BITEXTEDIT: Automatic Bitext Editing for Improved Low-Resource Machine Translation

Anonymous ACL submission

Abstract

Mined bitexts can contain imperfect translations that yield unreliable training signals for Neural Machine Translation (NMT). While filtering such pairs out is known to improve final model quality, we argue that it is suboptimal in low-resource conditions where even mined data can be limited. In our work, we propose instead, to refine the mined bitexts via automatic editing: given a sentence in a language $\mathbf{x_f}$, and a possibly imperfect translation of it $\mathbf{x_e}$, our model generates a revised version $\mathbf{x'_f}$ or $\mathbf{x'_e}$ that yields a more equivalent translation pair (i.e., $<\mathbf{x_f}, \mathbf{x'_e}>$ or $<\mathbf{x'_f}, \mathbf{x_e}>$). We use a simple editing strategy by (1) mining potentially imperfect translations for each sentence in a given bitext, (2) learning a model to reconstruct the original translations and translate, in a multi-task fashion. Experiments demonstrate that our approach successfully improves the quality of CCMatrix mined bitext for 5 low-resource language-pairs and 10 translation directions by up to $\sim$ 8 BLEU points, in most cases improving upon a competitive back-translation baseline.

1 Introduction

Neural Machine Translation (NMT) for low-resource languages is challenging due to the scarcity of bitexts (i.e., translated text in two languages) (Koehn and Knowles, 2017). Models are often trained on heuristically aligned (Resnik, 1999; Bañón et al., 2020; Esplà et al., 2019) or automatically mined data (Schwenk et al., 2021a,b), which can be low quality (Briakou and Carpuat, 2020; Caswell et al., 2021). This data can include errors that range from small meaning differences or partial mistranslations to major differences that yield completely incorrect translations and random noise (e.g., empty sequences, text in the wrong language, etc.).

Filtering out noisy samples from web-crawled bitexts is therefor standard practice for building high quality models (Koehn et al., 2018), and is particularly helpful in low-resource settings (Koehn et al., 2019, 2020). Despite the popularity of this approach, we argue it has two key limitations: First, partially correct translations provide signal that is lost if the entire example is dropped (see first sample bitext in Figure 1). Second, filtering out samples exacerbates the data scarcity problem for the long-tail of low-resource language-pairs.

Figure 1: Noisy bitexts consist of a mixture of good-quality, imperfect, and poor-quality translations. Filtering decreases the size of training samples which is crucial for low-resource NMT. Our approach, alternatively, revises noisy bitexts via utilizing imperfect translations in a more effective way, while keeps the size of training data untouched.

In this paper, we instead aim to make use of as much of the signal from the mined bitext as possible. We propose an **editing approach to bitext quality improvement**. Our model takes as input a bitext (i.e., $(\mathbf{x_f}, \mathbf{x_e})$), and edits one of the two sentences to generate a refined version of the original (i.e., $\mathbf{x'_f}$ or $\mathbf{x'_e}$) as necessary. By framing the problem as a bitext editing (BITEXTEDIT) task, we can perform a wide range of operations from *copying* good-quality bitext, to *partial editing* of small meaning mismatches, and *translating* from scratch

incorrect references. Following previous extrinsic evaluations of bitext quality (Koehn et al., 2019, 2020; Schwenk et al., 2021b,a), we compare NMT models trained on the original and revised versions of CCMatrix bitexts. Concretely, we report consistent improvements in translation quality for 10 low-resource NMT translation tasks: EN↔OC, IT↔OC, EN↔BE, EN↔MR, and EN↔SW, while in most cases we even improve upon a competitive back-translation baseline. Crucially, BITEXTEDIT yields from ~ 4, up to ~ 8 BLEU point improvements in the more data-scarce settings (i.e., EN-OC, IT-OC). Additionally, our quantitative and qualitative analysis indicate that BITEXTEDIT improves bitext quality in higher-resource settings with lighter editing that targets more fine-grained meaning differences.

2 Background

Bitext Mining The idea of using the web as a source of parallel texts has a long history (Resnik, 1999). Recent advances in multilingual representation learning (Artetxe and Schwenk, 2019; Liu et al., 2020) enable the curation of mined bitexts across multiple languages. For instance, combining LASER (Artetxe and Schwenk, 2019) embeddings with nearest neighbor search allows for effective bitext mining from Wikipedia (i.e., WikiMatrix (Schwenk et al., 2021a)) and CommonCrawl monolingual texts (i.e., CCMatrix (Schwenk et al., 2021b)). While the latter approach requires parallel text supervision to train the multilingual encoder, Tran et al. (2020) shows that it can be extended to an unsupervised framework via iterative self-supervised training (CRISS).

Issues in Bitext Quality Caswell et al. (2021) manually audit the quality of multilingual datasets in 205 language-specific corpora that result from automatic curation pipelines, including bitexts from CCAligned (El-Kishky et al., 2020), WikiMatrix (Schwenk et al., 2021a), and ParaCrawl (Bañón et al., 2020; Esplà et al., 2019). All have systematic issues, especially for low-resource languages. The vast majority of low-resource pairs contain less than 50% valid translations. However, they do often share structural similarity and partial content. Briakou and Carpuat (2020)—in a more fine-grained annotation study—highlight that small content mismatches are even found in high resource pairs: 40% of English-French WikiMatrix sentence-pairs have *small meaning mismatches*. Our work aims at improving bitext quality via eliminating their systematic issues via editing.

Bitext Quality vs. NMT Khayrallah and Koehn (2018) demonstrate the often significant impact of various types of noise on NMT, via increasing the percentage of 5 types of artificially injected errors on a clean English-German corpus—mimicking frequent issues in parallel texts (i.e., copying, wrong language, non-linguistic content, short segments, empty sequences). Ott et al. (2018) also argue that data uncertainty resulting from noisy references contributes to the miscalibration of NMT models. In contrast with prior studies that discuss how noise interacts with NMT training *solely* for high-resource pairs, we *primarily* focus on low-resource settings and improve NMT models by improving their training bitexts.

Bitext Filtering The most standardized approach to improving bitext either discards an example or treats it as a perfect training instance (Koehn et al., 2018). Past submissions to the "Parallel Corpus Filtering" WMT shared task employ a diverse set of approaches covering simple pre-filtering rules based on language identifiers and sentence features (Rossenbach et al., 2018; Lu et al., 2018; Ash et al., 2018), learning to weight scoring functions based on language models, extracting features from neural translation models and lexical translation probabilities (Sánchez-Cartagena et al., 2018), combining pre-trained embeddings (Papavassiliou et al., 2018), and dual-cross entropy (Chaudhary et al., 2019). Our work builds on top of prior work and instead of filtering out all the imperfect translation candidates, we selectively edit them and keep them in the pool of training data for NMT.

