Using Context in Neural Machine Translation Training Objectives

Danielle Saunders and **Felix Stahlberg**[*] and **Bill Byrne**

Department of Engineering, University of Cambridge, UK

ds636@cam.ac.uk   fstahlberg@google.com   wjb31@cam.ac.uk

Abstract

We present Neural Machine Translation (NMT) training using document-level metrics with batch-level documents. Previous sequence-objective approaches to NMT training focus exclusively on sentence-level metrics like sentence BLEU which do not correspond to the desired evaluation metric, typically document BLEU. Meanwhile research into document-level NMT training focuses on data or model architecture rather than training procedure. We find that each of these lines of research has a clear space in it for the other, and propose merging them with a scheme that allows a document-level evaluation metric to be used in the NMT training objective.

We first sample pseudo-documents from sentence samples. We then approximate the expected document BLEU gradient with Monte Carlo sampling for use as a cost function in Minimum Risk Training (MRT). This two-level sampling procedure gives NMT performance gains over sequence MRT and maximum-likelihood training. We demonstrate that training is more robust for document-level metrics than with sequence metrics. We further demonstrate improvements on NMT with TER and Grammatical Error Correction (GEC) using GLEU, both metrics used at the document level for evaluations.

1 Introduction

Neural Machine Translation (NMT) research has explored token-level likelihood functions (Sutskever et al., 2014; Bahdanau et al., 2015) and sequence-level objectives inspired by reinforcement learning (Ranzato et al., 2016; Bahdanau et al., 2016) or expected Minimum Risk Training (MRT) (Shen et al., 2016). A typical sequence objective in these cases is based on sentence-level BLEU (sBLEU) (Edunov et al., 2018). However

_____
[*]Now at Google

sBLEU, even if aggregated over sentences, is only an approximation of the desired metric, document-level BLEU. Beyond translation, many metrics for natural language tasks do not have robust sentence-level approximations. A logical progression is the extension of sequence-level NMT training objectives to include context from outside the sentence.

Document-based NMT, by contrast, aims to use out-of-sentence context to improve translation. Recent research explores lexical consistency by providing additional sentences during training (Maruf et al., 2019; Voita et al., 2018, 2019) or inference (Voita et al., 2019; Stahlberg et al., 2019), potentially with adjustments to model architecture. However, to the best of our knowledge, no attempt has been made to extend sequence-level neural training objectives to include document-level reward functions. This is despite document-level BLEU being arguably the most common NMT metric, and being the function originally optimised by Minimum Error Rate Training (MERT) for Statistical Machine Translation (SMT) (Och, 2003).

We propose merging lines of research on training objectives and document-level translation. We achieve this by presenting a document-level approach to sequence-level objectives which brings the training objective closer to the actual evaluation metric, using MRT as a representative example. We demonstrate MRT under document-level BLEU as well as Translation Edit Rate (TER) (Snover, 2006), which while decomposable to sentence level is less noisy when used over documents. We consider both pseudo-documents where sentences are assigned randomly to a mini-batch, and true document context where all sentences in the batch are from the same document.

We finally apply our scheme to supervised Grammatical Error Correction, for which using neural models is becoming increasingly popular (Xie et al., 2016; Sakaguchi et al., 2017; Stahlberg et al., 2019).

We show gains in GEC metrics GLEU (Napoles et al., 2015) and M2 (Dahlmeier and Ng, 2012).

1.1 Related Work

Minimum Error Rate Training was introduced for phrase-based SMT with document-level BLEU (Och, 2003). Shen et al. (2016) extend these ideas to NMT, using expected minimum risk at the sequence level with an sBLEU cost for end-to-end NMT training. Edunov et al. (2018) explore random and beam sampling for NMT sequence-MRT, as well as other sequence-level training losses.

Related developments in NMT include combined reinforcement-learning/cross-entropy approaches such as MIXER (Ranzato et al., 2016), which itself has origins in the REINFORCE algorithm described by Williams (1992). We do not explore such approaches, although our document-sampling and document-metric schemes could in principle be extended to them.

Sequence-level MRT has seen success outside NMT. Ayana et al. (2016) use sequence MRT for summarization, while Shannon (2017) uses a related approach for speech recognition. MRT can be seen as a special case of neural reinforcement learning, which Sakaguchi et al. (2017) apply to GEC with sequence-level costs. Closest to our approach is the work of Jean and Cho (2019) on NMT with a minibatch-context-sensitive training procedure. However, they do not optimize on document metrics over those contexts. They also sample contexts randomly, while we find diverse context sampling is important for the success of document-MRT.

