Synthesizing Parallel Data of User-Generated Texts with Zero-Shot Neural Machine Translation

Benjamin Marie **Atsushi Fujita**

National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
{bmarie, atsushi.fujita}@nict.go.jp

Abstract

Neural machine translation (NMT) systems are usually trained on clean parallel data. They can perform very well for translating clean in-domain texts. However, as demonstrated by previous work, the translation quality significantly worsens when translating noisy texts, such as user-generated texts (UGT) from online social media. Given the lack of parallel data of UGT that can be used to train or adapt NMT systems, we synthesize parallel data of UGT, exploiting monolingual data of UGT through crosslingual language model pre-training and zero-shot NMT systems. This paper presents two different but complementary approaches: One alters given clean parallel data into UGT-like parallel data whereas the other generates translations from monolingual data of UGT. On the MTNT translation tasks, we show that our synthesized parallel data can lead to better NMT systems for UGT while making them more robust in translating texts from various domains and styles.

1 Introduction

Neural machine translation (NMT) requires large parallel data for training. However, even when trained on large clean parallel data, NMT generates translations of very poor quality when translating out-of-domain or noisy texts. For instance, Michel and Neubig (2018) empirically showed that NMT systems trained on clean parallel data from the news and parliamentary debate domains perform reasonably well when translating news articles but poorly perform at translating user-generated texts (UGT) from a social media. UGT can be from various domains and manifest various forms of natural noise. For instance, they can exhibit spelling/typographical errors, words omission/insertion/repetition, grammatical/syntactic errors, or noise markers even more specific to the writing style of social media such as abbreviations, obfuscated profanities, inconsistent capitalization, Internet slang, and emojis. Normalizing and correcting them in a preprocessing step is a solution to facilitate translation (Gerlach et al., 2013; Matos Veliz et al., 2019), but it impedes the correct transfer of the style of the source text to its translation. In this paper, we posit that the NMT system should preserve the style during the translation. Another trend of work focuses on making NMT more robust in handling noisy tokens, such as tokens with spelling mistakes, which can greatly disturb NMT (Belinkov and Bisk, 2018). However, it has only a minimal impact in translating UGT (Karpukhin et al., 2019) that contains other types of noise/errors.

Whereas domain adaptation methods are helpful in improving NMT for UGT (Li et al., 2019), we do not usually have bilingual parallel data of UGT created by professional translators to train or adapt an NMT system. Consequently, previous work on NMT for UGT merely focused on scenarios for which we have UGT parallel data, such as the MTNT dataset (Michel and Neubig, 2018).

In contrast to previous work, we assume that parallel data of UGT are not available and that we can only rely on the formal and clean texts that are usually used to train NMT systems. In addition, we exploit UGT monolingual data that are publicly available in large quantity on the Internet for many languages. We propose to synthesize parallel data of UGT to train better NMT systems for UGT. For this purpose, we present two complementary approaches that associate a pre-trained crosslingual language model with zero-shot NMT systems. Our contributions are as follows:

- A method for altering clean parallel data into UGT parallel data

Transactions of the Association for Computational Linguistics, vol. 8, pp. 710–725, 2020. https://doi.org/10.1162/tacl_a_00341
Action Editor: Philipp Koehn. Submission batch: 4/2020; Revision batch: 7/2020; Published 11/2020.

Ex1	source	Et vlà après l'Italie le #COVID19 arrive en France !
	vanilla NMT	And that is what will happen after Italy's # 8 D19, which is coming to France!
	our work	And so, after Italy, # COVID19 arrives in France!
Ex2	source	Et voilà après l'Italie le COVID19 arrive est en France!
	vanilla NMT	Now, after Italy, CO6D19 has arrived in France !
	our work	Now, after Italy, COVID19 arrive is in France!
Ex3	source	Et voilà après l'Italie le COVID19 arrive en France !
	vanilla NMT	And now, after Italy, CO6D19 comes to France!
	our work	Now, after Italy, COVID19 is coming to France!
reference		And then after Italy the COVID19 arrives in France!

Figure 1: Examples of the impact of noise in NMT. The NMT systems are presented in Table 2. Vanilla NMT is trained on clean parallel data, whereas ''our work'' refers to the configuration #1+#2 presented in Section 5.4 trained on synthetic parallel data of UGT.

- A method for synthesizing parallel data of UGT from monolingual data

- An empirical evaluation, in four translation directions, of our methods that shows consistent improvements in translation quality over previous work for UGT but also on various domains and styles

The remainder of this paper is organized as follows. In Section 2, we present the research problem and questions that we answer in this work. Then, in Section 3, we present a zero-shot NMT framework that we use to synthesize parallel data of UGT by our two methods presented in Section 4. We evaluate the usefulness of our approaches for better translating UGT in Section 5. In Sections 6 and 7, we evaluate alternative configurations for our zero-shot NMT systems, and in Section 8 we verify whether our NMT systems trained on the synthetic parallel data are more robust to changes of domain and style. We analyze the synthetic sentences and present examples in Section 9 to better understand why our data lead to better NMT systems. Following the presentation of related work in Section 10, we conclude the paper in Section 11.

2 Motivation

UGT contains many different types of noise that can also differ from one type of UGT to another. For instance, posts on Twitter contain many spelling errors intentionally introduced for text compression, whereas this kind of error is rather marginal in the discussions from Reddit (Michel and Neubig, 2018).

Figure 1 shows the impact on MT of two different types of noise: spelling (Ex1) and syntactic (Ex2) errors, compared to the translation of the same but clean sentence (Ex3). Ex1 has an intentional spelling error ''vlà'' (instead of ''voilà'') and a UGT-specific symbol, ''#.'' Comparison with Ex3 suggests that they have negative effects on the vanilla NMT system and eventually lead to an incorrect translation largely different from the translation of the clean source of Ex3. In Ex2, a syntactic error ''arrive est'' instead of ''arrive'' has also an impact, but to a lesser extent, by inducing the past tense in English. Vanilla NMT gives the best translation for the clean source sentence (Ex3) only failing in translating ''COVID19.'' For indicative purpose, we present in the row ''our work'' translations generated by our work. These examples highlight the inability of vanilla NMT in translating sentences with various types of noise.

In conducting the research to better translate UGT, we answer the following research questions:

Q1 How can we generate synthetic parallel data for UGT in a specific domain/style without relying on any manually produced parallel data of UGT?

Q2 Do the synthetic parallel data lead to a better NMT system for the targeted UGT and do they make it more robust to the change of domain or style?

3 Zero-Shot NMT for Synthesizing Parallel Data

We describe in this section our zero-shot NMT system used to synthesize parallel data of UGT.

3.1 Objective and Prerequisites

Let L1 and L2 be two languages for clean texts and R1 and R2 for the same languages, respectively, but for UGT. The data prerequisites for our NMT system described in Section 3.2 are as follows:

- $P_{L1\text{-}L2}$ parallel data of clean and formal texts that are usually used for training NMT,
- M_{L1} and M_{L2} monolingual data from any domains, and
- M_{R1} and M_{R2} monolingual data of UGT.

Unlike previous work on NMT for UGT, we do not assume any $P_{R1\text{-}R2}$ parallel data for training or validating NMT systems, except for evaluation. $P_{L1\text{-}L2}$, M_{L1}, and M_{L2}, parallel and monolingual data, are usually used to build state-of-the-art NMT systems. M_{R1} and M_{R2} monolingual data are obtained by crawling social media.

