# On Exposure Bias, Hallucination and Domain Shift in Neural Machine Translation

**Chaojun Wang**[1] **Rico Sennrich**[2,1]

[1]School of Informatics, University of Edinburgh
[2]Department of Computational Linguistics, University of Zurich
zippo_wang@foxmail.com, sennrich@cl.uzh.ch

## Abstract

The standard training algorithm in neural machine translation (NMT) suffers from exposure bias, and alternative algorithms have been proposed to mitigate this. However, the practical impact of exposure bias is under debate. In this paper, we link exposure bias to another well-known problem in NMT, namely the tendency to generate hallucinations under domain shift. In experiments on three datasets with multiple test domains, we show that exposure bias is partially to blame for hallucinations, and that training with Minimum Risk Training, which avoids exposure bias, can mitigate this. Our analysis explains why exposure bias is more problematic under domain shift, and also links exposure bias to the beam search problem, i.e. performance deterioration with increasing beam size. Our results provide a new justification for methods that reduce exposure bias: even if they do not increase performance on in-domain test sets, they can increase model robustness to domain shift.

## 1 Introduction

Neural Machine Translation (NMT) has advanced the state of the art in MT (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017), but is susceptible to domain shift. Koehn and Knowles (2017) consider out-of-domain translation one of the key challenges in NMT. Such translations may be fluent, but completely unrelated to the input (*hallucinations*), and their misleading nature makes them particularly problematic.

We hypothesise that *exposure bias* (Ranzato et al., 2016), a discrepancy between training and inference, makes this problem worse. Specifically, training with teacher forcing only exposes the model to gold history, while previous predictions during inference may be erroneous. Thus, the model trained with teacher forcing may over-rely on previously predicted words, which would exacerbate error propagation. Previous work has sought to reduce exposure bias in training (Bengio et al., 2015; Ranzato et al., 2016; Shen et al., 2016; Wiseman and Rush, 2016; Zhang et al., 2019). However, the relevance of error propagation is under debate: Wu et al. (2018) argue that its role is overstated in literature, and that linguistic features explain some of the accuracy drop at higher time steps.

Previous work has established a link between domain shift and hallucination in NMT (Koehn and Knowles, 2017; Müller et al., 2019). In this paper, we will aim to also establish an empirical link between hallucination and exposure bias. Such a link will deepen our understanding of the hallucination problem, but also has practical relevance, e.g. to help predicting in which settings the use of sequence-level objectives is likely to be helpful. We further empirically confirm the link between exposure bias and the 'beam search problem', i.e. the fact that translation quality does not increase consistently with beam size (Koehn and Knowles, 2017; Ott et al., 2018; Stahlberg and Byrne, 2019).

We base our experiments on German→English IWSLT'14, and two datasets used to investigate domain robustness by Müller et al. (2019): a selection of corpora from OPUS (Lison and Tiedemann, 2016) for German→English, and a low-resource German→Romansh scenario. We experiment with Minimum Risk Training (MRT) (Och, 2003; Shen et al., 2016), a training objective which inherently avoids exposure bias.

Our experiments show that MRT indeed improves quality more in out-of-domain settings, and reduces the amount of hallucination. Our analysis of translation uncertainty also shows how the MLE baseline over-estimates the probability of random translations at all but the initial time steps, and how MRT mitigates this problem. Finally, we show that the beam search problem is reduced by MRT.

*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

## 2 Minimum Risk Training

The de-facto standard training objective in NMT is to minimize the negative log-likelihood $\mathcal{L}(\boldsymbol{\theta})$ of the training data $D$[1]:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{(\mathbf{x},\mathbf{y}) \in D} \sum_{t=1}^{|\mathbf{y}|} -\log P\left(\mathbf{y}_t | \mathbf{x}, \mathbf{y}_{<t}; \boldsymbol{\theta}\right) \quad (1)$$

where $\mathbf{x}$ and $\mathbf{y}$ are the source and target sequence, respectively, $\mathbf{y}_t$ is the $\mathrm{t}^{\mathrm{th}}$ token in $\mathbf{y}$, and $\mathbf{y}_{<t}$ denotes all previous tokens. MLE is typically performed with teacher forcing, where $\mathbf{y}_{<t}$ are ground-truth labels in training, which creates a mismatch to inference, where $\mathbf{y}_{<t}$ are model predictions.

