Addressing Exposure Bias With Document Minimum Risk Training: Cambridge at the WMT20 Biomedical Translation Task

Danielle Saunders and **Bill Byrne**

Department of Engineering, University of Cambridge, UK

Abstract

The 2020 WMT Biomedical translation task evaluated Medline abstract translations. This is a small-domain translation task, meaning limited relevant training data with very distinct style and vocabulary. Models trained on such data are susceptible to exposure bias effects, particularly when training sentence pairs are imperfect translations of each other. This can result in poor behaviour during inference if the model learns to neglect the source sentence.

The UNICAM entry addresses this problem during fine-tuning using a robust variant on Minimum Risk Training. We contrast this approach with data-filtering to remove 'problem' training examples. Under MRT fine-tuning we obtain good results for both directions of English-German and English-Spanish biomedical translation. In particular we achieve the best English-to-Spanish translation result and second-best Spanish-to-English result, despite using only single models with no ensembling.

1 Introduction

Neural Machine Translation (NMT) in the biomedical domain presents challenges in addition to general domain translation. Text often contains specialist vocabulary and follows specific stylistic conventions. For this task fine-tuning generic pre-trained models on smaller amounts of biomedical-specific data can lead to strong performance, as we found in our 2019 biomedical submission (Saunders et al., 2019). For our WMT 2020 submission we start with strong single models from that 2019 submission and fine-tune them exclusively on the small Medline abstracts training sets (Bawden et al., 2019). This allows fast training on very relevant training data, since the test set is also made up of Medline abstracts.

However, fine-tuning on relevant but small corpora has pitfalls. The small number of training examples exacerbates the effect of any noisy or poorly aligned sentence pairs. We treat this as a form of exposure bias, in that model overconfidence in training data results in poor translation hypotheses at test time.

Our contributions in this system paper are:

- A discussion of exposure bias in the form of imperfect training data, focusing on the biomedical domain.
- An exploration of straightforward ways to mitigate exposure bias via data preparation and training objective.
- A discussion of our 2020 Biomedical task results for single models fine-tuned on small, domain-specific data sets.

1.1 Exposure bias in the biomedical domain

Exposure bias for an autoregressive sequence decoder refers to a discrepancy between decoder conditioning during training and inference (Bengio et al., 2015; Ranzato et al., 2016). During training the decoder generates a hypothesis for the t^{th} output token $\hat{y}_t$ conditioned on $y_{1:t-1}$, the gold target sequence prefix. During inference, the gold target y is unavailable, and $\hat{y}_t$ is conditioned instead on the hypothesis prefix $\hat{y}_{1:t-1}$.

Previous work has interpreted the risk of exposure bias primarily in terms of the model over-relying on correct gold target translations, resulting in error propagation when mistakes are made during inference. We take a different view, focusing on mistakes in the training data which harm the model through teacher-forcing exposure and cause it to make related mistakes during inference.

We identify a specific feature of the Medline abstract training data which caused noticeable translation errors. The data contains instances in which either the source or target sentence contains the

Proceedings of the 5th Conference on Machine Translation (WMT), pages 862–869
Online, November 19–20, 2020. ©2020 Association for Computational Linguistics

English source	[Associations of work-related strain with subjective sleep quality and individual daytime sleepiness].
Human translation	[Zusammenhang von arbeitsbezogenen psychischen Beanspruchungsfolgen mit subjektiver Schlafqualität und individueller Tagesschläfrigkeit.]
MLE	Zusammenfassung.
MRT	[Assoziationen arbeitsbedingter Belastung mit subjektiver Schlafqualität und individueller Tagesschläfrigkeit].
English source	[Effectiveness of Upper Body Compression Garments Under Competitive Conditions: A Randomised Crossover Study with Elite Canoeists with an Additional Case Study].
Human translation	[Effektivität von Oberkörperkompressionsbekleidung unter Wettkampfbedingungen: eine randomisierte Crossover-Studie an Elite-Kanusportlern mit einer zusätzlichen Einzelfallanalyse.]
MLE	Eine randomisierte Crossover-Studie mit Elite-Kanuten mit einer Additional Case Study wurde durchgeführt.
MRT	Eine randomisierte Crossover-Studie mit Elite-Kanüsten mit einer Additional Case Study hat zur Wirksamkeit von Oberkörperkompressionsbekleidung unter kompetitiven Bedingungen geführt.