Our objective is to synthesize parallel data of UGT, which we henceforth denote as $P^{S}_{R1\text{-}R2}$. To this end, we propose the following two approaches:

#1 Alter a clean parallel data $P_{L1\text{-}L2}$ into $P^{S}_{R1\text{-}R2}$

#2 Synthesize $P^{S}_{R1\text{-}R2}$ parallel data by translating M_{R2} monolingual data into R1

These approaches must regard L1 and R1, and similarly L2 and R2, as two different languages. For #1, we alter the $P_{L1\text{-}L2}$ parallel data by performing L1→R2 and L2→R1 translations.[1] For #2, we generate the data via R2→R1 translation. Note that L1→R2, L2→R1, and R2→R1 are all zero-shot translation tasks, because we do not assume any $P_{L1\text{-}R2}$, $P_{L2\text{-}R1}$, $P_{R1\text{-}R2}$ parallel data, nor any parallel data using a pivot language.

3.2 Zero-Shot NMT

For a given language pair L1-L2, we require only one multilingual and multidirectional NMT system to synthesize parallel data. The components of this system are presented in Figure 2. Inspired by previous work in unsupervised NMT (Conneau and Lample, 2019), we first pre-train a cross-lingual language model to initialize the NMT system. We use the XLM approach (Conneau and Lample, 2019) trained with the combination of the following two different objectives:

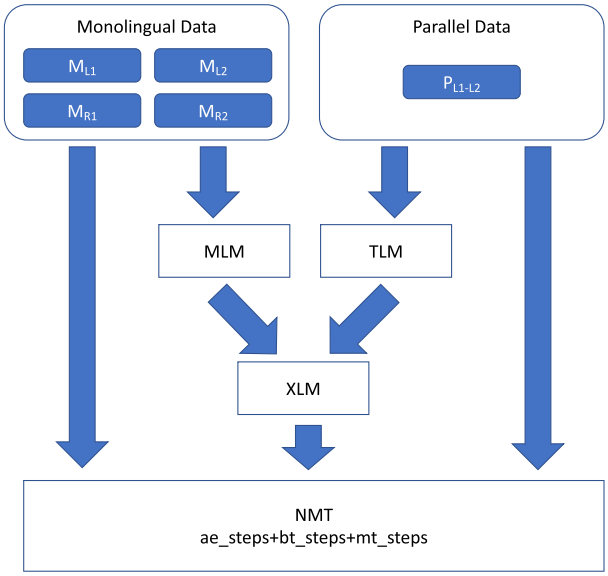

Figure 2: Our zero-shot NMT framework.

Masked Language Model (MLM): MLM has a similar objective to BERT (Devlin et al., 2019) but uses text streams for training instead of pairs of sentences. We optimize the MLM objective on the M_{L1}, M_{L2}, M_{R1}, and M_{R2} monolingual data.

Translation Language Model (TLM): TLM is an extension of MLM where parallel data are leveraged so that we can rely on context in two different languages to predict masked words. We optimize the TLM objective on $P_{L1\text{-}L2}$ parallel data, alternatively exploiting both translation directions.

The XLM approach alternates between MLM and TLM objectives to train a single model. By sharing a single vocabulary for all of L1, L2, R1, and R2, we expect XLM to implicitly model translation knowledge for our zero-shot translation directions, namely, L1→R2, L2→R1, and R2→R1, thanks to the joint training of MLM and TLM, also maximally exploiting the similarity between L1 and R1, and between L2 and R2.

Then, the embeddings from the XLM model are used to initialize the encoder and decoder embeddings of the NMT system instead of the standard

[1] We do not consider L1→R1 and L2→R2 (see Section 4.1).

random initialization. We exploit unsupervised NMT objectives (Lample et al., 2018) to which we associate a supervised NMT objective as follows:

Auto-encoder (AE) Objectives: Using a noise model that drops and swaps words, the objective is to reconstruct the original sentences. We use AE objectives for L1, L2, R1, and R2.

Back-translation (BT) Objectives: For training translation directions for which we do not have parallel data, a round-trip translation is performed during training in which a sentence s from monolingual data is translated, and its translation back-translated, with the objective of generating s. We use the BT objectives corresponding to our targeted zero-shot translation directions: L1→R2 →L1, R2→L1→R2, L2→R1→L2, R1→L2→R1, R1→R2→R1, and R2→R1→R2.

Machine Translation (MT) Objectives: we use this objective for L1→L2 and L2→L1, for which we have parallel data.

AE and BT are unsupervised NMT objectives used to train our zero-shot translation directions. However, using only these objectives would result in very poor performance, especially for distant and difficult language pairs. We thus also use MT objectives for the necessary supervision.

To alter $P_{L1\text{-}L2}$ into $P^{S}_{R1\text{-}R2}$ by our method #1, we could have trained an NMT system for L1→R1 and L2→R2 with the BT objectives L1→R1→L1 and L2→R2→L2. However, due to the similarity between L1 and R1, the NMT system would often perform a copy of M_{L1} to M^{S}_{R1}. Therefore, as done by previous work in paraphrase generation (Bannard and Callison-Burch, 2005; Mallinson et al., 2017), we instead rely on pivot languages, for instance, by translating the L1 side of $P_{L1\text{-}L2}$ parallel data into R2 as a translation of L2.

4 Synthesizing Parallel Data of UGT

This section presents our two approaches to synthesize parallel data of UGT mentioned in Section 3.1: #1 alters existing parallel data and #2 generates translations of UGT monolingual data.

4.1 Parallel Data Alteration

There exist several methods to synthesize parallel data of UGT from existing parallel data in various style or domains, but mostly requiring the use of UGT parallel data. Vaibhav et al. (2019) proposed a synthetic noise induction (SNI) that applies manually defined editing operations, such as adding/dropping characters from a word or adding emojis, to introduce noise into existing parallel data. The resulting data were used for adapting an NMT system for translating UGT. They also proposed a tag-based method given a small $P_{R1\text{-}R2}$ parallel data: concatenate $P_{R1\text{-}R2}$ and $P_{L1\text{-}L2}$ parallel data, prepend a tag onto each source sentence to indicate whether the sentence pair is from $P_{R1\text{-}R2}$ or $P_{L1\text{-}L2}$, and train NMT systems on that data. Then, they used this NMT system to translate the L1 side of another $P_{L1\text{-}L2}$ parallel data prepended with the tag for $P_{R1\text{-}R2}$ so that the system is forced to translate L1 sentences as R1 sentences. The resulting parallel data are noisier than the original data and potentially more suitable to train NMT systems for UGT. The data are used to fine-tune NMT systems trained on $P_{L1\text{-}L2}$ parallel data.

