MALM: Mixing Augmented Language Modeling for Zero-Shot Machine Translation

Kshitij Gupta
Department of Electrical and Electronics Engineering
BITS Pilani, Pilani Campus
Rajasthan, India
mailguptakshitij@gmail.com

Abstract

Large pre-trained language models have brought remarkable progress in NLP. Pre-training and Fine-tuning have given state-of-art performance across tasks in text processing. Data Augmentation techniques have also helped build state-of-art models on low or zero resource tasks. Many works in the past have attempted at learning a single massively-multilingual machine translation model for zero-shot translation. Although those translation models are producing correct translations, the main challenge is those models are producing the wrong languages for zero-shot translation. This work and its results indicate that prompt conditioned large models do not suffer from off-target language errors i.e. errors arising due to translation to wrong languages. We empirically demonstrate the effectiveness of self-supervised pre-training and data augmentation for zero-shot multi-lingual machine translation.

1 Introduction

Machine Translation is one of the classic problems in Natural Language Processing(NLP). Several products like Google Translate, Bing Translate provide services to millions of translation requests across a diversity of language pairs. While the requests for these services come in almost all language pairs imaginable, the quality of translation for low-resource language pairs like German-Arabic is especially low. This is prompted due to a lack of quality training data for such locale pairs compared to high-resource languages like English-French etc.

Specifically, for few-shot machine translation, there have been many successful techniques proposed. Zoph et al. (2016) demonstrated that transfer learning from high-resource languages to low-resource languages can be used to achieve remarkably high BLEU scores. Building on top of it, Gu et al. (2018) showed that universal lexical representations achieve better alignment of lexical and syntactic relations between languages. Similarly, (Fadaee et al., 2017) have been successfully used to utilize computer vision leanings in augmentation to low-resource translation.

Success due to techniques like transfer-learning, data augmentation, etc. has also provided great progress in building large multi-lingual neural machine translation models (Johnson et al., 2017). The objective here is to build a single high-capacity model that is able to generate translations for any language pair and can be trained at the same time. Zhang et al. (2021) have used conditional specific language routing for achieving impressive performance across low resource language pairs. Similar to this work, Xia et al. (2019) utilized data augmentation strategies and Zhang et al. (2020) used random online back translation to achieve state-of-the-art performance for low-resource machine translation. In their work, Arivazhagan et al. (2019a) encourage parameter sharing across language by implementing an auxiliary loss function. Similarly, de-noising objective (Liao et al., 2021) and distillation techniques (Sun et al., 2020) have also been shown to have boosted zero translation learning.

Recently, research direction in massively multilingual translation models(MMT) (Aharoni et al., 2019) has also been popularized to build zero-shot translation systems. Arivazhagan et al. (2019b) provides a survey of challenges associated with MMT models, while also emphasizing the importance of preprocessing and vocabulary in knowledge transfer across language pairs. Although, Gonzales et al. (2020) detail the lack of robustness of zero-shot models across training runs, we do not notice it in our training runs and find that augmentation techniques help stabilize the training process.

Neural models like Transformer (Vaswani et al.,

The 2nd International Workshop on Natural Language Processing for Digital Humanities (NLP4DH), pages 53–58
November 20, 2022. ©2022 Association for Computational Linguistics

2017) have brought significant advances in tasks across NLP. Pre-trained language models like BERT (Devlin et al., 2018), BART (Lewis et al., 2019), T5 (Raffel et al., 2019) have achieved state-of-art performance across the NLP spectrum. Similarly, generative models like GPT-2 (Brown et al., 2020) have shown few and zero-shot abilities on many tasks. The wide success across tasks has not only been limited to high-resource languages like English, and French but has indeed been shared with low-resource languages like Azerbaijani, Belarussian, Galician, Urdu, etc (Lakew et al., 2021). It has also brought significant progress in single large multi-lingual models mBERT (Devlin et al., 2018), mBART (Liu et al., 2020), mT5 (Xue et al., 2020) that learn universal representation across languages. This is responsible for remarkable zero and few-shot performance across tasks for languages that lack supervised training data.

In this paper, we will investigate language modeling pre-training and data augmentation strategy for zero-shot translation Our work provides 2 major and concise pieces of contributions:

1. *Prompt Conditioned* models like mT5 do not suffer from off-target translation and a language tag in the task prompt is sufficient for the model to generate output in the right language.

