Altering Parallel Data into User-Generated Texts with Zero-Shot Neural Machine Translation

Benjamin Marie Atsushi Fujita
National Institute of Information and Communications Technology
{bmarie, atsushi.fujita}@nict.go.jp

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 [1] empirically showed that NMT systems trained on clean parallel data 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 which are characteristics of their style. In this paper, we posit that the NMT system should preserve the style during the translation.

A major difficulty in training NMT for UGT is that we do not usually have bilingual parallel data of UGT created by professional translators to train or adapt an NMT system. Nevertheless, previous work on NMT for UGT merely focused on scenarios for which we have UGT parallel data, such as the MTNT dataset [1].

In this work, we do not assume the availability of parallel data of UGT. We propose to synthesize parallel data of UGT from monolingual data, through a zero-shot NMT system, to train better NMT systems for UGT.

2 Zero-Shot NMT for Synthesizing Parallel Data

2.1 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 2.2 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.

$P_{L1\text{-}L2}$, 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 for UGT, which can be obtained for instance by crawling social media.

Our objective is to synthesize parallel data of UGT, which we henceforth denote $P^{S}_{R1\text{-}R2}$. To this end, we propose to alter a clean parallel data $P_{L1\text{-}L2}$ into $P^{S}_{R1\text{-}R2}$. We alter the $P_{L1\text{-}L2}$ parallel data by performing L1→R2 and L2→R1 translations.

Note that L1→R2 and L2→R1 are both zero-shot translation tasks, since we do not assume any $P_{L1\text{-}R2}$ or $P_{L2\text{-}R1}$ parallel data, nor any parallel data using a pivot language.

2.2 Zero-Shot NMT

To synthesize parallel data of UGT, i.e., $P^{S}_{R1\text{-}R2}$, we build only one multilingual and multidirectional NMT system (see Figure 1). Inspired by previous work in unsupervised NMT [2], we first pre-train a cross-lingual language model to initialize the NMT system. We use the XLM approach [2] trained with the combination of the following two different objectives:

Masked Language Model (MLM): MLM has a similar objective to BERT [3] 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 XLM model. We then train

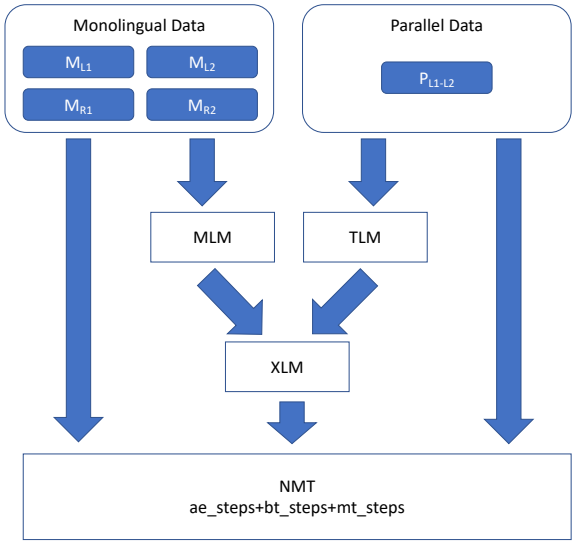

Figure 1 Our zero-shot NMT framework.

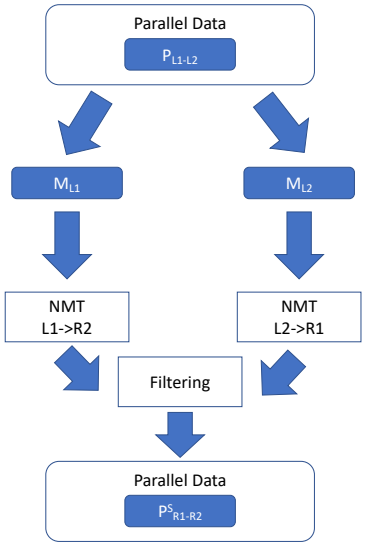

Figure 2 Alteration of $P_{L1\text{-}L2}$ parallel data to synthesize $P^{S}_{R1\text{-}R2}$ parallel data.

an NMT model, initializing its encoder and decoder embeddings with those of the pre-trained XLM model, and exploiting unsupervised NMT objectives [4] 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, and R1→L2→R1.

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. We also use MT objectives for the necessary supervision.

3 Parallel Data Alteration

As illustrated in Figure 2, 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, i.e., $P^{S}_{R1\text{-}R2}$, 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 [5] (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 pre-determined threshold T, we filter out the sentence pair.

4 Experiments

4.1 Data

We conducted experiments for two language pairs, English–French (en-fr) and English–Japanese (en-ja), with the MTNT translation tasks [1]. The test sets were made from posts extracted from an online discussion website, Reddit.

For parallel data, we did not use any of the Reddit parallel data of the MTNT. 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 sentences pairs consisting of the news-commentary (news commentaries) and Europarl (parliamentary debates) corpora provided by WMT15 [6]. 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 giving in a total of 3.9M 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[1] 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. For English and French, we tokenized and truecased all the data with the `Moses` tokenizer. As for Japanese, we only tokenized the data with MeCab.[2]

For English, we selected the noisiest part of Reddit, 25M sentences, similarly to [1] when they built the MTNT dataset. Since there are significantly less Japanese and French Reddit data, 0.8M and 1.2M sentences, respectively, we used all the French and Japanese sentences.