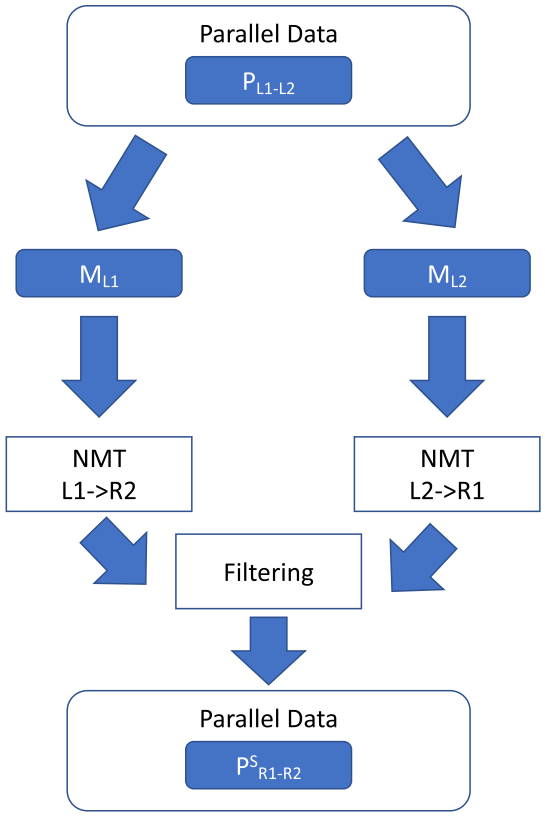

Figure 3: Alteration of $P_{L1\text{-}L2}$ parallel data to synthesize $P^{S}_{R1\text{-}R2}$ parallel data.

In contrast, as illustrated in Figure 3, our approach uses a zero-shot NMT system that does not require any manually produced $P_{R1\text{-}R2}$ nor relies on manually defined editing operations. Given $P_{L1\text{-}L2}$, we perform L1→R2 and L2→R1

translation for each of L1 and L2 sentences, respectively, to obtain a synthetic R1-R2 version, that is, $P^{S}_{R1\text{-}R2}$, of the original $P_{L1\text{-}L2}$. The resulting $P^{S}_{R1\text{-}R2}$ can be too noisy to be used to train NMT. To filter $P^{S}_{R1\text{-}R2}$, we evaluate the similarity between original L1 and L2 sentences with their respective R1 and R2 versions using sentence-level BLEU (Lin and Och, 2004) (sBLEU). Given a sentence pair in $P^{S}_{R1\text{-}R2}$, if either sBLEU of L1 with respect to R1 or sBLEU of L2 with respect to R2 is below a predetermined threshold T, we filter out the sentence pair, consider that it has been too much altered. T can be set empirically: Create several version of $P^{S}_{R1\text{-}R2}$ using different T values, train an NMT system for each version, and choose the value that leads to the NMT system achieving the best BLEU score on some $P_{L1\text{-}L2}$ validation data.

Finally, after filtering, we exploit the resulting $P^{S}_{R1\text{-}R2}$ by concatenating it to the original $P_{L1\text{-}L2}$ parallel data and train a new NMT system for translating UGT, or by using it for fine-tuning an NMT system trained on $P_{L1\text{-}L2}$ parallel data.

4.2 Translation of Monolingual Data

Previous work also proposed to synthesize parallel data from monolingual data using NMT (Sennrich et al., 2016a): An L1→L2 NMT system is used to translate M_{L1} monolingual data into L2, and then the synthesized $P^{S}_{L1\text{-}L2}$ parallel data are concatenated to original parallel data and used to train new L2→L1 (back-translation) or L1→L2 (forward translation) NMT systems. However, to the best of our knowledge, nobody has studied the use of large UGT monolingual data, without any manually produced $P_{R1\text{-}R2}$ parallel data, and its impact on translation quality.[2]

In our scenario, translating R1 texts with an L1→L2 would lead to translations of R1, that we can denote R2, of a very poor quality (see Section 2). Consequently, back-translations or forward translations generated this way would be too noisy to train R1↔R2 NMT systems. We verify this assumption in Section 5.2.1. Instead, as illustrated in Figure 4, we use R1→R2 and R2→R1 zero-shot NMT to synthesize parallel data from M_{R1} and M_{R2} monolingual data, respectively. Because our NMT system uses a pre-trained language model for R1 and R2, we can expect it to generate better translation than

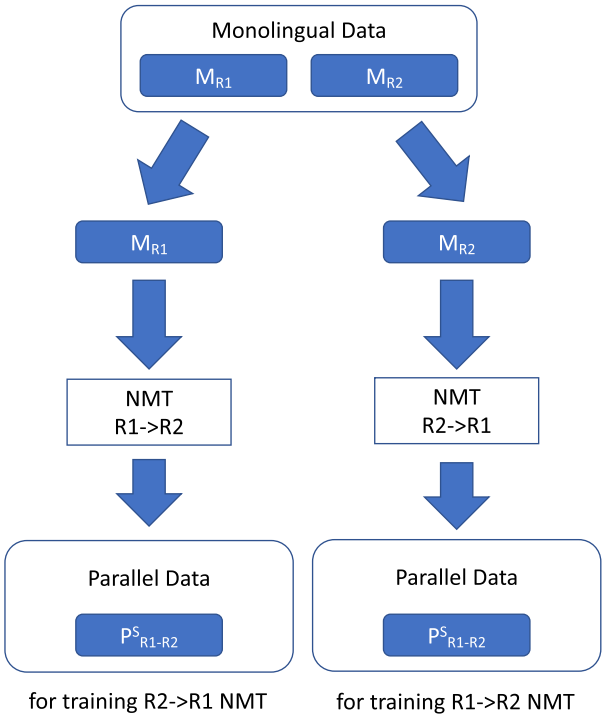

Figure 4: Translation of monolingual data M_{R1} and M_{R2} to synthesize $P^{S}_{R1\text{-}R2}$ parallel data.

a standard NMT system trained only on $P_{L1\text{-}L2}$ parallel data, (i.e., that never saw UGT during training). As in Section 4.1, the resulting $P^{S}_{R1\text{-}R2}$ parallel data can be used for fine-tuning or concatenated with the original $P_{L1\text{-}L2}$ parallel data for training.

In this work, we only examine the use of $P^{S}_{R1\text{-}R2}$ parallel data with their synthetic part on the source side, as back-translations, because in our preliminary experiments we have consistently observed better results than when $P^{S}_{R1\text{-}R2}$ is used as forward translations.[3] Note also that we do not filter the synthesized data and use all the data generated from the monolingual data, in contrast to another approach presented in Section 4.1. We could potentially obtain better results by filtering synthetic parallel data with some existing methods proposed, for instance, for filtering back-translations (Imankulova et al., 2019). We leave the investigation of such filtering techniques for future work.

5 Experiments

In this section, we empirically evaluate the usefulness of the parallel data synthesized by

[2]Berard et al. (2019a) showed that a large monolingual corpus of UGT can be successfully back-translated with a system trained on $P_{R1\text{-}R2}$ parallel data.

[3]Li and Specia (2019) observed improvements using forward translations but only in combination with manually produced $P_{R1\text{-}R2}$ parallel data.

our proposed approaches in training better NMT systems for translating UGT.

5.1 Data

We conducted experiments for two language pairs, English–French (en-fr) and English–Japanese (en-ja), with the MTNT translation tasks (Michel and Neubig, 2018). The test sets were made from posts extracted from an online discussion Web site, Reddit. Translations in the MTNT test sets were produced by professional translators with the instructions of keeping the style. Errors in the source texts were also preserved. In the four test sets, one for each translation direction, the source side contains original texts, that is, our systems will not have to translate translationese.

