The Highs and Lows of Simple Lexical Domain Adaptation Approaches for Neural Machine Translation

Nikolay Bogoychev[*]   **Pinzhen Chen**[*]
School of Informatics, University of Edinburgh
{n.bogoych, pinzhen.chen}@ed.ac.uk

Abstract

Machine translation systems are vulnerable to domain mismatch, especially in a low-resource scenario. Out-of-domain translations are often of poor quality and prone to hallucinations, due to exposure bias and the decoder acting as a language model. We adopt two approaches to alleviate this problem: lexical shortlisting restricted by IBM statistical alignments, and hypothesis re-ranking based on similarity. The methods are computationally cheap, widely known, but not extensively experimented on domain adaptation. We demonstrate success on low-resource out-of-domain test sets, however, the methods are ineffective when there is sufficient data or too great domain mismatch. This is due to both the IBM model losing its advantage over the implicitly learned neural alignment, and issues with subword segmentation of out-of-domain words.

1 Introduction

Neural Machine translation (NMT) has achieved state-of-the-art performance in a variety of language pairs and settings (Bahdanau et al., 2015; Vaswani et al., 2017), but it is vulnerable to domain mismatch, where the test set differs significantly from the training data in terms of vocabulary, genre, length, etc. This issue is exacerbated in a low-resource condition (Koehn and Knowles, 2017).

Teacher forcing is used during traditional maximum likelihood neural network training, leading to a strong exposure bias, and model confusion when presented with unexpected sequences. This typically results in hallucinations in the output (Müller et al., 2020), because the overly zealous language model component prefers a fluent translation, as opposed to an adequate one. A number of methods have been proposed in order to tackle the issue: exposing the model to its predictions during training (Ranzato et al., 2015; Shen et al., 2016; Zhang et al.,

---

* Equal contribution.

2019; Wang and Sennrich, 2020); tuning directly towards BLEU (Wiseman and Rush, 2016) or using minimum Bayes risk decoding (Kumar and Byrne, 2004; Stahlberg et al., 2017). A common weakness of such methods is that they are computationally expensive.

In this paper, we adopt and experiment with two approaches inspired by previous research. We use lexical shortlisting to interpolate a statistical alignment model with NMT; on top of it, we perform n-best list re-ranking by hypothesis agreement. Our aim is to constrain the lexical choice of the decoder, to prevent hallucinations from being generated. The methods are computationally simpler, as they require no change to the model or training. We analyse the effectiveness of these methods in different scenarios of domain adaptation. We show BLEU gains on a variety of out-of-domain datasets in a low-resource English-German setting. However, the methods show no improvements once the datasets are large, or the domains are too distant.

2 Methodology

We purse and analyse two separate strategies for improving neural machine translation's performance in low-resource domain mismatched settings: lexical shortlisting and n-best list re-ranking based on inter hypothesis agreement.

2.1 Lexical shortlisting

Neural machine translation systems have a vocabulary size of tens of thousands, but for every single translation, most of the vocabulary items are improbable choices. Many researchers attempt to improve on this to speed up the computation in the output layer (Schwenk et al., 2007; Le et al., 2012; Devlin et al., 2014). This is done by preparing a list of likely word-level translations for each sentence (commonly known as lexical shortlist), by using IBM alignment models such as fast-align (Dyer et al., 2013) and limiting the output layer choices

to it. This has been implemented in NMT frameworks for efficient systems (Junczys-Dowmunt et al., 2018b) and since widely used (Bogoychev et al., 2020, inter alia).

Recently, Li et al. (2019) showed that even in a high-resource scenario, the quality of the IBM alignments can outperform those learned by a neural model. Therefore, we find it sensible to incorporate word alignments into neural models, especially in a low-resource situation. Although shortlists have a negligible impact on the quality of strong neural systems, we can successfully limit the output layer to the likely tokens according to an IBM model trained on a particular domain, and improve out-of-domain BLEU scores in a low-resource setting. This suggests that the IBM model provides complementary information to the neural model in this scenario.