3 Approach: BITEXTEDIT

We frame bitext refinement as an editing task (i.e, BITEXTEDIT) that takes two *input sentences*: a sentence $\mathbf{x_f}$ in language $\mathbf{f}$ and a sentence $\mathbf{x_e}$ in language $\mathbf{e}$, and aims at editing *one* of them (i.e., it *outputs* $\mathbf{x'_f}$ or $\mathbf{x'_e}$) with the goal of yielding a more equivalent translation pair (i.e., <$\mathbf{x_f}$, $\mathbf{x'_e}$> or <$\mathbf{x'_f}$, $\mathbf{x_e}$>). Figure 2 gives an overview of our approach while below we describe the bitext refinement model (§3.1) and the curation of data needed to train our model (§3.2).

3.1 Bitext Editing

Architecture Our bitext editing model is a transformer seq2seq architecture. Each bitext $(\mathbf{x_f}, \mathbf{x_e})$ is encoded via adding position embeddings that are reset for each input sentence to facilitate their alignment (Conneau and Lample, 2019) and two language embeddings, initialized at random, to indicate the two languages for the editing model. The decoder generates autoregressively a refined version of $\mathbf{x_f}$ or $\mathbf{x_e}$, where the first generated token indicates which of the two input sentences is edited, as described below.

Learning During training, we optimize the multi-task loss presented in Equation 1, which has two components. The first represents a *edit-based reconstruction loss* (i.e., $\mathcal{L}_{\text{EDIT}}$) that reconstructs one of the two sentences e.g., $\mathbf{x_f}$ started from a noised version of the original bitexts e.g., $\mathbf{x'_f}$ and $\mathbf{x_e}$. We make this loss bi-directional via adding a symmetrical loss that reconstructs $\mathbf{x_e}$ from $\mathbf{x_f}$ and $\mathbf{x'_e}$, respectively. The second component, is implemented as a bi-directional *translation loss* (i.e., $\mathcal{L}_{\text{MT}}$) via masking the inputs of the target translation directions (e.g., generate $\mathbf{x_e}$ given $\mathbf{x_f}$ and <MASK>). Finally, in both losses a language identification symbol (i.e., <f> or <e>) is used as the initial token to predict the language of the output text.

$$\mathcal{L} = \sum_{(\mathbf{x_f},\mathbf{x_e})} \Big(\underbrace{\text{logp}\big([\text{<e>}\ \mathbf{x_e}] \mid (\mathbf{x_f}, \mathbf{x'_e})\big) + \text{logp}\big([\text{<f>}\ \mathbf{x_f}] \mid (\mathbf{x'_f}, \mathbf{x_e})\big)}_{\mathcal{L}_{\text{EDIT}}} + \underbrace{\text{logp}\big([\text{<e>}\ \mathbf{x_e}] \mid (\mathbf{x_f}, \text{<MASK>})\big) + \text{logp}\big([\text{<f>}\ \mathbf{x_f}] \mid (\text{<MASK>}, \mathbf{x_e})\big)}_{\mathcal{L}_{\text{MT}}} \Big) \tag{1}$$

Inference At test time, the language identification token is used to infer which of the two inputs gets revised. Finally, we pair the edited output sequence with the original input that does not get revised, yielding a refined bitext.

3.2 Bitext Mining

Our model requires access to $\mathbf{x'_f}$ and $\mathbf{x'_e}$ training instances that are treated as noised versions of $\mathbf{x_f}$ and $\mathbf{x_e}$, respectively. Since our goal is to develop a model that can refine mismatches found in mined bitexts at inference time, we want our noised training instances to share similar properties with the mined ones (e.g., fluent text in the target language, possibly imperfect translations of the source text). To this direction, we take a distance-based mining approach to construct the noised samples similar to Schwenk (2018).[1] Concretely, given the mined bitext $(\mathbf{x_f}, \mathbf{x_e})$ and two pools of monolingual sentences $\mathcal{F}$ and $\mathcal{E}$, in language $\mathbf{f}$ and $\mathbf{e}$, we extract $\mathbf{x'_f}$ and $\mathbf{x'_e}$ as follows:

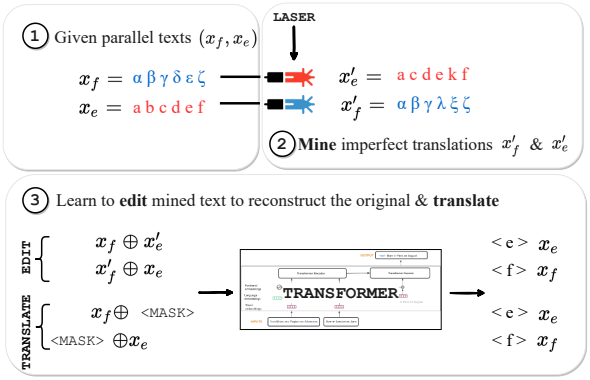


Figure 2: BITEXTEDIT training strategy.

$$\begin{aligned} \mathbf{x'_f} &= \text{argmax}_{\mathbf{z}\in\mathcal{F}}\ \cos(\text{LASER}(\mathbf{z}), \text{LASER}(\mathbf{x_e})) \\ \mathbf{x'_e} &= \text{argmax}_{\mathbf{z}\in\mathcal{E}}\ \cos(\text{LASER}(\mathbf{x_f}), \text{LASER}(\mathbf{z})) \end{aligned} \tag{2}$$

where LASER (Artetxe and Schwenk, 2019) represents a multilingual encoder used to extract sentence embeddings for each sentence, while the most similar sentence is returned based on nearest neighbor retrieval. Furthermore, this formula is extended to retrieval of top k sentences, while we also allow mining of the original CCMatrix translations. The latter happens to expose the model to good translations at training time, that should not be edited.

4 Experimental Setting

Bitexts We focus on CCMatrix data for two main reasons: a) it constitutes the only large-scale available resource for a lot of low-resource language pairs and b) recent efforts of auditing this corpus raise concerns regarding the quality of mined bitext of low-resource pairs. CCMatrix is mined using LASER embeddings following the "max-strategy" approach: a margin score is computed for all monolingual sentences in two languages, then the union of forward and backward candidates is build and pairs that score above a pre-defined threshold are treated as translations. Schwenk et al. (2021b) set the threshold grobally for all languages at 1.06.

Our primary goal is to explore whether bitexts that are typically discarded by filtering can be re-

[1] Note that unlike Artetxe and Schwenk (2019) we do not use a margin score on the *normalized* cosine distance of sentence-pairs to keep the computation cost low and encourage mining of more diverse imperfect translations.

PAIR	SCRIPTS	Pool A	Pool B
EN-OC	Latin-Latin	0.2M	0.1M
IT-OC	Latin-Latin	0.3M	0.1M
EN-BE	Latin-Cyrillic	0.7M	1.1M
EN-MR	Latin-Devanagari	1.5M	2.1M
EN-SW	Latin-Latin	1.7M	0.9M

Table 1: Statistics of CCMatrix bitexts.

fined by our model and thus benefit low-resource NMT. For this purpose, we define two pools of CCMatrix data: Pool A corresponds to CCMatrix data with LASER scores greater than 1.06, while Pool B contains bitexts with scores lower than 1.06 and greater than 1.05. The latter threshold is primarily chosen since CCMatrix bitexts is only available above this value. Editing bitexts with even smaller scores is an interesting area for future work.