For parallel data, we did not use any of the Reddit parallel data of the MTNT, since our approach is supposed to be agnostic of manually produced $P_{R1\text{-}R2}$ translations. To make our settings comparable with previous work, we used only the clean parallel data in MTNT as $P_{L1\text{-}L2}$ data for training and validating our NMT systems. For the en-fr pair, $P_{L1\text{-}L2}$ data contain 2.2M sentence pairs consisting of the news-commentary (news commentaries) and Europarl (parliamentary debates) corpora provided by WMT15 (Bojar et al., 2015). For the en-ja pair, $P_{L1\text{-}L2}$ data consist of the KFTT (Wikipedia articles), TED (transcripts of online conference talks), and JESC (subtitles) corpora, resulting in a total of 3.9 M sentence pairs. All $P_{L1\text{-}L2}$ parallel data can be considered rather clean and/or formal in contrast to Reddit data.

As monolingual data, M_{L1} and M_{L2}, we used the entire News Crawl provided for WMT20[4] for Japanese, 3.4M lines, and a sample of 25M lines for English and French. As M_{R1} and M_{R2}, we crawled data using the Reddit API and applied fastText[5] for language identification.[6] As preprocessing steps for English and French, we first normalized the punctuation of all the data, except for the reference translations in the test sets, with the `Moses` (Koehn et al., 2007)[7] punctuation normalizer, and then tokenized all the data with the `Moses` tokenizer. Finally, we truecased the data with the `Moses` truecaser trained on the Reddit monolingual data. As for Japanese, we only tokenized the data with MeCab.[8] We removed all empty lines and lines longer than 120 tokens from the monolingual and parallel data. Because we could crawled plenty of English data (595M lines) on Reddit, we only selected its noisiest part, similarly to Michel and Neubig (2018) when they built the MTNT dataset. We trained a language model on the English News Crawl monolingual data using `LMPLZ` (Heafield et al., 2013), scored all lines of English Reddit data with the language model, normalized the score by the number of tokens in each line, and kept only the 25M lines with the lowest score. Because there are significantly less Japanese and French Reddit data, 0.8M and 1.2M sentences, respectively, we did not apply this filtering for these two languages. English Reddit data are thus much larger and can also be considered noisier than French and Japanese Reddit data.

For validation, we used the $P_{L1\text{-}L2}$ validation data from the MTNT dataset: Newsdiscuss-dev2015 for en-fr and the concatenation of the validation data provided with the KFTT, TED, and JESC corpora.

For evaluation, we used `SacreBLEU` (Post, 2018) that includes the MTNT test sets. For en→ja, we report on scores using the character-level metric chrF (Popović, 2015) instead of BLEU (Papineni et al., 2002) to avoid any tokenization mismatch with previous/future work.[9] We tested the significance of our results via bootstrap re-sampling and approximate randomization with `MultEval` (Clark et al., 2011).[10]

5.2 Baselines Systems

To train NMT systems, we first segmented tokens into sub-words using a BPE segmentation (Sennrich et al., 2016b) with 32k operations

[4] `http://www.statmt.org/wmt20/translation-task.html`.

[5] `https://fasttext.cc/`.

[6] In our preliminary experiments, we observed large improvements in translation quality (beyond 5.0 BLEU points) with our approaches when the crawled M_{R1} contains the source side of the test sets. We rather chose to experiment without the knowledge of the source side of the test set and carefully removed it from the monolingual data.

[7] `https://github.com/moses-smt/mosesdecoder`.

[8] `https://taku910.github.io/mecab/`.

[9] The sacreBLEU signatures, where xx is among {en,fr,ja} are as follows: BLEU+case.mixed+lang.xx-xx+numrefs.1 +smooth.exp+test.mtnt1.1/test+tok.13a+version.1.4.2; chrF2+ case.mixed+lang.en-ja+numchars.6+numrefs.1 +space.False+ test.mtnt1.1/test+version.1.4.2.

[10] `https://github.com/jhclark/multeval`.

System	BLEU			chrF
	fr→en	en→fr	ja→en	en→ja
vanilla	21.6	21.7	8.1	0.174
+ TBT News	25.8*	25.3*	8.6*	0.190*
+ TBT Reddit	22.9*	25.5*	0.5	0.181*
FT on SNI	23.1*	22.3*	8.2*	0.164
+ SNI	22.0	21.7	8.3*	0.158
FT on NGBT	0.2	17.3	0.5	0.021

Table 1: Results for the MTNT test sets. Tagged back-translation systems (TBT) were trained on back-translations of News Crawl or Reddit monolingual data. ''+'' indicates that the generated data were concatenated to the original $P_{L1\text{-}L2}$ parallel data. ''FT'' denotes the fine-tuning of the vanilla NMT system. ''*'' denotes systems significantly better than the vanilla NMT system with a p-value < 0.05.

jointly learned for each language pair on the Reddit monolingual data.

We used the Transformer (Vaswani et al., 2017) implementation in `Marian` (Junczys-Dowmunt et al., 2018) with standard hyper-parameters: 6 encoder and decoder layers, 512 dimensions for the embeddings and hidden states, 8 attention heads, and 2,048 dimensions for the feed-forward filter. During training, we evaluated the model using a mean cross-entropy score computed on the MTNT $P_{L1\text{-}L2}$ validation data after every 5k mini-batch updates and stopped training when it had not been improved for 5 consecutive times. We selected the model that yields the best BLEU, using the BLEU metric implemented in `Marian`, on the same validation data. We used the same training procedure for our vanilla NMT systems and all the NMT systems trained on synthetic parallel data.

Table 1 reports on the results for our vanilla NMT systems and other baseline systems described in Sections 5.2.1 and 5.2.2.

5.2.1 Tagged Back-translation

We generated back-translations from Reddit monolingual data, tagged (Caswell et al., 2019) and concatenated them to the original $P_{L1\text{-}L2}$ parallel data, and trained a new NMT system from scratch. Because Reddit data are noisy UGT, the generated back-translations may be of a very poor quality and harm the training of NMT. As contrastive experiments, we also evaluated the use of back-translations of News Crawl for which we can expect the system trained on $P_{L1\text{-}L2}$ to generate better but out-of-domain translations. In all experiments, we used as many monolingual sentences as in the $P_{L1\text{-}L2}$ parallel, or all of the Reddit data for French and Japanese since we do not have enough Reddit data to match the size of $P_{L1\text{-}L2}$.

As shown in Table 1, back-translations of Reddit are mostly useful, with up to 3.8 BLEU points of improvement, but dramatically failed for ja→en potentially due to the very low quality of the back-translations generated by the en→ja vanilla NMT system. Using back-translations of News Crawl is more helpful, especially for fr→en and ja→en.

Berard et al. (2019a) showed improvements when using back-translations of UGT. In contrast, we did not consistently observe improvements without using any manually produced $P_{R1\text{-}R2}$ to train the NMT systems for back-translation.

5.2.2 Synthetic Noise Generation

As potential baselines, we also evaluated the methods proposed by Vaibhav et al. (2019) for SNI, because it does not require any manually produced $P_{R1\text{-}R2}$. We applied their method to $P_{L1\text{-}L2}$ using their scripts[11] to create a noisy version of parallel data, namely, $P^{S}_{R1\text{-}R2}$. We also evaluated a similar approach to the tagged back-translations proposed by Vaibhav et al. (2019) (see Section 4.1). We used our systems trained on back-translations of Reddit to decode L1 sentences from $P_{L1\text{-}L2}$ parallel data, to which we added the back-translation tags to let the NMT system generate translation of L1 similar to UGT. We denote this noise generation from back-translation ''NGBT.'' As in Vaibhav et al. (2019), we introduced noise only to the source side of the parallel data performing L1→L2→L1 where the resulting L1 sentences comprise a noisy version of the original L1 sentences. We then replace L1 sentences in the $P_{L1\text{-}L2}$ parallel data with their noisy version. In addition to the use of the resulting $P^{S}_{R1\text{-}R2}$ data for fine-tuning as in Vaibhav et al. (2019), we also evaluated NMT systems trained from scratch on the concatenation of the $P^{S}_{R1\text{-}R2}$ and $P_{L1\text{-}L2}$.