Table 1: Two sentence from the English-German 2020 test set with hypothesis translations from various models, demonstrating the effects of exposure bias from training on imperfectly aligned training sentences. The first MLE example output is completely unrelated to the source sentence, but the second MLE translation is more misleading.

correct translation of the other sentence, but adds information that is not found in translation. For example, the following sentence appears in the English side of en-de Medline abstract training data:

[The effects of Omega-3 fatty acids in clinical medicine]. Effects of Omega-3 fatty acids (n-3 FA) in particular on the development of cardiovascular disease (CVD) are of major interest.

Its corresponding German sentence is

Der Nutzen von Omega-3-Fettsäuren (n-3-FS) in der Medizin, hauptsächlich in der Prävention kardio- und zerebrovaskulärer Erkrankungen, wird aktuell intensiv diskutiert. (Translated: ‘The uses of Omega-3 fatty acids in medicine, especially in prevention of cardiovascular and cerebrovascular diseases, are currently heavily discussed.’)

Some of the English sentence is present in the German translation, but the square-bracketed article title is not. In this example it might be possible to remove only the segment in square brackets, but in other examples there is even less overlap, while source and target sentences may still be related and therefore challenging to filter. For example, the following English and German sentences also correspond with still less overlap:

[Conflict of interest with industry–a survey of nurses in the field of wound care in Germany , Australia and Switzerland]. Background.

Hintergrund: Pflegende werden zunehmend von der Industrie umworben. (Translated: ‘Background: Nurses are being increasingly courted by industry.’)

These examples are quite frequent in Medline abstract data, especially in the form of titles. It is common to insert the English title of a non-English article into its translation, marked with square brackets (Patrias and Wendling, 2007). The marked title is not present in the original article. Consequently models trained on English source sentences with titles can behave erratically when given sentences with square-bracketed titles at test time: an exposure bias effect.

One possible approach to this problem is aggressively filtering sentences which may be poorly aligned. However, with such a small training set, this risks losing valuable examples of domain-specific source and target language. We hypothesise that such filtering is not the only way to reduce the effects during inference. Instead, we propose an approach in terms of the parameter fine-tuning scheme with Minimum Risk Training (MRT). Wang and Sennrich (2020) have recently shown MRT as effective for combating exposure bias in the context of domain shift – test sentences which are very different from the training data. We propose that MRT is also more robust against exposure to misaligned training data.

The examples in Table 1 show the different behaviour of MLE and MRT in such cases. In the first example, the MLE hypothesis is unrelated to the source sentence, while the MRT output is relevant. In the second example, the MLE output is more plausible and therefore misleading, as it still misses the first clause which the MRT hypothesis covers. Both MLE and MRT hypotheses are phrased like opening sentences rather than titles, and both feature the untranslated phrase ‘Additional Case Study’: while MRT may be more robust, it is not immune to exposure bias.

We note that title translations may not exist in the human reference. In these cases failure to translate the title will not negatively impact BLEU. However, we argue a biomedical translation model should be able to translate such sentences if required. It is also important to note that title translations are not the only case of inexact training pairs, but are simply easily identifiable.

1.2 Document MRT

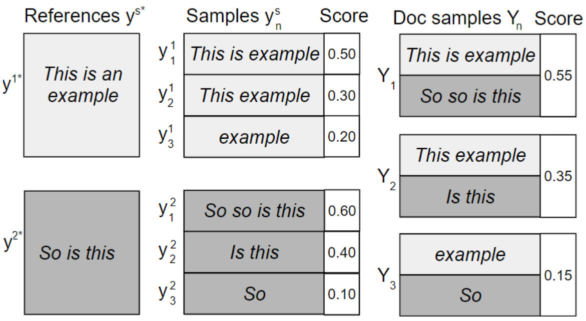

Figure 1: Two MRT schemes with an $S = 2$ sentence minibatch and $N = 3$ samples / sentence. In standard MRT (middle) each sample has a score, e.g. sBLEU. For doc-MRT (right) samples are sorted into minibatch-level 'documents', each with a combined score, e.g. document BLEU. Doc-MRT scores are less sensitive to individual samples, increasing robustness.