Minimum Risk Training (MRT) is a sequence-level objective that avoids this problem. Specifically, the objective function of MRT is the expected loss (*risk*) with respect to the posterior distribution:

$$\mathcal{R}(\boldsymbol{\theta}) = \sum_{(\mathbf{x},\mathbf{y}) \in D} \sum_{\tilde{\mathbf{y}} \in \mathcal{Y}(\mathbf{x})} P\left(\tilde{\mathbf{y}} | \mathbf{x}; \boldsymbol{\theta}\right) \Delta\left(\tilde{\mathbf{y}}, \mathbf{y}\right) \quad (2)$$

in which the loss $\Delta\left(\tilde{\mathbf{y}}, \mathbf{y}\right)$ indicates the discrepancy between the gold translation $\mathbf{y}$ and the model prediction $\tilde{\mathbf{y}}$. Due to the intractable search space, the posterior distribution $\mathcal{Y}(\mathbf{x})$ is approximated by a subspace $\mathcal{S}(\mathbf{x})$ by sampling a certain number of candidate translations, and normalizing:

$$\tilde{P}\left(\tilde{\mathbf{y}} | \mathbf{x}; \boldsymbol{\theta}, \alpha\right) = \frac{P\left(\tilde{\mathbf{y}} | \mathbf{x}; \boldsymbol{\theta}\right)^{\alpha}}{\sum_{\mathbf{y}' \in \mathcal{S}(\mathbf{x})} P\left(\mathbf{y}' | \mathbf{x}; \boldsymbol{\theta}\right)^{\alpha}} \quad (3)$$

where $\alpha$ is a hyperparameter to control the sharpness of the subspace. Based on preliminary results, we use random sampling to generate candidate translations, and following Edunov et al. (2018), do not add the reference translation to the subspace.

## 3 Experiments

### 3.1 Data

To verify the effectiveness of our MRT implementation on top of a strong Transformer baseline (Vaswani et al., 2017), we first conduct experiments on IWSLT'14 German→English (DE→EN) (Cettolo et al., 2014), which consists of 180 000 sentence pairs. We follow previous work for data splits (Ranzato et al., 2016; Edunov et al., 2018).

For experiments with domain shift, we use data sets and preprocessing as Müller et al. (2019)[2]. For DE→EN, data comes from OPUS (Lison and Tiedemann, 2016), and is comprised of five domains: *medical*, *IT*, *law*, *koran* and *subtitles*. We use *medical* for training and development, and report results on an in-domain test set and the four other domains (out-of-domain; OOD). German→Romansh (DE→RM) is a low-resource language pair where robustness to domain shift is of practical relevance. The training data is from the Allegra corpus (Scherrer and Cartoni, 2012) (*law* domain) with 100 000 sentence pairs. The test domain are *blogs*, using data from Convivenza[3]. We have access to 2000 sentences for development and testing, respectively, in each domain.

We tokenise and truecase data sets with Moses (Koehn et al., 2007), and use shared BPE with 32 000 units (Sennrich et al., 2016).

### 3.2 Model

We implement[4] MRT in the Nematus toolkit (Sennrich et al., 2017). All our experiments use the Transformer architecture (Vaswani et al., 2017). Following Edunov et al. (2018), we use 1-BLEU$_{\text{smooth}}$ (Lin and Och, 2004) as the MRT loss. Models are pre-trained with the token-level objective MLE and then fine-tuned with MRT. Hyperparameters mostly follow previous work (Edunov et al., 2018; Müller et al., 2019); for MRT, we conduct limited hyperparameter search on the IWSLT'14 development set, including learning rate, batch size, and the sharpness parameter $\alpha$. We set the number of candidate translations for MRT to 4 to balance effectiveness and efficiency. Detailed hyperparameters are reported in the Appendix.

### 3.3 Evaluation

For comparison to previous work, we report lowercased, tokenised BLEU (Papineni et al., 2002) with *multi-bleu.perl* for IWSLT'14, and cased, detokenised BLEU with SacreBLEU (Post, 2018)[5] otherwise. For settings with domain shift, we report average and standard deviation of 3 independent training runs to account for optimizer instability.

The manual evaluation was performed by two native speakers of German who completed bilin-

[1] This is equivalent to maximizing the likelihood of the data, hence *Maximum Likelihood Estimation* (MLE).

