Can Synthetic Translations Improve Bitext Quality?

Anonymous ACL submission

Abstract

Synthetic translations have been used for a wide range of NLP tasks primarily as a means of data augmentation. This work explores instead, how we can use synthetic translations to selectively replace potentially imperfect reference translations in mined bitext. We find that synthetic samples can improve bitext quality without any additional bilingual supervision, when they replace the originals based on a semantic equivalence classifier that helps mitigate NMT noise. The improved quality of the revised bitext is confirmed intrinsically via human evaluation and extrinsically through bilingual induction and MT tasks.

1 Introduction

While human-written data remains the gold standard to train Neural Machine Translation (NMT) and Multilingual NLP models, there is growing evidence that synthetic bitext samples—sentence-pairs that are translated by NMT—benefit a wide range of tasks. They have been used to enable semi-supervised MT training from monolingual data (Sennrich et al., 2016a; Zhang and Zong, 2016; Hoang et al., 2018), to induce bilingual lexicons (Artetxe et al., 2019; Shi et al., 2021), and to port models trained on one language to another (Conneau et al., 2018; Yang et al., 2019).

While synthetic bitexts are useful additions to original training data for downstream tasks, it remains unclear how they differ from naturally occurring data. Some studies suggest that synthetic samples might be simpler and easier to learn (Zhou et al., 2019a; Xu et al., 2021). Recognizing that naturally occurring bitext can be noisy, for instance when they are mined from comparable monolingual corpora (Resnik and Smith, 2003; Fung and Yee, 1998; Esplà et al., 2019; Schwenk et al., 2021), we hypothesize that synthetic bitext might also directly improve the equivalence of the two bitext sides. Thus synthetic samples might be useful not only for data augmentation, but also to revise potentially noisy original bitext samples.

In this paper, we present a controlled empirical study comparing the quality of bitext mined from monolingual resources with a synthetic version generated via MT. We focus on the widely used WikiMatrix bitexts for a distant (i.e, EN-EL) and a similar language-pair (i.e, EN-RO), since it has been shown that this corpus contains a significant proportion of erroneous translations (Caswell et al., 2021). We generate synthetic bitext by translating the original training samples using MT systems trained on the bitext itself, and therefore do not inject any additional supervision in the process. We also consider selectively replacing original samples with forward and backward synthetic translations based on a semantic equivalence classifier, which is also trained without additional supervision.

We show that the resulting synthetic bitext improves the quality of the original intrinsically using human assessments of equivalence, and extrinsically on bilingual induction (BLI) and MT tasks. We present an extensive analysis of synthetic data properties and of the impact of each step in its generation process. This study brings new insights in the use of synthetic samples in NLP. First, intrinsic evaluation shows that synthetic translations, in addition to "normalizing" the bitext as suggested by prior work (Zhou et al., 2019a; Xu et al., 2021), are of sufficient quality to improve over the original translations. Furthermore, the improved bitext provides more useful training signals for BLI tasks and NMT training in two settings (i.e., training from scratch and continued training), as confirmed by our extrinsic evaluations. Finally, ablations analysis that compare different ways to combine synthetic translations show that using *both translation directions* and *filtering using semantic equivalence* is key to improve bitext quality and calls for further exploration of best practices for using synthetic translations in NLP tasks.

2 Background

Synthetic Translations Generating synthetic translations has mainly been studied as a means of data augmentation for NMT through forward translation (Zhang and Zong, 2016) or back-translation (Sennrich et al., 2016a; Marie et al., 2020) of monolingual resources. Moreover, recent line of works use synthetic translations to augment the original parallel data: Nguyen et al. (2020) diversify the parallel data via translating both sides using multiple models and then merge them with the original to train a final NMT model; Jiao et al. (2020) employ a similar approach to rejuvenate inactive examples that contribute the least to the model performance. Sequence-level knowledge distillation (Kim and Rush, 2016) can also be viewed as replacing original bitext with synthetic translations. While its original goal was to guide the training of a student model of small capacity with the output of a teacher of high capacity, distillation is also necessary to effectively train some categories of MT architectures such as non-autoregressive models (Gu et al., 2017). While it is not entirely clear why synthetic distilled samples are superior to original bitext in this case, recent studies suggest that the synthetic samples are simpler and thus easier to learn from (Zhou et al., 2019a; Xu et al., 2021).

Synthetic Data Selection Prior work covers a wide spectrum of different selection strategies on top of synthetic translations generated from monolingual samples. Each of them focuses on identifying samples with specific properties: Axelrod et al. (2011) sample sentences that are most relevant to a target domain with the goal of creating pseudo in-domain bitext; Hoang et al. (2018) generate synthetic parallel data iteratively from increasingly better back-translation models for improving unsupervised NMT; Fadaee and Monz (2018) focus on the diversity of synthetic samples and sample synthetic translations containing words that are difficult-to-predict using prediction losses and frequencies of words. By contrast, our empirical study investigates whether synthetic translations can be used to ***selectively replace*** original references to improve bitext quality rather than augmenting it.

Bitext Quality Mining bitext from the web results in large-scale corpora that are usually collected without guarantees about their quality. For instance, they contain noisy samples, ranging from untranslated sentences to sentences with no linguistic content (Khayrallah and Koehn, 2018; Caswell et al., 2020). Some of this noise is typically filtered out automatically using heuristics (Ramírez-Sánchez et al., 2020) or NMT model scores (Junczys-Dowmunt, 2018; Koehn et al., 2019). Yet, even after this noise filtering, a wide range of the remaining samples contains small meaning mismatches (Briakou and Carpuat, 2020) that are, however, treated as exact equivalents. Our work explores whether synthetic translations can be used to replace potentially fine-grained divergences rather than coarse divergences and noise.

3 Approach

This section describes the methods and data we use to produce revised bitexts for our empirical study.

3.1 Methods for Revising Bitext

We rely on established techniques that can be applied using only the bitext that we seek to revise. First, we train NMT models on the original bitext to translate in both directions. For each original sentence-pair, we generate a pool of synthetic translations using NMT and apply a divergence ranking criterion to decide whether and how to replace the original references with a better translation. Algorithm 1 gives an overview of the process, and we describe each step below.

