# The Effects of Language Token Prefixing for Multilingual Machine Translation

**Rachel Wicks**[1,2] and **Kevin Duh**[1,2]
[1]Center for Language and Speech Processing
[2]Human Language Technology Center of Excellence
Johns Hopkins University
rewicks@jhu.edu, kevinduh@cs.jhu.edu

## Abstract

Machine translation traditionally refers to translating from a single source language into a single target language. In recent years, the field has moved towards large neural models translating from or into many languages. As the input and output languages vary, the model must be correctly cued to translate into the correct target language. This is typically done by prefixing *language tokens* onto the source or target sequence. A single token's content can denote the source language, target language, or language pair. The location and content of the prefix varies and many approaches exist without much justification towards one method or another. As guidance to researchers and directions for future work, we present a series of comprehensive experiments that show how the positioning and type of a target language prefix token affects translation performance. We show that source-side prefixes consistently improve performance. Further, we find that best language token content varies dependent on the supported language set.

## 1 Introduction

Machine translation (MT) started as a basic sequence-to-sequence problem. Confined to a single input and output language, the model was only responsible for learning the mapping between these two languages. Multilingual neural machine translation (MNMT) shifted the paradigm to consider many input and output languages (Ha et al., 2016). Language tokens, or tokens that signify the source language and the desired target language, became common prefixes on source and target sequences.

In Table 1, we display the typical combinations of prefixing techniques. In the simplest form, a neural multilingual model can be trained with the same pipeline as a bilingual model by prepending a single token to the source. One token can represent both the source and target in the language pair (as in S2T | ∅). Alternatively, the single token can be separated into two sequential tokens (S T | ∅). The model requires a signal for the target, but the source is optional so a single target-only token could be used (T | ∅). The same variety of tokens can also be prepended to the target sequence. It is also common to prepend the source language tag on the source and the target on the target (S | T).

| Label | Example (en-id) |
|---|---|
| S2T \| ∅ | **<en2id>** In the beginning, ...<br>Pada mulanya, waktu ... |
| S T \| ∅ | **<en>** **<id>** In the beginning, ...<br>Pada mulanya, waktu ... |
| T \| ∅ | **<id>** In the beginning, ...<br>Pada mulanya, waktu ... |
| ∅ \| S2T | In the beginning, ...<br>**<en2id>** Pada mulanya, waktu ... |
| ∅ \| S T | In the beginning, ...<br>**<en>** **<id>** Pada mulanya, waktu ... |
| ∅ \| T | In the beginning, ...<br>**<id>** Pada mulanya, waktu ... |
| S \| T | **<en>** In the beginning, ...<br>**<id>** Pada mulanya, waktu ... |

Table 1: Examples of using language tokens as prefixes to denote input and output languages. Blue (top sequence) tags denote the source and the red (bottom sequence) denote the target sequences.

Considerations for the placement of token may be convenience—prefixing on the source makes off-the-shelf training pipelines quickly deployable. Source-side prefixing obviously affects encodings, and there has been recent interest in making the encodings of a multilingual model language agnostic with evidence to suggest it makes the model more robust in zero-shot settings (Pan et al., 2021).

We focus on supervised directions—language pairs seen during training—which has not been thoroughly evaluated to the best of our knowledge.



*Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 148–153
November 20–23, 2022. 

We find differences in conclusions in supervised directions over previous results on zero-shot (Wu et al., 2021). In this work, we show that source side prefixing is preferable to target side prefixing, but the best token-type varies on language set. Adding source language information is beneficial for many language pairs—contrary to zero-shot conclusions. We also vary encoder and decoder dpeths to determine if the source-side tokens are successful as result of strong encodings and find similar results in both source and target side prefixing.

## 2 Related Work

Ha et al. (2016) introduced the first methodology to train a multilingual neural model that shared both encoder and decoder. They signaled source and target language to the model by prepending language tokens to each input (and output) token—creating inputs of the form "`@de@darum @de@geht @de@es @de@in @de@meinem @de@Vortrag`" to convey German (`de`) tokens. They also used prefixing and appending of the target language to "target-force" the language. Work compared these strategies (Ha et al., 2017) and subsequent work used single tokens as tags.