For validation, we used the $P_{L1\text{-}L2}$ validation data from the MTNT dataset. For evaluation, we used `SacreBLEU` [7].[3] We tested the significance of our results via bootstrap re-sampling and approximate randomization with `MultEval` [8].[4]

4.2 Baselines Systems

We evaluated vanilla NMT systems and other baseline systems exploiting tagged back-translation (TBT) and synthetic noise generation (SNI), using the Transformer [9] implementation in `Marian` [10] with standard hyper-parameters for all the NMT systems.

We generated back-translations from Reddit monolingual data and News Crawl, tagged [11] and concatenated them to the original $P_{L1\text{-}L2}$ parallel data, and trained a new NMT system from scratch. In all experiments, we used as much 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 $P_{L1\text{-}L2}$.

We also evaluated the methods proposed by [12] for SNI, since it does not require any manually produced $P_{R1\text{-}R2}$. We applied their method to $P_{L1\text{-}L2}$ using their scripts[5] to create a noisy version of parallel data, i.e., $P^{S}_{R1\text{-}R2}$. In addition to the use of the resulting $P^{S}_{R1\text{-}R2}$ data for fine-tuning as in [12], we also evaluated NMT systems trained from scratch on the concatenation of the $P^{S}_{R1\text{-}R2}$ and $P_{L1\text{-}L2}$.

1) http://www.statmt.org/wmt20/translation-task.html
2) https://taku910.github.io/mecab/
3) 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
4) https://github.com/jhclark/multeval
5) https://github.com/MysteryVaibhav/robust_mtnt

Table 1 Results for the MTNT test sets. TBT systems were trained on back-translations of News Crawl or Reddit monolingual data. "+" indicates that the synthesized parallel 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 ($p < 0.05$) than the vanilla NMT system.

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

Table 1 reports on the results. Back-translations of Reddit were mostly useful 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. Fine-tuning our vanilla NMT system on SNI improves translation quality for all the tasks, except en→ja. 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.

4.3 System Settings for our Approach

To train XLM, we used the data presented in Section 4.1 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[6] with the standard hyper-parameters proposed for unsupervised NMT. We used text streams of 256 tokens and a mini-batch size of 64. The Adam optimizer [13] with a linear warm-up [9] 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

6) 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.

En1	L1	Mr President, I think a situation in which **we are** [...]
	R1	Mr President, I believe a situation in which **we're** [...]
En2	L1	But **don't** count on a stable euro-dollar exchange rate [...]
	R1	But **dont** count on a stable euro-dollar exchange rate [...]
En3	L1	When **I** became a Commissioner at the end of 1999, **I** had to [...]
	R1	When **i** became European Commissioner, at the end of 1999 **i** had to [...]
En4	L1	The end result is always the same: **Nothing is done**.
	R1	The end result is always the same **lmao. Nothing ******* done**

Figure 3 Examples of English sentences from the Europarl and News Commentary corpora (L1) altered by our approach (R1). **Bold** indicates the alterations that we want to highlight for each example. We have manually masked a profanity in En4 with "*******".

Table 2 Results for the MTNT test sets using $P^{S}_{R1\text{-}R2}$ synthesized by our approach. "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 ($p < 0.05$) than the FT on SNI system.

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
$P^{S}_{R1\text{-}R2}$ 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*
With the Reddit training parallel data from MTNT				
FT on MTNT	29.0*	27.5*	9.9*	0.192*

it with the AE, BT, and MT objectives presented in Section 2.2, all having the same weights, using the same hyper-parameters as XLM. We evaluated the model every 200k sentences on the MTNT validation parallel data and stopped training when the averaged 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 approach using this system, 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. Then, we trained our final NMT models on the resulting $P^{S}_{R1\text{-}R2}$.

4.4 Results

The results of our models 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 [14] with almost no shared entries in the respective vocabulary of the two languages.

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 data to train NMT for UGT.

5 Example of Clean Sentences Altered into UGT

For a more concrete illustration of our synthetic data, we present in Figure 3 four English example sentences altered by our approach. 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 and En3 show spelling errors that may guide the system to make itself more robust. En4 introduces an instance of Internet slang with a profanity. 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 evidences that our approach successfully generates UGT in the style of Reddit.

6 Conclusion

We described our method for synthesizing parallel data to train better NMT systems for UGT. We successfully altered clean parallel data into parallel data that exhibit the characteristics of UGT of the targeted style. We showed that it improves translation quality for UGT.

Acknowledgments: This work was partly supported by JSPS KAKENHI grant numbers 20K19879 and 19H05660.

References

[1] Paul Michel and Graham Neubig. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 543–553, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

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

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186, Minneapolis, USA, June 2019. Association for Computational Linguistics.

[4] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5039–5049, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[5] Chin-Yew Lin and Franz Josef Och. ORANGE: A method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics*, pp. 501–507, Geneva, Switzerland, aug 23–aug 27 2004.

[6] 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. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 1–46, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

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

[8] Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 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*, pp. 176–181, Portland, USA, June 2011. Association for Computational Linguistics.

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

[10] 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. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pp. 116–121, Melbourne, Australia, July 2018. Association for Computational Linguistics.

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

[12] Vaibhav Vaibhav, Sumeet Singh, Craig Stewart, and Graham Neubig. 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)*, pp. 1916–1920, Minneapolis, USA, June 2019. Association for Computational Linguistics.

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

[14] Benjamin Marie, Haipeng Sun, Rui Wang, Kehai Chen, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita. 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)*, pp. 294–301, Florence, Italy, August 2019. Association for Computational Linguistics.