Generating synthetic translations We train NMT models $M_{S\to T}$ and $M_{T\to S}$ on the original bitext to translate in each direction (lines 2-3). For each sentence-pair, they are used to generate two candidates for replacement by forward and backward translation (lines 6-7): $(S_i, M_{S\to T}(S_i))$ and $(M_{T\to S}(T_i), T_i)$. As a result, NMT models translate the exact same data that they are trained on. We thus expect translation quality to be high , and that local errors in the original bitext might be corrected by the translation patterns learned by NMT models on the entire corpus.

Selective Replacement We propose to replace an original pair by a candidate *only if* the candidate is predicted to better convey the meaning of the source than the original, which we refer to as the *semantic equivalence condition*. We implement this by ranking the original sample (S_i, T_i), its revision by forward translation $(S_i, M_{S\to T}(S_i))$ and its revision by back-translation $(M_{T\to S}(T_i), T_i)$, according to their degree of semantic equivalence. If

Algorithm 1 Revising Bitext: Given a bitext $\mathcal{D} = (S, T)$, a divergent scorer $\mathbf{R}$, and a margin score t, return an equivalized bitext $\tilde{\mathcal{D}}$

1: **procedure** TRAIN($\mathcal{D} = (S, T)$)
2: Train $M_{S \to T}$ on $\mathcal{D}$ until convergence
3: **return** $M_{S \to T}$
4: **end procedure**
1: **procedure** EQUIVALIZE($\mathcal{D} = (S, T)$)
2: $M_{S \to T} \leftarrow$ TRAIN($\mathcal{D} = (S, T)$)
3: $M_{T \to S} \leftarrow$ TRAIN($\mathcal{D} = (T, S)$)
4: $\tilde{\mathcal{D}} \leftarrow \varnothing$
5: **for** $i \in 1,...,|\mathcal{D}|$ **do**
6: $(S_i, \hat{T}_i) \leftarrow (S_i, M_{S \to T}(S_i))$
7: $(\hat{S}_i, T_i) \leftarrow (M_{T \to S}(T_i), T_i)$
8: $d_F = \mathbf{R}(S_i, \hat{T}_i) - \mathbf{R}(S_i, T_i)$
9: $d_B = \mathbf{R}(\hat{S}_i, T_i) - \mathbf{R}(S_i, T_i)$
10: **if** $\max(d_F, d_B) > t$ **then**
11: **if** $\max = d_F$ **then**
12: $\tilde{\mathcal{D}} \leftarrow \tilde{\mathcal{D}} \cup \{(S_i, \hat{T}_i)\}$
13: **else**
14: $\tilde{\mathcal{D}} \leftarrow \tilde{\mathcal{D}} \cup \{(\hat{S}_i, T_i)\}$
15: **end if**
16: **else**
17: $\tilde{\mathcal{D}} \leftarrow \tilde{\mathcal{D}} \cup \{(S_i, T_i)\}$
18: **end if**
19: **end for**
20: **return** $\tilde{\mathcal{D}}$
21: **end procedure**

none of the synthetic samples score higher than the original, it is not replaced (line 17). Otherwise, the original is replaced by the highest scoring synthetic sample (lines 10-15). As a result the cardinality of the bitext remains constant. Semantic equivalence is predicted using a multilingual language model that we fine-tuned to generate the scores d_F and d_B (lines 8-9). Following Briakou and Carpuat (2020), fine-tuning is done on synthetic samples generated by perturbations of the original bitext (e.g., deletions, lexical or phrasal replacements) performed without any bilingual information.

3.2 Experimental Set-Up

Bitext We evaluate the use of synthetic translations for revising bitext on two language pairs of the WikiMatrix corpus (Schwenk et al., 2021). WikiMatrix consists of sentence-pairs mined from Wikipedia pages using language agnostic sentence embeddings (LASER) (Artetxe and Schwenk, 2018). Prior work indicates that, as expected, the corpus as a whole comprises many samples that are not exact translations: Caswell et al. (2021) report that for more than half of the audited low-resource language-pairs, mined pairs are on average misaligned; Briakou and Carpuat (2020) find that 40% of a random sample of the English-French bitext are not semantically equivalent, and include fine-grained meaning differences in addition to alignment noise. We focus on bitexts with fewer than one million sentence pairs in Greek↔English (EL↔EN, with 750,585 pairs) and Romanian↔English (RO↔EN, with 582,134 pairs), because improving bitext is particularly needed in this data regime. In much higher resource settings, filtering strategies might be sufficient as there might be more high quality samples overall. In much lower resource settings, the data is likely too noisy or too small to effectively revise bitexts using NMT. We filter out noisy pairs in the training data using `bicleaner` (Ramírez-Sánchez et al., 2020) so that our empirical study excludes the most obvious forms of noise, and focuses on the harder case of revising samples that standard preprocessing pipelines consider to be clean.[1]

Preprocessing We use the standard Moses scripts (Koehn et al., 2007) for punctuation normalization, true-casing, and tokenization. We learn 32K BPEs (Sennrich et al., 2016b) per language using `subword-nmt` [2].

NMT Models We use the base Transformer architecture (Vaswani et al., 2017) and include details on the exact architecture and training in Apendix C.

Selective Replacement The divergence ranking models are trained using the implementation of Briakou and Carpuat (2020).[3] Synthetic divergences are generated starting from the 5,000 top scoring WikiMatrix sentences based on LASER score (i.e., seed equivalents). We fine-tune the "BERT-Base Multilingual Cased" model (Devlin et al., 2019) and set the margin equal to 5 as per the original implementation. We use the same margin value for the margin score of Algorithm 1. The model is tested by prior work for the English-French language-pair yielding 84 F1 on a set of human-annotated fine-grained divergences in WikiMatrix.