Training data BITEXTEDIT models are trained based on procedures described in §3.2, where we use Pool A to seed the generation of noised training samples $\mathbf{x}'_\mathbf{f}$ and $\mathbf{x}'_\mathbf{e}$. We mine k samples $\mathbf{x}'_\mathbf{f}$ for each $\mathbf{x}_\mathbf{e}$ and k samples $\mathbf{x}'_\mathbf{e}$ for each $\mathbf{x}_\mathbf{f}$, respectively. We set k to 4 and include detailed statistics in F.

Language-pairs We experiment with the following languages: English-Occitan (EN-OC), Italian-Occitan (IT-OC), English-Belarusian (EN-BE), English-Marathi (EN-MR), and Englsh-Swahili (EN-SW). The 5 language pairs are chosen to include diverse low-resource pairs, which differ either in training data size or language similarity. Table1 summarizes the data.

Comparisons We run several extrinsic evaluations using NMT trained on different versions of CCMatrix data. First, we train NMT models on two versions of original CCMatrix data: Pool A (Schwenk et al., 2021b) and Pool $A \cup B$. Second, we aim at revising Pool B via a) a back-translation approach that revisits the source-side of the bitexts via back-translating their target-side with a model trained on original CCMatrix, (i.e., $b(.)$) and b) via editing either the source or the target side of it using our proposed approach (i.e, $r(.)$).

Model details Our models are implemented on top of `fairseq` (Ott et al., 2019).[2] We use the same Transformer architecture as in Schwenk et al. (2021b), with embedding size 512, 4,096 transformer hidden size, 8 attention heads, 6 transformer layers, and dropout 0.4. We train with 0.2 label smoothing and Adam optimizer with a batch size of 4,000 tokens per GPU.[3] We train for 100 epochs and select best checkpoint based on validation perplexity. We report single run results.

Data Preprocessing We use the standard Moses scripts (Koehn et al., 2007) for tokenization of EN, OC, IT, BE and SW and the Indic NLP library[4] for MR. For each language-pair, we learn 60K BPEs using `subword-nmt` (Sennrich et al., 2016).[5]

Evaluation We evaluate our models on the *devtest* of `flores` (Guzmán et al., 2019). We report `spm-bleu`[6] on the detokenized version of our outputs and chrF (Popović, 2015) as our second evaluation metric.[7]

5 Experimental Results

Bitext filtering revisited We first provide empirical evidence that bitext filtering might be a suboptimal solution to low-resource NMT. Table 2 shows that filtering out sentence pairs that score below the predefined threshold of 1.06 (i.e., Filtering) surprisingly hurts translation quality in almost all translation tasks (rows 2 vs. 1 and 8 vs. 7). This result is likely because the threshold was optimized for specific language-pairs, and the fact that—under low-resource regimes—increasing the amounts of *possibly imperfect* translation data might still benefit NMT. Furthermore, this experiment gives us insights on the quality of the training data our bitext editing model uses: for IT-OC, BE-EN, and EN-MR we expect Pool A to provide more noisy training signals (as BLEU scores of NMT models trained on it are $\sim$ 11), compared to EN-OC and EN-SW where the quality of the given bitext is expected to be significantly better (BLEU scores $\sim$ 18 and $\sim$ 37, respectively).

Editing Pool B Applying BITEXTEDIT to edit erroneous translations in Pool B (i.e., $A \cup r(B)$) improves the quality of NMT systems over the ones trained on the original CCMatrix corpus (rows 6 vs. 1 and 12 vs. 7). Among the language-pairs considered, the largest improvements are reported for IT-OC translation tasks (i.e., $+8.4/+6.7$), followed

[2] https://github.com/pytorch/fairseq

[3] More details are in Appendix D & G.

[4] https://anoopkunchukuttan.github.io/indic_nlp_library/

[5] https://github.com/rsennrich/subword-nmt

[6] https://github.com/facebookresearch/flores

[7] Results in Appendix A.

			EN→OC		IT→OC		EN→BE		EN→MR		EN→SW	
1 :	CCMatrix	$A \cup B$	20.5		11.5		11.0		12.2		38.1	
2 :	Filtering	A	18.1	−2.4	11.7	+0.2	9.8	−0.2	12.2	0.0	37.6	−0.5
3 :	Back-Translation	$b(A \cup B)$	20.8	+0.3	17.0	+5.5	12.3	+1.3	15.5	+3.2	37.6	−0.5
4 :	BITEXTEDIT	$r(A \cup B)$	25.4	+4.9	19.8	+8.3	12.8	+1.7	**15.8**	+3.6	37.8	−0.3
5 :	Back-Translation	$A \cup b(B)$	23.0	+2.5	17.0	+5.5	12.1	+1.1	15.4	+3.2	**38.8**	+0.7
6 :	BITEXTEDIT	$A \cup r(B)$	**26.0**	+5.5	**19.9**	+8.4	**13.0**	+2.0	15.3	+3.1	38.3	+0.2
			OC→EN		OC→IT		BE→EN		MR→EN		SW→EN	
7 :	CCMatrix	$A \cup B$	24.3		11.6		9.8		13.0		34.8	
8 :	Filtering	A	17.8	−6.5	11.1	−0.5	7.8	−2.0	11.3	−1.07	34.8	0.0
9 :	Back-Translation	$b(A \cup B)$	26.6	+2.3	17.3	+5.7	9.9	+0.1	13.6	+0.6	33.8	−1.0
10 :	BITEXTEDIT	$r(A \cup B)$	28.2	+3.9	**18.5**	+6.9	10.7	+0.9	16.4	+3.4	35.8	+1.1
11 :	Back-Translation	$A \cup b(B)$	27.7	+3.4	15.6	+4.0	9.6	−0.2	15.1	+2.1	**36.8**	+2.0
12 :	BITEXTEDIT	$A \cup r(B)$	**28.7**	+4.4	18.3	+6.7	**10.8**	+1.0	**16.7**	+3.7	36.2	+1.8

Table 2: Results on NMT tasks for models trained on different versions of CCMatrix. For each task the first column denotes spm-BLEU; the second columns (highlighted scores) give the difference of each row with the original CCMatrix. Models trained on the refined bitexts improve NMT for low-resource language-pairs.

by EN-OC (i.e., $+5.5/+4.4$). The magnitude of improvements might be explained by the relatedness of the two languages which facilitates editing with simpler operations (e.g., copying instead of translating).

Our approach also brings significant improvements over the original data for distant language-pairs written in different scripts, despite being trained on more noisy data, as discussed above. For example, we see improvements $+2.0/+1.0$ for EN-BE and $+3.1/+3.7$ for EN-MR. On the other hand, improvements on EN-SW are smaller (i.e., $+0.5/+1.8$). This is expected given the high BLEU scores that the original CCMatrix data yields.

