, 16c/5v
msa/zsm	Malay	Austronesian, Malayo-Polynesian	Latin	81.6m	Malaysia	SVO
cat	Catalan	Indo-European, Italic, Romance	Latin	9.2m	Spain	SVO, 22c/7v/4d
isl	Icelandic	Indo-European, Germanic	Latin	0.3m	Iceland	SVO, 4 cases, 20c/8v/5d
mkd	Macedonian	Indo-European, Balto-Slavic	Cyrillic	1.7m	North Macedonia	SVO, 26c/5v
mlt	Maltese	Afro-Asiatic, Semitic	Latin	0.5m	Malta	SVO, 23c10v8d
ben	Bengali	Indo-European, Indo-Iranian	Bengali	267.7m	Bangladesh	SOV, 5 cases, 35c/5v
afr	Afrikaans	Indo-European, Germanic	Latin	17.6m	South Africa	SVO, sometimes SOV, 20c/16v/9d
xho	Xhosa	Niger-Congo, Atlantic-Congo	Latin	19.2m	South Africa	SVO, 17 noun classes, 58c/10v, 2 tones
zul	Zulu	Niger-Congo, Atlantic-Congo	Latin	27.8m	South Africa	SVO, 13 noun classes, 30c/10v
sna	Shona	Niger-Congo, Atlantic-Congo	Latin	9.0m	Zimbabwe	SVO, 13 noun classes, 31c/5v/2d, 2 tones
gle	Irish	Indo-European, Celtic	Latin	1.2m	Ireland	VSO, 3 cases, 32c/11v/4d
hau	Hausa	Afro-Asiatic, Chadic	Latin	74.9m	Nigeria	SVO, 33c/10v/2d, 2 tones
tam	Tamil	Dravidian, Southern	Tamil	85.5m	India	SOV, 8 cases, 18c/10v/2d
urd	Urdu	Indo-European, Indo-Iranian	Arabic	230.1m	Pakistan	SOV, 30c/20v/2d
yor	Yoruba	Niger-Congo, Atlantic-Congo	Latin	43.0m	Nigeria	SVO, 17c/11v, 3 tones
kat	Georgian	Kartvelian, Georgian	Georgian	3.9m	Georgia	SOV, 18 cases 27c/5v
mal	Malayalam	Dravidian, Southern	Malayalam	37.9m	India	SOV, 7 cases 37c/11v/4d
azj	Azerbaijani	Turkic, Southern	Latin	9.2m	Azerbaijan	SOV, 6 cases, 24c/9v
jav	Javanese	Austronesian, Malayo-Polynesian	Latin	68.3m	Indonesia	SVO, 21c/8v
mar	Marathi	Indo-European, Indo-European	Devanagari	99.1m	India	SOV, 7 cases, 37c/8v/2d
nya	Nyanja	Niger-Congo, Atlantic-Congo	Latin	14.4m	Malawi	SVO
bel	Belarusian	Indo-European, Balto-Slavic, Slavic	Cyrillic	3.9m	Belarus	SVO, 6 cases, 37c/6v
hye	Armenian	Indo-European, Armenian	Armenian	3.8m	Armenia	SVO, 7 cases, 30c/7v
amh	Amharic	Afro-Asiatic, Semitic	Ethiopic	57.4m	Ethiopia	SOV 4 cases, 27c/7v
tel	Telegu	Dravidian, South Central	Telegu	95.6m	India	SOV, 7 cases, 21c/11v
npi	Nepali	Indo-European, Indo-Iranian	Devanagari	24.7m	Nepal	SOV, 11 noun classes, 4 cases, 29c/11v
som	Somali	Afro-Asiatic, Cushitic	Latin	21.9m	Somalia	SOV, 22c/10v, 3 tones
cym	Welsh	Indo-European, Celtic	Latin	0.6m	United Kingdom	VSO, 23c/12v/8d
lin	Lingala	Niger-Congo, Atlantic-Congo	Latin	2.3m	D.R. Congo	SVO, 12 noun classes, 16c/5v, 2 tones
lug	Ganda	Niger-Congo, Atlantic-Congo	Latin	11.0m	Uganda	SVO
mya	Burmese	Sino-Tibetan, Tibeto-Burman	Myanmar	43.0m	Myanmar	SOV, 31c/8v/4d, 3 tones
nso	Pedi	Niger-Congo, Atlantic-Congo	Latin	13.7m	South Africa	SVO
glg	Galician	Indo-European, Italic, Romance	Latin	3.1m	Spain	SVO
ceb	Cebuano	Austronesian, Malayo-Polynesian	Latin	15.9m	Phillippines	VSO, 16c/3v/4d
orm/gaz	Oromo	Afro-Asiatic, Cushitic	Latin	19.2m	Ethiopia	SOV, 7 cases, 25c/10v
kaz	Kazakh	Turkic, Western	Cyrillic	13.2m	Kazakhstan	SOV, 7 cases, 18c/9v
khm	Central Khmer	Austro-Asiatic, Mon-Khmer	Khmer	17.9m	Cambodia	SVO, 21c/17v/13d
ibo	Igbo	Niger-Congo, Atlantic-Congo	Latin	29.0m	Nigeria	SVO, 37c8v, 3 tones
mon/khk	Mongolian	Mongolic, Eastern	Cyrillic	2.7m	Mongolia	SOV, 7 cases, 29c/14v/4d
guj	Gujarati	Indo-European, Indo-Iranian	Gujarati	61.9m	India	SOV, 6 cases, 31c/8v/2d
kan	Kannada	Dravidian, South	Kannada	58.6m	India	SOV, 7 cases, 22c/20v/2d
tgk	Tajik	Indo-European, Indo-Iranian	Cyrillic	8.1m	Tajikistan	SOV, 27c/6v
pan	Panjabi	Indo-European, Indo-Iranian	Gurmukhi	52.2m	India	SOV, 7 cases, 15c/24v, 3 tones
kir	Kirghiz	Turkic, Western	Cyrillic	5.4m	Kyrgyzstan	SOV, 7 cases, 19c/8v