As shown in Table 1, fine-tuning our vanilla NMT system on SNI actually improves translation quality for all the tasks, except en→ja. These results are not in accordance with the results in

[11] `https://github.com/MysteryVaibhav/robust_mtnt`.

Vaibhav et al. (2019) that show a slight drop of the BLEU score for fr→en.[12] We speculate that the difference may come from the use of a different, better, vanilla NMT system for which we used a larger $P_{L1\text{-}L2}$ parallel data than in Vaibhav et al. (2019). Using the $P^{S}_{R1\text{-}R2}$ synthetic parallel data concatenated to the original $P^{S}_{L1\text{-}L2}$ leads to lower BLEU scores than fine-tuning, except for ja→en.

As expected, our adaptation of NGBT performed very poorly, showing that our systems trained on Reddit back-translations are not good enough to generate a useful noisy version of $P_{L1\text{-}L2}$ parallel data. We do not further explore this configuration in this paper.

5.3 System Settings for our Approaches

Our NMT systems used for synthesizing $P^{S}_{R1\text{-}R2}$ parallel data are initialized with XLM (Section 3.2). To train XLM, we used the data presented in Section 5.1 on which we applied the same BPE segmentation used by our vanilla NMT systems. For the MLM objectives, we used the News Crawl corpora as M_{L1} and M_{L2} and the Reddit corpora as M_{R1} and M_{R2} monolingual data. For the TLM objectives, we used the parallel data used to train our vanilla NMT system as $P_{L1\text{-}L2}$ parallel data. We used the publicly available XLM framework[13] with the standard hyperparameters proposed for unsupervised NMT: 6 layers for the encoder and the decoder, 1,024 dimensions for the embeddings, a dropout rate of 0.1, and the GELU activation. We used text streams of 256 tokens and a mini-batch size of 64. The Adam optimizer (Kingma and Ba, 2014) with a linear warm-up (Vaswani et al., 2017) was used. During training, the model was evaluated every 200k sentences on the MTNT validation parallel data for TLM and the monolingual validation data of MTNT for MLM. The training was stopped when the averaged perplexity of MLM and TLM had not been improved for 10 consecutive times.

We initialized our zero-shot NMT with XLM and trained it with the AE, BT, and MT objectives presented in Section 3.2, all having the same weights, using the same hyperparameters as XLM. We evaluated the model every 200k sentences on the MTNT validation parallel data and stopped training when the average BLEU of L1→L2 and L2→L1 had not been improved for 10 consecutive times.

Finally, we synthesized $P^{S}_{R1\text{-}R2}$ data with our approaches using this system and trained final NMT models on the resulting $P^{S}_{R1\text{-}R2}$.

System	BLEU			chrF
	fr→en	en→fr	ja→en	en→ja
zero-shot NMT	21.4	22.4	3.0	0.126
vanilla	21.6	21.7	8.1	0.174
FT on SNI	23.1	22.3	8.2	0.164
#1: $P^{S}_{R1\text{-}R2}$ synthesized from $P_{L1\text{-}L2}$				
FT on $P^{S}_{R1\text{-}R2}$	22.0	24.2*	9.0*	0.174
+ $P^{S}_{R1\text{-}R2}$	23.1	24.7*	9.5*	0.180*
#2: $P^{S}_{R1\text{-}R2}$ synthesized from M_{R2} monolingual data				
FT on $P^{S}_{R1\text{-}R2}$	26.5*	26.2*	9.1*	0.202*
+ $P^{S}_{R1\text{-}R2}$	29.3*	26.8*	10.0*	0.212*
$P^{S}_{R1\text{-}R2}$ synthesized by #1 and #2				
+ #1 + #2	29.0*	27.2*	10.4*	0.213*
With the Reddit training parallel data from MTNT				
FT on MTNT	29.0*	27.5*	9.9*	0.192*

Table 2: Results for the MTNT test sets using $P^{S}_{R1\text{-}R2}$ synthesized by our approaches. ''zero-shot NMT'' is the NMT system used for synthesizing $P^{S}_{R1\text{-}R2}$. ''FT on $P^{S}_{R1\text{-}R2}$'' are configurations for which we sampled 100k sentence pairs from $P^{S}_{R1\text{-}R2}$ to fine-tune the vanilla NMT system. The last row is given for reference: the vanilla NMT system fined-tuned on the official MTNT training parallel data. ''*'' denotes systems significantly better than the FT on SNI system with a p-value < 0.05.

5.4 Results

Our results are presented in Table 2. First, we checked the performance of our zero-shot NMT system. Whereas for fr↔en, it was comparable with the vanilla NMT system, for ja↔en, it performed much worse than the vanilla NMT model as expected. This is due to the use of unsupervised MT objectives that were shown to be very difficult to optimize for distant and difficult language pairs (Marie et al., 2019) with almost no shared entries in the respective vocabulary of the two languages.

[12] Vaibhav et al. (2019) observed improvements only when used in combination with an manually produced $P^{S}_{R1\text{-}R2}$.

[13] We refer the reader to the section III given at this URL to retrieve the complete settings of our training for XLM and unsupervised NMT: `https://github.com/facebookresearch/XLM`. The only difference is that we used our data in different languages, which is also used to train our own BPE vocabulary.

With approach #1, we synthesized $P^{S}_{R1\text{-}R2}$ from $P_{L1\text{-}L2}$ and filtered them with $T = 0.5$ for en-fr and $T = 0.25$ for en-ja, respectively, resulting 196,788 and 301,519 sentence pairs.[14] As shown in Table 2, fine-tuning on $P^{S}_{R1\text{-}R2}$ brings larger improvements than doing so on SNI, except for fr→en. Despite the small size of the $P^{S}_{R1\text{-}R2}$, concatenating it with $P_{L1\text{-}L2}$ achieves the best BLEU with up to 3.0 BLEU points of improvements. We conclude that our approach successfully alters $P_{L1\text{-}L2}$ into $P^{S}_{R1\text{-}R2}$ useful to train NMT for UGT.

We give an analysis of the altered sentences later in Section 9.

Our approach #2 to synthesize $P^{S}_{R1\text{-}R2}$ brought even larger improvements. In contrast to the back-translations of Reddit generated by the vanilla NMT system (see Table 1), $P^{S}_{R1\text{-}R2}$ synthesized by our zero-shot NMT systems from M_{R2} Reddit monolingual data (the same data used to generate ''TBT Reddit'') lead to larger improvements, especially when concatenated to $P_{L1\text{-}L2}$. For fr→en, for instance, the gain over the vanilla NMT system is 7.7 BLEU points. Note also that further gains may potentially be attainable by exploring upsampling or downsampling strategies to find the optimal ratio between the sizes of $P_{L1\text{-}L2}$ and $P^{S}_{R1\text{-}R2}$.

Finally, concatenating $P^{S}_{R1\text{-}R2}$ parallel data synthesized by #1 and #2 provides slightly better results than, or comparable to, the use of only parallel data synthesized by #2.