2.2 Hypothesis re-ranking

Fomicheva et al. (2020) estimate the quality of machine translation by measuring the agreement of the generated hypotheses for a given source sentence. The reason is that (higher) similarity between hypotheses reflects (higher) model confidence. In our problem, we assume that a hallucinated hypothesis will have a low agreement with the rest in the beam. Thus, we propose to re-rank the n-best list based on inter-hypothesis similarity and select the top one as the final translation.

For every source sentence, NMT generates b hypotheses where b is the beam size. For each hypothesis, we measure its similarity against others in the beam, then sort all hypotheses by their final aggregated similarity scores. The similarity score $score_i$ for a certain hypothesis hyp_i is calculated as Equation 1, where the similarity between hyp_i and every other hyp_j produced by the model is measured by an automatic metric $similarity$:

$$\text{score}_i = \sum_{j=0, i \neq j}^{b} \text{similarity}(\text{hyp}_i, \text{hyp}_j) \quad (1)$$

3 Experimental setup

For our experiments, we use OPUS English-German data (Lison and Tiedemann, 2016), with preprocessing performed as per the work of Müller et al. (2020). We follow the same data split as Wang and Sennrich (2020)'s low-resource, domain restricted setting: a model is trained on 1M sentences pairs from the *medical* domain, and is evaluated on *medical* (in-domain), *Koran*, *IT*, *subtitles* and *law* (out-of-domain). We use a joint vocabulary byte-pair encoding (BPE, Sennrich et al., 2016) trained on the *medical* domain with 32k merge operations. In this way, the vocabulary trained on the *medical* domain may be sub-optimal for the out-of-domain test sets. In reality, monolingual data from unknown domains might be available, which can reduce the bias of the BPE training towards the in-domain data. In order to simulate this scenario, we perform an analogous experiment where the BPE vocabulary is trained on all domains, except the *subtitles* corpus, as it is much larger than other domains and would dominate the vocabulary (reasons elaborated in Section 5).

For training, we used the Transformer-base preset of the Marian toolkit (Junczys-Dowmunt et al., 2018a) with transformer preprocessing normalisation as opposed to postprocessing normalisation, and additional attention and feed-forward layer dropout.

We performed ample hyperparameter search, to ensure that our neural network configuration achieves the best possible performance on this low-resource task. Decoding is always done with beam size 6 and length normalisation 0.6. On top of the baseline, we tried three different combinations of methods introduced in Section 2:

- A **shortlisting** configuration where we use a lexical shortlist generated by the *fast-align* model trained on the *medical* dataset. We try various configurations, and we settle on an optimal value of limiting the output layer to the 10 most probable unigram translations according to the IBM model.

- A **re-ranking** setup where we re-rank the n-best translation based on inter-hypothesis similarity. Like Fomicheva et al., we tested out a few different metrics to measure similarity: sentBLEU, ChrF, TER and METEOR. In an initial experiment on the *medical* domain, we found sentBLEU to have the best performance and stick to it for all experiments. We did not pick any neural metrics, as they may have a domain preference and are less interpretable.

- **Both**: shortlisting followed by re-ranking as specified above

Domain	BPE trained on *medical* only				BPE trained on all except *subtitles*			
	baseline	shortlist	re-rank	both	baseline	shortlist	re-rank	both
medical	*60.0*	*59.5*	***60.3***	*59.1*	***61.4***	*58.2*	*57.6*	*60.4*
Koran	0.9	1.0	0.7	**1.1**	0.8	0.9	0.9	**1.0**
law	19.6	**20.6**	16.6	17.8	17.8	19.3	19.8	**20.8**
IT	15.0	**16.3**	10.1	11.5	15.7	**18.0**	15.3	17.8
subtitles	2.8	**3.1**	1.4	1.9	2.6	**2.8**	2.4	**2.8**

Table 1: BLEU results on German-English systems trained on the *medical* domain, and tested on in-domain and out-domain datasets. In-domain results are in italics and the best BLEU on each domain dataset are in bold.