Comparison with Back-Translation Since Pool B bitexts are typically filtered out from the pool of NMT training instances, one reasonable way of incorporating them in NMT training is via treating them as monolingual samples. We experiment with back-translation which constitutes the most popular approach to employ data augmentation for NMT. Comparing NMT models trained on CCMatrix augmented with backtranslated Pool B against our revised Pool B version (i.e., rows 5 vs. 6 and 11 vs. 12) shows that editing outperforms back-translation for $7/10$ tasks, while it yields comparable results to it for the rest 3.

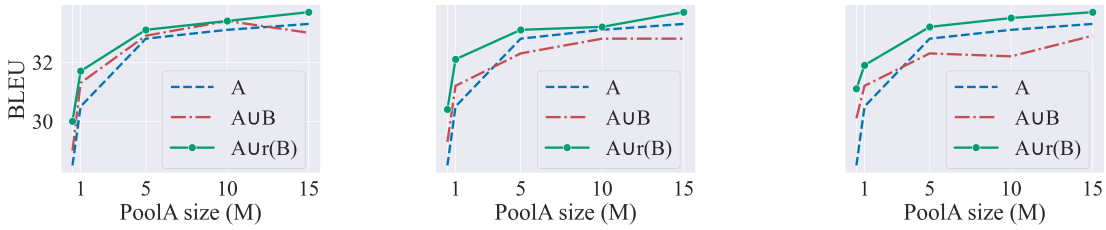
Editing Pool A and Pool B Since the editing framework gives us the potential to generalize all types of operations that *might* be needed to refine bitexts, it is also important that it does not perform *overediting* (i.e., editing already good quality bitexts). For this reason, we also attempt to revise the entire CCMatrix corpus (i.e., $r(A \cup B)$), using our bitext refinement models (i.e., rows 4 and 9). To better understand the importance of performing conservative editing on good quality bitexts, we also compare against back-translation (i.e, $b(A \cup B)$ in rows 3 and 9). First, we observe that our approach yields consistently significant improvements over CCMatrix with the exception of EN→SW where it performs comparably to it. Second, for most tasks the improvements are comparable to those reported when revising only Pool B, while it is consistently better than back-translation. It, overall, provides a universal method that works well in every case.

6 Analysis

We now turn into additional analysis with a focus on understanding the broader space where BITEXTEDIT can be applied. We experiment with scaling-up bitext refinement to higher-resource settings in §6.1, we perform qualitative analysis on the edited bitexts in §6.2, and quantitative analysis on the types and intensity of edits in different corpora in §6.3.

6.1 Scaling-up BITEXTEDIT

First, we examine how models trained only on good quality data (Figure 4) behave as we vary their quantity. We experiment with English-Greek EN-EL CCMatrix bitexts and simulate various resource settings via downsampling. In *low-resource* settings (i.e., $|A| < 1$M), translation quality exhibits rapid improvements, with an increase from 100K to 500K training samples boosting BLEU, by approximately 10 points. In *medium-resource* scenarios (i.e., $1 <$M$|A| < 5$M), a proportional in-



(a) $|B| = |A|/2$ (b) $|B| = |A|$ (c) $|B| = 2|A|$

Figure 3: Translation quality (i.e., BLEU) of EN→EL NMT models trained on different amounts of Pool A and Pool B data (i.e., $|A|$ given by x-axis). Across settings, bitext refinement (i.e., $A \cup r(B)$) performs better or comparably to training on the original CCMatrix (i.e., $A \cup B$) or its filtered version (i.e., A).

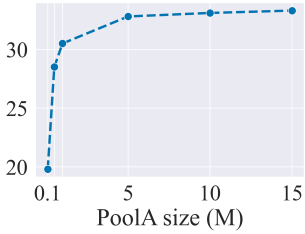


Figure 4: BLEU for EN→EL NMT trained on varying size of CCMatrix data (Pool A).

crease in the quantity of good quality bitexts yields smaller—yet, significant—translation quality improvements (i.e., moving from 1M to 5M bitexts yields +2 BLEU). Finally, in *high-resource* settings (i.e, $|A| > 5$), translation quality reaches a saturation point, with BLEU increases being small and insignificant (i.e., $\sim +0.2$) as we move from 10M to 15M training samples.

Second, we present a controlled analysis experiment on how bitext refinement impacts the translation quality of NMT systems under different resource settings (Figure 3). Concretely, starting from a high resource language-pair in CCMatrix (here, EN-EL) we sample good and poor quality bitexts (i.e, A and B, respectively) representing low- to high- data scenarios (e.g, 500K up to 15M sentence-pairs). Then, we train EN→EL NMT systems on $A \cup B$ while varying their distribution to represent three settings: (a) good quality bitexts overwhelm the training data (i.e, $|B| = |A|/2$), (b) good and poor quality bitext are equally represented (i.e, $|B| = |A|$), and (c) poor quality bitexts overwhelm the training data (i.e, $|B| = 2|A|/$).[8] Across distribution conditions, adding imperfect translations (i.e., B) to the original good quality data yields improvements for low-to-medium resource settings (i.e, $|A| < 5$). This results complement the earlier observations of §5 that question the appropriateness of a filtering framework in settings where data is scarce. On the other hand, when moving to high resource scenarios, the additional signal that results from imperfect references can have either insignificant (i.e., Figure 3a) or negative impact (i.e., Figures 3b and 3c) on translation quality. The latter depends on whether the good quality data is underrepresented in the training samples.

Third, starting from good quality bitexts of varying sizes, we train separate bitext refinement models and edit the corresponding poor quality samples (i.e., $r(.)$) defined earlier. Across the board, NMT models that are trained on $A \cup r(B)$ yield the best translation quality results compared to both filtering and training on original CCMatrix. However, we observe that the magnitude of the improvements depends on the data settings. Concretely, bitext refinement yields significant improvements on low-to-medium resource settings (i.e., $\sim +2$ BLUE points). On the other hand, in high resource scenarios bitext refinement helps mitigate the negative impact of overwhelming poor quality instances and performs comparably to filtering. The latter suggests that our refinement strategy *improves bitexts quality across low- to high- resource settings*.

6.2 Qualitative analysis

We conduct a qualitative study to confirm that BITEXTEDIT improves the quality of CCMatrix.[9] One of the authors manually evaluates a random sample of 200 EN-EL sentence-pairs where we compare the original bitexts against the refined ones.[10] As shown in Figure 5, our models performs edits that refine meaning mismatches found in the original CCMatrix data. While only $\sim 38\%$ of the orig-

[8] More details on experimental settings are in Appendix B.

[9] We include details on the annotation in Appendix C.

[10] Here, we present results on bitext refinement models that use 0.5M PoolA samples. Manual inspection on refined outputs of models trained on larger pools showed similar performance.