4 Intrinsic Evaluation of Bitext Quality

4.1 Human evaluation

We ask 3 bilingual speakers to evaluate the quality of the EN-EL bitexts. Given an original source sentence, they are asked to rank the original target and the candidate target in the order of their equivalence

[1] https://github.com/bitextor/bicleaner
[2] https://github.com/rsennrich/subword-nmt
[3] https://github.com/Elbria/xling-SemDiv

Candidate set	% Equivalized	Kendall's τ
ALL	60.0%	0.321
$d < 0$	26.4%	0.157
$0 \leq d \leq 5$	51.0%	0.234
$d > 5$	**87.5%**	**0.688**

Table 1: Human evaluation results for all evaluated pairs and ablation sets for different thresholds on divergent score differences between candidates and originals (i.e., d).

to the source. They are asked "Which sentence conveys the meaning of the source better?", and ties are allowed. A random sample of 100 pairs from forward and backward MT is annotated.

As can be seen in Table 1, 60% of ALL synthetic candidates are better translations of the WikiMatrix reference, which confirms the potential of NMT for improving over original translations. Further ablations confirm the benefits of selecting these synthetic candidates with the semantic equivalence condition. When the divergent scorer ranks a candidate higher than the original by a small margin (i.e., $0 \leq d \leq 5$ given $d = R(S_i, M_{S \to T}(T_i)) - R(S_i, T_i)$)), human evaluation shows that the candidate is actually better than the original only 51% of the times. When using our exact semantic equivalence condition ($d > 5$), candidates are judged as more equivalent that the original 87.5% of the times, and annotations within this set have stronger agreement (i.e, 0.688 Kendall's τ). This indicates that the condition $d > 5$ identifies more clear-cut examples of synthetic translations that fix semantic divergences in the original data and can be thus used for selective replacement of imperfect references by better quality translations.

Further inspection of the annotations reveals that most of the source-target WikiMatrix examples contain fine meaning differences (56%). In those cases we observe that most of the content between the sentences is shared but either small segments are mistranslated (e.g., "London" instead of "Athens" in the first example of Table 2), or some information is missing from either side of the pair (e.g., "all six" missing from the target side in the third example of Table 2). Furthermore, more coarse-grained divergences are found less frequently (12%)—in those cases we notice that sentences are usually either topically related or structurally similar (e.g., length, syntax) with a few anchor words (e.g, last example in Table 2). Finally, 32% of the times the original WikiMatrix pairs are perfect translation of each other.

WM-SRC	Απεβίωσε στην Αθήνα στις 5 Ιουνίου 1979.
GLOSS-SRC	*He died in Athens on 5 June 1979.*
WM-TGT	He died in London on 5 June 1979.
ST-TGT	He died in Athens on 5 June 1979.
WM-SRC	Ένας από τους οικισμούς που δημιούργησαν ήταν ο Καραβάς.
GLOSS-SRC	*Karavas was one of the first settlements they created.*
WM-TGT	One of the first towns to be created was Vila Barreto.
ST-TGT	One of the first settlements to be created was Karavas.
WM-SRC	Και οι έξι λέβητες κατασκευάστηκαν από την Waagner-Biro.
GLOSS-SRC	*All six boilers were manufactured by Waagner-Biro.*
WM-TGT	Boilers were supplied by Waagner-Biro.
ST-TGT	All six boilers were manufactured by Waagner-Biro.
WM-SRC	Το Διδακτικό προσωπικό της Σχολής είναι υψηλού επιπέδου.
GLOSS-SRC	*The school's teaching staff is of a high level.*
WM-TGT	The medical research level of the school is high.
ST-TGT	The teaching staff of the school is high.
WM-SRC	Ανήκει στο τριπλό αστρικό σύστημα του Άλφα Κενταύρου.
GLOSS-SRC	*It belongs to the Alpha Centauri triple star system.*
WM-TGT	This is the triple alpha process.
ST-TGT	It belongs to the triple star system of Alpha Centauri.
WM-SRC	Η εμφάνιση τυφώνων είναι σύνηθες φαινόμενο.
GLOSS-SRC	The occurrence of hurricanes is a common phenomenon.
WM-TGT	It is extremely rare: There were only 10 known cases in 1998.
ST-TGT	The appearance of hurricanes is a common phenomenon.

Table 2: Randomly sampled WikiMatrix (WM) pairs with synthetic translations (ST) that satisfy $d > 5$. ST successfully revise divergences of different granularities (highlighted segments) in the original references.

4.2 How do synthetic translations differ from originals?

Figure 1 presents the distribution of lexical differences (i.e., computed using LeD—a score that captures lexical differences based on the percentages of tokens that are not found in two sentences (Niu and Carpuat, 2020)) between original and synthetic translations (in EN) for candidates that replace and do not replace the originals. [4] First, we observe that a substantial amount of synthetic translations that do not replace original references (40%) corresponds to small LED scores (< 0.1), suggesting that the equivalence criterion could fall back to the original sentence not because of the poor quality of candidate references, but rather due to them being already close to the originals. Furthermore, all synthetic translated instances are represented in almost all bins, with fewer instances found on the extreme bins of > 0.7 LED scores. Finally, synthetic translations that replace original references are mostly concentrated within the range $[0.2, 0.6]$ of LeD scores. This indicates that they share lexical content with the original, which further supports the hypothesis that synthetic translations revise fine grained meaning differences in WikiMatrix in addition to alignment noise.

[4] LeD details are in Appendix A.

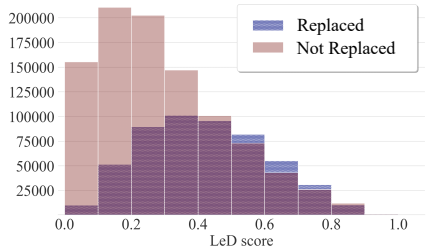

Figure 1: LeD differences of original vs. synthetic translations (EL→EN). Replaced candidates share lexical content with the originals.

4.3 How does the revised bitext differ from the original?

Table 3 presents differences in statistics of the original vs. revised WikiMatrix EN-EL bitexts to shed more light on the impact of selectively using synthetic translation for bitext quality improvement. [5] The refined bitext exhibits higher coverage (i.e., ratio of source words being aligned by any target words; rows 5 and 13) and smaller complexity (i.e., the diversity of target word choices given a source word (Zhou et al., 2019b)) compared to the original bitext. Moreover, the use of synthetic translations introduces small decreases in the lexical types covered in the final corpus (i.e., rows 3 and 11) which is expected as the additional coverage in the original corpus might be a result of divergent texts. Those observations are in line with prior work that seeks to characterize the nature of synthetic translations used in other settings, such as knowledge distillation (Zhou et al., 2019a; Xu et al., 2021).

