Language Tags Matter for Zero-Shot Neural Machine Translation

Liwei Wu, Shanbo Cheng, Mingxuan Wang, Lei Li

ByteDance AI Lab

{*wuliwei.000,chenshanbo,wangmingxuan.89,lileilab*}*@bytedance.com*

Abstract

Multilingual Neural Machine Translation (MNMT) has aroused widespread interest due to its efficiency. An exciting advantage of MNMT models is that they could also translate between unsupervised (zero-shot) language directions. Language tag (LT) strategies are often adopted to indicate the translation directions in MNMT. In this paper, we demonstrate that the LTs are not only indicators for translation directions but also crucial to zero-shot translation qualities. Unfortunately, previous work tends to ignore the importance of LT strategies. We demonstrate that a proper LT strategy could enhance the consistency of semantic representations and alleviate the off-target issue in zero-shot directions. Experimental results show that by ignoring the source language tag (SLT) and adding the target language tag (TLT) to the encoder, the zero-shot translations could achieve a +8 BLEU score difference over other LT strategies in IWSLT17, Europarl, TED talks translation tasks.

1 Introduction

Neural Machine Translation (NMT) based on the *encoder-decoder framework with attention mechanism* (Sutskever et al., 2014; Bahdanau et al., 2015; Luong et al., 2015; Wu et al., 2016; Gehring et al., 2017; Vaswani et al., 2017) has achieved state-of-the-art (SotA) results in many language pairs (Deng et al., 2018; Barrault et al., 2019). Pioneered by (Dong et al., 2015; Firat et al., 2016; Zoph and Knight, 2016; Ha et al., 2016; Johnson et al., 2017), researchers start to investigate the possibility of using a single model to translate between multiple languages, which is the Multilingual Neural Machine Translation (MNMT). Benefiting from the transferring ability of multilingual modeling, MNMT could achieve better translation quality between low-resource language directions than bilin-gual models (Gu et al., 2018; Wang et al., 2019). More exciting, MNMT could even translate between zero-shot language directions (Johnson et al., 2017; Gu et al., 2019; Pham et al., 2019; Kudugunta et al., 2019).

Unlike bilingual NMT, language-specific signals should be accessible to the MNMT model so that the model can distinguish the translation directions. Ha et al., (2016) first introduced a universal encoder-decoder framework for MNMT models with language-specific coded vocabulary to indicate different languages. The *encoder-decoder* architecture is identical to bilingual models (Bahdanau et al., 2015; Vaswani et al., 2017). To further simplify the MNMT models, Johnson et al., (2017) propose to add language tags (LTs) to the beginning of input data to indicate the target language. Then a shared vocabulary could be learned for all languages. The training data of different languages could thus be mixed-up to train the MNMT model. Such a strategy greatly simplifies the training and decoding procedure. We call it the *LT strategy*. This paper focuses on investigating the impact of LT strategies for zero-shot translation directions in MNMT (zero-shot MNMT). We conduct translation experiments (Section 3) and visualization analysis (Section 4) on several multilingual benchmarks with different LT strategies. We observe that:

- The TLT is more important than the SLT. The SLT even causes negative effects on the zero-shot translation.

- The placements of LTs have a surprisingly large impact on the translation quality. Placing different LTs on different parts of the NMT model lead to a +8 BLEU score difference in our experiments.

Our contributions are mainly twofold: (i) We

3001

Strategy	Source sentence	Target sentence
Original	`Hello World!`	`¡Hola Mundo!`
T-ENC	`__es__ Hello World!`	`¡Hola Mundo!`
T-DEC	`Hello World!`	`__es__ ¡Hola Mundo!`
S-ENC-T-ENC	`__en__ __es__ Hello World!`	`¡Hola Mundo!`
S-ENC-T-DEC	`__en__ Hello World!`	`__es__ ¡Hola Mundo!`

Table 1: Examples of modified input data by different LT strategies. The bold tokens are the SLT (`__en__`) or TLT (`__es__`). T-ENC is identical to (Johnson et al., 2017), which adds the TLT to the encoder (source) side. T-DEC means placing the TLT on the decoder (target) side of model. S-ENC-T-ENC and S-ENC-T-DEC place the SLT on the encoder side, but the former also places the TLT on encoder side, while the latter on the decoder side.

find that the LT strategies are crucial for the zero-shot MNMT translation quality. Ignoring SLTs and placing the TLTs on the encoder side could achieve the best performance during our experiments. (ii) We conduct extensive visualization analysis to demonstrate that the proper LT strategy could enhance the consistency of semantic representation and alleviate the off-target issue (Zhang et al., 2020), thus improving the translation quality. To the best of our knowledge, this is the first paper to systematically study the importance of LT strategies for zero-shot translation quality.