→	**[EN]** CCMATRIX	After that time the whole group would talk for 5 minutes.
	[EL] CCMATRIX	Αργότερα, η ομάδα μελέτης ζήτησε από όλους να διαλογιστούν για πέντε λεπτά.
	GLOSS	*Later, the study group asked everyone to meditate for 5 minutes.*
	[EN] BITEXTEDIT	Later, the study group asked everyone to meditate for five minutes.
→	**[EN]** CCMATRIX	We should, however, always be striving to live a sustainable and kind life.
	[EL] CCMATRIX	Πάντα πρέπει να παλεύουμε για δίκαιη και βιώσιμη ειρήνη.
	GLOSS	*We must always fight for a just and lasting peace.*
	[EN] BITEXTEDIT	We must always fight for just and sustainable peace.
	[EN] CCMATRIX	"The western influence came from film and television", he later explained.
→	**[EN]** CCMATRIX	«Η λογοκρισία εντείνεται όλο και περισσότερο στον κινηματογράφο και την τηλεόραση», εξήγησε ο ίδιος.
	GLOSS	*"Censorship is intensifying in cinema and television", he explained.*
	[EL] BITEXTEDIT	«Η δυτική επιρροή ήρθε από την ταινία και την τηλεόραση» , εξήγησε αργότερα.
	GLOSS	*"The western influence came from form and television", as their later explained.*
	[EN] CCMATRIX	I could work with a hospital specialist as a clinical assistant (as I have done).
→	**[EL]** CCMATRIX	Δούλευε ώς βοηθός ερευνητή παράλληλα με το διδακτορικό (όπως και εγώ)
	GLOSS	*They were working as an assistant researcher in parallel with their doctorate (as I have done).*
	[EL] BITEXTEDIT	Θα μπορούσα να δουλέψω με έναν ειδικό στο νοσοκομείο ως κλινικός βοηθός (όπως έχω κάνει).
	GLOSS	*I could work with a hospital specialist as a clinical assistant (as I have done).*

Table 3: Examples of CCMatrix bitexts along with refined sides generated by BITEXTEDIT. → denotes the side ([EL] or [EN]) that the model edits, while highlighted segments indicate the meaning mismatches in the original CCMatrix sentence that gets edited. Greek sentences are glossed to help understanding their meaning.

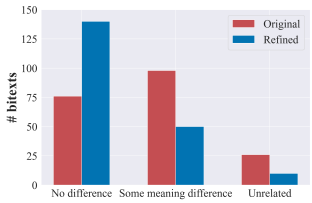


Figure 5: Number of bitexts manually rated as perfect translations (i.e., No difference), partial translations (i.e., some meaning difference), and wrong translations (i.e., unrelated) for a random sample of original vs. refined CCMatrix EN-EL data.

inal samples contain parallel texts that are perfect translations of each other, the revised sample contains $\sim 70\%$ perfect translations. Finally—apart from evaluating meaning differences—we also rate fluency of the edited translations. We find that our model does not suffer from major fluency issues with 84.5% of their outputs rated as *flawless* and 15.5% as *good*. Table 3 presents example outputs of our BITEXTEDIT approach for English-Greek.[11]

6.3 Quantitative analysis

Percentage of edited bitexts Table 5 presents coarse statistics on the percentage of refined bitexts that exhibit at least one edit compared to the original ones. First, we observe that the percentage of edited bitexts varies across the languages-pairs studied. This reflects the varying quality of PoolB samples in different languages and also connects to the varying magnitude of improvements we show in Table 2. The biggest improvements are given for IT-OC, where $\sim 76\%$ of the bitexts are edited by our refinement models. On the other hand, the smallest improvements are found for EN-SW, with only $\sim 36\%$ of its bitext being revised, probably due to the already good quality of the initial CCMatrix sentence pairs.

Editing EN-EL OPUS corpora Broadly speaking, a good bitext refinement model should be able to rewrite bitext in a way that improves potential errors in the original references. At the same time though, it should avoid over-editing (i.e., avoid editing an already good translation-pair). We perform a quantitative analysis on EN-EL corpora from OPUS that vary in their quality and extract Translation Error Rate (TER) label (Snover et al., 2006) token-level statistics to study both the *frequency* and the *types* of edits that our bitext refinement models perform. Table 4 presents results on random samples ($\sim$ 100K) of three popular corpora: (a) the Tatoeba corpus (Tiedemann, 2020) consisting of human translations, (b) the OpenSubtitles corpus (Lison and Tiedemann, 2016) consisting of sentence-aligned subtitles of movie series[12], and (c) the ParaCrawl corpus (Esplà et al., 2019) consisting of automatically crawled translations from document-level translations of European Parliament Proceedings.

As expected, our model performs minimal edit-

[11] More examples can be found in Appendix E.

[12] http://www.opensubtitles.org/, https://opus.nlpl.eu/OpenSubtitles-v2018.php

		C	S	D	I	C	S	D	I
CORPUS	EDITED SENT.	ALL (%)				ALL \ COPIES (%)			
Tatoeba	29.80%	97.47	1.88	0.29	0.34	86.38	10.16	1.56	1.88
OpenSubtitles	65.63%	90.46	5.53	1.27	2.73	74.51	14.79	3.39	7.29
ParaCrawl	88.11%	96.30	2.25	0.39	1.04	85.42	8.89	1.55	4.12

Table 4: TER statistics for bitext refinement of random samples of EN-EL OPUS bitexts. Second column gives the % of bitexts that get at least one edit operation; the last two columns present the percentage of correct (**C**), substituted (**S**), deleted (**D**), and inserted (**I**) tokens for all the bitexts (i.e., **ALL**) and the subset of bitexts that receive revisions compared to the original (i.e., **ALL**\ **COPIES**).

PAIR	SRC	TGT	BOTH
EN-OC	34.06%	66.58%	67.48%
IT-OC	34.76%	41.11%	75.78%
EN-MR	58.35%	19.90%	68.07%
BE-EN	21.01%	28.06%	49.06%
EN-SW	14.52%	21.05%	35.57%

Table 5: Percentage of sentences with at least one edit operation compared to the original for: source-side (SRC), target-side (TGT), and both sides (BOTH).

ing on the high-quality *Tatoeba* bitexts. Concretely, only $\sim 30\%$ of it gets revised, while as suggested by the token-level TER statistics even the revised sentence-pairs mostly consist of substituted tokens. Further manual inspection reveals that most of those tokens depict subtle spelling differences between Greek words. On the other hand, when editing the samples of automatically extracted bitexts our refinement model performs more frequent edits: it revises $\sim 65\%$ of OpenSubtitles and $\sim 88\%$ of ParaCrawl bitexts. Interestingly, although a greater amount of ParaCrawl texts get revised compared to OpenSubtitles, edits on the latter are more aggressive as it consists of at least 10% fewer "correct" (i.e., **C**) tokens than the former. A break down on the types of operations further reveals that editing OpenSubtitles requires more "deletion" (i.e., **D**) and "insertion" (i.e., **I**) operations compared to the other two. This observation connects to prior efforts on auditing OpenSubtitles that found sentence segmentation errors (i.e, added extra leading/trailing words in one side) to be a frequent type error for this corpus (Vyas et al., 2018).

