Simulated Multiple Reference Training Improves Low-Resource Machine Translation

Huda Khayrallah **Brian Thompson** **Matt Post** and **Philipp Koehn**
Johns Hopkins University
{huda, brian.thompson, phi}@jhu.edu, post@cs.jhu.edu

Abstract

Many valid translations exist for a given sentence, and yet machine translation (MT) is trained with a single reference translation, exacerbating data sparsity in low-resource settings. We introduce a novel MT training method that approximates the full space of possible translations by: *sampling* a paraphrase of the reference sentence from a paraphraser and training the MT model to predict the paraphraser's *distribution* over possible tokens. With an English paraphraser, we demonstrate the effectiveness of our method in low-resource settings, with gains of 1.2 to 7 BLEU.

1 Introduction

Variability and expressiveness are core features of language, and they extend to translation as well. Dreyer and Marcu (2012) showed that naturally occurring sentences have *billions* of valid translations. Despite this variety, machine translation (MT) models are optimized toward a single translation of each sentence in the training corpus.

Training high resource MT on millions of sentence pairs exposes it to similar sentences translated different ways, but training low-resource MT with a single translation for each sentence (out of potentially billions) exacerbates data sparsity. Despite active research in the area, low-resource settings remain a challenge for MT (Koehn and Knowles, 2017; Sennrich and Zhang, 2019).

A natural question is: To what extent does the discrepancy between linguistic diversity and standard single-reference training hinder MT performance? This was previously impractical to explore, since obtaining multiple human translations of training data is typically not feasible. However, recent advances in neural sentential paraphrasers produce fluent, meaning-preserving English paraphrases (Hu et al., 2019c). We introduce a novel

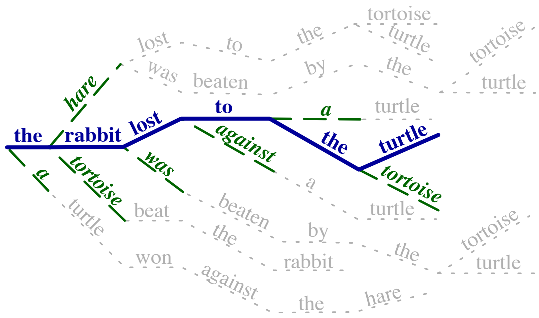

Figure 1: Some possible paraphrases of 'the turtle beat a hare' including a **sampled path** and some of the other ***tokens also considered in the training objective***

method that incorporates such a paraphraser directly in the training objective, and uses it to simulate the full space of translations.

We demonstrate the effectiveness of our method on two MATERIAL program low-resource datasets, and on publicly available data from GlobalVoices. We release data & code: `data.statmt.org/smrt`

2 Method

We propose a novel training method that uses a paraphraser to approximate the full space of possible translations, since explicitly training on billions of possible translations per sentence is intractable.

In standard neural MT training, the reference is: (1) used in the training objective; and (2) conditioned on as the previous target token.[1]

We approximate the full space of possible translations by: (1) training the MT model to predict the *distribution* of possible tokens from the paraphraser at each time step; and (2) *sampling* the previous target token from the paraphraser distribution. Figure 1 shows an example of possible paraphrases and highlights a sampled path and some of the other tokens used in the training objective distribution.

[1] In autoregressive NMT inference, predictions conditions on the previous target tokens. In training, predictions typically condition on the previous tokens in the reference, not the model's output (teacher forcing; Williams and Zipser, 1989).

We review the standard $\mathcal{L}_{\text{NLL}}$ training objective, and then introduce our proposed objective.

NLL Objective The standard negative log likelihood (NLL) training objective in NMT, for the i^{th} target word in the reference y is:

$$\mathcal{L}_{\text{NLL}} = -\sum_{v\in\mathcal{V}} \Big[\mathbb{1}\{y_i = v\} \times \log\, p_{\text{MT}}(y_i = v \,|\, x, y_{j<i})\Big] \tag{1}$$

where $\mathcal{V}$ is the vocabulary, $\mathbb{1}\{\cdot\}$ is the indicator function, and p_{MT} is the MT output distribution (conditioned on the source x, and on the previous tokens in the reference $y_{j<i}$). Equation 1 computes the cross-entropy between the MT model's distribution and the one-hot human reference.