Minimum Risk Training (MRT) aims to minimize the expected cost between N sampled target sequences $\boldsymbol{y_n^{(s)}}$ and the corresponding gold reference sequence $\boldsymbol{y}^{(s)*}$ for the S sentence pairs in each minibatch. For translation MRT is usually applied using a sentence-level BLEU (sBLEU) score corresponding to cost function $1 -$ sBLEU, and sentence samples are generated by auto-regressive sampling with temperature τ during training (Shen et al., 2016). Hyperparameter α controls sharpness of the distribution over samples. While MRT permits training from scratch, in practice it is exclusively used to fine-tune models.

Doc-MRT is a recently proposed MRT variant which changes sentence cost function to a document cost function, $D(.)$ (Saunders et al., 2020). D measures costs between minibatch-level 'documents' Y^* and Y_n. Y^* is formed of all S reference sentences in the minibatch, and Y_n is one of N sample 'documents' each formed of one sample from each sentence pair $(\boldsymbol{x}^{(s)}, \boldsymbol{y}^{(s)*})$. This permits MRT under document-level scores like BLEU, instead of sBLEU. The n^{th} sample for the s^{th} sentence in the minibatch-level document, $\boldsymbol{y}_n^{(s)}$, contributes the following term to the overall gradient:

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

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

Figure 1 gives a toy example of doc-MRT scoring samples in context. Document-level metrics aggregate scores across sentence samples, meaning a minibatch with some good samples and some poor samples will not have extreme score variation. Doc-MRT is therefore less sensitive than standard MRT to variation in individual samples.

Doc-MRT has been shown to give better performance than standard MRT for small datasets with a risk of over-fitting, as well as improved robustness to small N. More discussion of these results and a derivation of the document-level loss function can be found in Saunders et al. (2020). Since we are attempting fine-tuning on small datasets and since N is a limiting factor for MRT on memory-intensive large models, the biomedical task is an appropriate application for doc-MRT.

1.3 Related work

Fine-tuning general models on domain-specific datasets has become common in NMT. Simple transfer learning on new data can adapt a general model to in-domain data (Luong and Manning, 2015). Mixed fine-tuning where some original data is combined with the new data avoids reduced performance on the original data-set (Chu et al., 2017). We are only interested in performance on one domain, so use simple transfer learning.

For this task, we specifically fine-tune on a relatively small dataset. Adaptation to very small, carefully-chosen domains has been explored for speaker-personalized translation (Michel and Neubig, 2018) , and to reduce gender bias effects (Saunders and Byrne, 2020) while maintaining general domain performance. We wish to adapt to a very specific domain without need to maintain good general domain performance, but must avoid overfitting. Related approaches include fine-tuning a separate model for each test sentence (Li et al., 2018; Farajian et al., 2017) or test document (Xu et al., 2019; Kothur et al., 2018). We choose to train a single model for all test sentences in a language pair, but improve the robustness of that model to overfitting and exposure bias using MRT.

	Phase	Datasets	Sentence pairs	Dev datasets	Sentence pairs
en-es	Pre-training	UFAL Medical[1] Scielo[3] Medline titles[4] Medline abstracts Total	639K 713K 288K 83K 1723K / **1291K**	Khresmoi[2]	1.5K
	Fine-tuning	Medline abstracts	83K / **67.5K**	Biomedical19	800
en-de	Pre-training	UFAL Medical Medline abstracts Total	2958K 33K 2991K / **2156K**	Khresmoi Cochrane[5]	1.5K 467
	Fine-tuning	Medline abstracts	33K / **28.6K**	Biomedical19	800

Table 2: Biomedical training and validation data used in the evaluation task. For both language pairs identical data was used in both directions. Bolded numbers are totals after filtering

MRT has been widely applied to NMT in recent years (Shen et al., 2016; Neubig, 2016; Edunov et al., 2018). In particular, Wang and Sennrich (2020) recently highlighted the efficacy of MRT for reducing the effects of exposure bias.