Domain	System	1- to 4-gram precisions				Brevity penalty	BLEU ($\triangle$)	METEOR ($\triangle$)
law	baseline	53.0	27.5	16.9	11.0	0.778	17.8	0.36
	shortlist	**56.1**	**29.4**	**17.9**	**11.4**	0.804	19.3 (+1.5)	**0.39 (+0.03)**
	re-rank	51.4	26.4	16.1	10.5	0.906	19.8 (+2.0)	0.31 (−0.05)
	both	53.1	27.6	16.7	10.7	**0.919**	**20.8 (+3.0)**	0.35 (−0.01)
IT	baseline	34.6	18.8	13.1	9.5	0.930	15.7	0.16
	shortlist	**43.9**	**24.7**	**17.1**	**12.1**	0.828	**18.0 (+2.3)**	**0.18 (+0.02)**
	re-rank	33.5	17.2	11.7	8.1	**1.000**	15.3 (−0.4)	0.12 (−0.04)
	both	38.0	20.1	13.6	9.7	**1.000**	17.8 (+2.1)	0.09 (−0.07)

Table 2: Breakdown of BLEU scores for "BPE on all" experiments in Table 1 for *law* and *IT* domains.

4 Results and Analysis

We present in Table 1 our models' BLEU scores on in- and out-of-domain test sets, with both BPE segmentation schemes. When applying the more restrictive BPE trained on the *medical* domain, our shortlisting mechanism always yields a slight increase in BLEU on out-of-domain sets and a small drop on the in-domain set. The *law* and *IT* domains benefit the most, whereas *subtitles* and *Koran* are largely unaffected. Re-ranking is not helpful in this BPE setting.

When using the alternative BPE segmentation trained on all datasets, we see that the baseline scores are generally lower for the out-of-domain datasets, despite a stronger in-domain BLEU. This is potentially because the rare words from out-of-domain datasets get insufficient exposure during training (on *medical* only). However, when using a lexical shortlist in this setting, we see greater improvements in terms of BLEU on the *law* and *IT* domains compared to the *medical*-only BPE scenario. Re-ranking is also much more effective, performing similarly to shortlisting. The combination of re-ranking and shortlisting delivers the best BLEU scores in nearly all out-of-domain splits.

We see that the shortlisting method is always superior to the baseline method on all out-of-domain datasets, although the results vary with the data preprocessing. This is true even more so of re-ranking, which is much better when the BPE vocabulary is learned on all domains. Combining both shortlisting and re-ranking always brings in a slight improvement over just re-ranking on the out-of-domain datasets, but its effectiveness is again preprocessing-dependent.

This clearly shows that the IBM model implemented by fast-align can learn information complementary to NMT, and interpolating them is beneficial for achieving higher BLEU scores in out-of-domain settings. The lower BLEU scores on in-domain test sets are related to the aggressive shortlist we used: by increasing the output layer limit from 10 to 50 most probable tokens, we achieve identical BLEU scores as the baseline.

4.1 BLEU breakdown

According to BLEU scores reported earlier, shortlisting and re-ranking are beneficial to NMT domain adaptation (in a relatively low-resource condition). We try to understand what contributes to the increase in BLEU scores by breaking down BLEU scores into n-gram precisions and length (brevity) penalty. In Table 2 we list the numbers for *law* and *IT* domains, under the "BPE trained on all" setting in Table 1, on which we have seen the largest leap

Domain	law	medical	subtitles†	IT	Koran
Number of sentences	695k	1M	1M	372k	529k
Avg. original sentence length	22.1	12.5	8.0	7.5	20.4
Avg. BPE sentence length	30.4	14.3	11.1	12.7	24.1
Vocab size, appearing >20 times	34k	36k	30k	15k	20k
Vocab overlap with *medical*	11.5k	36k	9.0k	5.8k	5.1k

† The *subtitles* corpus was sampled down from 20M to 1M sentence pairs.

Table 3: Corpus statistics for the different domains.