While fixing divergent references contributes to this simplification effect, NMT translations might also reinforce unwanted biases from the original bitext. For instance, the distribution of two grammatical gender pronouns on the English side is a little more imbalanced in the improved bitext than in the original (rows 6-7 and 14-15), likely due to gender bias in NMT (Stanovsky et al., 2019). This calls for techniques to mitigate such biases (Saunders and Byrne, 2020; Stafanovičs et al., 2020) for NMT and other downstream tasks.

5 Extrinsic Evaluation of Bitext Quality

Our previous analysis suggests that selective replacement of divergent references with synthetic translations results in bitext of *improved quality*, with reduced level of noises and easier word-level mappings between the two languages, when compared to the original WikiMatrix corpus. To better understand how those differences impact downstream tasks we contrast the improved bitext with the original through a series of extrinsic evaluations for EN-EL and EN-RO languages that rely on parallel texts as training samples (see §5.2). First, we focus on the recent state-of-the-art unsupervised BLI approach of Shi et al. (2021) that relies on word-alignments of extracted bitexts. Second, we follow the recent bitext quality evaluation frameworks adopted by the "Shared Task on Parallel Corpus Filtering and Alignment" (Koehn et al., 2020) and built neural machine translation systems from scratch and by continued training on a multilingual pre-trained transformer model. Finally, we conduct extensive ablation experiments to test the impact of using synthetic translations without the semantic equivalence condition and contrast with familiar techniques used by prior work (see §5.3).

	Property	Original	Revised	δ
	EN			
1 :	#Sents	750,585	750,585	0.0%
2 :	#Tokens	15,244,413	15,239,474	−0.3%
3 :	#Types	358,681	350,224	−2.4%
4 :	Avg. Length	20.3	20.3	0%
5 :	Avg. Coverage	0.78	0.83	+6.0%
6 :	# SHE/HER/HERS Pronouns	45,028	43,629	−3.1%
7 :	# HE/HIS/HIM Pronouns	185,356	194,510	+4.7%
8 :	Complexity	63.03	53.61	−14.9%
	EL			
9 :	#Sents	750,585	750,585	0.0%
10 :	#Tokens	15,743,084	15,611,937	−0.8%
11 :	#Types	526,411	519,558	−1.3%
12 :	Avg. Length	21.0	20.8	−1.0%
13 :	Avg. Coverage	0.77	0.83	+7.0%
14 :	# Η/ΤΗΣ/ΤΗΝ Pronouns	792,005	776,947	−1.9%
15 :	# Ο/ΤΟΥ/ΤΟΝ Pronouns	799,249	794,275	−0.6%
16 :	Complexity	24.51	17.85	−27.0%

Table 3: Comparison of original vs. revised bitext. δ gives percentage differences between them.

5.1 Experimental Set-Up

BLI The task of BLI aims to induce a bilingual lexicon consisting of word translations in two languages. We experiment with the recently proposed method of Shi et al. (2021) that combines extracted bitext and unsupervised word alignment to perform fully unsupervised induction based on extracted statistics of aligned word-pairs. The induced lexicons are evaluated based on MUSE (Lample et al., 2018) consisting of 45,515 and 80,815 dictionary entries for EL-EN and EN-RO, respectively.[6] We extract word alignments using MBERT-based *Simalign*[7] (Jalili Sabet et al., 2020) and statistics based on the implementation of Shi et al. (2021).[8]

[5] Details on the metrics are in Appendix A.

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

[7] https://github.com/cisnlp/simalign

[8] https://github.com/facebookresearch/bitext-lexind

		All				*Low*	*Medium*	*High*
Pair	Bitext	Precision	Recall	F1	OOV rate		Precision	
EL-EN	Original	76.2	58.1	65.9	**6.7%**	59.4	76.6	81.4
	Revised	**77.6***	**58.6***	**66.8***	7.5%	**60.4***	**78.4***	81.6
EN-RO	Original	89.2	69.4	78.1	**15.8%**	78.6	86.9	87.1
	Revised	**90.8***	**71.3***	**79.8***	16.5%	**80.0***	**87.5***	86.9

Table 4: Results on MUSE for unsupervised BLI extrinsic evaluations. Revised bitexts yield statistically significant (∗) improvements over the original bitexts overall and for low-to-medium frequency dictionary entries.

MT We experiment with MT tasks following two approaches: (1) training standard transformer seq2seq models from scratch; (2) continued training for mT5 (Xue et al., 2021), a multilingual pretrained text-to-text transformer. We evaluate translation quality with BLEU (Papineni et al., 2002)[9] on the official development and test splits of the TED corpus (Qi et al., 2018). [10] For (1) we follow the experimental settings described in §3.2. For (2) we initialize the weights of transformer with "mT5-small" which consists of 300M parameters,[11]. We use the `simpletransformers` implementation.[12] We fine-tune for up to 5 epochs and include the parameter settings in Appendix D.

Ablation Settings We compare the NMT models trained on the variants of the synthetic bitext to isolate the impact of replacement criteria and of different candidates.[13] For the former, we experiment with the **rejuvenation** approach of Jiao et al. (2020) that replaces original references with forward translated candidates for the 10% least active original samples measured by NMT probability scores. Moreover, we experiment with **forward** and **backtranslation** baselines trained on bitexts that consist solely from target- or source-side candidate sentences (i.e., original references are entirely excluded) and with ablations that consider either forward or backward candidates for the proposed semantic equivalence condition. Finally, we consider two alternatives to the **semantic equivalence** condition based on divergent scores: the **ranking** condition replaces a candidate if it scores higher that the original (i.e., margin with $t = 0$) and the **thresholding** condition adds the additional constraint that candidates should rank higher than a threshold to replace the original pair.

[9] https://github.com/mjpost/sacrebleu

[10] Data statistics are found in Appendix E.

[11] https://github.com/google-research/multilingual-t5

[12] https://github.com/ThilinaRajapakse/simpletransformers

[13] Results on development sets are in Appendix B.