6 Impact of the Distinction Between L1/L2 and R1/R2 Monolingual Data

We empirically verified our assumption that M_{L1} and M_{R1}, M_{L2} and M_{R2}, must be distinguished in order to enforce our NMT systems to learn the difference between clean texts and UGT, while it also learns to translate between L1 and L2, and between R1 and R2. To this end, we set up two new configurations, #A and #B, where we have $P_{L1\text{-}L2}$ parallel data and only M_{L1} and M_{L2} monolingual data, that is, we do not define M_{R1} and M_{R2} monolingual data to train XLM and NMT systems used to synthesize parallel data.

#A We replace News Crawl for M_{L1} and M_{L2} monolingual data with those for Reddit.

System	BLEU			chrF
	fr→en	en→fr	ja→en	en→ja
#1: synthesized from $P_{L1\text{-}L2}$ parallel data				
+ original	23.1	24.7	9.5	0.180
+ A	21.5*	21.5*	8.0*	0.173*
+ B	21.9*	22.0*	8.3*	0.182
#2 TBT: generated from Reddit monolingual data				
+ original	29.3	26.8	10.0	0.212
+ A	21.3*	22.0*	8.1*	0.170*
+ B	22.0*	22.1*	8.7*	0.183*

Table 3: Results for the MTNT test sets using the configurations #A and #B. ''original'' denotes the system presented in Section 5.4. ''*'' denotes systems significantly worse than the the ''original'' configuration with a p-value < 0.05.

#B Because we have only few Reddit monolingual data for French and Japanese, #A is significantly disadvantaged by using much less monolingual data compared with our original system that also used News Crawl. In configuration #B, M_{L1} and M_{L2} are the concatenation of News Crawl and Reddit data with French and Japanese Reddit data upsampled to respectively match the size of the French and Japanese News Crawl corpora.

With #A and #B, we no longer have zero-shot translation directions for synthesizing $P^{S}_{R1\text{-}R2}$. Instead, we have an NMT system initialized using a pre-trained crosslingual language model also exploiting Reddit monolingual data.[15] With these configurations, we assume that the presence of a significant amount of Reddit data in the monolingual data may bias the NMT system in synthesizing Reddit-like texts.

The results of NMT systems trained on parallel data synthesized by #A and #B are presented in Table 3. With both our approaches #1 and #2, both configurations #A and #B perform significantly worse than our proposed NMT systems that exploit $P^{S}_{R1\text{-}R2}$ synthesized by zero-shot NMT systems. These results point out the necessity to set zero-shot NMT systems, differentiating clean texts from UGT, to synthesize useful parallel data of UGT.

[14] In terms of BLEU scores, we observed differences, in the range of 2.0 BLEU points, considering all the thresholds tested.

[15] We used the same framework used by our zero-shot NMT systems for #A and #B, also using the AE and BT objectives since removing them did not have a positive impact.

7 Ablation Study on Zero-Shot NMT's Objective

We performed an ablation study of the objectives exploited for training the zero-shot NMT presented in Section 3.2. We compared the following four combinations of objectives:

AE+BT+MT: The original combination used to train our zero-shot NMT system.

BT+MT: The AE objective is removed. This excludes any random noise in the source sentences. The system is no longer restricted to perform a simple copy of the source when performing round-trip BT.

AE+BT: Typical combination of objectives used for unsupervised NMT (Lample et al., 2018). Without the supervised MT objective, we expect a drop of the translation quality.

BT: Without AE and MT objectives, we can expect the system to be able to properly model neither languages nor translations.

Note that we cannot remove the BT objectives as this is the only objective that trained the system to translate, for instance, from L1 to R2 and from R2 to R1. We evaluated the zero-shot NMT itself and NMT systems exploiting the synthetic parallel data generated by the zero-shot NMT system using our approaches #1 without filtering[16] and #2.

The results are presented in Table 4. None of the alternative combinations performs better than AE+BT+MT in our original proposal. Removing AE (i.e., BT+MT) has a minimal impact but it is necessary to obtain the best results. In contrast, removing the MT objective (i.e., AE+BT) led to a significant drop of the translation quality as the zero-shot NMT is not supervised at all. Using only the BT objective led to extremely noisy synthetic data that cannot be used to train NMT.

Losses	fr→en	en→fr
Zero-Shot NMT		
AE+BT+MT	21.4	22.4
BT+MT	20.4*	22.3
AE+BT	19.2*	21.2*
BT	0.1*	0.4*
#1: only with $P^{S}_{R1\text{-}R2}$ synthesized from $P_{L1\text{-}L2}$		
AE+BT+MT	19.0	18.0
BT+MT	17.7*	17.0*
AE+BT	6.6*	0.7*
BT	0.0*	0.2*
#2: $P_{L1\text{-}L2}$ + $P^{S}_{R1\text{-}R2}$ synthesized from M_{R2}		
AE+BT+MT	29.3	26.8
BT+MT	28.5*	25.9*
AE+BT	28.3*	25.0*
BT	13.2*	12.1*

Table 4: BLEU scores for the MTNT test sets with some of the objectives deactivated for training the zero-shot NMT system that synthesizes $P^{S}_{R1\text{-}R2}$. The configurations using #1 synthetic data were trained exclusively on this data. ''*'' denotes systems significantly worse than using all the objectives with a p-value < 0.05.

8 Impact on the Robustness of NMT

Using extra test suites, we evaluated to what extent our NMT systems trained on synthetic parallel data of UGT are robust to domain/style changes or only adapted to better translate Reddit data.

Newstest2014 (en-fr): Translation task of WMT14 containing clean texts of news.

Newsdiscuss2015 (en-fr): Translation task of WMT15 containing UGT of discussions on news.

Foursquare (en-fr): A corpus of restaurant reviews (Berard et al., 2019a) that is another instance of UGT.

JESC, KFTT, and TED (en-ja): Test sets released with their respective training data in the MTNT dataset (see Section 5.1).

Twitter (en-ja): We collected 1,400 English tweets from the natural disaster domain and hired a translation firm to translate them into Japanese with specific instructions to preserve the style of the source texts. This test set is particularly noisy because it presents many tokens specific to tweets (user identifiers, hash tags, abbreviations, etc.).

For all these translation tasks, we experimented only with the original translation direction to avoid translating translationese, except for the cases

[16] We did not filter the data unlike our original proposal, because our goal is only to evaluate the quality of the data given the different systems used to generate them while saving the computational cost of finding a good threshold for filtering.

System	News2014		Newsdiscuss2015		Foursquare	KFTT	TED	JESC		Twitter
	fr→en	en→fr	fr→en	en→fr	fr→en	ja→en	en→ja	ja→en	en→ja	en→ja
vanilla	29.4	32.5	29.3	31.3	13.5	22.8	0.234	15.9	0.229	0.120
+ TBT News	**34.2***	**36.3***	**32.3***	33.7*	17.0*	22.2	0.226	14.1	0.217	**0.151***
+ TBT Reddit	27.4	33.1*	29.2	33.4*	15.4*	1.0	0.228	0.1	0.223	0.137*
+ #1	29.3	32.6	29.8*	31.6*	13.7*	22.9*	0.233	16.7*	0.232*	0.140*
+ #2	30.4*	34.2*	31.6*	33.0*	17.5*	**23.1***	0.236	17.5*	**0.240***	0.150*
+ #1 + #2	30.6*	34.5*	31.4*	33.3*	17.7*	23.0*	**0.238***	**17.6***	0.239*	0.150*
FT on MTNT	30.6*	33.4*	32.1*	**34.1***	**18.0***	22.9	0.228	15.0	0.229	0.127*

Table 5: BLEU ($*\rightarrow\{$en,fr$\}$) and chrf (en→ja) scores obtained on the extra test sets. Best scores are in **bold**. ''*'' denotes systems significantly better than the vanilla NMT system with a p-value < 0.05.

where the origin of the source texts is unknown or mixed. The results obtained with the same systems presented in Section 5.4 are presented in Table 5.