German	English
sein Pilot hat nicht die volle Kontrolle .	its p@@ il@@ ot is@@ n't in control .
und Z@@ eth@@ rid ? nur einen Strei@@ f@@ sch@@ uss .	and , Z@@ eth@@ rid , just gr@@ aze it .

Table 4: Two random German-English sentence pairs from the *subtitles* dataset after BPE.

of BLEU scores. Additionally, we include ME-TEOR which focuses on n-gram overlap and is not influenced by the output length.

From the table, shortlisting always significantly boosts n-gram precisions, whereas re-ranking alone decreases them. On the other hand, re-ranking "rectifies" the output length, leading to a better brevity penalty (closer to 1) compared to baseline or shortlisting. When using both together, we see slight improvement on both n-gram accuracies and length penalty over baseline. This is expected: shortlisting provides extra word alignment information which aids lexical accuracy; re-ranking favours the hypothesis with an average length, since too long or too short hypotheses will receive a lower similarity score. This implies that shortlisting and re-ranking enhance BLEU from different aspects.

On the contrary, according to METEOR, re-ranking is outperformed even by the baseline, leading to a negative METEOR change. This suggests that the bump in BLEU scores by re-ranking is only due to an improved brevity penalty. While it encourages a more desirable length, re-ranking does not produce better lexical choices. Adequacy wise, shortlisting is proven to be a more constructive method.

5 Limitations

Although our methods show promising gains on the low-resource domains, we found that they have limited application when the domain mismatch is too great or there is sufficient resource available.

5.1 Large domain mismatch

In order to better interpret why our methods are much more helpful on some domains than on others

(e.g. *law* versus *Koran*), we gather statistics of our test sets in different domains in Table 3, to reflect the distance between domains. We compute vocabulary overlap with the in-domain *medical* data, for each cleaned corpus prior to BPE encoding. We count only words that are seen at least 20 times, and we sample 5% of the *subtitles* corpus, since it is orders of magnitude larger than the rest. After down-sampling, it has a similar size to the *medical* corpora. We also compute the average sentence length before and after BPE to determine the extent to which BPE transforms the original sentences. Both vocabulary overlap and sentence lengths implicitly reflect the degree of domain mismatch.

For our most difficult datasets, *subtitles* and *Koran*, we see the lowest vocabulary overlap with the *medical* domain of just 22-25%. When BPE is applied to these corpora, the sentence length increases by 20-30%. In practice, extensive and uneven BPE segmentation on named entities makes it difficult for the IBM model to produce interpretable and meaningful alignments to aid a translation model, as exemplified in Table 4.

This means that lexical shortlisting is useful when we have a model trained on a relatively modest amount of data, and the out-of-domain dataset which we adapt to should share a reasonable amount of vocabulary. When the vocabulary overlap is too small, shortlisting drastically loses its effectiveness as a domain adaptation tool.

We confirm this by experimenting with another domain-distant, and extremely low-resource Burmese-English scenario. The training data consists 18k sentence pairs from *news* articles (Ding et al., 2019), and the out-of-domain test set is a *Bible* corpus (Christodouloupoulos and Steedman,

2015). The English side of the training set has 51k unique words, while the *Bible* set has 29k. The vocabulary overlap between the training and test sets is just 6k words, equivalent to 21% of the test vocabulary. After BPE, The average test sentence length increases from 25 to 38 tokens, seeing a 52% rise. We adopt the model trained by Aji et al. (2020) and add shortlisting during decoding. The results displayed in Table 5 show that out-of-domain performance does not improve with shortlisting, likely due to the high degree of vocabulary mismatch.

	baseline	shortlist
news (in-domain)	18.00	15.7
Bible	0.2	0.2

Table 5: Very low-resource Burmese-English results.

5.2 Availability of resources

Our previous German-English experiments are carried out under an artificially resource-constrained condition. We therefore verify the potential of our methods in a high-resource setting too, by applying shortlisting on a WMT19 German-English submission from Microsoft (Junczys-Dowmunt, 2019), and evaluating on the same out-of-domain datasets.