Proposed Objective We compute the cross entropy between the distribution of the MT model and the distribution from a paraphraser conditioned on the reference:[2]

$$\mathcal{L}_{\text{para}} = -\sum_{v\in\mathcal{V}} \Big[\, p_{\text{para}}(y'_i = v \,|\, y, y'_{j<i}) \times \log\, p_{\text{MT}}(y'_i = v \,|\, x, y'_{j<i})\Big] \tag{2}$$

where y is the single human reference, and y' is the paraphrase of that reference. p_{para} is the output distribution from the paraphraser (conditioned on the single human reference y and the previous tokens in the sentence produced by the paraphraser $y'_{j<i}$). p_{MT} is the MT output distribution (conditioned on the source sentence, x and the previous tokens in the sentence produced by the paraphraser, $y'_{j<i}$). At each timestep we sample a target token from the paraphraser's output distribution[3] to ensure coverage of the full space of translations.[4] We condition on this sampled y'_{i-1} as the previous target token for both the MT model and paraphraser.

3 Experimental Setup

3.1 Paraphraser

For our paraphraser we train a Transformer model (Vaswani et al., 2017) in FAIRSEQ (Ott et al., 2019) with an 8-layer encoder and decoder, 1024 dimensional embeddings, 16 encoder and decoder attention heads, and 0.3 dropout. We optimize using Adam (Kingma and Ba, 2014). We train on ParaBank2 (Hu et al., 2019c), an English paraphrase dataset.[5] ParaBank2 was generated by training an MT system on CzEng 1.7 (a Czech−English bitext with over 50 million lines (Bojar et al., 2016)), re-translating the Czech training sentences, and pairing the English output with the human English translation. Many potential candidates were generated from the translation model for each sentence, and high quality diverse paraphrases were selected.

3.2 NMT models

For both the baseline and our method, we train Transformer models in FAIRSEQ using parameters from the FLORES low-resource benchmark (Guzmán et al., 2019): 5-layer encoder and decoder, 512 dimensional embeddings, and 2 encoder and decoder attention heads. We regularize with 0.2 label smoothing, and 0.4 dropout. We optimize using Adam with a learning rate of 10^{-3}. We train for a maximum of 200 epochs, and model selection from checkpoints is based on validation set perplexity. We translate with a beam size of 5.

For our method we use the proposed objective $\mathcal{L}_{\text{para}}$ with probability $p = 0.5$ and standard $\mathcal{L}_{\text{NLL}}$ on the original reference with probability $1 - p$. We sample from only the 100 highest probability vocabulary items at a given time step when sampling from the paraphraser distribution to avoid very unlikely tokens (Fan et al., 2018).

Using our English paraphraser, we aim to demonstrate improvements in low-resource settings. We use Tagalog (tl) to English and Swahili (sw) to English bitext from the MATERIAL low-resource program (Rubino, 2018). We also report results on public data, using MT bitext from GlobalVoices, a non-profit news site that publishes in 53 languages.[6] We evaluate on the 10 lowest-resource settings that have at least 10,000 lines of parallel text with English: Hungarian (hu), Indonesian (id), Czech (cs), Serbian (sr), Catalan (ca), Swahili (sw),[7] Dutch (nl), Polish (pl), Macedonian (mk), Arabic (ar).

We use 2,000 lines each for: a validation set for model selection from checkpoints and a test set for reporting results. The approximate number of lines of training data is in Table 1.

[2] Note the paraphraser parameters are not modified when training the MT model.

[3] Graves (2013) introduced sampling in sequence to sequence models for variety in handwriting generation.

[4] We resample every time a sentence is observed in training.

[5] Parabank2 also released a trained SOCKEYE paraphrase model but we are using FAIRSEQ, so we retrain it.