2 Background and Notations

Improving the consistency of semantic representations and alleviating the off-target issue (Zhang et al., 2020) are effective ways to improve the zero-shot translation quality (Al-Shedivat and Parikh, 2019; Arivazhagan et al., 2019; Zhu et al., 2020). The semantic representations of different languages should be close to each other to get better translation quality (Ding et al., 2017). The off-target issue indicates that the MNMT model tends to translate input sentences to the wrong languages, which leads to low translation quality.

Due to its simplicity and efficiency, LT strategy has become a fundamental strategy for MNMT (Dabre et al., 2020). Though previous work adopted different LT strategies (Wang et al., 2018; Blackwood et al., 2018; Conneau and Lample, 2019; Liu et al., 2020b), the usages of LT strategies are intuitive and lack systematic study. In this paper, we investigate 4 popular LT strategies, namely `T-ENC`, `T-DEC`, `S-ENC-T-ENC` and `S-ENC-T-DEC`. Each of them only requires simple modifications to the input data. Table 1 comprehensively illustrates the strategies with an English to Spanish translation pair (`Hello World!`

$\rightarrow$ `¡Hola Mundo!`).

3 Experiments

3.1 Experiment Settings

Datasets We carry out our experiments on the publicly available IWSLT17 (Cettolo et al., 2017), TED talks (Qi et al., 2018) and Europarl v7 (Koehn, 2005) datasets. Table 2 shows an overview of the datasets. We choose four different languages (English included) for both IWSLT17 and Europarl, and 20 languages for TED talks. All the training data are English-centric parallel data, which means either the source-side or target-side of the sentence pair is English. We have 6, 6, and 342 zero-shot translation directions and an average of 145k, 1.96M (M = million), and 187k sentence pairs per direction for the three datasets respectively. We choose the official tst2017, WMT `newstest08`, and the TED talks testsets (Qi et al., 2018) as our test sets, respectively. We learned a joint SentencePiece model (Kudo and Richardson, 2018) for sub-word training on all languages with 40,000 merge operations for each dataset. We limit the size of joint vocabulary to 40,000 for all three datasets.

Settings We use the open-source implementation (Ott et al., 2019) of Transformer model (Vaswani et al., 2017). Following the settings of (Liu et al., 2020a), we use a 5-layer encoder and 5-layer decoder variation of Transformer-base model (Vaswani et al., 2017) for TED and IWSLT17. For Europarl v7, we use a standard Transformer-big model (Vaswani et al., 2017). Sentence pairs are batched together by approximate sentence length. Each batch has approximately 30,000 source tokens and 30,000 target tokens. We use the Adam (Kingma and Ba, 2015) optimizer to update the parameters and

Dataset	languages	#zero-shot directions	#training sents per direction	#sents per testset
IWSLT17	en, it, nl, ro	6	145k	1144
Europarl	en, fr, de, es	6	1.96m	2000
TED	en, ar, he, ru, ko, it, ja zh, es, fr, pt, nl, tr, ro pl, bg, vi, de, fa, hu	342	187k	4507

Table 2: An overview of the datasets. The second column is the languages the training data contains. The third column denotes the number of zero-shot translation directions. The fourth and fifth column denote the averaged number of training data and test data per language direction, respectively.

