An Empirical Investigation of Word Alignment Supervision for Zero-Shot Multilingual Neural Machine Translation

Alessandro Raganato, Raúl Vázquez, Mathias Creutz and **Jörg Tiedemann**
University of Helsinki,
{name.surname}@helsinki.fi

Abstract

Zero-shot translations is a fascinating feature of Multilingual Neural Machine Translation (MNMT) systems. These MNMT models are usually trained on English-centric data, i.e. English either as the source or target language, and with a language label prepended to the input indicating the target language. However, recent work has highlighted several flaws of these models in zero-shot scenarios where language labels are ignored and the wrong language is generated or different runs show highly unstable results. In this paper, we investigate the benefits of an explicit alignment to language labels in Transformer-based MNMT models in the zero-shot context, by jointly training one cross attention head with word alignment supervision to stress the focus on the target language label. We compare and evaluate several MNMT systems on three multilingual MT benchmarks of different sizes, showing that simply supervising one cross attention head to focus both on word alignments and language labels reduces the bias towards translating into the wrong language, improving the zero-shot performance overall. Moreover, as an additional advantage, we find that our alignment supervision leads to more stable results across different training runs.

1 Introduction

Multilingual Neural Machine Translation (MNMT) focuses on translation between multiple language pairs through a single optimized neural model, and has been explored from different angles witnessing a rapid progress in recent years (Arivazhagan et al., 2019b; Wang et al., 2020; Dabre et al., 2020; Lin et al., 2021). Besides the great flexibility MNMT models offer, they are also highlighted by their so called zero-shot translation capabilities, i.e., translating between all combinations of languages available in the training data, including those with no parallel data seen at training time (Ha et al., 2016; Firat et al., 2016; Johnson et al., 2017). Many studies have investigated this feature, focusing on the impact of both, the model architecture design (Arivazhagan et al., 2019a; Pham et al., 2019) and data pre-processing (Lee et al., 2017; Wang et al., 2019; Rios et al., 2020; Wu et al., 2021). Broadly speaking, MNMT architectures are categorized according to their degree of parameter sharing, from fully shared (Johnson et al., 2017) to the use of language-specific components (Vázquez et al., 2020; Escolano et al., 2021; Zhang et al., 2021). The Johnson et al. (2017) MNMT model is widely used, due to its simplicity and good translation quality. It uses the fully shared parameters setting, and relies on appending an artificial language label to each input sentence to indicate the target language. While this method allows for zero-shot translation, several works have highlighted two major flaws: i) its failure to reliably generalize to unseen language pairs, ending up with the so called *off-target* issue, where the language label is ignored and the wrong target language is produced as a result (Zhang et al., 2020), ii) its lack of stability in translation results between different training runs (Rios et al., 2020).

In this work, we investigate the role of guided alignment in the Johnson et al. (2017) setting, by jointly training one cross attention head to explicitly focus on the target language label. We show that alignment supervision mitigates the *off-target* translation issue in the zero-shot case. Our method improves the zero-shot translation performance and results in more stable results across different training runs.

2 Methodology

Alignment Methods. Given a bitext $B_{src} = (s_1, ..., s_j, ..., s_N)$ and $B_{trg} = (t_1, ..., t_i, ..., t_M)$ where B_{src} is a sentence in the source language and B_{trg} is its translation in the target language, an alignment A is a mapping of words between B_{src} and B_{trg} (Tiedemann, 2011), formally defined as

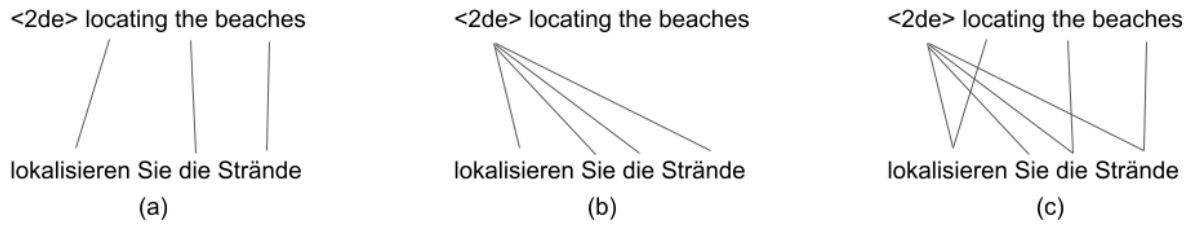

Figure 1: English → German example sentence with different alignment methods. Alignments in (a) show word alignments between corresponding words in the two languages, (b) our introduced alignments between all target words and the input language label, and (c) the union of the two.