Johnson et al. (2017) use target language tags on the source sentence while focusing on low-resource and zero-shot directions. M2M100 (Fan et al., 2021), a pre-trained multilingual model, use a source-side source token and a target-side target token (S | T). mBART (Liu et al., 2020) uses a similar method, but *appends* the token after the `</s>` at the end of the sequence rather than prepending it. The new T5 models (Raffel et al., 2019) leverage a natural language structure and train for many tasks. mT5 (Xue et al., 2020) supports multilingual machine translation and uses an approach similar to "S T | ∅" by prepending *phrases* such as "translate German to English:" to the source.

Investigation in these techniques has been limited to studying the effects on zero-shot translation. Ha et al. (2017) considered combinations of these techniques to target zero-shot translation but ultimately found that constraining the decoding by filtering for the target language is more productive. Conversely, Wu et al. (2021) has investigated zero-shot translation and found that "T | ∅" outperforms other approaches. N ElNokrashy et al. (2022) find that "S T | T" can beat "T | ∅" in zero-shot settings. The preferred prefixing technique may be dependent on use-case and the set of supported languages. We focus on supervised settings to complement these works in search of a more thorough understanding of prefixing tokens.

| | Family | Script | ISO | Sentences |
|---|---|---|---|---|
| TASK$_1$ | Indo-European | Latin | en | 107M |
| | | | hr | 23.7M |
| | | Cyrillic | mk | 1.4M |
| | | | sr | 11.3M |
| | Uralic | Latin | et | 20.4M |
| | | | hu | 50.1M |
| TASK$_2$ | Indo-European | Latin | en | 18.0M |
| | Malayo-Polynesian | Latin | id | 12.7M |
| | | | jv | 1.4k |
| | | | ms | 3.3M |
| | | | tl | 1.1M |
| | Dravidian | Tamil | ta | 879k |

Table 2: Amount of training data used for the two tracks, broken down by individual language, script, and language family.

Token prefixing pitfalls can be mitigated by having multiple decoders responsible for a subset of languages. Shallow decoders have been shown to be ineffective in MNMT compared to bilingual equivalents but multiple shallow decoders can compensate for these differences (Kong et al., 2021; Sen et al., 2019). We use a single unified decoder.

## 3 Experimental Design

Language tokens are typically additional vocabulary items where the content designates the source language, the target language or a combination of the two (i.e., `<src>`, `<tgt>`, and `<src2tgt>`, respectively). Designating the target language is necessary and many choose to add source information as well as an additional signal to the encoder.

These tokens can be prepended onto either the source or target—directly affecting the encodings of either the encoder or decoder. In order to compare across these techniques, we train models with seven prefixing strategies outlined in Table 1 in three different datasets (described in Section 3.1).

### 3.1 Data

We consider the two small tracks for the Workshop on Machine Translation's (WMT21) Large-Scale Multilingual Shared Task. The small tracks focus on regional language groups which covers linguistically diverse languages and are relatively balanced