5.2 Extrinsic Evaluation Results

BLI Table 4 presents results for unsupervised BLI on the MUSE gold-standard dictionaries, for EL-EN and EN-RO. Across languages, the revised bitexts induce better lexicons compared to the original WikiMatrix. Crucially, improvements are reported both in terms of Recall—which connects to the observation that the revised bitext exhibits higher coverage than the original, and in terms of Precision—which connects to the noise reduction effect that impacts the extracted word-alignments. Additionally, a break-down on the Precision of the induced lexicons binned by the frequency of MUSE source-side entries (i.e., last 3 columns in Table 4) reveals that the improvements come from better induction of low- and medium-frequency words which we expect are more sensitive to noisy misalignments that result from divergent bitext. Finally, those improvements are reported despite the small increase of the OOV rate in the revised lexicons that results from decrease in the lexical types covered in it, as mentioned in the analysis (i.e., §4.3).

Furthermore, following the advice of Kementchedjhieva et al. (2019) who raise concerns on BLI evaluations based on gold-standard pre-defined dictionaries, we accompany our evaluation with manual verification to confirm that our conclusions are consistent with those of the automatic evaluation. Concretely, we manually check the *false positives* induced translation pairs from the original vs. the improved bitext. We found that 65/80 are *false false positives* (due to incompleteness of pre-defined dictionaries) for the improved bitext and 51/80 for the original (see Appendix F for the complete list). This confirms that the metric improvements we observe are meaningful and suggests that the improved bitext help learn better mappings between source and target words.

Pair	Original	Revised
EL→EN	28.15 ± 0.13	$\mathbf{29.63 \pm 0.29}$
EN→EL	27.08 ± 0.18	$\mathbf{27.89 \pm 0.05}$
RO→EN	23.68 ± 0.12	$\mathbf{24.54 \pm 0.06}$
EN→RO	20.65 ± 0.10	$\mathbf{20.84 \pm 0.04}$

Table 5: BLEU on NMT training from scratch.

MT Table 5 presents translation quality (BLEU) on EN↔RO and EN↔EL tasks for MT training from scratch and Figure 2 shows translation quality of mT5 continued training across epochs. Across tasks and settings, the revised bitext yields better translation quality than the original WikiMatrix data. The consistent improvements we observe across the two settings suggest that the properties of the synthetic translations that replace original samples and bring those improvements are invariant to specific models. Moreover, the magnitude of improvements is larger in the continued training setting compared to training from scratch (e.g., $\sim$ +0.8 vs. $\sim$ +1.5, for EN→EL; $\sim$ +0.2 vs. $\sim$ +1.5, for RO→EN). The latter suggests that improvements from using synthetic samples do not only come from the normalization effect (i.e., synthetic samples are easier to model by NMT) but also connect to the reduced noise in the training samples. This further complements our hypothesis that synthetic translations can improve the quality of imperfect references that should, in principle, yield noisy training signals—and thus impact the resulting quality—of different MT models.

5.3 Ablation Study

Table 6 compares the translation quality (BLEU) of NMT systems trained on different synthetic translations. By forcing the semantic equivalence condition when deciding whether a synthetic translation replaces an original, we revise 50% of the latter yielding the best results across directions with significant improvements (i.e, increases do not lie within 1 stdev of the original's bitext performance) of +0.81 (EN→EL, row 9) and +1.49 (EL→EN, row 18) points over the original bitext.

Impact of semantic equivalence condition Table 6 shows that naively disregarding the original references and training only on synthetic translations gives mixed results: training on *forward-translated* references only (i.e., row 2) gives small improvements (+0.36) over the model trained on WikiMatrix for EN→EL, while it performs comparably to it for EL→EN (i.e., row 11). On the other hand, training on *backward* data only (i.e., row 12) improves BLEU by a small margin (+0.23) for MT into EN while it hurts BLEU when translating into EL (i.e., row 3). This indicates that the good quality of the synthetic translations cannot be taken for granted and motivates replacing original pairs under conditions that account for semantic controls.

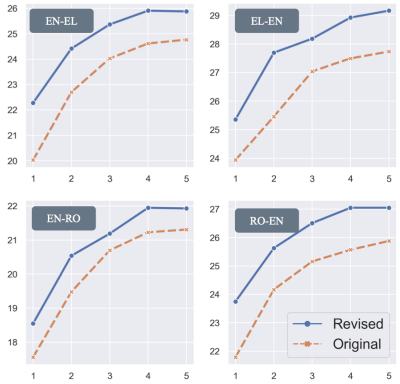

Figure 2: BLEU scores across epochs (x-axis) for continued training on mt5. The revised bitext improves translation quality compared to the original for all epochs and translation tasks.

The latter is further confirmed by results on the rejuvenation baseline: replacing candidates for the 10% of the most inactive WikiMatrix samples results in small and insignificant increases in BLEU when compared to models trained on original WikiMatrix data (i.e., rows 1-4 and 10-13). This indicates that rejuvenation might not be well-suited to lower resource settings than the ones it was originally tested on (Jiao et al., 2020). The rejuvenaton technique might be affected by the decreased NMT quality and calibration in lower resource settings. By contrast using synthetic translations with semantic control mitigates their impact.

Finally, all three semantic control variants based on divergent scores yield bitexts that improve BLEU compared to the original WikiMatrix (i.e., rows 5-8 and 14-18). Among them, the *margin* condition is the most successful, followed by the *thresholding* variant. The break down of training statistics reveals the reason behind their differences: the *thresholding* condition is a more strict constraint as it only allows synthetic candidates to replace the original pairs if they are predicted as exact equivalents, allowing for fewer revision of divergent pairs in WikiMatrix. By contrast, the condition based on *margin* is a contrastive approach that allows for more revisions of the original data (i.e., a candidate might be a more fine-grained divergent of the source). The *ranking* criterion is the least suc-