a subset of the Cartesian product of the word positions (Och and Ney, 2003):

$$\text{A} \subseteq \{(j, i) : j = 1, ..., N; i = 1, ..., M\} \quad (1)$$

We study three different settings: (a) standard word alignment between corresponding words, (b) alignments between all target words and the language label in the input string, and (c) the union between the former two. Figure 1 shows an example of those approaches. To produce word alignments between parallel sentences, i.e., Figure 1 (a), we use the `awesome-align` tool (Dou and Neubig, 2021), a recent work that leverages multilingual BERT (Devlin et al., 2019) to extract the links.[1]

Models. To train Many-to-Many MNMT models, we use a 6-layer Transformer architecture (Vaswani et al., 2017), prepending a language label in the input to indicate the target language (Johnson et al., 2017). Following Garg et al. (2019), given an alignment matrix $AM_{M,N}$ and an attention matrix computed by a cross attention head $AH_{M,N}$, for each target word i, we use the following cross-entropy loss $\mathcal{L}_a$ to minimize the Kullback-Leibler divergence between AH and AM:

$$\mathcal{L}_a(AH, AM) = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{N} AM_{i,j}\log(AH_{i,j}) \quad (2)$$

The overall loss $\mathcal{L}$ is:

$$\mathcal{L} = \mathcal{L}_t + \gamma\mathcal{L}_a(AH, AM) \quad (3)$$

where $\mathcal{L}_t$ is the standard NLL translation loss, and γ is a hyperparameter. We use $\gamma = 0.05$, supervising only one cross attention head at the third last layer.[2] Given the sparse nature of the alignments, we replace the softmax operator in the cross attention head with the α-entmax function (Peters et al., 2019; Correia et al., 2019). Entmax allows sparse attention weights for any $\alpha > 1$. Following Peters et al. (2019), we use α=1.5.

3 Experimental Setup

We use three highly multilingual MT benchmarks:

- **TED Talks** (Qi et al., 2018). An English-centric parallel corpus with 10M training sentences across 116 translation directions. Following Aharoni et al. (2019), we evaluate on a total of 16 language directions, while as zero-shot test we evaluate on 4 language pairs.

- **WMT-2018** (Bojar et al., 2018).[3] A parallel dataset provided by the WMT-2018 shared task on news translation. We use all available language pairs, i.e. 14, up to 5M training sentences for each language pair. We evaluate the models on the test sets of the shared task, i.e. newstest2018. As there are no zero-shot test sets provided by the competition, we use the test portion from the Tatoeba-challenge (Tiedemann, 2020),[4] in all possible language pair combinations included in the challenge.

- **OPUS-100** (Zhang et al., 2020). An English-centric multi-domain benchmark, built upon the OPUS parallel text collection (Tiedemann, 2012). It covers a total of 198 language directions, with up to 1M training sentence per

[1] We use the `bert-base-multilingual-cased` checkpoint, without fine-tuning, and with `softmax` as a extraction function.

[2] As we use the `OpenNMT-py` (Klein et al., 2017) toolkit, it is recommended to supervise the third last layer. See https://github.com/OpenNMT/OpenNMT-py/issues/1843.

[3] http://data.statmt.org/wmt18/translation-task/preprocessed/

[4] release v2020-07-28.