[6] We use v2017q3 released on Opus (`opus.nlpl.eu/GlobalVoices.php`). Not all 53 languages have MT bitext.

[7] Swahili is in both. MATERIAL data is not widely available, so we separate them to keep GlobalVoices reproducible.

dataset	GlobalVoices										MATERIAL	
* → en	hu	id	cs	sr	ca	sw	nl	pl	mk	ar	sw	tl
train lines	8k	8k	11k	14k	15k	24k	32k	40k	44k	47k	19k	46k
baseline	2.3	5.3	3.4	11.8	16.0	17.9	22.2	16.0	27.0	12.7	37.8	32.5
this work	**5.4**	**12.3**	**6.6**	**16.1**	**20.0**	**20.5**	**24.8**	**18.0**	**28.2**	**14.9**	**39.0**	**33.7**
Δ	+3.1	+7.0	+3.2	+4.3	+4.0	+2.6	+2.6	+2.0	+1.2	+2.2	+1.2	+1.2

Table 1: Test set results translating to English. 'train lines' indicates amount of training bitext. We **bold** the best value; all improvements are statistically significant at the 95% confidence level.

We train an English SentencePiece model (Kudo and Richardson, 2018) on the paraphraser data, and apply it to the target (English) side of the MT bitext, so that the paraphraser and MT models have the same output vocabulary. We also train SentencePiece models on the source-side of the bitexts. We use a subword vocabulary size of 4,000 for each.

4 Results

Results are shown in Table 1. We improve over the baseline in all settings, by 1.2 to 7 BLEU (all statistically significant at the 95% confidence level (Koehn, 2004)).[8] We see more pronounced improvements in the the lower-resource settings.[9]

5 Analysis

In this section, we analyze our method to explore: (1) How it performs at a variety of resource levels; and (2) How it compares to the popular data augmentation method of back-translation.

5.1 MT Data Ablation

In order to better understand how this method performs across data sizes on the same data set, we ablate Bengali-English bitext from GlobalVoices.[10] After reserving validation and test sets (as in § 3.2), approximately 132k lines are left for training; we ablate this to 100k, 50k, 25k, and 15k lines.

Figure 2 plots the performance of our method and the baseline against the log of the data amount. Our improvements of 2.7, 3.7, 1.6, and 0.8 BLEU at the 15k, 25k, 50k, and 100k subsets are statistically significant at the 95% confidence level; the 0.1 improvement for the full 132k data amount is not. Similar to Table 1, we see more pronounced improvements in lower-resource ablations.

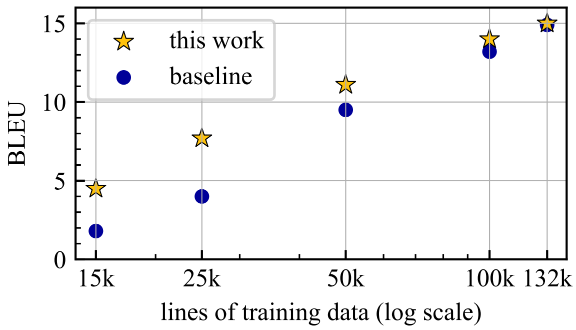

Figure 2: Bengali-English data ablation.

Neural paraphrasers are very rapidly improving in both adequacy and diversity (Wieting et al., 2017, 2019b; Li et al., 2018; Wieting and Gimpel, 2018; Hu et al., 2019a,b,c); as they continues to improve our method will likely provide larger improvements across the board, including for higher-resource MT.

5.2 Back-translation

Back-translation (Sennrich et al., 2016) is the most common method for incorporating non-parallel data in NMT. We investigate how our method interacts with it. Table 2 shows the results for back-translation, our work, and the combination of back-translation and our work.[11] Adding our method to the strong data augmentation baseline of back-translation improves performance by 0.5 to 5.7 BLEU[12] over back-translation alone.

For all our settings, the best performance either comes from our method combined with back-translation, or our method alone. In the lowest-

[8] All BLEU scores are SacreBLEU (Post, 2018).