	Microsoft WMT19		low-resource
	baseline	shortlist	baseline
medical	14.4	14.4	61.4
Koran	0.0	0.0	0.8
law	8.7	8.7	17.8
IT	15.4	15.4	15.7
subtitles	1.0	1.0	2.6

Table 6: High-resource German-English results.

Results in Table 6 show that shortlisting has no impact on BLEU comparing to the baseline. We conclude that given a high-resource setting and the apparent large domain mismatch, the IBM model's alignments do not contribute additional information to the model. This finding is corroborated by Li et al. (2019): IBM model alignments are mostly better at capturing function words, not content words, compared to a neural model. Furthermore, unnatural and aggressive BPE segmentation on out-of-domain text (e.g. Table 4) could result in a lexical shortlist not capturing any meaningful alignment.

We add our low-resource baseline trained on *medical* with BPE on all data (Section 4 Table 1)

to comparison, and find that it surpasses the huge WMT model on each domain. Two reasons account for this: a WMT model is heavily biased to the news domain only; and the BPE scheme learned from news data is inferior to one learned from the out-of-domain datasets, when being evaluated on these domains.

6 Conclusion

We explore computationally cheap methods to improve neural machine translation performance in out-of-domain settings. We suggest that adding a lexical shortlist trained on the same data is always beneficial. While re-ranking also improves BLEU, it targets the BLEU brevity penalty, and does not produce better word choices. Although our results are promising in a low-resource condition, they do not transfer well to a scenario with very distant domains or sufficient resources. Our analysis shows that this is due to little vocabulary overlap, and the limited contribution from the IBM model under out-of-domain BPE segmentation.

References

Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich. 2020. In neural machine translation, what does transfer learning transfer? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7701–7710, Online. Association for Computational Linguistics.

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*.

Nikolay Bogoychev, Roman Grundkiewicz, Alham Fikri Aji, Maximiliana Behnke, Kenneth Heafield, Sidharth Kashyap, Emmanouil-Ioannis Farsarakis, and Mateusz Chudyk. 2020. Edinburgh's submissions to the 2020 machine translation efficiency task. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 218–224, Online. Association for Computational Linguistics.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: The bible in 100 languages. *Lang. Resour. Eval.*, 49(2):375–395.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1370–1380, Baltimore, Maryland. Association for Computational Linguistics.

Chenchen Ding, Hnin Thu Zar Aye, Win Pa Pa, Khin Thandar Nwet, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita. 2019. Towards Burmese (Myanmar) morphological analysis: Syllable-based tokenization and part-of-speech tagging. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(1):5.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Marina Fomicheva, Lucia Specia, and Francisco Guzmán. 2020. Multi-hypothesis machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1218–1232, Online. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at wmt 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018a. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Kenneth Heafield, Hieu Hoang, Roman Grundkiewicz, and Anthony Aue. 2018b. Marian: Cost-effective high-quality neural machine translation in C++. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 129–135, Melbourne, Australia. 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.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Hai Son Le, Alexandre Allauzen, and François Yvon. 2012. Continuous space translation models with neural networks. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 39–48, Montréal, Canada. Association for Computational Linguistics.

Xintong Li, Guanlin Li, Lemao Liu, Max Meng, and Shuming Shi. 2019. On the word alignment from neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1293–1303, Florence, Italy. Association for Computational Linguistics.

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'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the*

Association for Machine Translation in the Americas (AMTA 2020), pages 151–164, Virtual. Association for Machine Translation in the Americas.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks.

Holger Schwenk, Marta R. Costa-jussà, and Jose A. R. Fonollosa. 2007. Smooth bilingual n-gram translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 430–438, Prague, Czech Republic. 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, Adrià de Gispert, Eva Hasler, and Bill Byrne. 2017. Neural machine translation by minimising the Bayes-risk with respect to syntactic translation lattices. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 362–368, Valencia, Spain. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

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.

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.

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.