ID	Model	#Param.	EN → X (16)	X → EN (16)	$BLEU_{zero}$ (4)	ACC_{zero} (4)
	Aharoni et al. (2019)-103	473M	**20.11**	**29.97**	9.17	-
	Aharoni et al. (2019)	93M	19.54	28.03	-	-
1	Transformer	93M	18.93 ±0.15	27.56 ±0.25	6.81 ±0.86	72.38 ± 7.18
2	1 + 1.5-entmax	93M	18.90 ±0.25	27.21 ±0.38	10.02 ±1.50	87.81 ± 8.80
3	2 + (a)	93M	18.99 ±0.07	27.58 ±0.12	8.38 ±5.37	73.12 ±41.14
4	2 + (b)	93M	18.98 ±0.08	27.48 ±0.13	6.35 ±0.87	65.01 ± 6.10
5	2 + (c)	93M	19.06 ±0.11	27.37 ±0.19	**11.94** ±0.86	**97.25** ± 2.66

Table 1: Results on the Many-to-Many TED Talks benchmark. The baselines consist of (1) our replication of the standard 6-layer Transformer model by Aharoni et al. (2019), and (2) its variant with a 1.5-entmax function on the cross attention heads as in Correia et al. (2019). The labels (a), (b), (c) denote the use of different alignment supervision (see Section 2). "#Param.": trainable parameter number. "EN -> X (16)" and "X-> EN (16)": average BLEU scores for English to Non-English languages and for Non-English languages to English on 16 language pairs respectively. "$BLEU_{zero}$ (4)" and "ACC_{zero} (4)": average BLEU scores and target language identification accuracy over 4 zero-shot language directions. We report average BLEU and accuracy scores, plus the standard deviation over 3 training runs with different random seeds.

language pair. It provides supervised translation test data for 188 language pairs, and zero-shot evaluation data for 30 pairs.

Following related work (Aharoni et al., 2019; Zhang et al., 2020), we apply joint Byte-Pair Encoding (BPE) segmentation (Sennrich et al., 2016; Kudo and Richardson, 2018), with a shared vocabulary size of 32K symbols for TED Talks and 64K for WMT-2018 and OPUS-100. As evaluation measure, we use tokenized BLEU (Papineni et al., 2002) to be comparable with Aharoni et al. (2019) for the TED Talks benchmark, and SACREBLEU [5] (Post, 2018) for WMT-2018 and OPUS-100.[6] As an additional evaluation, we report the target language identification accuracy score for the zero-shot cases (Zhang et al., 2020), called ACC_{zero}. We use fasttext as a language identification tool (Joulin et al., 2017), counting how many times the translation language matches the reference target language.

The Transformer models follow the *base* setting of Vaswani et al. (2017), with three different random seeds in each run. All of them are trained on the Many-to-Many English-centric scenario, i.e., on the concatenation of the training data having English either as the source or target language. Details about data and model settings in the Appendix.

[5]Signature: BLEU+case.mixed+numrefs.1+smooth.exp+tok.{13a,ja-mecab-0.996-IPA,zh}+version.1.5.0

[6]We report average BLEU over all test sets. Scores for each language pair are available in the supplementary material.

4 Results and Discussion

Throughout this section we refer to our baseline MNMT models by the labels (1) and (2), while (3), (4), and (5) mark the models trained with the auxiliary alignment supervision task, (a), (b), (c) from Figure 1 respectively (see Section 2).

TED Talks. Table 1 shows the results on the TED Talks benchmark. Regarding translation quality on the language pairs seen during training (EN → X and X → EN columns), average BLEU scores from all models end up in the same ballpark. In contrast, zero-shot results vary across the board, with (5) attaining the best performance, with almost 2 BLEU points better than its baseline (2). Moreover, (5) considerably improves target language identification accuracy (ACC_{zero}), with more stable results, i.e. lower standard deviation, than counterparts. Surprisingly, the addition of alignment supervision (a) and (b) as an auxiliary task has an overall detrimental effect on the zero-shot performance, even though model (4) results in more stable results than (2).

WMT-2018. Table 2 reports the results on the WMT-2018 benchmark. As expected, in a high-resource scenario bilingual baselines are hard to beat. Among multilingual models, the overall performance follows a similar trend as before. Enriching the model with alignment supervision (c) results in the best system overall, with an improvement of more than 3 BLEU points in the zero-shot