Table 8: Properties of the 60 languages investigated in this paper, according to Ethnologue (Eberhard et al., 2021). **Code** is the ISO 639-3 language code. When two ISO codes are given (e.g. lav/lvs), the first is the macrolanguage code used in FLORES101 and the second refers to the individual language code we used to look up the language properties: e.g. Latvian (lav) is considered a macrolanguage that includes both Standard Latvian (lvs) and Latagalian (ltg), and the table focuses on one (lvs) for concreteness. A macrolanguage is defined by ISO 639-3 as consisting of "multiple, closely related individual languages that are deemed in some usage contexts to be a single language." **Family** indicates the language family classification. **Script** indicates the writing system used in our data, but note that in practice these languages could be written by other scripts (e.g. Kazahk may be written in both Cyrillic and Latin alphabets). **Speaker** is the estimated worldwide population of L1 and L2 speakers (in millions). Rows in this table are sorted by the amount of bitext resources, so note that #speaker does not correlate with resource availability, as is true in many "low-resource" problems. **Region** indicates an example geographic region where the language is recognized as well-established: languages in Asia, Europe, and Africa are represented, but missing are those from Oceania and the Americas. The last column lists a few interesting linguistic properties, when available. For **morphosyntax**, we list the common word order (e.g. SVO, SOV, VSO), number of noun classes, or number of cases. For **phonology**, the format "22c/5v/4d" in e.g. Hebrew means 22 consonants, 5 vowels, and 4 dipthongs.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations section (unnumbered, before references).

☑ A2. Did you discuss any potential risks of your work?
Ethics section (unnumbered, before references).

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract & 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

2

☑ B1. Did you cite the creators of artifacts you used?
2

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
2

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We are not releasing any artifacts. Our use of open source software was consistent with its intended use.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We did not collect any data. We used existing open source datasets (i.e., OPUS bitext).

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Appendix F.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
2

C ☑ Did you run computational experiments?

3, 5, & 6.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix E.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
2

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
2 & 4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.