A Study of Multilingual versus Meta-Learning for Language Model Pre-Training for Adaptation to Unseen Low Resource Languages

Jyotsana Khatri jyotsanak@cse.iitb.ac.in
Department of Computer Science and Engineering, IIT Bombay, India.
Rudra Murthy rmurthyv@in.ibm.com
IBM Research India
Amar Prakash Azad amarazad@in.ibm.com
IBM Research India
Pushpak Bhattacharyya pb@cse.iitb.ac.in
Department of Computer Science and Engineering, IIT Bombay, India.

Abstract

In this paper, we compare two approaches to train a multilingual language model: (i) simple multilingual learning using data-mixing, and (ii) meta-learning. We examine the performance of these models by extending them to unseen language pairs and further finetune them for the task of unsupervised NMT. We perform several experiments with varying amounts of data and give a comparative analysis of the approaches. We observe that both approaches give a comparable performance, and meta-learning gives slightly better results in a few cases of low amounts of data. For *Oriya-Punjabi* language pair, meta-learning performs better than multilingual learning when using 2M, and 3M sentences.

1 Introduction

Neural Machine Translation (NMT) works well with large amounts of parallel data (Vaswani et al., 2017). For many language pairs, such data is not available. Unsupervised NMT has achieved performance comparable to supervised NMT for a few European language pairs; however, it only works well for languages that have a good amount of monolingual data available. The current state-of-the-art approaches of unsupervised NMT have a language model pre-training phase and a finetuning phase based on iterative back-translation (Conneau and Lample, 2019; Song et al., 2019a; Lewis et al., 2019).

Translation involving low-resource languages (for which monolingual data is also scarce) is very difficult. The current state-of-the-art approaches of unsupervised NMT perform poorly for such language pairs (Kim et al., 2020; Marchisio et al., 2020). To utilize the benefit of high-resource language pairs, multilingual language model pre-training has been utilized (Lewis et al., 2020; Siddhant et al., 2020). Chronopoulou et al. (2020) proposed to train a language model for high-resource language pair and then use it as initialization for the low-resource language pair.

Dou et al. (2019) explored the use of meta-learning after pretraining with high-resource languages for low resource natural language understanding tasks. They claim that multi-task

Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track, pages 26–34
September 4–8, 2023, Macau SAR, China.

learning might favor high-resource tasks, while meta-learning learns a good initialization that can be adapted to any task with a small number of iterations.

In this paper, we use a meta-learning framework for multilingual language model pretraining and compare it with a multilingual learning paradigm based on data-mixing and finetuning it for unseen language pairs. We use these finetuned models to further training for the task of unsupervised NMT. Specifically, we utilize MAML (Model Agnostic Meta-Learning) Finn et al. (2017) which is a meta-learning algorithm based on gradient descent and is used to get good generalizations for multiple tasks. When using meta-learning, each language is considered a task in the pretraining phase. Our goal is to find a method to efficiently learn parameters in a shared parameter space across multiple languages in the language model pretraining, which works as good initialization for the language model training for unseen language pairs and improves the performance of unsupervised NMT. A good pretrained multilingual language model should be able to adjust to newer language pairs (unseen languages) using a limited amount of training data. Our contributions are:

- Comparison of two approaches of multilingual language model pre-training: (i) simple multilingual learning using data-mixing, and (ii) meta-learning. We compare these two approaches by extending them for unseen language pairs and further finetuning them for unsupervised NMT.
- We perform experiments with varying amounts of data for unseen language pairs and analyze the impact of different pretraining mechanisms.

2 Related Work

2.1 Unsupervised NMT

The initial works on unsupervised MT were based on statistical decipherment (Ravi and Knight, 2011; Dou and Knight, 2012, 2013; Dou et al., 2015, 2014). Decipherment assumes one language as cipher text and tries to generate the text in other languages.

Unsupervised NMT gained popularity after the initial proposals of Artetxe et al. (2018) and Lample et al. (2018) to train an NMT system without using any parallel data. These systems are majorly based on three things: unsupervised bilingual embeddings, denoising auto-encoders, and iterative back-translation. The first step is to learn bilingual embeddings in an unsupervised way by training two pretrained monolingual embedding spaces and aligning them using a linear transformation based on Procrustes refinement. Denoising auto-encoder aims to make the decoder learn to generate sentences. The Back-translation step involves generating synthetic parallel sentences using the current state of the machine translation model and using them to train the model in the opposite direction. This process of generation of synthetic parallel corpus and training is performed iteratively.