ID	Model	#Param.	EN → X (7)	X → EN (7)	$BLEU_{zero}$ (24)	ACC_{zero} (24)
	Transformer, Bilingual	127M	**18.28**	**19.25**	-	-
1	Transformer	127M	15.18 ±0.54	18.39 ±0.65	9.78 ±0.61	74.17 ±4.78
2	1 + 1.5-entmax	127M	15.17 ±0.41	18.33 ±0.56	8.55 ±0.61	65.31 ±4.46
3	2 + (a)	127M	11.99 ±0.37	16.42 ±0.73	6.38 ±0.83	73.78 ±7.84
4	2 + (b)	127M	15.46 ±0.16	18.66 ±0.31	11.72 ±0.76	85.64 ±3.37
5	2 + (c)	127M	15.50 ±0.18	18.70 ±0.23	**11.98** ±0.12	**85.68** ±0.82

Table 2: Results on the Many-to-Many WMT-2018 benchmark. Average BLEU, target language identification accuracy and standard deviation of 3 training runs.

ID	Model	#Param.	EN → X (94)	X → EN (94)	EN → X (4)	X → EN (4)	$BLEU_{zero}$ (30)	ACC_{zero} (30)
	Transformer, Bilingual†	110M	-	-	**20.28**	21.23	-	-
	Transformer+MATT†	141M	20.77	29.15	16.08	24.15	4.71	39.40
	MATT+LALN+LALT†	173M	**22.86**	**29.49**	19.25	24.53	**5.41**	**51.40**
1	Transformer	142M	18.50 ±0.08	26.85 ±0.13	18.37 ±0.39	**25.70** ±0.05	4.59 ±0.21	30.91 ±2.05
2	1 + 1.5-entmax	142M	18.47 ±0.15	26.83 ±0.14	18.42 ±0.38	25.67 ±0.10	4.39 ±0.86	30.51 ±5.62
3	2 + (a)	142M	17.80 ±0.23	26.21 ±0.40	17.53 ±0.34	25.18 ±0.39	3.96 ±0.43	28.95 ±2.61
4	2 + (b)	142M	18.56 ±0.04	26.91 ±0.18	18.32 ±0.36	25.47 ±0.10	4.63 ±0.48	31.05 ±5.93
5	2 + (c)	142M	18.63 ±0.07	26.69 ±0.09	18.51 ±0.18	25.39 ±0.01	4.73 ±0.16	32.00 ±0.96

Table 3: Results on the Many-to-Many OPUS-100 benchmark. Results marked with † are taken from Zhang et al. (2020). MATT denotes the use of merged attention (Zhang et al., 2019). LALN and LALT indicate the use of language-aware components. Average BLEU, target language identification accuracy and standard deviation of 3 training runs.

testbed compared to baseline (2), and with stable results across three training runs (standard deviations of 0.12 and 0.82).

OPUS-100. As one can see from Table 3, we confirm the positive effect of adding the alignment strategy (c) both as translation quality and as a mechanism to produce stable results even in a highly multilingual setup, i.e., training on 198 language directions. The average score over 30 zero-shot language pairs is low but the individual results range from 0.3 to 17.5 BLEU showing the potentials of multilingual models in this challenging data set as well.[7] Even though the results from our best model still lag behind models with language-specific components, i.e. MATT+LALN+LALT from Zhang et al. (2020), we note that our results demonstrate the positive effect of alignment on zero-shot translation.[8]

Overall, our experiments show consistent results across different benchmarks, providing quantitative evidence on the utility of guided alignment in highly multilingual MT scenarios. Supervising a single cross attention head with the alignment method (c) substantially reduces the instability between training runs, mitigating the off-target translation issue in the zero-shot evaluation. Zero-shot improvements, i.e. $BLEU_{zero}$ and ACC_{zero}, are large in two benchmarks out of three, i.e. Ted Talks and WMT-2018, and with a similar trend in OPUS-100. We also note that performance differences may be related to the different data sizes (see Appendix A). TED Talks is a rather small and imbalanced multilingual dataset with 116 language directions with a total of 10M training sentences, while WMT-2018 and OPUS-100 comprise 14 language pairs for a total of 47.8M training sentences, and 110M training sentences for 198 language pairs, respectively. We plan on investigating the impact of the training size and the resulting alignments on the zero-shot test sets further in future work.