| | s2t \| ∅ | s t \| ∅ | t \| ∅ | ∅ \| s2t | ∅ \| s t | ∅ \| t | s \| t | s2t \| ∅ | s t \| ∅ | t \| ∅ | ∅ \| s2t | Bi. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\text{Task}_1$ | $\text{Task}_1$ | $\text{Task}_1$ | $\text{Task}_1$ | $\text{Task}_1$ | $\text{Task}_1$ | $\text{Task}_1$ | $\text{Task}_{1,2}$ | $\text{Task}_{1,2}$ | $\text{Task}_{1,2}$ | $\text{Task}_{1,2}$ | |
| en-et | 19.9 | **20.1** | 19.9 | 19.6 | 18.5 | 19.5 | 19.5 | 19.1 | **19.4** | 18.9 | 18.4 | 21.3 |
| en-hr | 24.3 | 24.3 | **24.6** | 24.4 | 23.5 | 23.4 | 24.3 | **24.0** | 23.5 | 23.6 | 23.3 | 25.7 |
| en-hu | 21.5 | **22.4** | 21.7 | 22.1 | 21.9 | 21.3 | 22.0 | **21.8** | 21.5 | 21.5 | 21.4 | 22.4 |
| en-mk | 21.9 | **22.9** | 22.6 | 22.6 | 21.9 | 21.6 | 22.4 | **22.4** | 21.4 | 21.4 | 20.7 | 30.3 |
| en-sr | 14.3 | **16.4** | 15.3 | 15.2 | 12.9 | 12.0 | 11.9 | **15.7** | 14.1 | 12.6 | 13.1 | 21.8 |
| et-en | 27.7 | 27.9 | 28.1 | **28.3** | 28.0 | 26.9 | **28.3** | **27.9** | 27.1 | 27.2 | 27.3 | 30.6 |
| hr-en | 29.7 | 30.7 | 30.2 | **30.9** | 29.5 | 29.5 | 30.1 | **29.8** | 29.7 | 29.7 | **29.8** | 31.3 |
| hu-en | 27.7 | **28.4** | 28.0 | 28.2 | 27.8 | 27.4 | 28.2 | **27.8** | 27.7 | 27.6 | 27.6 | 28.6 |
| mk-en | 28.9 | **29.9** | 29.5 | 29.7 | 29.2 | 29.1 | 29.6 | **29.5** | 28.7 | 28.3 | 29.0 | 24.4 |
| sr-en | 29.9 | **31.0** | 30.7 | 30.8 | 30.1 | 29.7 | 30.2 | **30.7** | 30.3 | 30.0 | 29.7 | 35.6 |
| AVG. | 24.6 | **25.4** | 25.1 | 25.2 | 24.3 | 24.0 | 24.7 | - | - | - | - | - |
| | $\text{Task}_2$ | $\text{Task}_2$ | $\text{Task}_2$ | $\text{Task}_2$ | $\text{Task}_2$ | $\text{Task}_2$ | $\text{Task}_2$ | - | - | - | - | - |
| en-id | 42.4 | 43.2 | **44.0** | 43.1 | 43.3 | 43.4 | 43.1 | **39.5** | 38.9 | 39.2 | 38.6 | 43.6 |
| en-jv | 1.3 | 0.9 | 3.9 | 4.0 | 4.0 | 3.2 | **4.2** | **4.1** | 1.7 | 3.9 | 3.2 | 0.1 |
| en-ms | 37.6 | 38.0 | **38.9** | 37.9 | 38.2 | 38.3 | 38.1 | **34.2** | 34.1 | 33.6 | 33.0 | 37.5 |
| en-ta | 8.8 | 9.5 | **9.7** | 8.0 | 7.7 | 8.2 | 8.7 | **5.4** | 5.0 | 5.1 | 4.4 | 11.2 |
| en-tl | 27.8 | 27.9 | **28.4** | 27.5 | 27.4 | 27.8 | 28.2 | **24.1** | 23.1 | 23.3 | 23.2 | 29.2 |
| id-en | 35.7 | **37.2** | 36.9 | 36.5 | 36.5 | 36.7 | 37.0 | **33.9** | 33.5 | 33.4 | 33.3 | 36.4 |
| jv-en | **8.6** | **8.6** | 8.3 | 6.9 | 7.9 | 8.3 | 8.4 | **8.6** | 6.3 | 6.4 | 5.9 | 0.1 |
| ms-en | 34.8 | 35.9 | **36.2** | 35.4 | 35.8 | 35.7 | 35.6 | **33.2** | 32.6 | 32.3 | 32.2 | 33.4 |
| ta-en | 15.5 | 16.5 | **16.7** | 15.7 | 15.8 | 15.2 | 15.8 | **13.1** | 12.2 | 12.5 | 12.5 | 18.2 |
| tl-en | 30.8 | 32.5 | **33.2** | 31.8 | 31.4 | 32.2 | 32.5 | **27.8** | 26.7 | 26.9 | 26.7 | 35.1 |
| AVG. | 24.3 | 25.0 | **25.6** | 24.7 | 24.8 | 24.9 | 25.2 | **23.6** | 22.9 | 22.9 | 22.7 | |

Table 3: BLEU scores for each $\text{Task}_1$ (top left), $\text{Task}_2$ (bottom left) and $\text{Task}_{1,2}$ (right) with each prefixing technique. Bold indicates highest score; green highlighting indicates models are not statistically worse compared to best model. We include bilingual models' scores (right-most column) to help contextualize these scores.

in data quantity. We use language pairs containing English for training. Each track contains five languages from the same region which gives significant overlap between language families making them ideal candidates for MNMT.

The first task ($\text{Task}_1$) contains Croatian (hr), Hungarian (hu), Estonian (et), Serbian (sr), Macedonian (mk), and English (en). This set is comprised of two Uralic languages and four Indo-European languages. Despite some language pairs with significant similarity, a mixture of both Latin and Cyrillic script across the languages confounds the problem. The second task ($\text{Task}_2$) contains Javanese (jv), Indonesian (id), Malay (ms), Tagalog (tl), Tamil (ta), and English. With the exception of Tamil, the remaining languages are all part of the Malayo-Polynesian language family (subfamily of Austronesian) written with a Latin script. Tamil is a Dravidian language written with Tamil script. We also consider a combined set ($\text{Task}_{1,2}$) of all languages from both tasks. The breakdown of languages, size, family, and script is in Table 2.