Current state-of-the-art approaches to unsupervised NMT involve a language model pretraining and a finetuning phase based on iterative back-translation. Different kinds of language modeling objectives have been proposed for the pretraining (Conneau and Lample, 2019; Song et al., 2019a; Lewis et al., 2019). Conneau and Lample (2019) (XLM) uses the Masked Language Modeling (MLM) objective, whereas Song et al. (2019b) (MASS) uses the Masked Sequence Generation objective. Lewis et al. (2020) proposed a language modeling objective similar to Song et al. (2019b), but it predicts the entire sentence on the decoder side and uses a different masking strategy. The architecture is based on a shared encoder and a shared decoder.

The success of unsupervised NMT depends on the model's capability to learn effective multilingual representations in the pretraining stage. Existing unsupervised NMT approaches fail for distant languages and languages with low amounts of data (Marchisio et al., 2020). Recently, many multilingual pretraining mechanisms have been proposed using similar masking objec-

tives but involving multiple languages, which were shown to perform better for low-resource languages (Liu et al., 2020; Conneau et al., 2019; Siddhant et al., 2020).
Recently few papers have also explored the use of in-context learning, instruction tuning with large language models (Chowdhery et al., 2022; Brown et al., 2020; Zhang et al., 2023; Moslem et al., 2023; Lyu et al., 2023; Peng et al., 2023; Karpinska and Iyyer, 2023; Wang et al., 2023; Jiao et al., 2023a; Zhu et al., 2023; Hendy et al., 2023; Garcia et al., 2023; Pilault et al., 2023; Vilar et al., 2022; Jiao et al., 2023b; Agrawal et al., 2022). Our work is not in the direction of in-context learning rather we are trying to find an optimal way of training a multilingual model based on its capabilities to be able to extend to unseen languages.

2.2 Meta-learning

Meta-learning solves the problem of fast adaptation to new training data. Gu et al. (2018) proposed an approach to apply meta-learning in NMT for low-resource language pairs. They use MAML (model agnostic meta-learning) to train a multilingual model that can be finetuned for new language pairs, this finetuning requires very few numbers of iterations, which is referred to as fast-adaptation. Sharaf et al. (2020) proposed an approach for domain adaptation based on a meta-learning framework, they use MAML and reptile for meta-learning. Qian and Yu (2019) propose to use meta-learning for domain adaptation. Nooralahzadeh et al. (2020) proposed to introduce MAML for cross-lingual language understanding tasks to effectively utilize training data of high resource and other auxiliary languages. The approach is to first train XLM using a high-resource language, followed by meta-learning using the low-resource languages, and final few-shot finetuning using low resource target language for the target task. Dou et al. (2019) explores the use of MAML for low-resource natural language understanding tasks.

3 Approach

We compare two multilingual language model pretraining approaches: (i) multilingual learning based on simple data mixing and (ii) other based on a meta-learning framework. We try to find a good set of initialization for language model pretraining for unseen language pairs using many high-resource languages. In multilingual learning, the training simply iterates between different languages. For meta-learning, we utilize MAML together with MASS Song et al. (2019b) objective to train a multilingual language model. The main aim of MAML is to find a good initialization from which a target task learning requires fewer iterations. It uses many other source tasks related to the target task to learn this initialization. We try to meta-learn using the source tasks and then continue to learn for the target tasks. This process is different than a simple multilingual learning framework. Algorithm 1 shows the training algorithm for the meta-learning framework. We extend both the models to finetune them for unseen language pairs and use the vocabulary extension method proposed in Chronopoulou et al. (2020) to extend the vocabulary of the multilingual model.

$$\theta = \theta - \alpha \sum_{T_i} \nabla L_i(f_{\theta_i^k}) \tag{1}$$

α is a hyperparameter, which represents the learning rate. The model is represented by a function f_θ with parameters θ. θ_i^k represents the state of the parameters when adapting to task T_i and here gradient update is performed using k examples. L represents the loss function.