2 Background

2.1 Sequence-level MRT

Sentence-level MRT for NMT aims to minimize the expected loss on training data with a loss function between sampled target sentences $\boldsymbol{y}$ and gold reference sentences $\boldsymbol{y}^*$. For NMT a common sentence-level cost function $\Delta(\boldsymbol{y}, \boldsymbol{y}^*)$ is 1 - sBLEU, where sBLEU is smoothed by setting initial n-gram counts to 1 (Edunov et al., 2018).

We take N samples for each of the S sentences in a mini-batch. We write the cost function between the sth reference in a mini-batch, $\boldsymbol{y}^{(s)*}$, and its nth sample, $\boldsymbol{y}_n^{(s)}$, as $\Delta_n^{(s)} = \Delta(\boldsymbol{y}_n^{(s)}, \boldsymbol{y}^{(s)*})$. The risk gradient for end-to-end NMT with MRT as in Shen

et al. (2016), with sample-count scaling, is then:

$$\nabla_\theta R(\theta) = \frac{1}{N} \sum_{s=1}^{S} \sum_{n=1}^{N} \Delta_n^{(s)} \frac{\partial}{\partial \theta} \log P(\boldsymbol{y}_n^{(s)} | \boldsymbol{x}^{(s)}; \theta)$$
(1)

2.2 Document-level MRT

By analogy with sequence-level MRT, we consider MRT over batches of S sentence pairs, which we treat as a pseudo-document. In practice we experiment both with sentences chosen randomly from all training data, and with true context where all sentences per batch are from a single document.

Let $X = [\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(S)}]$ be the source document, $Y = [\boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(S)}]$ be a document of candidate translations, and $Y^* = [\boldsymbol{y}^{(1)*}, \dots, \boldsymbol{y}^{(S)*}]$ be the reference translations. Document-level metric $D(Y, Y^*)$, which may be non-differentiable, replaces the sequence-level metric $\Delta(\boldsymbol{y}, \boldsymbol{y}^{(s)*})$. We define the document-level risk:

$$R(\theta) = \sum_Y D(Y, Y^*) P(Y|X; \theta)$$

Using $p_\theta \nabla_\theta \log p_\theta = \nabla p_\theta$:

$$\nabla_\theta R(\theta) = \sum_Y D(Y, Y^*) P(Y|X; \theta) \nabla_\theta \log P(Y|X; \theta)$$
$$= \mathbb{E}\big[D(Y, Y^*) \nabla_\theta \log P(Y|X; \theta) | X; \theta\big]$$
(2)

Using simple Monte-Carlo, after Shannon (2017), we replace the expectation by an average taken over N sampled translation documents $Y_n \sim P(Y|X; \theta)$

$$\nabla_\theta R(\theta) \approx \frac{1}{N} \sum_{n=1}^{N} D(Y_n, Y^*) \nabla_\theta \log P(Y_n|X; \theta)$$

The nth sample for the sth sentence in the batch-level document, $\boldsymbol{y}_n^{(s)}$, contributes the following to the overall gradient:

$$\nabla_\theta R(\theta) \approx \frac{1}{N} \sum_{Y : \boldsymbol{y}^{(s)} = \boldsymbol{y}_n^{(s)}} D(Y, Y^*) \nabla_\theta \log P(\boldsymbol{y}_n^{(s)} | \boldsymbol{x}^{(s)}; \theta)$$

In other words the gradient of each sample is weighted by the aggregated document-level scores for documents in which the sample appears.
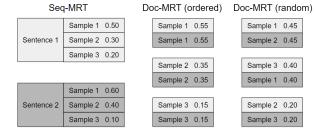
Seq-MRT

Sentence 1	Sample 1	0.50
	Sample 2	0.30
	Sample 3	0.20

Sentence 2	Sample 1	0.60
	Sample 2	0.40
	Sample 3	0.10

Doc-MRT (ordered)

Sample 1	0.55
Sample 1	0.55
Sample 2	0.35
Sample 2	0.35
Sample 3	0.15
Sample 3	0.15

Doc-MRT (random)

Sample 1	0.45
Sample 2	0.45
Sample 3	0.40
Sample 1	0.40
Sample 2	0.20
Sample 3	0.20

Figure 1: Sample-ordering schemes for MRT with $S = 2$ sentences / batch and $N = 3$ samples / sentence, showing sample costs. In sequence-MRT each sample has its own cost (e.g. sBLEU). For doc-MRT (ordered), samples are ordered and sorted into N-wise 'documents', each with a combined cost (e.g. document BLEU). The ordered assignment enforces an extreme range of combined costs. In doc-MRT (random), samples are randomly assigned, making documents on average less diverse with less distinct scores, with a low likelihood of extreme distributions.