These results point out that our approaches did not only adapt NMT systems to the domain and style of Reddit but also improved them overall. NMT systems trained on the parallel data synthesized by our approaches perform better than the vanilla NMT systems irrespective of the domain and style of the text to translate. In contrast, exploiting the Reddit monolingual data through tagged back-translation consistently led to lower BLEU scores (except for en→fr Newsdiscuss2015), highlighting the ability of our framework in producing better synthetic parallel data. The configuration ''TBT News,'' which exploits tagged back-translation from News Crawl, is as expected the best system for translating Newstest2014, Newsdiscuss2015, and tweets, since some of the tweets have been posted by news agencies, but performed lower than our system for translating UGT from Foursquare.

With these results and the results obtained on the MTNT test sets (see Section 5.4), we conclude that our approaches improve translation quality for UGT in general and did not only adapt the NMT system to translate a specific type of UGT.

9 Analysis of Clean Sentences Altered into UGT

This section takes a closer look at the parallel data synthesized by approach #1 to observe how the clean sentences from $P_{L1\text{-}L2}$ parallel data were altered and to better understand why the use of synthetic data leads to a better NMT system for UGT.

We first focus on some of the characteristics of the MTNT datasets and compare how well these characteristics are exhibited in $P^S_{R1\text{-}R2}$. For this analysis, we mainly relied on the scripts and resources provided by Michel and Neubig (2018).[17] We randomly sampled source sentences from $P_{L1\text{-}L2}$ and $P^S_{R1\text{-}R2}$ as much as there are in the MTNT test sets, and performed our analysis on them.[18] We counted the occurrences of profanities in the English, French, and Japanese. For English, we also counted the number of word contractions[19] and Internet slang expressions. We also counted words ending by ''-ise'' and ''-ize'' to account for some of the differences between US English and UK English word spellings. Because $P_{L1\text{-}L2}$ is mainly made of Europarl, we can expect that UK English spelling is mainly used, whereas we expect to find a higher ratio of US English spelling in the Reddit data, since Reddit is an American platform. For Japanese, we counted the numbers of formal and informal pronouns, assuming that MTNT datasets contain more informal pronouns than $P_{L1\text{-}L2}$. Michel and Neubig (2018) also counted spelling and grammar errors, and emojis. We did not count spelling and grammar errors, expecting that they are artificially numerous in our synthetic data, since they had been automatically generated. As for the emojis, both $P_{L1\text{-}L2}$ and $P^S_{R1\text{-}R2}$ did not contain any.

Table 6 demonstrates that according to all the indicators, $P^S_{R1\text{-}R2}$ exhibits more of the characteristics of MTNT datasets than $P_{L1\text{-}L2}$. For instance, $P^S_{R1\text{-}R2}$ is in more US English, contains more Internet slang, and uses significantly more

[17] https://github.com/pmichel31415/mtnt.

[18] For this analysis, the sentences sampled from $P^S_{R1\text{-}R2}$ are the synthetic versions of the sentences sampled from $P_{L1\text{-}L2}$.

[19] We searched for the tokens: 're, 's, 't, 'd, 'll, and 've.

Dataset	English				French	Japanese	
	Profanities %	Slang %	Contractions %	-ise/-ize Ratio	Profanities %	Profanities %	Formal/Informal Pronouns Ratio
MTNT	0.27	0.21	1.90	40.00/60.00	0.90	0.01	68.75/31.25
$P_{L1\text{-}L2}$	0.01	0.00	0.03	92.00/8.00	0.45	0.00	96.88/3.12
$P^{S}_{L1\text{-}L2}$	0.06	0.04	0.21	41.03/58.97	0.57	0.01	83.01/16.99

Table 6: Quantitative analysis of the generated data. ‘‘%’’ indicates the number for occurrences per 100 tokens. For English, we compute the statistics on the en-fr data. For the MTNT test sets, the statistics are computed on the source side. $R^{S}_{L1\text{-}L2}$ has been generated by the alteration of $P_{L1\text{-}L2}$ by our approach #1.

En1	R1	Mr President, I believe a situation in which **we’re** [...]
	L1	Mr President, I think a situation in which **we are** [...]
En2	R1	But **dont** count on a stable euro-dollar exchange rate [...]
	L1	But **don’t** count on a stable euro-dollar exchange rate [...]
En3	R1	When **i** became European Commissioner, at the end of 1999 **i** had to [...]
	L1	When **I** became a Commissioner at the end of 1999, **I** had to [...]
En4	R1	The end result is always the same **lmao. Nothing ******* done**
	L1	The end result is always the same: **Nothing is done**.
Fr1	R1	Vice-président de la Commission. - Je serai **très chiante**.
	L1	Vice-présidente de la Commission. - Je serai **franche**.
Fr2	R1	**Et** la réponse officielle est probablement fausse.
	L1	**En outre,** la réponse officielle est probablement fausse.
Fr3	R1	**Ca** aurait été une grande idée il y a quatre ans et demi.
	L1	**Cela** aurait été une bonne idée il y a quatre ans et demi.
Fr4	R1	Après, les déficits commerciaux **ça veux dire** des pertes d’emplois.
	L1	Après tout, les déficits commerciaux **impliquent** bien des pertes d’emplois.

Figure 5: Examples of French and English original sentence from the Europarl and News Commentary corpora (M_{L1}) altered by our approach #1 (M_{R1}). **Bold** indicates the alterations that we want to highlight for each example. We have manually masked a profanity in En4 with ‘‘*******’’.

English contractions. This partly explains the usefulness of $P^{S}_{R1\text{-}R2}$ as NMT training data for the MTNT translation tasks, but most indicators show that $P^{S}_{R1\text{-}R2}$ is still far from perfectly matching with the characteristics of Reddit data, suggesting some room for improvement.

For a more concrete illustration of our synthetic data, we present in Figure 5 four English and four French example sentences altered by our approach #1. These examples are all instances of a successful alteration of clean texts into UGT. En1 introduces an English contraction ‘‘we’re’’ that is a characteristic of less formal English. En2, En3, and Fr3 show spelling errors (for Fr3, ‘‘Ca’’ should be written ‘‘Ça’’) that may guide the system to make itself more robust. En4 introduces an instance of Internet slang with a profanity, as in Fr1 where ‘‘très chiante,’’ a vulgar translation of ‘‘very annoying’’ diverges from the original meaning of ‘‘franche’’ that can be translated by ‘‘frank.’’ Fr2, Fr3, and Fr4 are simplifications that make the sentences less formal: ‘‘en outre’’ and ‘‘impliquent’’ are usually used in texts that perform a formal demonstration, while ‘‘ça veux dire’’ is a more familiar turn of phrase for ‘‘impliquent’’ in this context. We also observed many instances of person names written with Reddit syntax for referring to a Reddit user account by prepending ‘‘/u/,’’ e.g., ‘‘Berlusconi’’ becomes ‘‘/u/Berlusconi.’’ All these examples are evidence that our approach successfully generates UGT in the style of Reddit.