Dataset	LT Strategy	Supervised	Zero-Shot	Off-Target (%)
IWSLT17	T-ENC	32.30	**16.00** (+14.02)	**9.16**
	T-DEC	32.43	10.44	29.50
	S-ENC-T-ENC	32.56	1.98	94.14
	S-ENC-T-DEC	32.39	7.67	48.87
Europarl	T-ENC	35.55	**32.25** (+24.24)	1.18
	T-DEC	35.49	30.73	**1.13**
	S-ENC-T-ENC	35.53	8.01	79.53
	S-ENC-T-DEC	35.53	29.81	2.26
TED talks	T-ENC	25.63	**10.69** (+8.78)	**12.63**
	T-DEC	25.58	3.11	58.47
	S-ENC-T-ENC	25.84	4.07	65.03
	S-ENC-T-DEC	25.63	1.91	77.02

Table 3: Translation results on 3 datasets. The **supervised** and **zero-shot** column denote the averaged BLEU score of supervised or zero-shot directions. The **off-target (%)** denotes the averaged percentage of sentences being translated to wrong languages in zero-shot directions.

train each model for 100,000 steps to make sure it converges. We use beam search for heuristic decoding, and set the beam size to 4. We use SacreBLEU (Papineni et al., 2002; Post, 2018) to evaluate the translation results. To calculating the percentage of off-target translations, we use the langdetect[1] tool to detect the language of the translated sentences.

3.2 Experimental Results

We show the translation results on the IWSLT17, Europarl, and TED talks datasets in Table 3. For all three datasets, different strategies achieve comparable BLEU score on supervised directions. However, for the zero-shot directions, the BLEU score varies significantly using different LT strategies. One observation is that the T-ENC strategy consistently outperforms the other three strategies on all

datasets in terms of BLEU score with large margin, regardless of the corpus size and number of languages. In terms of off-target issue, T-ENC achieves the best performance in most cases.

Besides, ignoring the SLT (T-ENC v.s. S-ENC-T-ENC) also helps the zero-shot BLEU score. The percentage of off-target translations reaches 94.14% in the **IWSLT17** dataset by S-ENC-T-ENC strategy, while only 9.16% by T-ENC strategy. It indicates that the model translates almost all the sentences to the wrong languages in S-ENC-T-ENC, while to the right languages in T-ENC. It proves again the SLT hurts the zero-shot translation.

Another interesting observation is that placing the TLT on the encoder side also helps the zero-shot performance. Compared with T-ENC, both the translation quality and off-target performance are significantly worse in T-DEC. We will study

the reasons behind the above observations by visualization analysis in Section 4.

4 Visualization Analysis

We conduct the visualizations on the TED talks data to analyze the impact of different LT strategies on the semantic representation consistency and the off-target issue in MNMT.
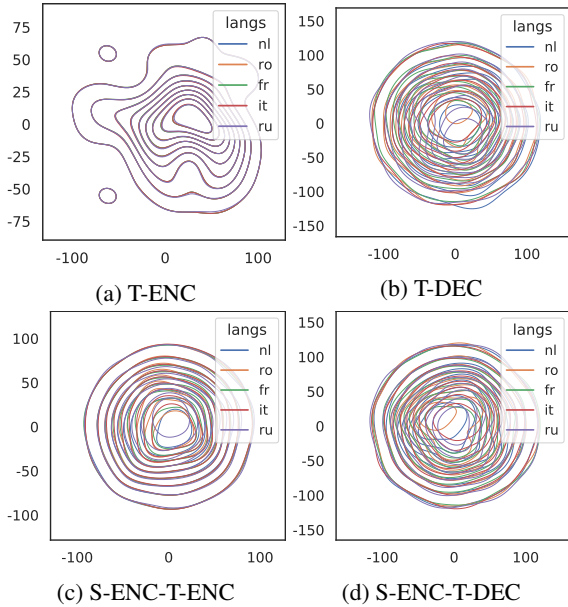

Figure 1: KDE visualization of encoder output using different LT strategies. 5 languages (nl, ro, fr, it, ru → zh) are randomly chosen for better readability.

Enhancing the Semantic Representation Consistency Figure 1 shows the kernel density estimation (KDE) (Parzen, 1962) of t-SNE (Van der Maaten and Hinton, 2008) reduced average encoder output on different languages. We randomly chose 5 source languages (nl, ro, fr, it, ru → zh) instead of all languages for clearer visualization. We choose 100 sentences for each language, and each sentence has its corresponding translation in the other 4 languages. The contour lines drawn by Kernel Density Estimation tools [2] was used to estimate the semantic distribution of the encoder outputs. The contour lines visualize the semantic representation of different languages. The representations are more consistent if the contour lines of different languages overlap more with each other.