2.3 Mini-batch level document sampling

To generate sample documents we first sample sentences. Sentence sampling for NMT generates new tokens in a left-to-right manner (Shen et al., 2016). In left-to-right generation each token is sampled from a distribution conditioned on previously sampled tokens, minimizing exposure bias to gold references which the model is unlikely to see at inference time (Ranzato et al., 2016). Sampling can be via beam search, or random sampling from the model distribution given previously sampled tokens. Beam search produces more likely samples which may be less diverse compared to random sampling (Edunov et al., 2018).

Here we only consider sampling during training. While samples can be more easily generated offline with respect to fixed model parameters, such samples are not representative of the current model.

With N sample translations for each of the S sentence pairs per batch we can construct N^S possible sample documents as sequences of S sentences. Considering all possible documents is intractable unless N and S are small. It also carries the risk that a single sentence will appear in multiple sampled documents, giving it undue weight.

Instead we propose creating N documents by first ordering samples for each sentence (e.g. by sBLEU), then creating the nth sample document Y_n by concatenating the nth sample from each sentence. This gives a set of N diverse documents sampled from N^S possibilities. We expect the sampled documents to be diverse in contents, since a given sentence will only ever occur in a single document context, and diverse in score. We refer to this scheme as ordered document sampling.

Figure 1 illustrates ordered document sampling by comparison to a scheme which randomly samples sentences to form documents.

3 Experiments

We report on English-German NMT. We initialize with a baseline trained on 17.5M sentence pairs from WMT19 news task datasets (Barrault et al., 2019), on which we learn a 32K-merge joint BPE vocabulary (Sennrich et al., 2016). We validate on newstest2017, and evaluate on newstest2018.

We apply MRT only during fine-tuning, following previous work (Edunov et al., 2018; Shen et al., 2016). In early experiments, we found that training from scratch with discriminative objectives (sequence- or document-based) is ineffective. We suspect samples produced early in training are so unlike the references that the model never receives a strong enough signal for effective training.

We fine-tune on old WMT news task test sets (2008-2016) in two settings. With **random batches** sentences from different documents are shuffled randomly into mini-batches. In this case doc-MRT metrics are over pseudo-documents. With **document batches** each batch contains only sentences from one document, and doc-MRT uses true document context. We use the same sampling temperatures and the same risk sharpness factors for both forms of MRT for each experiment.

For Grammatical Error Correction (GEC) we train on sentences from NUCLE (Dahlmeier et al., 2013) and Lang-8 Learner English (Mizumoto et al., 2012) with at least one correction, a total of 660K sentences. We evaluate on the JFLEG (Napoles et al., 2017) and CoNLL 2014 (Ng et al., 2014) sets. For GEC experiments we use random batching only.

For all models we use a Transformer model (Vaswani et al., 2017) with the 'base' Tensor2Tensor parameters (Vaswani et al., 2018).

We train to validation set BLEU convergence on a single GPU. The batch size for baselines and MLE is 4096 tokens. For MRT, where each sentence in the batch is sampled N times, we reduce batch size by N while delaying gradient updates by the same factor to keep the effective batch size constant (Saunders et al., 2018). At inference time we decode using beam size 4. All BLEU scores

are for cased, detokenized output, calculated using SacreBLEU (Post, 2018).

3.1 Computation and sample count

Our proposed document-MRT approach is more complex than sequence-MRT due to the additional score-aggregation and context-sampling steps. In practice we find that the extra computation of ordering and aggregating sequence scores is negligible when compared to the computational cost of sentence sampling, required for all forms of MRT.

Our MRT experiments use $N = 8$ random samples per sentence unless otherwise stated. In this we choose the highest N we can practically experiment with, since previous work finds MRT performance increasing steadily with more samples per sentence (Shen et al., 2016).

That we see improvements with so few samples is in contrast to previous work which finds BLEU gains only with 20 or more samples per sentence for sequence-MRT (Shen et al., 2016; Edunov et al., 2018). However, we find that document-MRT allows improvements with far fewer samples, perhaps because the aggregation of scores over sentences in a context increases robustness to variation in individual samples.

Relatedly, we find that add-one BLEU smoothing (Lin and Och, 2004) is required for sequence-MRT as in Shen et al. (2016). However we find that doc-MRT can achieve good results without smoothing, perhaps because n-gram precisions are far less likely to be 0 when calculated over a document.