4 Experiments

We experiment with *Hindi*, *Bengali*, *Gujarati* as our high resource languages to train a multilingual model using masked sequence to sequence pretraining objective. We use *Oriya-Punjabi*

Algorithm 1 Multilingual LM pretraining with MAML

Source tasks: L_1,L_2,...L_n
Target tasks: T_1,T_2
while true **do**
    for all Source tasks L_i **do**
        Compute θ_i^k using MASS objective
    end for
    Update θ as per MAML objective as per equation 1
end while

and *Assamese-Nepali* as our unseen language pairs. The details of the data are given in Section 4.1.

4.1 Dataset

We experimented using monolingual data provided by the AI4Bharat Kunchukuttan et al. (2020) dataset for the Indic languages, *viz, Hindi*, *Bengali*, *Gujarati*, *Punjabi*, and *Assamese*. We use *Nepali* monolingual dataset from common crawl corpus [1] Wenzek et al. (2020), and use the same amount of sentences equal to *Assamese*. The size of the data is given in Table 1. Our test data is taken from WAT2021 multi-indic-nmt shared task. The details of the dev and test data in Table 2. The dev and test data of *as-ne* is taken from FLORES-2021 dataset (Guzmán et al., 2019; Goyal et al., 2022). We convert all language data to same script (we choose devnagri as the common script which is an arbitrary choice) to reduce the vocabulary mismatch and have same lexical representations (Khatri et al., 2021).

Language	Number of Sentences
Bengali (bn)	7.21 M
Gujarati (gu)	7.89 M
Hindi (hi)	63.00 M
Oriya (or)	3.59 M
Punjabi (pa)	6.55 M
Assamese (as)	1.38M
Nepali (ne)	1.38M

Table 1: Monolingual data

4.2 Results

We train 3 types of models:

- **Bilingual**: Bilingual language model pretraining using only monolingual data of target language pair, followed by finetuning using iterative back-translation.

- **Multilingual**: Multilingual pretraining using masked sequence to sequence pretraining using high resource languages, followed by training for unseen language pair using same language modeling objective and then final finetuning using iterative back-translation.

[1] https://metatext.io/redirect/cc100-nepali

Language pair	Validation data	Test data
or-pa	1000	2390
as-ne	997	1012

Table 2: Validation and Test data

Data Size	Bilingual		Multilingual		Meta-learning	
	or → pa	**pa → or**	**or → pa**	**pa → or**	**or → pa**	**pa → or**
1M	1.2 ± 0.2	0.6 ± 0.1	6.9 ± 0.4	**3.3 ± 0.3**	**7.1 ± 0.4**	3.2 ± 0.3
2M	3.5 ± 0.3	2.3 ± 0.3	7.7 ± 0.4	4.1 ± 0.4	**8.5 ± 0.4**	**4.4 ± 0.4**
3M	4.6 ± 0.3	3.4 ± 0.3	8.3 ± 0.4	4.4 ± 0.4	**9.0 ± 0.5**	**4.9 ± 0.4**
Full data	5.2 ± 0.4	4.2 ± 0.4	9.8 ± 0.5	5.3 ± 0.4	9.8 ± 0.5	5.3 ± 0.5
Data Size	**Bilingual**		**Multilingual**		**Meta-learning**	
	as → ne	**ne → as**	**as → ne**	**ne → as**	**as → ne**	**ne → as**
0.5M	1.1 ± 0.3	1.0 ± 0.3	**2.2 ± 0.3**	**2.2 ± 0.3**	2.0 ± 0.4	2.1 ± 0.3
1M	2.5 ± 0.4	2.4 ± 0.4	3.0 ± 0.4	**3.0 ± 0.4**	3.0 ± 0.4	2.9 ± 0.4
Full data	2.6 ± 0.4	2.5 ± 0.4	3.0 ± 0.4	3.2 ± 0.4	**3.1 ± 0.4**	3.2 ± 0.4

Table 3: Test set BLEU scores for *Oriya-Punjabi* and *Assamese-Nepali* using Bilingual, Multilingual and Meta-learning approaches for language model pretraining

- **Meta-learning**: Multilingual pretraining using masked sequence to sequence pretraining with meta-learning framework explained in Algorithm 1, followed by the same process described in multilingual learning.

Our multilingual models are trained using *Hindi*, *Bengali*, and *Gujarati* for two approaches of multilingual language model pretraining one is based on data-mixing, and another one utilizes meta-learning. We use six layers in the transformer encoder and decoder, which is shared across all languages. The number of attention heads is 8. We use the toolkit provided by Song et al. (2019a) [2], and modify it for using MAML in the language model pretraining phase.