The contour lines are nearly perfectly overlap with each other in T-ENC (Figure 1a), while they do not for the other strategies. Comparing Figure 1a and Figure 1c, we can see that ignoring SLT

---

[2]https://seaborn.pydata.org/generated/seaborn.kdeplot.html

greatly helps the model to learn more consistent representations. Comparing Figure 1a and Figure 1b, placing TLT to encoder side instead of the decoder side also helps the semantic consistency. Both comparisons validate that T-ENC could learn the most consistent and different semantic representations, thus achieves the best BLEU score. It might be why the shape of contour lines in T-ENC is significantly different from other strategies.

Alleviating the off-target Issue Figure 2 shows the attention visualization of a Russian to Italian translation example using different LT strategies. The x-axis is the Italian translation.

In Figure 2a1, T-ENC strategy pays attention to the TLT (in this case, the token ⎵it⎵ in the red background) during the whole translation procedure (left-to-right). Compared to T-ENC, both T-DEC and S-ENC-T-DEC pay less attention to the TLT after a few tokens are generated. It validates that placing the TLT on the encoder side would also help the model distinguish the target languages. The S-ENC-T-ENC pays nearly equal attention to both SLT and TLT, which might make the model confused about which one is the target language. Both comparisons prove that the T-ENC strategy has the best ability to distinguish the target languages, thus alleviates the off-target issue.

Combining Both Semantic Consistency and off-target Issue Figure 3 visualizes the cosine similarity of the layer-wise encoder and decoder output of different languages in zero-shot setting (English excluded). We sampled 100 muti-way data from the test set and averaged the cosine similarity between each language.

In the many-to-one setting, we randomly select Russian as the target language and translate the other 18 languages to Russian to obtain the model outputs. The similarity improves from encoder layer 0 to 3 and decoder layer 0 to layer 4, which indicates that the semantic consistency improves as the layer goes up. Interestingly, the similarity drops from encoder layer 3 to layer 4. It might be because the decoder interacts with the encoder directly between encoder layer 4 and decoder layer 0, thus interferes with the top-layer encoder output. But the dropping trend is less rapid in T-ENC than in other strategies. The T-ENC achieves the highest similarity on the last layer, which shows that the T-ENC learns more consistent semantics representations.

(a1) T-ENC   (a2) S-ENC-T-ENC

(a) The Decoder Cross Attention

(b1) T-DEC   (b2) S-ENC-T-DEC
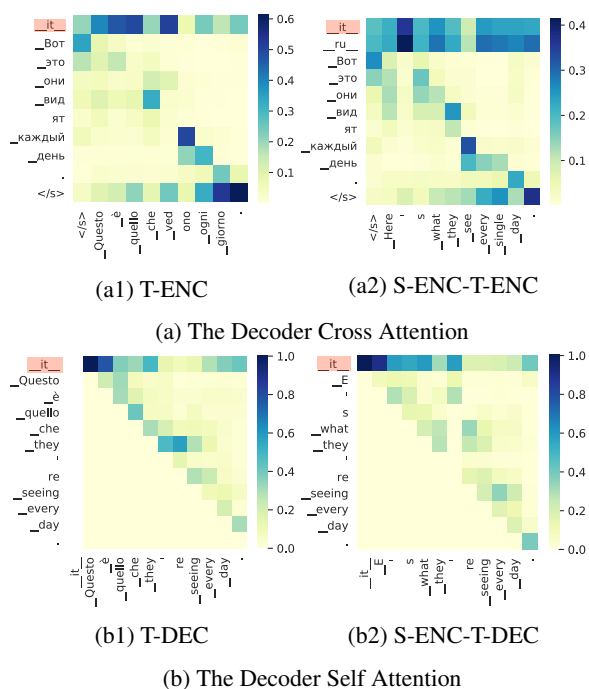
(b) The Decoder Self Attention

Figure 2: Attention visualization on a Russian to Italian translation example using different LT strategies. Note that we present the cross-attention for `T-ENC` and `S-ENC-T-ENC`, the decoder self-attention for `T-DEC` and `S-ENC-T-DEC` to visualize the TLT token.