[2] https://github.com/ZurichNLP/domain-robustness

[3] https://www.suedostschweiz.ch/blogs/convivenza

[4] Code available at https://github.com/zippotju/Exposure-Bias-Hallucination-Domain-Shift

[5] Signature: BLEU+c.mixed+#.1+s.exp+tok.13a+v.1.4.2

|  | **inter-annotator** |  |  | **intra-annotator** |  |  |
|---|---|---|---|---|---|---|
| annotation | $P(A)$ | $P(E)$ | $K$ | $P(A)$ | $P(E)$ | $K$ |
| fluency | 0.66 | 0.38 | 0.44 | 0.87 | 0.42 | 0.77 |
| adequacy | 0.82 | 0.61 | 0.54 | 0.93 | 0.66 | 0.79 |

Table 1: Inter-annotator (N=250) and intra-annotator agreement (N=617) of manual evaluation.

| system | BLEU |
|---|---|
| ConvS2S (MLE) (Edunov et al., 2018) | 32.2 |
| ConvS2S (MRT) (Edunov et al., 2018) | 32.8 (+**0.6**) |
| Transformer (MLE) (Wu et al., 2019) | 34.4 |
| DynamicConv (MLE) (Wu et al., 2019) | 35.2 |
| MLE | 34.7 |
| MRT | 35.2 (+**0.5**) |

Table 2: Results for IWSLT'14 DE→EN with MLE and MRT (in brackets, improvement over MLE).

gual (German/English) high school or University programs. We collected ~3600 annotations in total, spread over 12 configurations. We ask annotators to evaluate translations according to fluency and adequacy. For fluency, the annotator classifies a translation as fluent, partially fluent or not fluent; for adequacy, as adequate, partially adequate or inadequate. We report kappa coefficient ($K$) (Carletta, 1996) for inter-annotator and intra-annotator agreement in Table 1, and assess statistical significance with Fisher's exact test (two-tailed).

### 3.4 Results

Table 2 shows results for IWSLT'14. We compare to results by Edunov et al. (2018), who use a convolutional architecture (Gehring et al., 2017), and Wu et al. (2019), who report results with Transfomer-base and dynamic convolution.

With 34.7 BLEU, our baseline is competitive. We observe an improvement of 0.5 BLEU from MRT, comparable to Edunov et al. (2018), although we start from a stronger baseline (+2.5 BLEU).

Table 3 shows results for data sets with domain shift. To explore the effect of label smoothing (Szegedy et al., 2016), we train baselines with and without label smoothing. MLE with label smoothing performs better by itself, and we also found MRT to be more effective on top of the initial model with label smoothing. For DE→EN, MRT increases average OOD BLEU by 0.8 compared to the MLE baseline with label smoothing; for DE→RM the improvement is 0.7 BLEU. We note that MRT does not consistently improve in-domain performance, which is a first indicator that exposure bias may be more problematic under domain shift.

Our OOD results lag slightly behind those of Müller et al. (2019), but note that the techniques employed by them, namely reconstruction (Tu et al., 2017; Niu et al., 2019), subword regularization (Kudo, 2018), and noisy channel modelling (Li and Jurafsky, 2016) are orthogonal to MRT. We leave the combination of these approaches to future work.

## 4 Analysis

BLEU results indicate that MRT can improve domain robustness. In this section, we report on additional experiments to establish more direct links between exposure bias and domain robustness, hallucination, and the beam search problem. Experiments are performed on DE→EN OPUS data.

### 4.1 Hallucination

We manually evaluate the proportion of hallucinated translations on out-of-domain and in-domain test sets. We follow the definition and evaluation by Müller et al. (2019), considering a translation a hallucination if it is **(partially) fluent**, but unrelated in content to the source text (**inadequate**). We report the proportion of such hallucinations for each system.

Results in Table 4 confirm that hallucinations are much more pronounced in out-of-domain test sets (33–35%) than in in-domain test sets (1–2%). MRT reduces the proportion of hallucinations on out-of-domain test sets (N=500 for each system; reductions statistically significant at $p < 0.05$) and improves BLEU. Note that the two metrics do not correlate perfectly: MLE with label smoothing has higher BLEU (+1) than MRT based on MLE without label smoothing, but a similar proportion of hallucinations. This indicates that label smoothing increases translation quality in other aspects, while MRT has a clear effect on the number of hallucinations, reducing it by up to 21% (relative).