7 Related Work

Automatic Post-Editing APE aims at automatically correcting the output of a "black-box" MT system. Recent approaches on APE (Chatterjee et al., 2019, 2020) fine-tune pre-trained multilingual models models (Lopes et al., 2019) or translation models (Yang et al., 2020) on a combination of gold-standard APE data and artificially augmented candidates resulting from external translations.

Low-resource MT Haddow et al. (2021) structure the diverse set of approaches to low-resource MT to (a) efforts for increasing the amounts of available bitexts (i.e., *data collection*; Schwenk et al. (2021a,b)), (b) methods that explore how other types of data can be incorporated into MT (i.e., *data exploitation*; Baziotis et al. (2020); Zoph et al. (2016); García-Martínez et al. (2017)), and (c) advances in modeling (i.e., *model choices*; Vaswani et al. (2017)). BITEXTEDIT is an alernative *data exploitation* approach that does not require further bilingual data or other sources of supervision.

Retrieve & Edit Approaches Retrieve and edit approaches have been integrated at inference time for several tasks, such as NMT (Gu et al., 2018; Zhang et al., 2018; Cao and Xiong, 2018; Hossain et al., 2020), APE (Hokamp, 2017), dialogue generation (Weston et al., 2018), among others.

8 Conclusion

We introduce an alternative approach for bitext quality improvement that we show is better suited for low-resource language pairs. Instead of filtering out imperfect translation references that result from automatic bitext mining, we instead edit them with the goal of improving their quality. Our editing models are trained using only synthetic supervision, which can be gathered at scale for any language pair that support bitext mining. Extensive quantitative analysis suggests that our approach successfully improves bitext quality for a variety of language-pairs and different resource conditions. Furthermore, extrinsic experiments on 10 low-resource NMT tasks suggest that bitext refinement constitutes a successful approach to improving NMT translation quality in low data regimes. Those findings highlight the importance of the good *quality* bitexts in scenarios where large *quantities* cannot be guaranteed and motivate future research on improving low-resource NMT further.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Tom Ash, Remi Francis, and Will Williams. 2018. The speechmatics parallel corpus filtering system for WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 853–859, Belgium, Brussels. Association for Computational Linguistics.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Christos Baziotis, Barry Haddow, and Alexandra Birch. 2020. Language model prior for low-resource neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7622–7634, Online. Association for Computational Linguistics.

Eleftheria Briakou and Marine Carpuat. 2020. Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1563–1580, Online. Association for Computational Linguistics.

Qian Cao and Deyi Xiong. 2018. Encoding gated translation memory into neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3042–3047, Brussels, Belgium. Association for Computational Linguistics.

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Javier Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Balli, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2021. Quality at a glance: An audit of web-crawled multilingual datasets. *CoRR*, abs/2103.12028.

Rajen Chatterjee, Christian Federmann, Matteo Negri, and Marco Turchi. 2019. Findings of the WMT 2019 shared task on automatic post-editing. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 11–28, Florence, Italy. Association for Computational Linguistics.

Rajen Chatterjee, Markus Freitag, Matteo Negri, and Marco Turchi. 2020. Findings of the WMT 2020 shared task on automatic post-editing. In *Proceedings of the Fifth Conference on Machine Translation*, pages 646–659, Online. Association for Computational Linguistics.

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. Low-resource corpus filtering using multilingual sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 261–266, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

Mercedes García-Martínez, Loïc Barrault, and Fethi Bougares. 2017. Neural machine translation by generating multiple linguistic factors. *CoRR*, abs/1712.01821.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. 2018. Search engine guided neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

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. 2021. Survey of low-resource machine translation.

Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. 2014. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 174–180, Baltimore, Maryland. Association for Computational Linguistics.

Chris Hokamp. 2017. Ensembling factored neural machine translation models for automatic post-editing and quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 647–654, Copenhagen, Denmark. Association for Computational Linguistics.

Nabil Hossain, Marjan Ghazvininejad, and Luke Zettlemoyer. 2020. Simple and effective retrieve-edit-rerank text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2532–2538, Online. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online. Association for Computational Linguistics.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72, Florence, Italy. 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.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

António V. Lopes, M. Amin Farajian, Gonçalo M. Correia, Jonay Trénous, and André F. T. Martins. 2019. Unbabel's submission to the WMT2019 APE shared task: BERT-based encoder-decoder for automatic post-editing. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 118–123, Florence, Italy. Association for Computational Linguistics.

Jun Lu, Xiaoyu Lv, Yangbin Shi, and Boxing Chen. 2018. Alibaba submission to the WMT18 parallel corpus filtering task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 917–922, Belgium, Brussels. Association for Computational Linguistics.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3956–3965. PMLR.

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 *Proceedings of NAACL-HLT 2019: Demonstrations*.

Vassilis Papavassiliou, Sokratis Sofianopoulos, Prokopis Prokopidis, and Stelios Piperidis. 2018. The ILSP/ARC submission to the WMT 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 928–933, Belgium, Brussels. Association for Computational Linguistics.

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.

Philip Resnik. 1999. Mining the web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 527–534, College Park, Maryland, USA. Association for Computational Linguistics.

Nick Rossenbach, Jan Rosendahl, Yunsu Kim, Miguel Graça, Aman Gokrani, and Hermann Ney. 2018. The RWTH Aachen University filtering system for the WMT 2018 parallel corpus filtering task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 946–954, Belgium, Brussels. Association for Computational Linguistics.

Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez. 2018. Prompsit's submission to WMT 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 955–962, Belgium, Brussels. Association for Computational Linguistics.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. 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 6490–6500, Online. Association for Computational Linguistics.

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

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

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

Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual retrieval for iterative self-supervised training. In *Advances in Neural Information Processing Systems*, volume 33, pages 2207–2219. Curran Associates, Inc.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yogarshi Vyas, Xing Niu, and Marine Carpuat. 2018. Identifying semantic divergences in parallel text without annotations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1503–1515, New Orleans, Louisiana. Association for Computational Linguistics.

Jason Weston, Emily Dinan, and Alexander Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 87–92, Brussels, Belgium. Association for Computational Linguistics.

Hao Yang, Minghan Wang, Daimeng Wei, Hengchao Shang, Jiaxin Guo, Zongyao Li, Lizhi Lei, Ying Qin,

Shimin Tao, Shiliang Sun, and Yimeng Chen. 2020. HW-TSC's participation at WMT 2020 automatic post editing shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 797–802, Online. Association for Computational Linguistics.

Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura. 2018. Guiding neural machine translation with retrieved translation pieces. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1325–1335, New Orleans, Louisiana. Association for Computational Linguistics.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

A Results on Second Evaluation Metric

Table A presents results on NMT tasks for a second evaluation metric.