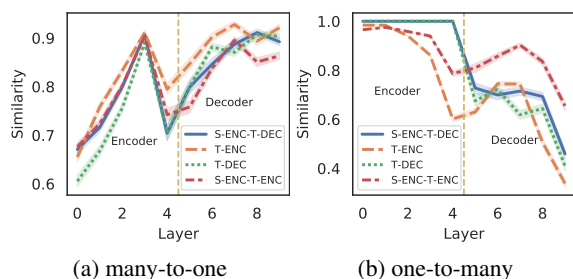(a) many-to-one   (b) one-to-many

Figure 3: The cosine similarity of layer-wise output using different LT strategies. Note that layer 0 to 4 are the encoder layers, layer 5 to 9 are the decoder layers.

In the one-to-many setting, we treat Russian as the source language and translate Russian to the other 18 languages to get the model output. The semantic similarity drops as the layer goes up in all four strategies. It indicates that the model can distinguish different target languages as the layer goes up. `T-ENC` achieves the lowest similarity at the last layer output among all strategies. It shows again that the `T-ENC` has the best ability to alleviate the off-target issue.

5 Conclusion

We show that the language tags in MNMT are not just indicators for translation directions but also

significantly impact the zero-shot translation quality. By extensive experiments and visualization analysis, we found that (i) ignoring the SLTs could help the models learn consistent semantic representations. (ii) Placing the TLTs on the encoder side could help the decoder pay more attention to the target language, thus alleviating the off-target issue. Zero-shot translation quality could be improved by investigating how to enhance the semantic representation consistency further and alleviate the off-target issue by optimizing LT strategies. We will conduct methods to optimize the LT strategy in our future work.

References

Maruan Al-Shedivat and Ankur P Parikh. 2019. Consistency by agreement in zero-shot neural machine translation. *arXiv preprint arXiv:1904.02338*.

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

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*.

Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussa, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61.

Graeme W. Blackwood, Miguel Ballesteros, and Todd Ward. 2018. Multilingual neural machine translation with task-specific attention. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3112–3122. Association for Computational Linguistics.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Niehues Jan, Stüker Sebastian, Sudoh Katsuitho, Yoshino Koichiro, and Federmann Christian. 2017. Overview of the iwslt 2017 evaluation campaign. In *International Workshop on Spoken Language Translation*, pages 2–14.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing*

Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 7057–7067.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A comprehensive survey of multilingual neural machine translation. *arXiv preprint arXiv:2001.01115*.

Yongchao Deng, Shanbo Cheng, Jun Lu, Kai Song, Jingang Wang, Shenglan Wu, Liang Yao, Guchun Zhang, Haibo Zhang, Pei Zhang, et al. 2018. Alibaba's neural machine translation systems for wmt18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 368–376.

Yanzhuo Ding, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Visualizing and understanding neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1150–1159.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 866–875. The Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pages 1243–1252. PMLR.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O. K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 344–354. Association for Computational Linguistics.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O. K. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1258–1268. Association for Computational Linguistics.

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 B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Trans. Assoc. Comput. Linguistics*, 5:339–351.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

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.

Sneha Reddy Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 1565–1575. Association for Computational Linguistics.

Danni Liu, Jan Niehues, James Cross, Francisco Guzmán, and Xian Li. 2020a. Improving zero-shot translation by disentangling positional information. *CoRR*, abs/2012.15127.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020b. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421. The Association for Computational Linguistics.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of*

the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. Association for Computational Linguistics.

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.

Emanuel Parzen. 1962. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076.

Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, and Alexander H. Waibel. 2019. Improving zero-shot translation with language-independent constraints. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 1: Research Papers*, pages 13–23. 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.

Ye Qi, Devendra Singh 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, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 529–535. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Xinyi Wang, Hieu Pham, Philip Arthur, and Graham Neubig. 2019. Multilingual neural machine translation with soft decoupled encoding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yining Wang, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2018. Three strategies to improve one-to-many multilingual translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2955–2960.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

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, ACL 2020, Online, July 5-10, 2020*, pages 1628–1639. Association for Computational Linguistics.

Changfeng Zhu, Heng Yu, Shanbo Cheng, and Weihua Luo. 2020. Language-aware interlingua for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1650–1655.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 30–34. The Association for Computational Linguistics.