A closer inspection of segments where the MLE system was found to hallucinate shows that some segments were scored higher in adequacy with MRT, others lower in fluency. One example for each case is shown in Table 5. Even the example where MRT was considered disfluent and inadequate actually shows an attempt to cover the source sentence: the source word 'Ableugner' (denier) is

|  | DE→EN |  | DE→RM |  |
|---|---|---|---|---|
| system | in-domain | average OOD | in-domain | average OOD |
| SMT (Müller et al., 2019) | 58.4 | 11.8 | 45.2 | 15.5 |
| NMT (Müller et al., 2019) | 61.5 | 11.7 | 52.5 | 18.9 |
| NMT+RC+SR+NC (Müller et al., 2019) | 60.8 | 13.1 | 52.4 | 20.7 |
| MLE w/o LS | 58.3 (±0.53) | 9.7 (±0.25) | 52.2 (±0.19) | 15.8 (±0.39) |
| +MRT | 58.4 (±0.39) | 10.2 (±0.26) | 52.1 (±0.08) | 15.9 (±0.28) |
| MLE w/ LS | 58.9 (±0.45) | 11.2 (±0.16) | 53.9 (±0.16) | 18.0 (±0.17) |
| +MRT | 58.8 (±0.36) | 12.0 (±0.29) | 53.9 (±0.12) | 18.7 (±0.09) |

Table 3: Average BLEU and standard deviation on in-domain and out-of-domain test sets for models trained on OPUS (DE→EN) and Allegra (DE→RM). RC: reconstruction; SR: subword regularization, NC: noisy channel.

|  | % hallucinations (BLEU) |  |
|---|---|---|
| system | out-of-domain | in-domain |
| MLE w/o LS | 35% (9.7) | 2% (58.3) |
| +MRT | 29% (10.2) | - |
| MLE w/ LS | 33% (11.2) | 1% (58.9) |
| +MRT | 26% (12.0) | - |

Table 4: Proportion of hallucinations and BLEU on out-of-domain and in-domain test sets. DE→EN OPUS.

|  |  |
|---|---|
| source | **Wir haben** ihn **gefunden**. |
| reference | **We found** him. |
| MLE | Do not pass it. |
| MRT | **We have found** it. |
| source | So höre **nicht** auf die **Ableugner**. |
| reference | So hearken **not** to those who **deny** (the Truth). |
| MLE | **Do not** drive or use machines. |
| MRT | **Do not** apply to **dleugner**. |

Table 5: Out-of-domain translation examples. MLE hallucinates in both examples; MRT was rated more adequate in top example, less fluent in bottom one.

mistranslated into ‘dleugner’. We consider this preferable to producing a complete hallucination.

### 4.2 Uncertainty Analysis

Inspired by Ott et al. (2018), we analyse the model’s uncertainty by computing the average probability at each time step across a set of sentences. Besides the reference translations, we also consider a set of ‘distractor’ translations, which are random sentences from the in-domain test set which match the corresponding reference translation in length.

In Figure 1, we show out-of-domain results for an MLE model and multiple checkpoints of MRT fine-tuning. The left two graphs show probabilities for references and distractors, respectively. The right-most graph shows a direct comparison of probabilities for references and distractors for the MLE baseline and the final MRT model. The MLE baseline assigns similar probabilities to tokens in the references and the distractors. Only for the first time steps is there a clear preference for the references over the (mostly random!) distractors. This shows that error propagation is a big risk: should the model make a wrong prediction initially, this is unlikely to be penalised in later time steps.

MRT tends to increase the model’s certainty at later time steps[6], but importantly, the increase is sharper for the reference translations than for the distractors. The direct comparison shows a widening gap in certainty between the reference and distractor sentences.[7] In other words, producing a hallucination will incur a small penalty at each time step (compared to producing the reference), presumably due to a higher reliance on the source signal, lessening the risk of error propagation and hallucinations.

Our analysis shows similar trends on in-domain references. However, much higher probabilities are assigned to the first few tokens of the references than to the distractors. Hence, it is much less likely that a hallucination is kept in the beam, or will overtake a good translation in overall probability, reducing the practical impact of the model’s over-reliance on its history.[8]

### 4.3 Beam Size Analysis

Figure 1 shows that with MLE, distractor sentences are assigned lower probabilities than the references at the first few time steps, but are assigned similar, potentially even higher probabilities at later time steps. This establishes a connection between exposure bias and the beam search problem, i.e. the problem that increasing the search space can lead

[6] The uncertainty of the baseline is due to label smoothing.