	EN→OC	IT→OC	EN→BE	EN→MR	EN→SW
1 :	41.59	30.92	29.28	31.19	59.17
2 :	39.73	32.26	28.24	31.90	58.76
3 :	42.34	40.62	30.96	35.41	58.60
4 :	47.40	42.83	31.21	35.01	59.02
5 :	44.66	39.01	30.66	35.20	59.50
6 :	47.74	43.03	31.08	34.65	59.49
	OC→EN	**OC→IT**	**BE→EN**	**MR→EN**	**SW→EN**
7 :	48.04	32.90	37.13	37.84	57.10
8 :	42.10	33.73	33.51	36.55	57.07
9 :	50.99	42.42	37.13	39.99	56.74
10 :	52.13	42.42	39.20	42.45	57.96
11 :	51.63	38.99	36.99	40.29	58.74
12 :	53.86	44.05	39.18	42.71	58.29

Table 6: Results on NMT tasks for the chrF metric (rows follow the enumeration of Table 2).

B Scaling-Up Settings

Tables 7, 8, and 9 present training data sizes for experiments in Figure 3.

A	0.5M	1.0M	5.0M	10.0M	15.0M
A∪B	0.75M	1.5M	7.5M	15.0M	22.5M
A∪r(B)	0.75M	1.5M	7.5M	15.0M	22.5M

Table 7: Training data size for experiments in Figure 3(a), where $|B| = |A|/2$.

A	0.5M	1.0M	5.0M	10.0M	15.0M
A∪B	1.0M	2.0M	10.0M	20.0M	30.0M
A∪r(B)	1.0M	2.0M	10.0M	20.0M	30.0M

Table 8: Training data size for experiments in Figure 3(b), where $|B| = |A|$.

A	0.5M	1.0M	5.0M	10.0M	15.0M
A∪B	1.5M	3.0M	15.0M	30.0M	70M
A∪r(B)	1.0M	2.0M	10.0M	20.0M	70.0M

Table 9: Training data size for experiments in Figure 3(c), where $|B| = 2|A|$.

C Manual Annotation Details

For each bitext (i.e., original CCMatrix sample or refined sample edited by a bitext refinement model) we rate the **degree of equivalence** between the two sentences following the protocol of semantic divergences (Briakou and Carpuat, 2020). Concretely, a bitext is annotated as having *no meaning difference* if it corresponds to perfect translations, *some meaning differences* if the sentences share important content in common but differ by few tokens (e.g., small added content, or phrasal mistranslation), and *unrelated* if the sentences are only topically or structurally related. For rating **fluency** we evaluate the output sentence of the bitext refinement models in isolation on a discrete scale of 1 to 5, following Heilman et al. (2014) (Other → Incomprehensible → Somewhat Comprehensible → Comprehensible → Perfect).

D Fairseq configuration details

Table 10 presents details of NMT training with fairseq. The same parameters are used to train BITEXTEDIT models.

```
-arch transformer
-share-all-embeddings
-encoder-layers 6
-decoder-layers 6
-encoder-embed-dim 512
-decoder-embed-dim 512
-encoder-ffn-embed-dim 4096
-decoder-ffn-embed-dim 4096
-encoder-attention-heads 8
-decoder-attention-heads 8
-encoder-normalize-before
-decoder-normalize-before
-dropout 0.4
-attention-dropout 0.2
-relu-dropout 0.2
-weight-decay 0.0001
-label-smoothing 0.2
-criterion label smoothed cross entropy
-optimizer adam
-adam-betas '(0.9, 0.98)'
-clip-norm 0
-lr-scheduler inverse sqrt
-warmup-updates 4000
-warmup-init-lr 1e-7
-lr 1e-3
-max-tokens 4000
-update-freq 4
-max-epoch 100
-save-interval 10
```

Table 10: Fairseq configuration used for NMT training.

E BITEXTEDIT: Model outputs

Table 11 presents model outputs samples edited by our model for EN-EL CCMatrix instances.

→	[EN] CCMATRIX	Respect the dignity of all people, regardless of their age.
	[EL] CCMATRIX	Πιστεύω στην αναγκαιότητα αξιοποίησης όλων των άξιων ανθρώπων ανεξάρτητα από την ηλικία τους.
	GLOSS	*I believe in the importance of using all skilled people, regardless of their age.*
	[EL] BITEXTEDIT	Σεβασμός στην αξιοπρέπεια όλων των ανθρώπων , ανεξάρτητα από την ηλικία τους.
→	[EN] CCMATRIX	After that time the whole group would talk for 5 minutes.
	[EL] CCMATRIX	Αργότερα, η ομάδα μελέτης ζήτησε από όλους να διαλογιστούν για πέντε λεπτά.
	GLOSS	*Later, the study group asked everyone to meditate for 5 minutes.*
	[EN] BITEXTEDIT	Later, the study group asked everyone to meditate for five minutes.
	[EN] CCMATRIX	Say no to fake products and scams.
→	[EL] CCMATRIX	Είπατε όχι στις ψεύτικες υποσχέσεις και στη συναλλαγή.
	GLOSS	*You said no to fake products and transactions.*
	[EL] BITEXTEDIT	Πείτε όχι στα ψεύτικα προϊόντα και απάτες.
→	[EN] CCMATRIX	We're all part of a larger system.
	[EL] CCMATRIX	Τα πάντα είναι μέρος ενός μεγαλύτερου Συστήματος.
	GLOSS	*Everying is part of a larger System.*
	[EN] BITEXTEDIT	Everything is part of a larger system.
→	[EN] CCMATRIX	Currently, no equivalent technology exists on the market .
	[EN] CCMATRIX	Δεν υπάρχει αντίστοιχη ανταγωνιστική τεχνολογία στον κόσμο αυτή τη στιγμή.
	GLOSS	*There is no corresponding competing technology in the word right now.*
	[EN] BITEXTEDIT	There is no competitive technology in the world right now.
	[EN] CCMATRIX	"The western influence came from film and television", he later explained.
→	[EN] CCMATRIX	«Η λογοκρισία εντείνεται όλο και περισσότερο στον κινηματογράφο και την τηλεόραση», εξήγησε ο ίδιος.
	GLOSS	*"Censorship is intensifying in cinema and television", he explained.*
	[EL] BITEXTEDIT	«Η δυτική επιρροή ήρθε από την ταινία και την τηλεόραση» , εξήγησε αργότερα.
	[EN] CCMATRIX	Then he paused, surveying the surreal scene.
→	[EN] CCMATRIX	Και πράγματι έφυγε , προσπερνώντας τον έκπληκτο Κέλι .
	GLOSS	*And indeed he left, passing Keli, who was surprised.*
	[EL] BITEXTEDIT	Στη συνέχεια σταμάτησε, επιθεωρώντας την σουρεαλιστική σκηνή.
→	[EN] CCMATRIX	Device installation error is a frequent error.
	[EL] CCMATRIX	Η ακατάλληλη φόρμα βιογραφικού, είναι ένα πολύ συχνό λάθος.
	GLOSS	*An improper resume form, is a very frequent mistake.*
	[EN] BITEXTEDIT	The inappropriate biographical form is a very frequent mistake.
	[EN] CCMATRIX	I could work with a hospital specialist as a clinical assistant (as I have done).
→	[EL] CCMATRIX	δούλευε ώς βοηθός ερευνητή παράλληλα με το διδακτορικό (όπως και εγώ)
	GLOSS	*They were working as an assistant researcher in parallel with their doctorate (as I have done).*
	[EL] BITEXTEDIT	Θα μπορούσα να δουλέψω με έναν ειδικό στο νοσοκομείο ως κλινικός βοηθός (όπως έχω κάνει).
→	[EN] CCMATRIX	We should, however, always be striving to live a sustainable and kind life.
	[EL] CCMATRIX	Πάντα πρέπει να παλεύουμε για δίκαιη και βιώσιμη ειρήνη.
	GLOSS	*We must always fight for a just and lasting peace.*
	[EN] BITEXTEDIT	We must always fight for just and sustainable peace.