We also modify the codebase for vocabulary extension when finetuning a pretrained multilingual model for unseen languages. We use IndicNLP [3] library for tokenization and script conversion. The multilingual models are trained for 150 epochs, where epoch size is 0.2M sentences. The multilingual model is finetuned for 100 epochs using the data of unseen low resource language pair for MASS objective and then finetuned for 50 epochs using iterative back-translation. We report results in the form of BLEU score for our experiments in Table 3. The BLEU score is calculated using sacreBLEU (Post, 2018).

5 Discussion

Pretrained multilingual models help in improving the performance for unseen languages, which is clear from Table 3; all bilingual models have lower BLEU scores compared to models which have been initialized using multilingual pretrained models. When we use 2M, and 3M sentences for *or-pa*, we see minor improvements when using meta-learning over our baseline model.

[2] https://github.com/microsoft/MASS

[3] https://github.com/anoopkunchukuttan/indic_nlp_library

When we utilize full available data of *Oriya* and *Punjabi*, meta-learning performs similar to multilingual learning. But when we use 0.5M sentences, multilingual learning is working better than meta-learning for *or-pa*. For *as-ne* multilingual learning and meta-learning both give similar performance.
For *or-pa*, after the language model pretraining phase is complete for the unseen language pair, the cross-lingual perplexity is higher for meta-learning than the multilingual model but the BLEU score is better, which indicates that fluency is not getting better but the translation is getting improved indicating better learning of shared representations. We also observe that the ratio of source words is 3.27% for multilingual and 4.27% for meta-learning when experimenting with 2M sentences for *or* to *pa* translation even without finetuning it for iterative back-translation.

6 Conclusion and Future Work

In this paper, we perform a comparison of two approaches to train a multilingual language model: (i) simple multilingual learning, and (ii) meta-learning. We conduct experiments to extend these models for unseen language-pair and then finetune them for unsupervised NMT to compare the performance. We observe that both approaches give a comparable performance. In a few cases of low amounts of data, meta-learning gives slightly better results. In the future, we would like to explore the performance of both approaches to train the multilingual language model for other tasks.

References

Agrawal, S., Zhou, C., Lewis, M., Zettlemoyer, L., and Ghazvininejad, M. (2022). In-context examples selection for machine translation. *arXiv preprint arXiv:2212.02437*.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2018). Unsupervised neural machine translation. In *ICLR 2018, Proceedings of the Sixth International Conference on Learning Representations*, Vancouver, Canada. 12pp.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2022). Palm: Scaling language modeling with pathways.

Chronopoulou, A., Stojanovski, D., and Fraser, A. (2020). Reusing a Pretrained Language Model on Languages with Limited Corpora for Unsupervised NMT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2703–2711, Online. Association for Computational Linguistics.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems, Proceedings*, pages 7057–7067, Vancouver, Canada.

Dou, Q. and Knight, K. (2012). Large scale decipherment for out-of-domain machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 266–275, Jeju Island, Korea. Association for Computational Linguistics.

Dou, Q. and Knight, K. (2013). Dependency-based decipherment for resource-limited machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1668–1676, Seattle, Washington, USA. Association for Computational Linguistics.

Dou, Q., Vaswani, A., and Knight, K. (2014). Beyond parallel data: Joint word alignment and decipherment improves machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 557–565, Doha, Qatar. Association for Computational Linguistics.

Dou, Q., Vaswani, A., Knight, K., and Dyer, C. (2015). Unifying Bayesian inference and vector space models for improved decipherment. 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 836–845, Beijing, China. Association for Computational Linguistics.

Dou, Z.-Y., Yu, K., and Anastasopoulos, A. (2019). Investigating meta-learning algorithms for low-resource natural language understanding tasks. 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 1192–1197.

Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR.

Garcia, X., Bansal, Y., Cherry, C., Foster, G., Krikun, M., Feng, F., Johnson, M., and Firat, O. (2023). The unreasonable effectiveness of few-shot learning for machine translation. *arXiv preprint arXiv:2302.01398*.

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., and Fan, A. (2022). The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Gu, J., Wang, Y., Chen, Y., Li, V. O., and Cho, K. (2018). Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631.