Limitations Finally, we highlight that we have focused on a quantitative evaluation on English-centric MNMT benchmarks only, therefore we lack a comprehensive evaluation on complete MNMT benchmarks including training data without English as source and target language (Freitag and Firat, 2020; Rios et al., 2020; Tiedemann, 2020;

[7] Individual scores available in the supplementary material.

[8] Also note that Zhang et al. (2020) average the last 5 checkpoints whereas we report single checkpoints per run.

Goyal et al., 2021).

5 Conclusions and Future Work

In this work we present an empirical comparative evaluation of integrating different alignment methods in Transformer-based models for highly multilingual English-centric MT setups. Our extensive evaluation over three alignment variants shows that adding alignment supervision between corresponding words and the language label consistently improves the stability of the models, resulting in stable performance across different runs and mitigating the off-target translation issue in the zero-shot scenario. We believe that our work will pave the way for designing new and better multilingual MT models to improve their generalization in zero-shot setups.

As future work, we intend to analyze the quality of the learned alignments and their effect on the other attention weights in both supervised and zero-shot evaluation data (Raganato and Tiedemann, 2018; Tang et al., 2018; Mareček and Rosa, 2019; Voita et al., 2019). Finally, we plan to explore other mechanisms to inject prior knowledge to better handle zero-shot translations (Deshpande and Narasimhan, 2020; Raganato et al., 2020; Song et al., 2020).

Acknowledgments

This work is part of the FoTran project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 771113).

The authors gratefully acknowledge the support of the CSC – IT Center for Science, Finland, for computational resources. Finally, We would also like to acknowledge NVIDIA and their GPU grant.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual 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 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019a. The missing ingredient in zero-shot neural machine translation. *arXiv preprint arXiv:1903.07091*.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019b. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Gonçalo M. Correia, Vlad Niculae, and André F. T. Martins. 2019. Adaptively sparse transformers. 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 2174–2184, Hong Kong, China. Association for Computational Linguistics.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Computing Surveys (CSUR)*, 53(5):1–38.

Ameet Deshpande and Karthik Narasimhan. 2020. Guiding attention for self-supervised learning with transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4676–4686, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Carlos Escolano, Marta R. Costa-jussà, José A. R. Fonollosa, and Mikel Artetxe. 2021. Multilingual machine translation: Closing the gap between shared and language-specific encoder-decoders. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 944–948, Online. Association for Computational Linguistics.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Markus Freitag and Orhan Firat. 2020. Complete multilingual neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 550–560, Online. Association for Computational Linguistics.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. 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 4453–4462, Hong Kong, China. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *arXiv preprint arXiv:2106.03193*.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *arXiv preprint arXiv:1611.04798*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

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

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378.

Zehui Lin, Liwei Wu, Mingxuan Wang, and Lei Li. 2021. Learning language specific sub-network for multilingual machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 293–305, Online. Association for Computational Linguistics.

David Mareček and Rudolf Rosa. 2019. From balustrades to pierre vinken: Looking for syntax in transformer self-attentions. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 263–275, Florence, Italy. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

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. Association for Computational Linguistics.

Ben Peters, Vlad Niculae, and André F. T. Martins. 2019. Sparse sequence-to-sequence models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1504–1519, Florence, Italy. Association for Computational Linguistics.

Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, and Alexander Waibel. 2019. Improving zero-shot translation with language-independent constraints. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 13–23, Florence, Italy. Association for Computational Linguistics.

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

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2020. Fixed encoder self-attention patterns in transformer-based machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 556–568, Online. Association for Computational Linguistics.

Alessandro Raganato and Jörg Tiedemann. 2018. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297, Brussels, Belgium. Association for Computational Linguistics.

Annette Rios, Mathias Müller, and Rico Sennrich. 2020. Subword segmentation and a single bridge language affect zero-shot neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 528–537, Online. 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.

Kai Song, Kun Wang, Heng Yu, Yue Zhang, Zhongqiang Huang, Weihua Luo, Xiangyu Duan, and Min Zhang. 2020. Alignment-enhanced transformer for constraining nmt with pre-specified translations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8886–8893.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2018. An analysis of attention mechanisms: The case of word sense disambiguation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 26–35, Brussels, Belgium. Association for Computational Linguistics.