[9] We acknowledge our three lowest-resource baselines (hu-en, id-en, cs-en) have very low BLEU scores and indicate very poor translations, and our even our large improvements may not be enough to make those systems practically usable. However, based on manual inspection, the improvement from 5.3 to 12.3 for id-en makes that system useful for gisting.

[10] We choose bn-en for its relatively large size while still containing dissimilar languages, as ablating French-English (another similarly-sized option from GlobalVoices) does not reflect typical low-resource MT performance.

[11] We use a 1:1 ratio of bitext to back-translated bitext. We use newscrawl2016 (`data.statmt.org/news-crawl`) as our monolingual text. When combining with our work, we run our method on both the original and back-translation data.

[12] All statistically significant at the 95% confidence level.

dataset	GlobalVoices										MATERIAL	
* → en	hu	id	cs	sr	ca	sw	nl	pl	mk	ar	sw	tl
train lines	8k	8k	11k	14k	15k	24k	32k	40k	44k	47k	19k	46k
baseline	2.3	5.3	3.4	11.8	16.0	17.9	22.2	16.0	27.0	12.7	37.8	32.5
baseline w/ back-translation	2.8	7.1	4.6	17.6	20.1	20.7	26.9	19.3	29.1	16.0	38.8	33.0
this work	**5.4**	12.3	**6.6**	16.1	20.0	20.5	24.8	18.0	28.2	14.9	39.0	**33.7**
this work w/ back-translation	4.9	**12.8**	**6.6**	**19.6**	**23.4**	**23.0**	**27.5**	**20.2**	**29.7**	**16.8**	**39.3**	**33.7**

Table 2: Comparison between back-translation and this work on the test set. We **bold** the best value as well as any result where the difference from it is not statistically significant at the 95% confidence level.

resource setting (hu-en) our method alone outperforms the baseline by 3.1 BLEU, but adding back-translation reduces the improvement by 0.5 BLEU. For cs-en and tl-en adding back-translation to our method does not change performance. In the remaining 9 (of 12) settings, back-translation and our proposed method are complementary and we see improvements of 1.2 to 7.8 BLEU[12] over the baseline when combining the two.

6 Related Work

Knowledge Distillation Our proposed objective is similarly structured to word-level knowledge distillation (KD; Hinton et al., 2015; Kim and Rush, 2016), where a student model is trained to match the output distribution of a teacher model. In KD both models are translation models trained on the same data, have the same input and output languages, and use the human reference as the previous token. In contrast, we train toward the distribution of the paraphraser, which takes as input the human reference sentence (in the target language), with the sampled paraphrase as the previous token. KD is usually used to train smaller models and does not incorporate additional data sources, like we do.

Integrating Paraphrases in MT Hu et al. (2019a) present case studies on paraphrasing as data augmentation for NLP tasks, including an appendix on NMT, where they show small gains. They generate paraphrases as an offline preprocessing step using heuristic constraints on the model's output, and train on the synthetic and original data. They then also find it necessary to fine tune on only the original data. Our work differs in that we train toward the paraphraser *distribution*, and we *sample* from the distribution rather than using heuristics.

Wieting et al. (2019a) used a paraphrase-similarity metric for minimum risk training (MRT; Shen et al., 2016) in NMT. They note MRT is expensive, and, following prior work, use it for fine-tuning after maximum likelihood training. While our method is ∼3 times slower than standard $\mathcal{L}_{\text{NLL}}$, this is not prohibitive in low-resource settings.

Paraphrasing was explored in the context of statistical machine translation (SMT) too. Callison-Burch et al. (2006) and Marton et al. (2009) used paraphrases to augment the phrase table directly, focusing on *source-side* paraphrasing to improve test set coverage. Madnani et al. (2007, 2008) used a coverage-focused paraphrasing technique to augment the set of references used during SMT tuning.