2 Experimental setup

2.1 Data

We report on two language pairs: English-Spanish (en-es) and English-German (en-de). Table 2 lists the data used to train our biomedical domain evaluation systems. For each language pair we use the same training data in both directions, and preprocess all data with Moses tokenization, punctuation normalization and truecasing. We use a 32K-merge joint source-target BPE vocabulary (Sennrich et al., 2016) learned on the pre-training data.

All of our submitted approaches involve fine-tuning pre-trained models. We initialise fine-tuning with the strong biomedical domain models that formed our 'run 1' submission for the WMT19 biomedical translation task. Details of data preparation and training for these models are discussed in Saunders et al. (2019).

We fine-tune these models on Medline abstracts data, validating on test sets from the 2019 Biomedical task. For these we concatenate the src-trg and trg-src 2019 test sets for each language pair, and select only the 'OK' aligned sentences as annotated by the organizers.

Before fine-tuning we carry out detected language filtering on the Medline abstracts fine-tuning data using the Python LangDetect package[6]. We find LangDetect has a tendency to incorrectly label short sentences or those with rare vocabulary (very common in Medline) as a random language. For each language pair we therefore filter out only sentences where LangDetect identifies the source sentence as belonging to the target language, and vice versa.

We then use a series of simple heuristics to further filter the parallel datasets, removing duplicate sentence pairs, those with source/target length ratio of $<$ 1:3.5 or $>$ 3.5:1, and sentences with $>$ 120 tokens. For the more aggressively-filtered 'no-title' experiments we additionally remove all lines containing multiple tokens in square brackets, which in medical writing are used to denote the English translation of a non-English article's title (Patrias and Wendling, 2007). This leaves 27.3K sentence pairs for en-de and 64.8K for en-es: about 96% of the filtered data in both cases.

2.2 Model hyperparameters and training

We use the Tensor2Tensor implementation of the Transformer model with the `transformer_big` setup for all NMT models (Vaswani et al., 2018). We use the same effective batch size of 4k tokens for both MLE and doc-MRT. Because of model size constraints and the need to sample multiple targets for doc-MRT, we achieve the 4k effective batch size by accumulating gradients (Saunders et al., 2018) over every 4 batches of 1k tokens for MLE and every 16 batches of 256 tokens for doc-MRT.

For doc-MRT we use sampling temperature $\tau = 0.3$, smoothing parameter $\alpha = 0.6$ and $N = 8$ samples per sentence, which gave the best results for our doc-MRT experiments in Saunders et al. (2020).

[1] https://ufal.mff.cuni.cz/ufal_medical_corpus

[2] Dušek et al. (2017)

[3] Neves et al. (2016)

[4] https://github.com/biomedical-translation-corpora/medline (Yepes et al., 2017)

[5] http://www.himl.eu/test-sets

[6] https://pypi.org/project/langdetect/

		de2en	en2de	es2en	en2es
1	Baseline	38.8	30.6	48.5	46.6
2	MLE fine-tuning from 1	40.9	32.5	48.5	46.0
3	Checkpoint averaging 2 (en-de) / 1 (en-es)	41.1	32.2	48.5	47.1
4	MRT from 1	40.0	31.1	**49.0**	47.4
5	MRT from 2 (en-de only)	**41.3**	32.9	-	-
6	Checkpoint averaging 5 (en-de) / 4 (en-es)	**41.3**	**33.0**	48.9	**47.7**

Table 3: Validation BLEU developing models used in English-German and English-Spanish language pair submissions. Scores for single checkpoints unless indicated. MLE fine-tuning did not improve over the en-es baselines, so we do not use these models to initialise MRT.

	de2en	en2de	es2en	en2es
MLE from baseline	41.1	32.2	-	-
MLE from baseline, no-title	41.4	31.8	-	-
MRT from: MLE (en-de) / baseline (en-es)	41.3	**33.0**	48.9	**47.7**
MRT no-title from: MLE no-title (en-de) / baseline (en-es)	**41.9**	32.6	**49.0**	47.2

Table 4: Validation BLEU developing models used in English-German and English-Spanish language pair submissions. Scores for averaged checkpoints. MLE fine-tuning with either dataset did not improve over the en-es baselines.