Jörg Tiedemann. 2011. Bitext alignment. *Synthesis Lectures on Human Language Technologies*, 4(2):1–165.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. 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 *Advances in Neural Information Processing Systems*, pages 5998–6008.

Raúl Vázquez, Alessandro Raganato, Mathias Creutz, and Jörg Tiedemann. 2020. A systematic study of inner-attention-based sentence representations in multilingual neural machine translation. *Computational Linguistics*, 46(2):387–424.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Xinyi Wang, Hieu Pham, Philip Arthur, and Graham Neubig. 2019. Multilingual neural machine translation with soft decoupled encoding. In *International Conference on Learning Representations*.

Yiren Wang, ChengXiang Zhai, and Hany Hassan. 2020. Multi-task learning for multilingual neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1022–1034, Online. Association for Computational Linguistics.

Liwei Wu, Shanbo Cheng, Mingxuan Wang, and Lei Li. 2021. Language tags matter for zero-shot neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3001–3007, Online. Association for Computational Linguistics.

Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. 2021. Share or not? learning to schedule language-specific capacity for multilingual translation. In *International Conference on Learning Representations (ICLR) 2021*.

Biao Zhang, Ivan Titov, and Rico Sennrich. 2019. Improving deep transformer with depth-scaled initialization and merged attention. 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 898–909, Hong Kong, China. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

A Data and Model details

A.1 Data

TED Talks (Qi et al., 2018). This parallel corpus includes 59 language pairs from and to English. It is a highly imbalanced benchmark, ranging from less than 4K up to 215K training sentences. We use the same languages as Aharoni et al. (2019) for both supervised testing and zero-shot evaluation. As supervised test sets, we use {Azerbeijani, Belarusian, Galician, Slovak, Arabic, German, Hebrew, Italian}↔English. As zero-shot test sets, we use Arabic↔French, and Ukrainian↔Russian.

WMT-2018 (Bojar et al., 2018). We use training and testing data as provided by the WMT 2018 news translation task organizers. The benchmark contains a total of 14 language pairs: {Chinese, Czech, Estonian, Finnish, German, Russian, Turkish}↔English. For training, we use up to 5M parallel sentences per language pair, with Turkish↔English, Estonian↔English, and Finnish↔English, having only 200K, 1M, and 2.7M training sentences, respectively. For zero-shot test sets, we use the test data from Tiedemann (2020), using the following 24 language directions:

Czech ↔ German,	German ↔ Russian,
German ↔ Chinese,	Finnish ↔ German,
Finnish ↔ Turkish,	Russian ↔ Finnish,
Russian ↔ Chinese,	Turkish ↔ Chinese,
Czech ↔ Russian,	German ↔ Turkish,
Estonian ↔ Russian,	Russian ↔ Turkish

OPUS-100 (Zhang et al., 2020). OPUS-100 is a recent benchmark consisting of 55M English-centric sentence pairs covering 100 languages. The data is collected from movie subtitles, GNOME documentation, and the Bible. Out of 99 language pairs, 44 have 1M sentences, 73 have at least 100K sentences, and 95 at least 10K. It provides also zero-shot test sets, pairing the following languages: Arabic, Chinese, Dutch, French, German, and Russian.

A.2 Model hyperparameters

We use the OpenNMT-py framework (Klein et al., 2017), and the Transformer *base* model setting (Vaswani et al., 2017). Specifically, we use 6 layers for the encoder and the decoder, 512 as model dimension, and 2048 as hidden dimension.

	#Lang. pairs	#Train. sent.	#Zero-shot lang. pairs
TED Talks	116	10M	4
WMT-2018	14	47M	24
OPUS-100	198	110M	30

Table 4: Benchmark statistics: number of language pairs used for training, total number of training sentences, and number of language pairs for zero-shot evaluation.

We applied 0.1 as dropout for both residual layers and attention weights, using the Adam optimizer (Kingma and Ba, 2015) with $\beta 1 = 0.9$, and $\beta 2 = 0.998$, with learning rate set at 3 and 40K warmup steps as in Aharoni et al. (2019). We train the models with three random seeds each, for 200K training steps for the TED Talks and WMT-2018 benchmarks, while for 500K training steps for the OPUS-100. To speed up training, we use half-precision, i.e., FP16.