Guzmán, F., Chen, P., Ott, M., Pino, J., Lample, G., Koehn, P., Chaudhary, V., and Ranzato, M. (2019). Two new evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. arxiv 2019. *arXiv preprint arXiv:1902.01382*.

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., and Awadalla, H. H. (2023). How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Jiao, W., Huang, J.-t., Wang, W., Wang, X., Shi, S., and Tu, Z. (2023a). Parrot: Translating during chat using large language models. *arXiv preprint arXiv:2304.02426*.

Jiao, W., Wang, W., Huang, J.-t., Wang, X., and Tu, Z. (2023b). Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Karpinska, M. and Iyyer, M. (2023). Large language models effectively leverage document-level context for literary translation, but critical errors persist. *arXiv preprint arXiv:2304.03245*.

Khatri, J., Murthy, R., Banerjee, T., and Bhattacharyya, P. (2021). Simple measures of bridging lexical divergence help unsupervised neural machine translation for low-resource languages. *Machine Translation*, 35(4):711–744.

Kim, Y., Graça, M., and Ney, H. (2020). When and why is unsupervised neural machine translation useless? *arXiv preprint arXiv:2004.10581*.

Kunchukuttan, A., Kakwani, D., Golla, S., N.C., G., Bhattacharyya, A., Khapra, M. M., and Kumar, P. (2020). AI4Bharat-IndicNLP Corpus: Monolingual Corpora and Word Embeddings for Indic Languages. In *5th Workshop on Representation Learning for NLP (RepL4NLP-2020)*, Online. 7pp.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2018). Unsupervised machine translation using monolingual corpora only. In *Proceedings of the Sixth International Conference on Learning Representations*, Vancouver, Canada. 14pp.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Lyu, C., Xu, J., and Wang, L. (2023). New trends in machine translation using large language models: Case examples with chatgpt. *arXiv preprint arXiv:2305.01181*.

Marchisio, K., Duh, K., and Koehn, P. (2020). When does unsupervised machine translation work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583, Online. Association for Computational Linguistics.

Moslem, Y., Haque, R., and Way, A. (2023). Adaptive machine translation with large language models. *arXiv preprint arXiv:2301.13294*.

Nooralahzadeh, F., Bekoulis, G., Bjerva, J., and Augenstein, I. (2020). Zero-shot cross-lingual transfer with meta learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4547–4562.

Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., Ouyang, Y., and Tao, D. (2023). Towards making the most of chatgpt for machine translation. *arXiv preprint arXiv:2303.13780*.

Pilault, J., Garcia, X., Bražinskas, A., and Firat, O. (2023). Interactive-chain-prompting: Ambiguity resolution for crosslingual conditional generation with interaction. *arXiv preprint arXiv:2301.10309*.

Post, M. (2018). A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Qian, K. and Yu, Z. (2019). Domain adaptive dialog generation via meta learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2639–2649.

Ravi, S. and Knight, K. (2011). Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 12–21, Portland, Oregon, USA. Association for Computational Linguistics.

Sharaf, A., Hassan, H., and Daumé III, H. (2020). Meta-learning for few-shot nmt adaptation. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 43–53.

Siddhant, A., Bapna, A., Cao, Y., Firat, O., Chen, M. X., Kudugunta, S., Arivazhagan, N., and Wu, Y. (2020). Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2827–2835.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. (2019a). Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. (2019b). MASS: Masked Sequence to Sequence Pre-training for Language Generation. In *ICML 2019, Thirty-sixth International Conference on Machine Learning, Proceedings*, pages 5926–5936, Long Beach, California, USA.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vilar, D., Freitag, M., Cherry, C., Luo, J., Ratnakar, V., and Foster, G. (2022). Prompting palm for translation: Assessing strategies and performance. *arXiv preprint arXiv:2211.09102*.

Wang, L., Lyu, C., Ji, T., Zhang, Z., Yu, D., Shi, S., and Tu, Z. (2023). Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., and Grave, É. (2020). Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012.

Zhang, B., Haddow, B., and Birch, A. (2023). Prompting large language model for machine translation: A case study. *arXiv preprint arXiv:2301.07069*.

Zhu, W., Liu, H., Dong, Q., Xu, J., Kong, L., Chen, J., Li, L., and Huang, S. (2023). Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.