For each approach we fine-tune on a single GPU, saving checkpoints every 1K updates, until fine-tuning validation set BLEU fails to improve for 3 consecutive checkpoints. Generally this took about 5K updates. We then perform checkpoint averaging (Junczys-Dowmunt et al., 2016) over the final 3 checkpoints to obtain the final model.

2.3 Inference

For the 2020 submissions, we additionally split any test lines containing multiple sentences before inference using the Python NLTK package[7], translate the split sentences separately, then remerged. We found this gave noticeable improvements in quality for the few sentences it applied to. In all cases we decode with beam size 4 using SGNMT (Stahlberg et al., 2017). Test scores are as provided by the organizers for ”OK” sentences using Moses tokenization and the multi-eval tool. Validation scores are for case-insensitive, detokenized text obtained using SacreBLEU[8] (Post, 2018).

2.4 Results

We first assess the impact of small-domain adaptation to the full title-included Medline training set. Results in Table 3 show that small-domain MLE can lead to over-fitting and reduced performance (en-es) but also significant gains (en-de). Further fine-tuning with doc-MRT improved performance relative to the best MLE model for all translation directions by up to 0.8 BLEU when comparing with or without checkpoint averaging. While checkpoint averaging slightly decreased validation set performance for en2de MLE, we use it in all cases since it reduces sensitivity to randomness in training (Popel and Bojar, 2018).

In Table 4 we explore the impact of fine-tuning only on aggressively filtered ‘no-title’ data. This does noticeably improve performance for de2en, with a very small improvement for es2en. Since the added information in ‘title’ sentences is on the English side, this suggests that target training sentence quality impacts both MLE and MRT performance. However, removing these sentences entirely results in a noticeable performance decrease for the en2de and en2es models, demonstrating that they can be valuable training examples.

We submitted three runs to the WMT20 biomedical task for each language pair. For en-de run 1 was the baseline model fine-tuned on MLE with all data, while for en-es we submitted the checkpoint averaged baseline as MLE fine-tuning did not improve dev set performance. Run 2 was the run 1 model fine-tuned with doc-MRT on no-title data. Run 3 was the run 1 model fine-tuned with doc-MRT on all Medline abstract data. Table 5 gives scores for these submitted models.

Our best runs achieve the best and second-best results among all systems for en2es and es2en respectively as reported by the organizers. For en-de our test scores are further behind other systems, perhaps indicating that the baseline system could have been stronger before fine-grained adaptation.

[7] `https://pypi.org/project/nltk/` sentence splitter

[8] SacreBLEU signature: `BLEU+case.lc+numrefs.1+smooth.exp+tok.13a+version.1.2.11`

	de2en		en2de		es2en		en2es	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
MLE (all data) (en-de) / Baseline (en-es)	41.1	39.6	32.2	32.9	48.5	**46.6**	47.1	45.7
MRT (no-title data)	**41.9**	39.6	32.6	32.8	**49.0**	46.4	47.2	**46.7**
MRT (all data)	41.3	**39.8**	**33.0**	**33.2**	48.9	**46.6**	47.7	46.6

Table 5: Validation and test BLEU for models used in English-German and English-Spanish language pair submissions. Test results are for ”OK sentences” as scored by the organizers.

This is also indicated by the strong improvement of these models under simple MLE.

We submitted the MRT model on no-title data instead of the MLE on no-title data because MLE optimization did not improve over the baseline for en-es or en-es, with or without title lines, whereas MRT fine-tuning did. We also wanted to further examine whether MRT was robust enough to benefit from ‘noisy’ data like the title lines, or whether cleaner no-title training data was more useful. In fact both forms of doc-MRT performed similarly on the test data, except in the case of en2de, where ‘no-title’ MRT scored 0.4 BLEU worse – further confirmation that source sentences with more information than the gold target can benefit MRT. We note that a MRT run was the best run or tied best run in all cases.

For the test runs, we additionally experimented with simply removing square bracket tokens from source sentences, since these could act as ‘triggering’ tokens for title sentences. This did seem to improve translations for the sentences it applied to, but is clearly not applicable to all forms of exposure bias, since it requires knowledge of all behaviours that could trigger exposure bias. MRT does not require such knowledge, but still reduces the effects of exposure bias.