Data Augmentation in NMT Back-translation (BT) translates target-language monolingual text to create synthetic source sentences (Sennrich et al., 2016). BT needs a reverse model for each *language pair*. In contrast, our work needs a paraphraser only for each *target language*. Zhou et al. (2019) found BT is harmful in some low-resource language pairs, but a strong paraphraser can be trained as long as the target language is sufficiently high resource.

Fadaee et al. (2017) insert low frequency words in novel contexts in the existing bitext, using automatic word alignment and a language model. RAML (Norouzi et al., 2016) and SwitchOut (Wang et al., 2018) randomly replace words with another word from the vocabulary. In contrast to random or targeted word replacement, we generate semantically similar sentential paraphrases. Label smoothing (which we use with $\mathcal{L}_{\text{NLL}}$) spreads probability mass over all non-reference tokens equally (Szegedy et al., 2016); in $\mathcal{L}_{\text{para}}$ the paraphraser places more mass on semantically plausible tokens.

7 Conclusion

In this work we find that our novel method for simulating multiple references in the MT training leads to significantly improved performance in low-resource settings, with gains of 1.2 to 7 BLEU.

References

Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. 2016. Czeng 1.6: Enlarged Czech-English parallel corpus with processing tools dockered. In *Text, Speech, and Dialogue*, pages 231–238, Cham. Springer International Publishing.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 17–24, Stroudsburg, PA, USA. Association for Computational Linguistics.

Markus Dreyer and Daniel Marcu. 2012. HyTER: Meaning-equivalent semantics for translation evaluation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 162–171, Montréal, Canada. Association for Computational Linguistics.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Alex Graves. 2013. Generating sequences with recurrent neural networks. *CoRR*, abs/1308.0850.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6100–6113, Hong Kong, China. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019a. Improved lexically constrained decoding for translation and monolingual rewriting. 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 839–850, Minneapolis, Minnesota. Association for Computational Linguistics.

J. Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme. 2019b. ParaBank: Monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. In *Proceedings of AAAI*.

J. Edward Hu, Abhinav Singh, Nils Holzenberger, Matt Post, and Benjamin Van Durme. 2019c. Large-scale, diverse, paraphrastic bitexts via sampling and clustering. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 44–54, Hong Kong, China. 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. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

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

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. 2018. Paraphrase generation with deep reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3865–3878, Brussels, Belgium. Association for Computational Linguistics.

Nitin Madnani, Necip Fazil Ayan, Philip Resnik, and Bonnie Dorr. 2007. Using paraphrases for parameter tuning in statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 120–127, Prague, Czech Republic. Association for Computational Linguistics.

Nitin Madnani, Philip Resnik, Bonnie J. Dorr, and Richard Schwartz. 2008. Are multiple reference translations necessary? investigating the value of paraphrased reference translations in parameter optimization. In *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*, Waikiki, Hawaii.

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 381–390, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mohammad Norouzi, Samy Bengio, zhifeng Chen, Navdeep Jaitly, Mike Schuster, Yonghui Wu, and Dale Schuurmans. 2016. Reward augmented maximum likelihood for neural structured prediction. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1723–1731. Curran Associates, Inc.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Carl Rubino. 2018. Keynote: Setting up a machine translation program for IARPA. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Papers)*, Boston, MA. Association for Machine Translation in the Americas.

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

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692, Berlin, Germany. Association for Computational Linguistics.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018. SwitchOut: an efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 856–861, Brussels, Belgium. Association for Computational Linguistics.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019a. Beyond BLEU:training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

John Wieting, Kevin Gimpel, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019b. Simple and effective paraphrastic similarity from parallel translations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4602–4608, Florence, Italy. Association for Computational Linguistics.

John Wieting, Jonathan Mallinson, and Kevin Gimpel. 2017. Learning paraphrastic sentence embeddings from back-translated bitext. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 274–285, Copenhagen, Denmark. Association for Computational Linguistics.

Ronald J. Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280.

Chunting Zhou, Xuezhe Ma, Junjie Hu, and Graham Neubig. 2019. Handling syntactic divergence in low-resource machine translation. 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 1388–1394, Hong Kong, China. Association for Computational Linguistics.