[7] For intermediate checkpoints, see Appendix, Figure 2.

[8] Figures are shown in the Appendix (Figure 3).

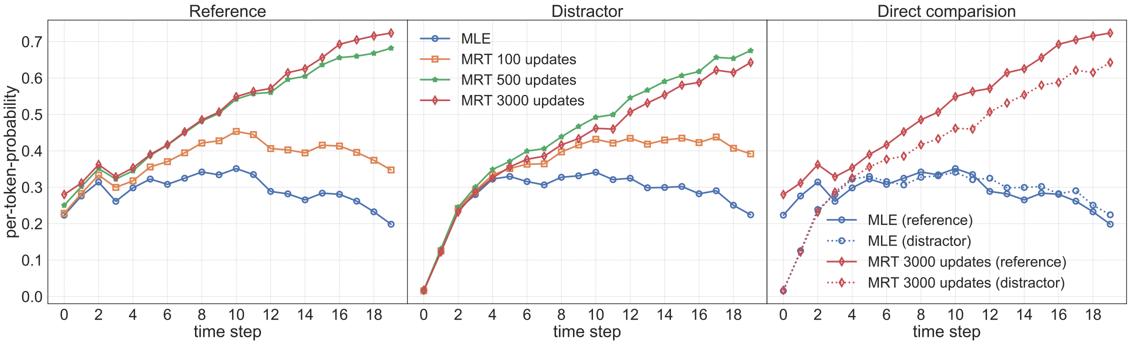

Figure 1: Per-token probability of out-of-domain reference translations and in-domain distractors (first two graphs share legend). Rightmost plot shows direct comparison for MLE baseline and final MRT model. DE→EN OPUS .

to worse model performance.[9] With larger beam size, it is more likely that hallucinations survive pruning at the first few time steps, and with high probabilities assigned to them at later time steps, there is a chance that they become the top-scoring translation.

We investigate whether the beam search problem is mitigated by MRT. In Table 6, we report OOD BLEU and the proportion of hallucinations with beam sizes of 1, 4 and 50. While MRT does not eliminate the beam search problem, performance drops less steeply as beam size increases. With beam size 4, our MRT models outperform the MLE baseline by 0.5-0.8 BLEU; with beam size 50, this difference grows to 0.6-1.5 BLEU. Our manual evaluation (N=200 for each system for beam size 1 and 50) shows that the proportion of hallucinations increases with beam size, and that MRT consistently reduces the proportion by 11-21% (relative). For the system with label smoothing, the relative increase in hallucinations with increasing beam size is also smaller with MRT (+33%) than with MLE (+44%).

| | BLEU (% hallucinations) | | |
|---|---|---|---|
| system | $k = 1$ | $k = 4$ | $k = 50$ |
| MLE w/o LS | 8.9 (28%) | 9.7 (35%) | 9.3 (37%) |
| +MRT | 9.1 (24%) | 10.2 (29%) | 9.9 (33%) |
| MLE w/ LS | 10.6 (27%) | 11.2 (33%) | 9.4 (39%) |
| +MRT | 11.3 (24%) | 12.0 (26%) | 10.9 (32%) |

Table 6: Average OOD BLEU and proportion of hallucinations with different beam sizes $k$. DE→EN OPUS.

## 5 Conclusions

Our results and analysis show a connection between the exposure bias due to MLE training with teacher forcing and several well-known problems in neural machine translation, namely poor performance under domain shift, hallucinated translations, and deteriorating performance with increasing beam size. We find that Minimum Risk Training, which does not suffer from exposure bias, can be useful even when it does not increase performance on an in-domain test set: it increases performance under domain shift, reduces the number of hallucinations substantially, and makes beam search with large beams more stable.

Our findings are pertinent to the academic debate how big of a problem exposure bias is in practice – we find that this can vary substantially depending on the dataset –, and they provide a new justification for sequence-level training objectives that reduce or eliminate exposure bias. Furthermore, we believe that a better understanding of the links between exposure bias and well-known translation problems will help practitioners decide when sequence-level training objectives are especially promising, for example in settings where the test domain is unknown, or where hallucinations are a common problem.

## Acknowledgments

Chaojun Wang was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) fellowship grant EP/S001271/1 (MTStretch). Rico Sennrich acknowledges support of the Swiss National Science Foundation (MUTAMUR; no. 176727). This project has received support from Samsung Electronics Polska sp. z o.o. - Samsung R&D Institute Poland.