2. *SeqMix* style data augmentation technique on top of large pre-trained language models like mT5 is a simple yet competitive approach against a strong baseline on zero-shot translation.

2 Problem

We will first look at the challenge of off-target translation. For all the zero-shot language pairs, we construct a random test dataset of 1000 examples[1] from the source language that may or may not be part of the original test dataset. We then run the translation system over those examples and identify all the output that corresponds to the wrong target language.

Say, a translation model(M) generates for data instances($x_1 ... x_n$) translations as ($y'_1 ... y'_n$) and reference translations as ($y_1 ... y_n$) for source language(s) and target language(t). Also, given a language identification oracle as L, where $L(x)$ is the predicted language M for data instance$x$. In this work, we will utilize Salcianu et al. (2016)'s Language Identification system to measure language performance. We then describe *off-target translation error rate*(OTTER) as:

$$OTTER(M) = \frac{\sum_i L(y'_i) \neq t}{\sum_i L(y_i) = t} \quad (1)$$

In the original paper, Zhang et al. (2020) used the accuracy of translation language as a metric to compare. We argue that any language identification system is noisy and thus accuracy on just translation output doesn't take into account errors of the language identification system. OTTER, on the other hand, is a noisy measure that measures language accuracy over both reference and translation output text of the translation system.

The main question that we investigate is improving the quality of zero-resource translation. The problem at hand is learning a single model that is able to learn translation across language pairs that are unseen during the training time. This is motivated by human language learning experience, that if a person knows German, English, and Arabic and can translate over German $\rightarrow$ English and Arabic $\rightarrow$ English then they should be able to translate with sufficiently good quality between German $\rightarrow$ Arabic without any formal training. This is true for us because beneath all the lexical and grammatical differences across languages, we share the grounding of various concepts in the same representation. Basically, the representation of the concept 'cat' is the same as the word 'Katze' in German.

Traditionally, a pivot language has been prominently used to achieve this task. For example; if a German-to-Arabic translation is required then the original text is passed through the first German-to-English translation engine, and then the output English sentence is passed through the English-to-Arabic translation system.

This simple yet effective strategy has been shown to achieve state-of-art performance across zero-resource settings in many language pairs. Currey and Heafield (2019) use data augmentation with pivot language to generate pseudo-parallel data across zero-shot language pairs and then re-train a system. Recently, Dabre et al. (2021) have even utilized multi-pivot languages and simultaneous translation as a method to improve zero-shot performance. While Kim et al. (2019) combined pivoting with transfer learning and an adapter module, Siddhant et al. (2020) leveraged monolingual data

[1]We choose 1000 as a good compromise across languages irrespective of their original test size

with self-supervision for low resource languages to achieve impressive performance.

3 Experimental Setup

For the purpose of this study, we will constrain the study and experiments to OPUS-100 by Zhang et al. (2020). OPUS-100 is an English-centric dataset with over 100 language pairs that have either source/target language as English. It also consists of several other non-English-centric pairs that are available for zero-shot translation objectives. We should emphasize that while previous and related work on this dataset has been centered around massively-multilingual translation as well as zero-shot translation, the objective of the current work is only on zero-shot translation. As part of zero-shot languages, OPUS-100 provides 15 language pairs that are combinations of French, German, Arabic, Russian, Chinese and Dutch. For the evaluation of our translation model, we shall use the BLEU score, which is a standard metric of automatic evaluation across machine translation.

We run our experiments using the mT5 implementation available in the Transformers library provided by HuggingFace (Wolf et al., 2019) and use pre-trained mT5 models(small, large, and xx-large) from Xue et al. (2020). For reproducibility purposes, we use the Adam optimizer and run our experiments on a Google TPU v2-32 instance for 64,000 steps with 256 max length, 512 batch size, 0.0001 learning rate and we use a beam size of 4 at inference time.

For baseline experiments, we consider 2 strong baselines – (i) First, for any language pair say XX-YY, we train 2 Transformer models XX-En and En-YY, and run zero shot inference for any new sentence from language XX through XX-En and then through En-YY, note that we do not pre-train these models (ii) Second, we consider the model from Zhang et al. (2020) which implements random online back translation to recover from off-target translation.