3.2 MRT for NMT

Model	Random batches		Document batches	
Baseline	42.7			
MLE	40.0		41.0	
	$N = 4$	$N = 8$	$N = 4$	$N = 8$
Seq-MRT	42.6	43.5	42.6	43.5
Doc-MRT (random)	41.7*	43.1*	43.1	43.0
Doc-MRT (ordered)	**43.4**	**43.7**	**43.4**	**43.9**

Table 1: BLEU on en-de after MLE and MRT under $1-$sBLEU (seq-MRT) and $1-$doc BLEU (doc-MRT). Results indicated by $*$ are averages over 3 runs with the same settings, which all came within 0.2 BLEU.
.

In Table 1, we fine-tune an en-de baseline on documents from past news sets. We compare sentence-BLEU and document-BLEU MRT to fine-tuning with Maximum Likelihood Estimation (MLE).

Model	Random batches	Document batches
Baseline	39.2	39.2
MLE	41.2	40.0
Seq-MRT	39.4	40.5
Doc-MRT (ordered)	**39.0**	**38.9**

Table 2: TER on en-de after MLE and MRT under sentence-TER (seq-MRT) and doc-TER (doc-MRT). Lower TER is better.

MLE fine-tuning degrades the baseline. This suggests the baseline is well-converged, as is desirable for applying MRT (Shen et al., 2016). The degradation is smaller with batches containing only sentences from the same document. We connect this to the idea that NMT batches with fewer sentence pairs have noisier estimated gradients, harming training (Saunders et al., 2018). We expect batches of sentences from a single document to be similar and therefore give less noisy gradient estimates.

Both seq-MRT and doc-MRT improve over the baseline with random sampling and $N = 8$. We also explore MRT at $N = 4$, with batch size adjusted as described in section 3 for the same effective batch size per update, and with fewer training steps such that the model 'sees' a similar proportion of the overall dataset. We do not report beam sampling results as early experiments indicate beam sampling gives similarly poor results for both seq-MRT and doc-MRT. This may be because beam search produces insufficiently diverse samples for this task (Freitag and Al-Onaizan, 2017).

Sequence-MRT gives a 0.8 BLEU gain over the baseline with both batching schemes using $N = 8$ samples, but starts to degrade the baseline with $N = 4$ samples. With document batches and $N = 8$ Doc-MRT (ordered) outperforms seq-MRT by a further 0.4 BLEU. With $N = 4$ doc-MRT (ordered) still achieves a 0.7 BLEU improvement over the baseline, or a 0.8 BLEU improvement over seq-MRT. We suggest therefore that doc-MRT (ordered) may be a computationally more efficient alternative to seq-MRT when large sample counts are not practical.

For contrast with the ordered document sampling approach of Section 2.3, we give results for doc-MRT (random), which uses randomly sampled contexts. This approach falls significantly behind doc-MRT (ordered) with either batching scheme. Since doc-MRT (random) with random batches is exposed to randomness at the batch construction,

Model	JFLEG				CONLL2014			
	P	**R**	**M2**	**GLEU**	**P**	**R**	**M2**	**GLEU**
Baseline	**67.3**	38.2	**58.4**	50.4	**54.4**	21.8	41.9	67.3
MLE	64.7	37.7	56.6	50.1	51.4	20.9	39.8	67.1
Seq-MRT	62.7	39.1	56.0	50.0	52.4	24.5	42.7	67.1
Doc-MRT (ordered)	64.4	**41.0**	57.8	**51.4**	53.2	**24.6**	**43.2**	**67.5**

Table 3: GEC Precision, Recall, M2, and GLEU after MLE and MRT. MRT is under $1-$sentence-GLEU for seq-MRT and $1-$doc-GLEU for doc-MRT. Both MRT schemes uses random batches and random sentence sampling. Higher scores are better for all metrics.

sentence sampling and document sampling stages, these results are averages over 3 experimental runs, which gave fairly consistent results (<0.2 BLEU range). In general we do find that results with random batches and random ordering are variable and sensitive to batch size and batching scheme.

We interpret these results by considering the effect on the per-sentence cost for the different schemes. We find MRT works well when sample scores are different enough to be discriminated, but suffers if scores are too different. This is in line with the findings of Edunov et al. (2018) that including the gold reference causes the model to assign low relative probabilities to every other sample.

Doc-MRT aggregates scores over many samples, while seq-MRT uses individual scores. We believe this explains the stronger performance of doc-MRT for small values of N, especially for the ordered document scheme, which ensures scores are still different enough for MRT to discriminate.