[9] The beam search problem has previously been linked to length bias (Yang et al., 2018; Murray and Chiang, 2018) and the copy mode (Ott et al., 2018). We consider hallucinations another result of using large search spaces with MLE models.

## References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1171–1179. Curran Associates, Inc.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th iwslt evaluation campaign, iwslt 2014. In *Proceedings of the International Workshop on Spoken Language Translation, Hanoi, Vietnam*, page 57.

Sergey Edunov, Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 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)*, pages 355–364, New Orleans, Louisiana. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 1243–1252. JMLR.org.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

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. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Jiwei Li and Dan Jurafsky. 2016. Mutual information and diverse decoding improve neural machine translation.

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.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Mathias Müller, Annette Rios, and Rico Sennrich. 2019. Domain robustness in neural machine translation.

Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223, Belgium, Brussels. Association for Computational Linguistics.

Xing Niu, Weijia Xu, and Marine Carpuat. 2019. Bidirectional differentiable input reconstruction for low-resource 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 442–448, Minneapolis, Minnesota. Association for Computational Linguistics.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan. Association for Computational Linguistics.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. 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 *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Yves Scherrer and Bruno Cartoni. 2012. The trilingual ALLEGRA corpus: Presentation and possible use for lexicon induction. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2890–2896, Istanbul, Turkey. European Languages Resources Association (ELRA).

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nădejde. 2017. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain. 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.

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.

Felix Stahlberg and Bill Byrne. 2019. On NMT search errors and model errors: Cat got your tongue? 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 3347–3353, Hong Kong, China. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

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.

Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. 2017. Neural machine translation with reconstruction. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pages 3097–3103. AAAI Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz 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.

Sam Wiseman and Alexander M. Rush. 2016. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1296–1306, Austin, Texas. Association for Computational Linguistics.

Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. In *International Conference on Learning Representations*.

Lijun Wu, Xu Tan, Di He, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. Beyond error propagation in neural machine translation: Characteristics of language also matter. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3602–3611, Brussels, Belgium. Association for Computational Linguistics.

Yilin Yang, Liang Huang, and Mingbo Ma. 2018. Breaking the beam search curse: A study of (re-)scoring methods and stopping criteria for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3054–3059, Brussels, Belgium. Association for Computational Linguistics.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. 2019. Bridging the gap between training and inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4343, Florence, Italy. Association for Computational Linguistics.

# A Appendix

| | IWSLT | OPUS/Allegra |
|---|---|---|
| **General hyperparameters** | | |
| embedding layer size | 512 | |
| hidden state size | 512 | |
| tie encoder decoder embeddings | yes | |
| tie decoder embeddings | yes | |
| loss function | per-token-cross-entropy (MRT) | |
| label smoothing | 0.1 | |
| optimizer | adam | |
| learning schedule | transformer (constant) | |
| warmup steps | 4000 | 6000 |
| gradient clipping threshold | 1 | 0 |
| maximum sequence length | 100 | |
| token batch size | 4096 | |
| length normalization alpha | 0.6 | 1 |
| encoder depth | 6 | |
| decoder depth | 6 | |
| feed forward num hidden | 1024 | 2048 |
| number of attention heads | 4 | 8 |
| embedding dropout | 0.3 | 0.1 |
| residual dropout | 0.3 | 0.1 |
| relu dropout | 0.3 | 0.1 |
| attention weights dropout | 0.3 | 0.1 |
| beam size | 4 | |
| | beam search sampling | random sampling |
| **MRT-revelant hyperparameters** | | |
| learning rate | 0.00003 | 0.00001 |
| batch size | 8192 (tokens) | 10 (sentences) |
| sharpness alpha | 0.005 | 0.005 |

Table 7: Configurations of NMT systems used to pre-train and fine-tune over three datasets. Note in general hyperparameters, the items in brackets denote the options that will be used in MRT fine-tuning.