	SELECTIVE REPLACEMENT	DATA TYPES	BLEU	δ	BITEXT STATISTICS (O)	(F)	(B)	VISUALIZATION
			EN→EL					
1 :	✗	O	27.08 ± 0.18	–	100%	0%	0%	
2 :	✗	F	27.45 ± 0.06	+0.36	0%	100%	0%	
3 :	✗	B	26.22 ± 0.26	−0.86	0%	0%	100%	
4 :	Rejuvenation	O & F	27.24 ± 0.11	+0.16	90%	10%	0%	
5 :	Ranking	O & F	27.21 ± 0.43	+0.13	22%	78%	0%	
6 :	Thresholding	O & F	27.56 ± 0.11	+0.48	78%	21%	0%	
7 :	Semantic equivalence (margin)	O & F	27.64 ± 0.22	+0.56	63%	37%	0%	
8 :	Semantic equivalence (margin)	O & B	27.61 ± 0.09	+0.52	66%	0%	34%	
9 :	Semantic equivalence (margin)	O & F & B	$\mathbf{27.89 \pm 0.05}$	**+0.81**	50%	23%	27%	
			EL→EN					
10 :	✗	O	28.15 ± 0.13	–	100%	0%	0%	
11 :	✗	F	28.16 ± 0.17	+0.01	0%	100%	0%	
12 :	✗	B	28.38 ± 0.09	+0.23	0%	0%	100%	
13 :	Rejuvenation	O & F	28.27 ± 0.12	+0.12	90%	10%	0%	
14 :	Ranking	O & F	28.81 ± 0.13	+0.67	26%	74%	0%	
15 :	Thresholding	O & F	28.79 ± 0.17	+0.64	81%	19%	0%	
16 :	Semantic equivalence (margin)	O & F	29.00 ± 0.15	+0.85	66%	34%	0%	
17 :	Semantic equivalence (margin)	O & B	29.19 ± 0.25	+1.05	63%	0%	37%	
18 :	Semantic equivalence (margin)	O & F & B	$\mathbf{29.63 \pm 0.29}$	**+1.49**	50%	27%	23%	

Table 6: BLEU results (averages of 3 seeds) on EN↔EL NMT. δ denotes average improvements over the original bitext. Bitext statistics give percentage of original (O), forward (F), and backward (B) translated candidates. First column shows the selective replacement condition for candidate replacement (when applicable).

cessful method—this is expected as the divergence ranker is not trained as a regression model.

Impact of bi-directional candidates Considering both forward (F) and backward (B) translated candidates during selective replacement, yields to further improvements (0.22-0.44 points) over bitext induced by the semantic equivalence condition with candidates from a single NMT model (i.e., rows 7-9 and 16-18). When forward and backward candidates are considered independently, they replace $34 - 37\%$ of the original pairs; in contrast, when considered together they replace 50% of original WikiMatrix pairs. As a result, there is no perfect overlap between the original pairs replaced by the forward vs. backward model which motivates the use of both to revise more divergences in WikiMatrix. This finding raise the question of whether using synthetic translations from both directions might benefit other scenarios, such as knowledge distillation.

6 Conclusion

This paper explored how synthetic translations can be used to revise bitext, using NMT models trained on the exact same data we seek to revise. Our extensive empirical study surprisingly shows that, even without access to further bilingual data or supervision, this approach improves the quality of the original bitext, especially when synthetic translations are generated in both translation directions, and selectively replace the original using a semantic equivalence criterion. Specifically, intrinsic evaluation showed that, synthetic translations are of sufficient quality to improve over the original references, in addition to "normalizing" the bitext as suggested by prior work and corpus level statistics (Zhou et al., 2019a; Xu et al., 2021). . Extrinsic evaluations further show that the replaced synthetic translations provide more useful signal for BLI tasks and NMT training in two settings (i.e., training from scratch and continued training).

These findings provide a foundation for further exploration of the use of synthetic bitext. First, we focused our empirical study on language pairs and datasets where revising bitexts is the most needed and most likely to be useful: the resources available for these languages are not so large that mined bitext can simply be ignored or filtered with simple heuristics, yet there is enough data to build NMT systems of reasonable quality (i.e., $\sim$ 600K segments for EN-RO, and $\sim$ 750K for EN-EL). While in principle selective replacement of divergent references with synthetic translations should port to high-resource settings, where NMT is as good or better than for the languages considered in this work, other techniques are likely needed in low-resource settings where NMT quality is too low to provide reliable candidate translations. Second, having established that the revised bitext improves the quality of the original bitext in isolation, it remains to be seen how to best revise bitexts in more heterogeneous scenarios with diverse sources of parallel or monolingual corpora. Overall, as synthetic data generated by NMT is increasingly used to improve cross-lingual transfer in multilingual NLP, our study motivates taking a closer look at the properties of synthetic samples, to better understand how they might impact downstream tasks beyond raw performance metrics.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. Bilingual lexicon induction through unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5002–5007, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *CoRR*, abs/1812.10464.

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

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on 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.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. 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 Volume 2: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

Marzieh Fadaee and Christof Monz. 2018. Back-translation sampling by targeting difficult words in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 436–446, Brussels, Belgium. Association for Computational Linguistics.

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 414–420, Montreal, Quebec, Canada. Association for Computational Linguistics.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. 2017. Non-autoregressive neural machine translation. *CoRR*, abs/1711.02281.

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.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Wenxiang Jiao, Xing Wang, Shilin He, Irwin King, Michael Lyu, and Zhaopeng Tu. 2020. Data Rejuvenation: Exploiting Inactive Training Examples for Neural Machine Translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2255–2266, Online. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Yova Kementchedjhieva, Mareike Hartmann, and Anders Søgaard. 2019. Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. 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 3336–3341, Hong Kong, China. 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.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

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.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

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.

Xuan-Phi Nguyen, Shafiq Joty, Kui Wu, and Ai Ti Aw. 2020. Data diversification: A simple strategy for neural machine translation. In *Advances in Neural Information Processing Systems*, volume 33, pages 10018–10029. Curran Associates, Inc.

Xing Niu and Marine Carpuat. 2020. Controlling neural machine translation formality with synthetic supervision. *Proceedings of the AAAI Conference on Artificial Intellige nce*, 34(05):8568–8575.

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.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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.

Haoyue Shi, Luke Zettlemoyer, and Sida I. Wang. 2021. Bilingual lexicon induction via unsupervised bitext construction and word alignment. 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 813–826, Online. Association for Computational Linguistics.