4 Methodology

4.1 Large Pre-trained Model with Prompt Conditioning

First, we provide a brief introduction to the T5 architecture. T5 or Text-to-Text Transfer Transformer is a recently introduced framework that frames all the NLP tasks as a text-to-text problem. While the model architecture is vanilla Transformer architecture (Vaswani et al., 2017), it has been pretrained on the C4 dataset (Raffel et al., 2019) on a Masked Language Modeling objective (Devlin et al., 2018). Any new task could be provided as a brief prompt to the model along with the input, for example, translation from German to English could be specified as `translate German to English: This is a test input sentence`, while the output is generally not formatted. This is the default prompt style used by (Raffel et al., 2019) as well as by this work.

Recently, Xue et al. (2020) introduced the multilingual version of this model which is pre-trained on the mC4 dataset. They have shown that mT5 exhibits zero-shot capabilities, as learning a task in one language is directly transferable to the same task in a different language without any further training. They also highlight that the model suffers from unexpected translation in the output space. For example, a model trained on English Part-of-speech and when inference is run on French input outputs an English translation of the French input. This issue is similar to what (Zhang et al., 2020) has suggested that the massively multilingual translation models suffer from. The lack of language signals to the model results in although correct output but in an incorrect target language.

4.2 Data Augmentation

We run experiments on 2 main techniques in this work:

- *Sentence Concatenation*
- *Seq2Mix*

In the first set, if our objective is to run translation from German to Arabic, then at each training time, we choose a data point from the German-English dataset and a random data point from the English-Arabic dataset. We then concatenate the source sentence and target sentence for both language pairs with a simple `<sep>` token. This results in a training sentence whose input has the first half as a German sentence and the second half as an English sentence. Similarly, the output has the first half in English and the Second half as Arabic sentences. We modify the input prompt slightly here to identify the languages present in the new input and output, as `translate German and English to English and Arabic`. We hypothesize that the model is intelligent enough to pick up the right words and context from the

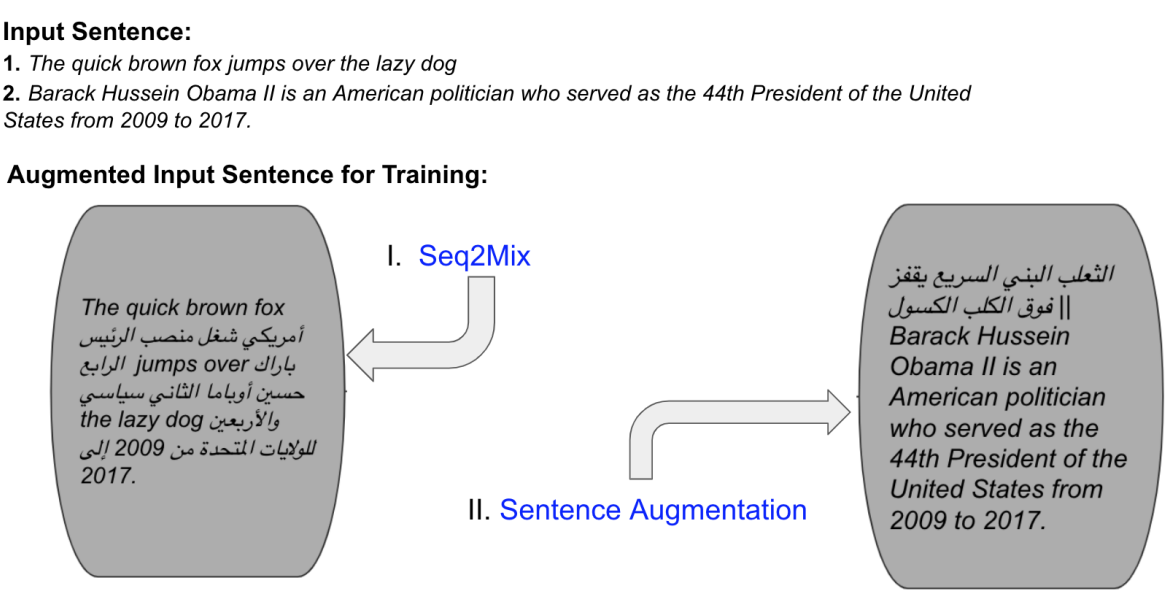

Figure 1: Example of Data Augmentation

prompt to produce the right output.

Secondly, we used Seq2Mix as introduced by Guo et al. (2020). They propose 2 variants of the Seq2Mix algorithm – hard and soft Seq2Mix. For the purpose of this work, we will only focus on the hard version. In essence, for 2 input sentences of equal sequence length[2], from German $(Gx_1, ..., Gx_n)$ and English $(Ex_1, ..., Ex_n)$, we construct a German-English sentence $(GEx_1, ..., GEx_n)$, where GEx_i is a token taken randomly from either (Gx_i, Ex_i) with a sample probability from Binomial(λ).[3] A similar process is run over to obtain an English-Arabic sentence which serves as output to the German-English input.