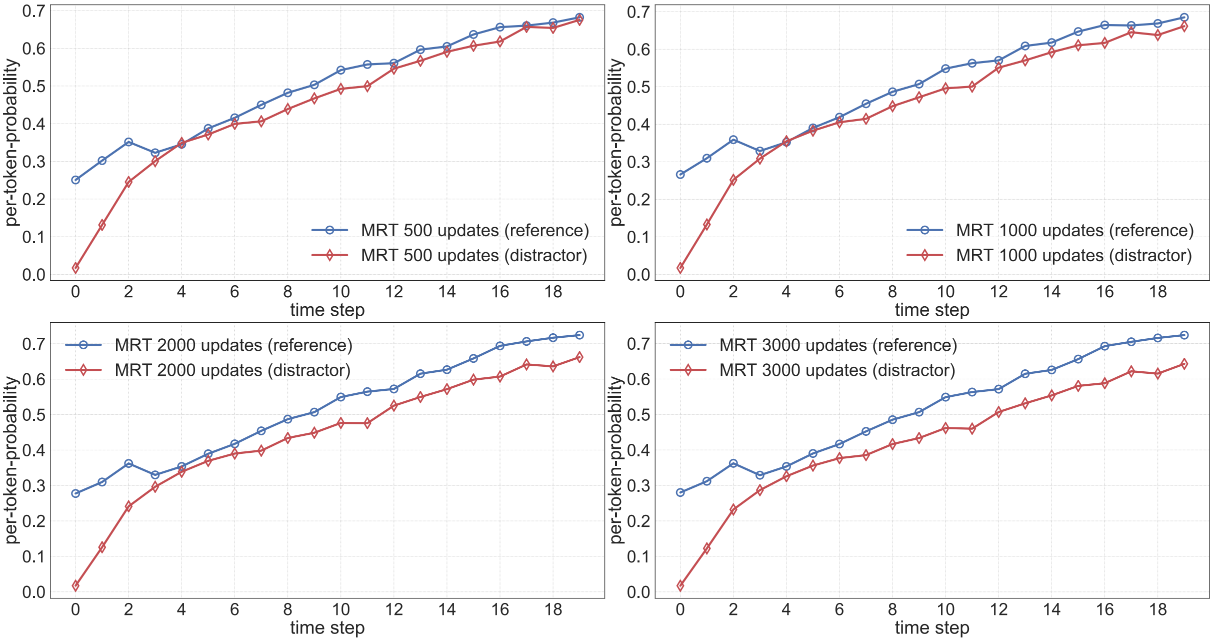

Figure 2: Per-token probability of **out-of-domain** reference translations and in-domain distractors for different checkpoints in MRT training, showing a widening gap between references and distractors. DE→EN OPUS.

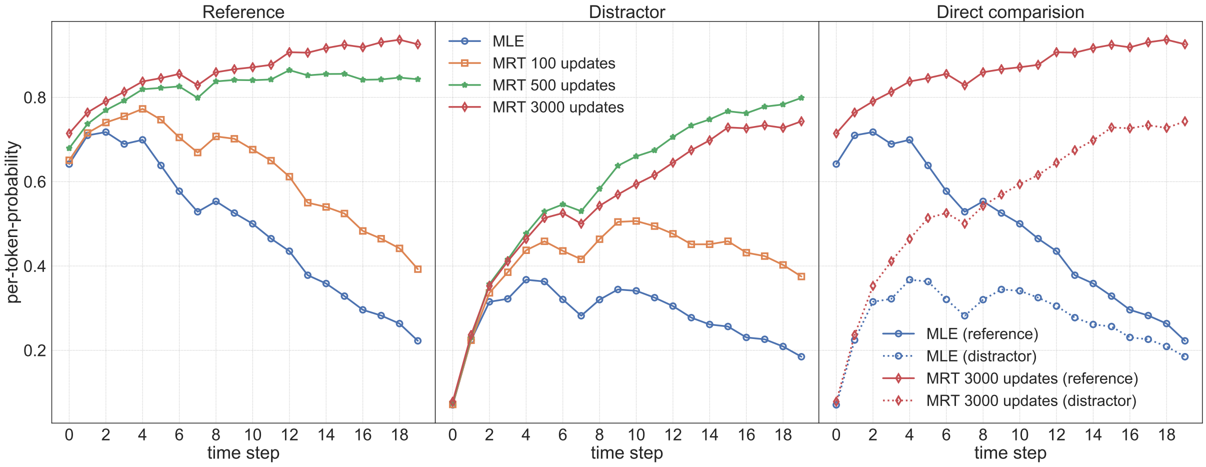

Figure 3: Per-token probability of **in-domain** reference translations and distractors. Rightmost plot shows direct comparison for MLE baseline and final MRT model. DE→EN OPUS.