10 Related Work

Several approaches for better translating UGT have been proposed taking advantage of the parallel data of UGT in the MTNT datasets (Michel and Neubig, 2018). Because of their relatively small size, they have been mostly used for fine-tuning (Li et al., 2019) and designing specific pre- and post-processing rules to improve translation quality (Berard et al., 2019b). Vaibhav et al. (2019) also proposed to generate synthetic parallel data of UGT through back-translation by exploiting the parallel data in MTNT. Monolingual data of UGT have been exploited to a lesser extent through forward translation (Li and Specia, 2019) or back-translation (Berard et al., 2019a) and always with NMT systems trained on parallel data of UGT. To the best of our knowledge, Vaibhav et al. (2019) proposed the only approach that synthesizes parallel data of UGT without relying on existing parallel data of UGT. Having obtained texts in the target style of UGT, they designed editing operations to make existing parallel data in other styles more similar to the targeted style.

Another line of work exploits NMT to perform style transfer across texts, that is, applying some characteristics of one text to another, without exploiting any parallel data of UGT, but has never been applied to NMT for UGT. Prabhumoye et al. (2018) performed style transfer through back-translation to preserve the meaning of the text while reducing its stylistic properties and then exploit adversarial generation algorithms to apply the desired style to the back-translated texts, assuming that meaning and style can be disentangled. Their approach also requires a classifier that can accurately predict the style of a given text. Zhang et al. (2018) proposed a three-step pipeline combining unsupervised statistical and neural MT to generate instances of texts in the targeted style that is then evaluated by a given style classifier as in Prabhumoye et al. (2018).

11 Conclusion

We described two new methods for synthesizing parallel data to train better NMT systems for UGT. Both methods work through a zero-shot NMT system, initialized with a pre-trained crosslingual language model that exploits monolingual corpora of UGT. Our first method (#1) successfully alters clean parallel data into parallel data that exhibit the characteristics of UGT of the targeted style. Our second method (#2) uses the same zero-shot NMT system to translate monolingual corpora of UGT for synthesizing parallel data useful to train NMT. We showed that both methods, separately or combined, improve translation quality for UGT.

For future work, we will study the use of manually produced UGT parallel data to better train our NMT system that synthesizes the parallel data. We will also explore other applications for this framework, such as paraphrase generation. We will also investigate the use of the recently proposed mirror-generative NMT (Zheng et al., 2020), a semi-supervised architecture that exploits jointly large source and target monolingual corpora, such as those of UGT, during training using source and target language models in the same latent space.

Acknowledgments

We would like to thank the action editor, Philipp Koehn, and reviewers for their useful comments and suggestions. A part of this work was conducted under the program ‘‘Research and Development of Enhanced Multilingual and Multipurpose Speech Translation System’’ of the Ministry of Internal Affairs and Communications (MIC), Japan. Benjamin Marie was partly supported by JSPS KAKENHI grant number 20K19879 and the tenure-track researcher start-up fund in NICT. Atsushi Fujita was partly supported by JSPS KAKENHI grant number 19H05660.

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, page 597–604. Ann Arbor, MI, USA. Association for Computational Linguistics. **DOI:** https://doi.org/10.3115/1219840.1219914

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations*. Vancouver, Canada.

Alexandre Berard, Ioan Calapodescu, Marc Dymetman, Claude Roux, Jean-Luc Meunier,

and Vassilina Nikoulina. 2019a. Machine translation of restaurant reviews: New corpus for domain adaptation and robustness. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 168–176. Hong Kong, China. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D19-5617

Alexandre Berard, Ioan Calapodescu, and Claude Roux. 2019b. Naver Labs Europe's systems for the WMT19 machine translation robustness task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 526–532. Florence, Italy. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/W19-5361

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46. Lisbon, Portugal. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/W15-3001, **PMID:** 25955892

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63. Florence, Italy. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/W19-5206

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176–181. Portland, MN, USA. Association for Computational Linguistics. **DOI:** https://dl.acm.org/doi/10.5555/2002736.2002774

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of Advances in Neural Information Processing Systems 32*, pages 7057–7067. Vancouver, Canada. Curran Associates, Inc.

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, MN, USA. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/N19-1423

Johanna Gerlach, Victoria Porro Rodriguez, Pierrette Bouillon, and Sabine Lehmann. 2013. Combining pre-editing and post-editing to improve SMT of user-generated content. In *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*, pages 45–53. Nice, France.

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

Aizhan Imankulova, Takayuki Sato, and Mamoru Komachi. 2019. Filtered pseudo-parallel corpus improves low-resource neural machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 19(2):24:1–16. **DOI:** https://doi.org/10.1145/3341726

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121. Melbourne, Australia. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/P18-4020

Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. 2019. Training on synthetic noise improves robustness to natural noise in machine translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 42–47. Hong Kong, China. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D19-5506

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

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. **DOI:** https://dl.acm.org/doi/10.5555/1557769.1557821

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049. Brussels, Belgium. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D18-1549

Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan Pino, and Hassan Sajjad. 2019. Findings of the first shared task on machine translation robustness. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 91–102. Florence, Italy. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/W19-5303

Zhenhao Li and Lucia Specia. 2019. Improving neural machine translation robustness via data augmentation: Beyond back-translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 328–336. Hong Kong, China. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D19-5543

Chin-Yew Lin and Franz Josef Och. 2004. ORANGE: A method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 501–507. Geneva, Switzerland. **DOI:** https://dl.acm.org/doi/10.3115/1220355.1220427

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893. Valencia, Spain. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/E17-1083

Benjamin Marie, Haipeng Sun, Rui Wang, Kehai Chen, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita. 2019. NICT's unsupervised neural and statistical machine translation systems for the WMT19 news translation task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 294–301. Florence, Italy. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/W19-5330

Claudia Matos Veliz, Orphee De Clercq, and Veronique Hoste. 2019. Benefits of data augmentation for NMT-based text normalization of user-generated content. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 275–285. Hong Kong, China. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D19-5536

Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*,

pages 543–553. Brussels, Belgium. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D18-1050, **PMID:** 29565364

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, PA, USA. Association for Computational Linguistics. **DOI:** https://doi.org/10.3115/1073083.1073135

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. **DOI:** https://doi.org/10.18653/v1/W15-3049

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. Brussels, Belgium. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/W18-6319

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876. Melbourne, Australia. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/P18-1080

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. **DOI:** https://doi.org/10.18653/v1/P16-1009

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. **DOI:** https://doi.org/10.18653/v1/P16-1162

Vaibhav Vaibhav, Sumeet Singh, Craig Stewart, and Graham Neubig. 2019. Improving robustness of machine translation with synthetic noise. 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 1916–1920. Minneapolis, USA. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/N19-1190

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems 30*, pages 5998–6008. Long Beach, USA. Curran Associates, Inc.

Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou, and Enhong Chen. 2018. Style transfer as unsupervised machine translation. *CoRR*, abs/1808.07894.

Zaixiang Zheng, Hao Zhou, Shujian Huang, Lei Li, Xin-Yu Dai, and Jiajun Chen. 2020. Mirror-generative neural machine translation. In *Proceedings of the 8th International Conference on Learning Representations*. Virtual.