Figure 1 depicts both the augmentation techniques for the input sentences in the same. It is noteworthy that we need not merge sentences that are similar to each other, thus we could select any two sentences from our dataset for creating the augmented data. All of these data augmentation strategies work to create synthetic datasets that are employed along with the original bilingual datasets at training time with equal weighting. We hypothesize that while the model learns translation from both synthetic and original datasets, the prompt conditioning along with the mixed vocab is learned better at training time.

[2]otherwise use padding to ensure the sequences are of equal length

[3]where λ itself is sampled from $\beta(0.5, 0,5)$

5 Results & Analysis

We find that a mixed data augmentation training regime helps bring down OTTER to an extremely low range as referenced in 1. We attribute this to the compositional relationships learned by the large language model on the mixed vocabulary as well as the language tag we use as part of the task prompt.

	OTTER	BLEU
Transformer + Pivot	-	12.98
Zhang et al. (2020)	-	14.78
Ours (small mT5)	27.1%	4.9
Ours (large mT5)	24.2%	5.1
Ours (XXL mT5)	19.4%	7.2
XXL mT5 + input concat	0.9%	15.4
XXL mT5 + Seq2Mix	**0.7%**	**15.7**

Table 1: OTTER and BLEU scores for zero-shot language pairs; results are average across all the 15 language pairs in the zero-shot setting

Similarly, we find that data augmentation techniques like Seq2Mix (Guo et al., 2020), can substantially improve zero-shot performance when used on top of large pre-trained language models. We explain the performance using the following reasoning:

- Mixing vocabulary in the same sentence during training force the internal representation of tokens to align themselves in similar clusters across languages.

- Using XX-English and English-YY translation objective along with data augmentation smoothens the loss landscape to facilitate better representation for zero-shot translation

6 Conclusion

In this paper, we utilized mixing augmentation techniques along with large sequence-to-sequence models to generate high-quality zero-shot translation models for language pairs that have no training data available. We successfully demonstrate that large pre-trained language models are able to learn the semantic spaces between languages and are already good at zero-shot machine translation. However, data augmentation techniques can further boost this performance to achieve impressive results on zero-shot translation.

7 Future Work

We plan on running further experiments with improved data augmentation strategies at pre-training time which we think will benefit downstream zero-shot translation.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. *arXiv preprint arXiv:1903.00089*.

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*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Anna Currey and Kenneth Heafield. 2019. Zero-resource neural machine translation with monolingual pivot data. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 99–107, Hong Kong. Association for Computational Linguistics.

Raj Dabre, Aizhan Imankulova, Masahiro Kaneko, and Abhisek Chakrabarty. 2021. Simultaneous multi-pivot neural machine translation. *arXiv preprint arXiv:2104.07410*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv:1705.00440*.

Annette Rios Gonzales, 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.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. 2018. Universal neural machine translation for extremely low resource languages. *arXiv preprint arXiv:1802.05368*.

Demi Guo, Yoon Kim, and Alexander Rush. 2020. Sequence-level mixed sample data augmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5547–5552, Online. Association for Computational Linguistics.

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

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-english languages. *arXiv preprint arXiv:1909.09524*.

Surafel M. Lakew, Matteo Negri, and Marco Turchi. 2021. Zero-shot neural machine translation with self-learning cycle. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 96–113, Virtual. Association for Machine Translation in the Americas.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Junwei Liao, Yu Shi, Ming Gong, Linjun Shou, Hong Qu, and Michael Zeng. 2021. Improving zero-shot neural machine translation on language-specific encoders-decoders. *arXiv preprint arXiv:2102.06578*.

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. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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. *arXiv preprint arXiv:1910.10683*.

Alex Salcianu, Andy Golding, and Anton Bakalov. 2016. Compact language detector v3. *https://github.com/google/cld3*.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. *arXiv preprint arXiv:2005.04816*.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2020. Knowledge distillation for multilingual unsupervised neural machine translation. *arXiv preprint arXiv:2004.10171*.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. *arXiv preprint arXiv:1906.03785*.

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. *arXiv preprint arXiv:2010.11934*.

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

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.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.