When training MNMT models , training data is often balanced via upsampling (Wang et al., 2020). Upsampling helps improve performance in low-resource pairs. We are concerned with differences between techniques overall rather than optimizing model performance across pairs so we do not upsample the bitext and acknowledge that the model will underperform with some pairs.

### 3.2 Training

We train bilingual Transformer (Vaswani et al., 2017) models with 16k vocabularies to contextualize BLEU score ranges. The vocabularies are trained using SentencePiece[1] BPE (Sennrich et al., 2016). Multilingual vocabularies have been studied to optimize performance, manage model capacity, and help under-resourced languages (Chung et al., 2020; Zheng et al., 2021). These tasks have some differences in script and data balance so we used both a traditional BPE training method with no sampling and also used the union of the bilingual models as the vocabulary for the multilingual models[2]. The union of these vocabularies results in a combined 65k and 75k for the $\text{Task}_1$ and $\text{Task}_2$ languages respectively. Using these numbers, we choose to train the multilingual models with a 64k vocabulary. For hyperparameters, please see Table 5 in the Apendix.

[1] https://github.com/google/sentencepiece

[2] We do not find significant differences between the unioned vocabulary and the regular vocabulary with respect to prefixes so we only present the traditional vocabulary models here.

| | T \| ∅ | | | | | ∅ \| T | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 10E-2D | 8E-4D | 6E-6D | 4E-8D | 2E-10D | 10E-2D | 8E-4D | 6E-6D | 4E-8D | 2E-10D |
| en-et | **20.1** | 19.2 | 19.9 | 19.3 | 19 | 17.9 | 18.9 | **19.5** | 18.7 | 18.9 |
| en-hr | 24.1 | 23.3 | **24.6** | 24 | 23.1 | 22.5 | **23.8** | 23.4 | 23 | 23.4 |
| en-hu | **22** | 20.9 | 21.7 | 21.3 | 20.5 | 21 | **21.3** | **21.3** | 20.7 | 21 |
| en-mk | 21.4 | 21 | **22.6** | **22.6** | 21.4 | 18 | 21.3 | 21.6 | 21.4 | **22.2** |
| en-sr | 14 | 13.1 | 15.3 | **15.4** | 14.7 | 11.4 | 13.1 | 12 | **13.9** | 13.6 |
| et-en | **28.2** | 27 | 28.1 | 27.2 | 25.8 | 27.5 | **28.7** | 26.9 | 27.2 | 26.9 |
| hr-en | **30.7** | 29.8 | 30.2 | 29.3 | 28.3 | 30.1 | **30.9** | 29.5 | 29.4 | 29.1 |
| hu-en | 27.9 | 27 | **28** | 27.4 | 26.9 | 27.5 | **28.4** | 27.4 | 27.6 | 27.2 |
| mk-en | **29.8** | 28.5 | 29.5 | 28.5 | 27.3 | 29 | **30.1** | 29.1 | 27.9 | 27.5 |
| sr-en | 30.4 | 29.4 | **30.7** | 29.4 | 28.3 | 30.3 | **31.3** | 29.7 | 28.7 | 28.7 |
| Seen LID | 90% | 90% | 91% | 91% | 91% | 90% | 90% | 90% | 91% | 91% |
| Unseen LID | 63% | 52% | 54% | 57% | 25% | 0.10% | 0.20% | 2% | 5% | 16% |

Table 4: BLEU scores of models trained with varying depths—the number of encoder and decoder layers. Correct LID reports the percent the output was in the correct language (based on a CLD3 LangID model) in seen (supervised) and unseen (zero-shot) directions. Zero-shot directions are all non-English language pairs in $\text{TASK}_1$.

## 4 Results

### 4.1 Prefixing

With the three data settings ($\text{TASK}_1$, $\text{TASK}_2$, and $\text{TASK}_{1,2}$), we train models for each prefixing techniques. In Table 3, we present the BLEU[3] scores for the individual tasks ($\text{TASK}_1$, $\text{TASK}_2$) and select prefixing techniques from the combined ($\text{TASK}_{1,2}$) setting. We also compute statistical significant tests using paired bootstrapping (Koehn, 2004).