Artūrs Stafanovičs, Mārcis Pinnis, and Toms Bergmanis. 2020. Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the Fifth Conference on Machine Translation*, pages 629–638, Online. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llio n 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.

Weijia Xu, Shuming Ma, Dongdong Zhang, and Marine Carpuat. 2021. How does distilled data complexity impact the quality and confidence of non-autoregressive machine translation? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4392–4400, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.

Chunting Zhou, Graham Neubig, and Jiatao Gu. 2019a. Understanding knowledge distillation in non-autoregressive machine translation. *CoRR*, abs/1911.02727.

Chunting Zhou, Graham Neubig, and Jiatao Gu. 2019b. Understanding knowledge distillation in non-autoregressive machine translation. *CoRR*, abs/1911.02727.

A Details on bitext analysis

Complexity We follow Zhou et al. (2019b) and compute the corpus complexity as a measure of translation uncertainty. Concretely, having access to an alignment model (here, `fast-align`), the complexity of a corpus d is computed by averaging the entropy of target words y conditioned on the source x, $L(d) = \frac{1}{|\mathcal{V}_x|} \sum_{x \in \mathcal{V}_x} H(y|x)$.

Coverage We follow Tu et al. (2016) and measure the coverage of each source-target parallel pair as the ratio of source words being aligned to target words, having access to an alignment model (here, `fast-align`). We compute the coverage for source-target and target-source bitexts separately. Corpus-level statistics correspond to average sentence-level results.

Grammatical Gender Pronouns The complete lists of grammatic gender pronouns we use for EL are: [ο, του, τον, αυτός, αυτού, αυτόν, εκείνος, εκείνου, εκείνον, οποίος, οποίου, οποίον] and [η, της, την, αυτήν, αυτής, αυτήν, εκείνη, εκείνης, εκείνην, οποία, οποίας, οποίαν].

Lexical Differences (LeD) We follow (Niu and Carpuat, 2020) and compute the Lexical Differences score between two sentences S_1 and S_2 as the percentage of tokens that are not found in both, LeD$= \frac{1}{2}(\frac{|S_1 / S_2|}{|S_1|}) + \frac{|S_2 / S_1|}{|S_2|}$.

B Result on development sets

Table 7 presents results on the main and secondary NMT tasks on TED developments sets. The refined bitext leads to consistent and significant improvements in BLEU across language-pairs and translation directions.

C Sockeye2 configuration details

We use the base Transformer architecture (Vaswani et al., 2017). with embedding size of 512, transformer hidden size of 2,048, 8 attention heads, 6 transformer layers, and dropout of 0.1. Target embeddings are tied with the output layer weights. We train with label smoothing (0.1). We optimize with Adam (Kingma and Ba, 2015) with a batch size of 4,096 tokens and checkpoint models every 1,000 updates. The initial learning rate is 0.0002, and it is reduced by 30% after 4 checkpoints without validation perplexity improvement. We stop training after 20 checkpoints without improvement. We select the best checkpoint based on validation BLEU (Papineni et al., 2002). All models are trained on a single GeForce GTX 1080 GPU. Tables 8 presents details of NMT training with `Sockeye2`.

Table 6

	EN→ EL		EL→ EN
1 :	25.50 ± 0.15	10 :	27.98 ± 0.18
2 :	25.52 ± 0.07	11 :	27.92 ± 0.15
3 :	24.55 ± 0.25	12 :	27.70 ± 0.15
4 :	25.35 ± 0.14	13 :	27.99 ± 0.15
5 :	25.27 ± 0.41	14 :	28.36 ± 0.13*
6 :	25.66 ± 0.05*	15 :	28.34 ± 0.18*
7 :	25.73 ± 0.14*	16 :	28.66 ± 0.14*
8 :	25.71 ± 0.19*	17 :	28.65 ± 0.27*
9 :	$\mathbf{25.91 \pm 0.09}$*	18 :	$\mathbf{29.00 \pm 0.26}$*

Table 5

	EN→ RO		RO→ EN
1 :	21.94 ± 0.11	3 :	24.98 ± 0.16
2 :	$\mathbf{22.05 \pm 0.03}$*	4 :	$\mathbf{26.11 \pm 0.20}$*

Table 7: BLEU results on the TED developments sets for each of the results of Tables 6 and 5 (enumeration follows the main text Tables). * denote one standard deviation improvements over the original bitexts.

```
-weight-tying-type="trg_softmax" #uni-NMT
-weight-tying-type="src_trg_softmax" #bi-NMT
-num-words 5000:5000
-label-smoothing 0.1
-encoder transformer
-decoder transformer
-num-layers 6
-transformer-attention-heads 84
-transformer-model-size 512
-num-embed 512
-transformer-feed-forward-num-hidden 2048
-transformer-preprocess n
-transformer-postprocess dr
-gradient-clipping-type none
-transformer-dropout-attention 0.1
-transformer-dropout-act 0.1
-transformer-dropout-prepost 0.1
-max-seq-len 80:80
-batch-type word
-batch-size 2048
-min-num-epochs 3
-initial-learning-rate 0.0002
-learning-rate-reduce-factor 0.7
-learning-rate-reduce-num-not-improved 4
-checkpoint-interval 1000
-keep-last-params 30
-max-num-checkpoint-not-improved 20
-decode-and-evaluate 1000
```

Table 8: NMT configurations on `Sockeye2`

D mt5 configuration details

Tables 9 presents details of continued training of mT5 on `SimpleTransformers`.

```
max-seq-length 100
train-batch-size 10
eval-batch-size 10
num-train-epochs 5
scheduler 'cosine schedule with warmup'
evaluate-during-training True
evaluate-during-training-steps 10000
learning-rate 0.0003
optimizer 'Adafactor'
use-multiprocessing False
save-model-every-epoch True
use-early-stopping False
do-lower-case True
```

Table 9: NMT configurations for continued training of mT5 on `SimpleTransformers`.

E Data Statistics

Table 10 presents data statistics for WikiMatrix training data, and TED evaluation sets.

LANGUAGE PAIR	TRAINING	DEV.	TEST
EL-EN	750,585	3,344	4,431
RO-EN	582,134	3,904	4,631

Table 10: Data statistics after pre-processing.