Table 11: Examples of CCMatrix bitexts along with refined sides generated by BITEXTEDIT. → denotes the side ([EL] or [EN]) that the model edits, while highlighted segments indicate the meaning mismatches in the original CCMatrix sentence that gets edited. Greek sentences are glossed to help understanding their meaning.

Corpus	Version	License	Citation	Link
CCmatrix	v2	-	Schwenk et al. (2021b)	*to be released*
FLORES	v1	CC-BY-SA	Guzmán et al. (2019)	https://github.com/facebookresearch/flores
OpenSubtitles	v2018	-	Lison and Tiedemann (2016)	https://opus.nlpl.eu/OpenSubtitles-v2018.php
Tatoeba	v2	CC–BY 2.0 FR	Tiedemann (2012)	https://opus.nlpl.eu/Tatoeba.php
ParaCrawl	v7.1	Creative Commons CC0	Esplà et al. (2019)	https://opus.nlpl.eu/ParaCrawl.php

Table 12: Additional documentation of scientific artifacts used in our paper.

F Details on Scientific Artifacts

Statistics on Training Examples Tables 13 and 14 include detailed statistics on training and dev samples used to train each of the NMT and BITEXTEDIT models discussed in the paper.

	Training		Dev	Test
Pair	$\|A\|$	$\|A \cup B\|$		
EN-OC	242,982	365,399	997	1,012
IT-OC	309,703	440,283	997	1,012
EN-BE	659,430	3,944,412	997	1,012
EN-MR	1,503,477	3,611,336	997	1,012
EN-SW	1,721,801	2,641,234	997	1,012

Table 13: Number of training/dev/test examples used to train NMT models in Table 2.

Pair (src-tgt)	All	Mined (src)	Mined (tgt)
Training samples			
EN-OC	3,822,800	965,184	946,216
IT-OC	4,743,350	1,228,328	1,143,347
EN-BE	10,152,596	2,637,575	2,544,460
EN-MR	17,764,241	5,640,928	5,991,336
EN-SW	16,232,991	6,734,355	6,859,214
Dev samples			
EN-OC	15,908	3,988	3,966
IT-OC	15,952	3,988	3,988
EN-BE	15,952	3,988	3,988
EN-MR	15,952	3,988	3,988
EN-SW	15,952	3,988	3,988

Table 14: Number of training/dev examples used to train BITEXTEDIT models in Table 2. The two last columns (i.e., *mined*) include further statistics on the number of mined bitexts consumed by the *edit-based reconstruction* loss; the rest of the training samples correspond to machine-translation samples upweighted to match the number of mined bitexts (i.e., equal contribution of two losses).

License details We use data derived from OPUS (https://opus.nlpl.eu/) corpora as summarized in Table 12. All data are solely used for research purposes.

	Original		Edited	
	# Tokens			
	SRC	TGT	SRC	TGT
EN-OC	3,591,876	3,995,351	3,601,179	3,978,861
OC-IT	5,717,341	5,496,704	5,763,860	5,428,767
BE-EN	97,172,691	43,007,326	16,806,977	19,002,607
EN-MR	36,468,349	32,934,460	36,411,035	3,2830,479
EN-SW	4,1855,796	40,666,513	41,978,701	40,472,724
	# Types			
	SRC	TGT	SRC	TGT
EN-OC	165,310	234,252	169,191	235,503
OC-IT	277,397	278,727	292,357	283,656
BE-EN	531,309	526,289	533,224	381,666
EN-MR	407,977	956,589	379,015	922,184
EN-SW	414,873	802,292	409,0224	791,853
	Type-Token ratio			
	SRC	TGT	SRC	TGT
EN-OC	4.6%	5.9%	4.7%	5.9%
OC-IT	5.9%	5.1%	5.1%	5.2%
BE-EN	0.5%	1.2%	3.2%	2.0%
EN-MR	1.1%	2.9%	1.0%	2.8%
EN-SW	2.0%	2.0%	1.0%	2.0%

Table 15: Lexical characteristics of Original vs. Edited version of CCMatrix bitexts.

G Compute Infrastructure & Run time

Each experiment runs on a single machine with 8 GPUs. NMT models require less than 3.5 hours (e.g., EN-OC on $A \cup B$ requires $\sim$ 20 minutes to train). Similarly, BITEXTEDIT models require less than 13.5 hours to train (e.g., EN-OC requires $\sim$ 5 hours). All models follow the transformer architecture detailed in Appendix D with a total of 165M parameters.

H Potential Risks

Hallucination detection Our approach introduces synthetic samples (i.e., edited references that replace the originally human generated samples) that are later consumed as training instances by NMT models. One concern of using synthetic instances highlighted by recent work (Zhou et al., 2021), is the generation of hallucinations (i.e., fluent text that is not tight to the source segment). To understand whether our method potentially contributes to the issue of hallucinations, one of the

authors examined a small sample of 20 outputs generated by three NMT models for EN→EL translation: 1. a model trained only on 1M of PoolA data; 2. a model trained on the concatenation of 1M PoolA and 2M PoolB data; 3. a model trained on the concatenation of 1M PooA and 2M edited PoolB data. The NMT outputs are annotated labeled as: incomprehensible, faithful, or contains hallucinations following the protocol of Zhou et al. (2021). All annotated instances are found to be faithful to the source.

Lexical Richness Synthetically generated data (e.g., machine-translated instances) are known to exhibit a decay in lexical richness when compared to human written texts (Vanmassenhove et al., 2019). To confirm that our approach does not potentially contribute to this issue, we report more detailed statistics on how the original and edited CCMatrix texts differ in terms of lexical features (i.e., #tokens, #types, and type-token ratio). As presented in Table 15 the edited text does exhibit a decrease in the type-token ratio percentage when compared to the original one.