3 Conclusions

Our WMT20 Biomedical submission investigates improvements on the English-German and English-Spanish language pairs under a single strong model. In particular, we focus on the behaviour of models trained on sentences with some predictable irregularities. We find that aggressively filtering target sentences can help overall performance, but that aggressively filtering source sentence tends to hurt performance. We also find that Minimum Risk Training can benefit from imperfectly aligned training examples while reducing the effects of exposure bias.

Acknowledgments

This work was supported by EPSRC grants EP/M508007/1 and EP/N509620/1 and has been performed using resources provided by the Cambridge Tier-2 system operated by the University of Cambridge Research Computing Service[9] funded by EPSRC Tier-2 capital grant EP/P020259/1.

References

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares, Amy Siu, Karin Verspoor, and Maika Vicente Navarro. 2019. Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 29–53, Florence, Italy. Association for Computational Linguistics.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391.

Ondřej Dušek, Jan Hajič, Jaroslava Hlaváčová, Jindřich Libovický, Pavel Pecina, Aleš Tamchyna, and Zdeňka Urešová. 2017. Khresmoi summary translation test data 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

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.

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

M. Amin Farajian, Marco Turchi, Matteo Negri, and Marcello Federico. 2017. Multi-domain neural machine translation through unsupervised adaptation. In *Proceedings of the Second Conference on Machine Translation*, pages 127–137, Copenhagen, Denmark. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Rico Sennrich. 2016. The AMU-UEDIN submission to the WMT16 news translation task: Attention-based NMT models as feature functions in phrase-based SMT. In *Proceedings of the First Conference on Machine Translation*, pages 319–325, Berlin, Germany. Association for Computational Linguistics.

Sachith Sri Ram Kothur, Rebecca Knowles, and Philipp Koehn. 2018. Document-level adaptation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 64–73, Melbourne, Australia. Association for Computational Linguistics.

Xiaoqing Li, Jiajun Zhang, and Chengqing Zong. 2018. One sentence one model for neural machine translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Minh-Thang Luong and Christopher D Manning. 2015. Stanford Neural Machine Translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 76–79.

Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–318, Melbourne, Australia. Association for Computational Linguistics.

Graham Neubig. 2016. Lexicons and minimum risk training for neural machine translation: NAIST-CMU at WAT2016. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 119–125, Osaka, Japan. The COLING 2016 Organizing Committee.

Mariana L Neves, Antonio Jimeno-Yepes, and Aurélie Névéol. 2016. The ScieLO Corpus: a Parallel Corpus of Scientific Publications for Biomedicine. In *LREC*.

Karen Patrias and Dan Wendling. 2007. Citing medicine: the nlm style guide for authors, editors. *and publishers. Bethesda, MD: National Library of Medicine. Retrieved June*, 27:2011.

Martin Popel and Ondřej Bojar. 2018. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.

Matt Post. 2018. A call for clarity in reporting BLEU scores. *CoRR*, abs/1804.08771.

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

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

Danielle Saunders, Felix Stahlberg, and Bill Byrne. 2019. UCAM biomedical translation at WMT19: Transfer learning multi-domain ensembles. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 169–174, Florence, Italy. Association for Computational Linguistics.

Danielle Saunders, Felix Stahlberg, and Bill Byrne. 2020. Using context in neural machine translation training objectives. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7764–7770, Online. Association for Computational Linguistics.

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.

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, pages 1715–1725.

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.

Felix Stahlberg, Eva Hasler, Danielle Saunders, and Bill Byrne. 2017. SGNMT–A Flexible NMT Decoding Platform for Quick Prototyping of New Models and Search Strategies. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 25–30.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for Neural Machine Translation. *CoRR*, abs/1803.07416.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual*

Meeting of the Association for Computational Linguistics, pages 3544–3552, Online. Association for Computational Linguistics.

Jitao Xu, Josep Crego, and Jean Senellart. 2019. Lexical micro-adaptation for neural machine translation. In *International Workshop on Spoken Language Translation*.

Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Karin Verspoor, Ondrej Bojar, Arthur Boyer, Cristian Grozea, Barry Haddow, Madeleine Kittner, Yvonne Lichtblau, et al. 2017. Findings of the wmt 2017 biomedical translation shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 234–247.