Our approach can also be used with document-level metrics that are not intended to be used with individual sentences. In Table 2 we demonstrate this with TER, which estimates the edit rate required to correct a set of translation hypotheses. Document-TER MRT improves over a strong baseline, although batching scheme has less of an impact here. Notably seq-level MRT does not improve TER over the baseline, indicating TER may be too noisy a metric for use at the sentence level.

3.3 MRT for GEC

Finally, we apply our MRT approach to the GEC GLEU metric (Napoles et al., 2015), an n-gram edit measure typically used at the document level. Table 3 shows that document MRT fine-tuning improves GLEU over the baseline, MLE fine-tuning, and a sequence-GLEU MRT formulation. Also notable is the change in M2, which finds the phrase-level edit sequence achieving the highest overlap with the gold-standard (Dahlmeier and Ng, 2012). MLE and sequence-MRT improve recall at a detri-

ment to precision, suggesting over-generation of spurious corrections. Document-MRT likewise improves recall, but with a precision score closer to the baseline for more balanced performance. There is clear indication of a tension between M2 and GLEU: a small increase in GLEU under doc-MRT on CONLL leads to a large increase in M2, while a large increase in GLEU under doc-MRT on JFLEG leads to a small decrease in M2.

We note that our improvements on JFLEG are similar to the improvements shown by Sakaguchi et al. (2017) for neural reinforcement learning with a sequence-GLEU cost metric. However, their results involve N=20 samples and 600k updates, compared to N=8 and 3k updates with our approach.

4 Conclusions and future work

We present a novel approach for structured loss training with document-level objective functions. Our approach relies on a procedure for sampling a set of diverse batch-level contexts using N-wise sample ordering. As well as randomly selecting training data, we assess training with mini-batches consisting only of single document contexts. While the scope of this work does not extend to sampling sentences given document context, this would be an interesting direction for future work.

We demonstrate improvements covering three document-level evaluation metrics: BLEU and TER for NMT and GLEU for GEC. We finish by noting that the original MERT procedure developed for SMT optimised document-level BLEU and with our procedure we reintroduce this to NMT.

---

[1] http://www.hpc.cam.ac.uk

References

Shiqi Shen Ayana, Zhiyuan Liu, and Maosong Sun. 2016. Neural headline generation with minimum risk training. *arXiv preprint arXiv:1604.01904*.

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016. An actor-critic algorithm for sequence prediction. *arXiv preprint arXiv:1607.07086*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR'15)*.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572. Association for Computational Linguistics.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, David Grangier, et al. 2018. Classical structured prediction losses for sequence to sequence learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 355–364.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60, Vancouver. Association for Computational Linguistics.

Sébastien Jean and Kyunghyun Cho. 2019. Context-aware learning for neural machine translation. *arXiv preprint arXiv:1903.04715*.

Chin-Yew Lin and Franz Josef Och. 2004. ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 501–507, Geneva, Switzerland. COLING.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. 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 3092–3102, Minneapolis, Minnesota. Association for Computational Linguistics.

Tomoya Mizumoto, Yuta Hayashibe, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2012. The effect of learner corpus size in grammatical error correction of ESL writings. In *Proceedings of COLING 2012: Posters*, pages 863–872. The COLING 2012 Organizing Committee.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 588–593. Association for Computational Linguistics.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14. Association for Computational Linguistics.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. 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, Belgium, Brussels. Association for Computational Linguistics.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *ICLR*.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2017. Grammatical error correction with neural reinforcement learning. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 366–372. Asian Federation of Natural Language Processing.

Danielle Saunders, Felix Stahlberg, Adrià de Gispert, and Bill Byrne. 2018. Multi-representation ensembles and delayed SGD updates improve syntax-based NMT. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 319–325, Melbourne, Australia. Association for Computational Linguistics.

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

Matt Shannon. 2017. Optimizing expected word error rate via sampling for speech recognition. *arXiv preprint arXiv:1706.02776*.

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, pages 1683–1692.

M Snover. 2006. A study of translation edit rate with targeted human annotation. *Proc. Association for Machine Translation in the Americas (AMTA2006)*.

Felix Stahlberg, Christopher Bryant, and Bill Byrne. 2019. Neural grammatical error correction with finite state transducers. 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 4033–4039.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 193–199, Boston, MA. Association for Machine Translation in the Americas.

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 *Advances in Neural Information Processing Systems*, pages 6000–6010.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. Context-aware monolingual repair for neural 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 876–885, Hong Kong, China. Association for Computational Linguistics.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Ziang Xie, Anand Avati, Naveen Arivazhagan, Dan Jurafsky, and Andrew Y Ng. 2016. Neural language correction with character-based attention. *arXiv preprint arXiv:1603.09727*.