F Manual inspection of BLI

Table 11 presents manual analysis results on False Positives entries.

Revised			Original		
αστεροειδές	star	?	απόστολος	apostolos	✗
προσφέρεται	offers	✓	βραχνό	raucous	✗
κεραυνός	keravnos	✗	μπανζούλ	bangaon	✗
συμπυκνώνει	encapsulates	?	βοηθητικές	auxiliary	✓
σεξτέτο	sexteto	✓	ομιλήτρια	spokesperson	✓
επιχειρηματολογία	argumentation	✓	πρωτεργάτη	forerunner	✗
επίπλωση	furniture	✓	αντιτρομοκρατική	anti-terrorist	✓
μπουγκ	bug	✗	πλεκτά	sweaters	✓
σχετικοί	related	✓	εμβολιαστεί	vaccinated	✓
δορυφόρους	moons	✗	αταξινόμητες	unclassified	✓
δειλή	timid	✓	στέιν	steen	✗
χάντινγκτον	huntingdon	✓	χιλιοστό	millimeter	✓
ποσότητες	amounts	✓	σελεστίν	célestine	✓
πλακέ	squamous	✓	κόβατς	kovács	✗
αποποίηση	relinquishing	?	σεμίνα	omni	✗
ατμούς	vapors	✓	σπάιντερμαν	spider-man	✓
τερματισμοί	endings	✓	πάνω	over	✓
αλεξανδρινό	alexandrine	✓	ενδιαφέρων	love	✗
σπασμοί	fits	?	αγριόγατες	cats	✗
σίδερα	sidelines	✗	αγορα	trade	✓
συνοδεύονται	are	✗	επικεφαλίδα	header	✓
διανέμονται	are	✗	μάσλοου	khan	✗
θραύση	fracturing	✓	τεχνητά	artificially	✓
κυβερνά	rule	✓	πέτροβιτς	petrović	✓
συνάξεις	meetings	✓	ανθίζει	flowers	✓
χριστιανία	christianity	✓	ζήτω	vive	✗
απειλούνται	are	✗	τυλίγει	picks	✗
ποινικοποίηση	penalize	✓	μπαέζ	ross	✗
στερέωμα	stardom	✗	φιλοδοξεί	is	✗
τζαπ	elford	✗	τρυφερή	loving	?
ταυρομαχία	bullfighting	✓	σωρός	remains	✗
χειρός	handbags	?	χαλυβουργεία	works	✗
κδ	cd	?	μάιρα	chloe	✗
τρομοκρατεί	terrorizes	✓	συγκλόνισε	shocked	✓
μακέι	mackey	✓	άτακτη	mischievous	✓
ζάκυνθο	zakynthos	✓	οταν	after	?
συμπτωματολογία	symptomology	✓	εντομοφάγα	insectivores	✓
πολυφυλετική	polyphyletic	✓	κραδασμούς	vibrations	✓
κούνια	cunha	✗	μπελάς	nuisance	✓
καταβεβλημένος	overcome	✓	πάστες	pastries	✓
απάτες	scams	✓	διασπαστική	divisive	✓
γιάννη	giannis	✓	κατάληψη	capture	✗
δηλητηριάσεις	poisonings	✓	παραδίδονται	surrender	✓
φιλόξενοι	colorful	✗	κλήρων	clergy	✓
φημισμένος	renowned	✓	σκεύη	vessels	✓
φουσκωμένα	filled	?	λεπτονίων	leptons	✓
υπονοούμενα	undertones	✓	εξάγονται	are	✗
όριο	boundary	✓	απότομο	abrupt	✓
χαλάρωσε	relaxed	✓	παρασυμπαθητικό	sympathetic	?
αισθητικός	aesthetic	✓	ταρίχευση	embalming	✓
ταμαντούα	tamanduas	✓	κεκτημένο	precedent	✗
εστίες	foci	?	καλκούτα	kolkata	✓
θεωρείται	is	✗	σίρι	sirri	✓
κορμό	trunk	✓	ξεπερασμένο	obsolete	✓
σπύρο	spyros	✓	ανώμαλος	bumpy	✓
αναισθητικά	anesthetics	✓	εξισορρόπησης	substance	✗
στρατηγικές	strategic	✓	πολυσακχαρίτης	polysaccharides	✓
αναπνέει	breathe	✓	επίπονες	persistent	✓
εξουδετερώσει	neutralize	✓	αμφιθέατρο	amphitheatre	✓
μελαγχολική	melancholic	✓	αναπληρωματικό	an	✗
θυμήθηκε	recalled	✓	εντελώς	entirely	✓
πασχαλίτσα	ladybird	✓	λιθόστρωτο	cobbled	✓
πυροκροτητές	caps	?	διοικητικοί	administrative	✓
κραυγαλέα	screaming	?	κομιστής	bearer	✓
μολδαβίας	moldavia	✓	συλλογικότητες	competitions	✗
σαλιγκάρι	shilling	✗	χουλιγκανισμού	micromanagement	✗
ενισχυθεί	enhance	✓	τσάρους	tsars	✓
πρεσβυτέριο	presbytery	✓	ντόνελ	dorff	✗
μάγιστρος	master	✓	κίραν	kiran	✓
αλτ	alt	✓	πρωτοποριακή	pioneering	✓
χρονολογία	date	✓	λένοξ	brookline	✗
κανένα	any	✓	λείπουν	are	✗
κορμός	road	✗	εξάντα	astronomy	✗
καθαριστήριο	cleanup	✗	πτωτική	downward	✓
ανατεθεί	assigned	✓	αρχιτεκτονικές	architectural	✓
εξοικονόμηση	save	✓	γαλλόφωνο	french-speaking	✓
μπαρακούντα	barracudas	✓	μέντε	mede	✗
ταυτοποίησης	identification	✓	εκθρονίζοντας	deposing	✓

Table 11: Manually labeled acceptability judgments for random 80 error cases made by lexicons induced using the original and revised bitexts. ✓and ✗ denote acceptable and unacceptable translation, respectively. ? denotes word pairs that may be acceptable in rare or specific contexts.