Prior work on zero-shot translations found that *only* "T | ∅" improved performance Wu et al. (2021). In supervised settings, we find that "S T | ∅" often performs as well if not better than "T | ∅." As the number of languages scale, "S2T | ∅" takes a remarkable edge over both of these methods—though this prefix has no equivalent in zero-shot translation. In general, the model benefits from source language tokens in supervised settings. It is logical that specifying both the source and target is better in supervised settings as the model has already seen these combinations of language tokens during training.

This all supports that source-side prefixing performs better than target-side. In $\text{TASK}_1$ (the upper-left section of the table), we see the source-side "S T | ∅", and "T | ∅" perfoming well with "∅ | S2T" being the only target-side equivalent. In $\text{TASK}_2$ (bottom left section), none of the target-side prefixes are competitive with "T | ∅" or "S T | ∅." In $\text{TASK}_{1,2}$ (right section), we display the source-side prefixes against the best-performing target-side prefix (∅ | S2T) which underperforms all source-side methods. Beyond performance, source-side prefixing is also desirable for speed as Transformer decoding times increase with target sequence length.

[3]scored using SacreBLEU

Lastly, we not that the form of the token (whether it denotes source, target, or language pair) depends on language set. "S2T | ∅" significantly outperformed alternatives in the $\text{TASK}_{1,2}$ setting but was outperformed by both "S T | ∅" and "T | ∅" in the single tasks. This effect may be due to the increased number of languages which are more diverse in both family and script than the original sets. Future work should consider how prefixing scales language sets increase to different quantities of languages.

### 4.2 Encoder and Decoder Depths

As the source-side prefixing techniques have an advantage, we additionally study whether these effects are multiplied by a strong decoder. We train additional models with twelve total layers, varying the depth of encoders and decoders with one source-side (T | ∅) and one target-side (∅ | T) prefixing strategy. Results are in Table 4.

We find that models with deeper encoders or an even-balance do better with both prefixes. Both prefixes benefited from deeper encoders, though depth varied. Neither benefited from deeper decoders—implying the prefixing technique is not heavily dependent on the depth of the encoder/decoder.

## 5 Conclusion

Prefixing strategies are wide and varied. Previous work focused on zero-shot settings while our work complements that by investigating supervised performance. Source-side prefixing performs better than target-side irrespective of encoder/decoder depth. Further, researchers should consider the number of languages in their set as the quantity, diversity, and balance of pairs may make some

prefixes more beneficial than others. Future work should consider more forceful prompting methodologies and experiment with how prefixes function with respect to language set scaling.

## References


Hyung Won Chung, Dan Garrette, Kiat Chuan Tan, and Jason Riesa. 2020. Improving multilingual models with language-clustered vocabularies. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4536–4546, Online. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Thanh-Le Ha, Jan Niehues, and Alex Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2017. Effective strategies in zero-shot neural machine translation. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 105–112, Tokyo, Japan. International Workshop on Spoken Language Translation.

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.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Xiang Kong, Adithya Renduchintala, James Cross, Yuqing Tang, Jiatao Gu, and Xian Li. 2021. Multilingual neural machine translation with deep encoder and multiple shallow decoders. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1613–1624, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *CoRR*, abs/2001.08210.

Muhammad N ElNokrashy, Amr Hendy, Mohamed Maher, Mohamed Afify, and Hany Hassan. 2022. Language tokens: Simply improving zero-shot multi-aligned translation in encoder-decoder models. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 70–82, Orlando, USA. Association for Machine Translation in the Americas.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural 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 244–258, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multilingual unsupervised NMT using shared encoder and language-specific decoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3083–3089, Florence, Italy. 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Xinyi Wang, Yulia Tsvetkov, and Graham Neubig. 2020. Balancing training for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8526–8537, 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.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *CoRR*, abs/2010.11934.

Bo Zheng, Li Dong, Shaohan Huang, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021. Allocating large vocabulary capacity for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3203–3215, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

# A Appendix

| Parameter | Value |
|---|---|
| Framework | Sockeye 2 |
| Encoder Layers | 6 |
| Decoder Layers | 6 |
| Model Size | 512 |
| Feed Forward | 1024 |
| Attention Heads | 8 |
| Dropout | 0.1 |
| Label Smoothing | 0.1 |
| Update Interval | 5 batches |
| Validation Interval | 750 updates |
| Early Stopping | 10 validations |

Table 5: Hyperparameters. We use Sockeye Recipes 2 to create reproducible training scripts. Recipes will be released upon publication.