Breaking the Curse: Is Few-Shot Semi-Supervised Learning the Answer to Multilingual Neural Machine Translation?

1 Introduction

Multilingual Neural Machine Translation (MNMT), where a single model can translate from multiple source languages into multiple target languages, is an active area of research with the end goal of removing global language barriers. To put this into perspective, there are 7000 different languages worldwide, roughly 50% of the Internet's content is in English, and only 1 in 5 of the global population are English speakers. Thanks to recent advancements in MNMT, models have achieved impressive performance following supervised and unsupervised training. However, both approaches suffer significant drawbacks. Supervised methods rely on large amounts of parallel corpora (direct translations between all languages and directions) which take a long time to train and are often non-existent between extreme languages (e.g. Portuguese to Vietnamese). On the contrary, unsupervised methods require large amounts of domain-matched monolingual data which is scarce.

More significantly, both methods suffer from capacity bottlenecks when supporting multiple languages [13]. Such phenomenon has been termed as the *Curse of Multilinguality*, first coined by Conneau et al. [16]. This is because, for each language added, there is increased interference and less capacity available to learn representations for each language. Whilst most MNMT literature attempts to scale English-centric models into the hundreds of billions of parameters, the thousands of low-resource languages (LRLs) [1] are being left behind. For example, Google Translate supports Icelandic which has roughly 330k L1 speakers, but it doesn't support Kirundi which is spoken by over 11 million people globally. Additionally, in cases where a LRL *is* supported, the resulting translations are often inaccurate as their representations are dominated by HRLs, such as English [22].

This paper proposes that this problem can be solved by using a hybrid approach (i.e. semi-supervised) to produce an MNMT model that could a) translate into any language, even if it hasn't been seen during training, b) provide some level of control over the style of the resulting translation, and c) doesn't fall subject to the aforementioned curse.

2 Related Work

2.1 Before NMT

Rule-based machine translation (RBMT) was the earliest major attempt at machine translation which used human-written rule-based methods that were labour-intensive and lacked transferability between languages, leaving little opportunity to scale to multiple languages. The next major milestone was Statistical Machine Translation (SMT) which was trained by finding basic statistical patterns from parallel data [36]. It wasn't until 2014 that today's widely adopted NMT sequence-sequence representation was put into practice, thanks to the advent of deep learning.

2.2 Sequence-to-Sequence Representation

seq2seq [62] was the first paper to demonstrate the potential of deep NMT, setting the foundation for the sequence-to-sequence representation of inputs. At the time, it consisted of two Recurrent Neural Networks (RNNs): an encoder to encode the input sentence into a single vector and a decoder that generates the output sequences conditioned on the encoder's output. These models were trained on pairs of sequences (supervised), conditioning on an input sequence to produce an output. This meant that sentences of varying lengths could be mapped to one another. This was superior to RBMT and SMT as it removed the need for any human-designed rules or feature extraction. However, this approach was flawed by the fact that the encoding of the source sentence needed to capture all information about it in one vector so that it could be moved to the decoder, which proved challenging for long sentences (commonly known as *information bottleneck*). Variations such as Long Short-Term Memory (LSTMs) and Gated Recurrent Units (GRUs) proposed improved long-term dependencies, but this didn't avoid the fact that the most recent inputs were always prioritised which was undesirable.

[1] Defined here as languages with $<$1M publicly available translations. High-resource languages (HRLs) are defined as having $>$100M.

This issue gave rise to Attention [43] providing a direct connection to the encoder on each step of the decoder, allowing for focus on a particular section of the source sequence. This meant that all words in a source translation had equal contribution to the target translation and it was the model's job to learn the importance of each word. This worked effectively on models that covered <6 languages. However, when dealing with more languages, the models became overwhelmed because this naive approach would necessitate training $O(N^2)$ individual models, where N represents the number of supported languages.

2.3 Scaling to a Multilingual Level

This was until 2017 when Johnson et al. [33] took their original LSTM model [68] and appended a label to all source sentences with the desired target language, forcing the decoder to translate to that language. For example, a sentence to be translated to Spanish would be labelled as <2es>. There was no label for the source language so that the model could adapt to any input with code-switching. Using an LSTM architecture, this shared-parameter model using just a single encoder-decoder architecture [30], was able to scale to 12 languages through transfer learning. Thanks to zero-shot learning (ZSL), it could also translate between language pairs that were never seen during training through pivoting/bridging (i.e. two-step supervised translation through the pivot language). Currey et al. [19] improved the accuracy of this approach by adding monolingual data to the pivot language, but expressed that this would not be suitable on a multilingual level.

Unlike previous attempts which had limited scalability due to having a dedicated encoder/decoder for each language [23], Google's ZSL paper [33] was the first to explore using a single encoder/decoder architecture across multiple languages, promoting cross-lingual transfer. Their attempt relied on English as the "centre" language (i.e. training on data which has translations from or to English) given that it has the most available data. Whilst this saw good performance into English, the reverse was not as promising due to an imbalanced dataset. The authors therefore suggested over-sampling the non-English target pairs for future work. The most influential attempt at resolving this imbalance was using Temperature Sampling where the probability of a language p_l is replaced with $p_l^{\frac{1}{T}}$ [3]. Alternatively, one could sample sentences according to a multinomial distribution, offering an increased number of tokens associated with LRLs and bias-alleviation towards HRLs [41].

2.4 Transformers & Self-Attention

Whilst this early MNMT attempt [33] was highly successful, it was released just before the substantial introduction of Transformers [67], which quickly outperformed LSTMs [38, 40, 56], given their ability to understand context and give equal prioritisation to all tokens being passed in. This was in combination with a single self-attention mechanism which is shared across both the encoder and decoder. Within this, layers were stacked to improve translation quality. They also have multiple attention heads which is crucial for understanding the context of the entire sentence being passed in.

Another important consideration was deciding how to share scripts across languages. Google's ZSL attempt [33] used a single script to cover all languages (therefore all parameters are shared) as opposed to the other end of the spectrum where you have a separate script for each language [30]. Both approaches have their drawbacks where a single script tries to merge unrelated languages, but separate scripts don't permit the sharing of embeddings between similar languages. Alternatively, a middle-ground approach where similar languages are clustered together and trained separately has been found to improve translation accuracy [63], especially for LRLs [47]. This ensures that the word embeddings of the two source languages are put into similar vector spaces. Note that "similar" in this context can refer to the similarities identified by linguistic experts or the embedding similarities. This benefits LRLs in situations where vocabularies are shared (e.g. there are about 100k Nepali sentences on Wikipedia but 600k in Hindi). Grouping the two means that their 80% token crossover can be merged into a shared vocabulary of 100k subword units [41].

Fan et al. [22] built on this with a novel random rerouting scheme where, after grouping similar languages into their own sublayers, they randomly picked another sublayer instead of the designated one to encourage sharing of information between languages, in turn benefiting LRLs by training on similar HRLs. This saw an improvement of +0.8 BLEU on low & mid resource languages.

2.5 Tokenization

The method of tokenization concerns choosing how to represent a language in an embedding space. These include splitting by word [62, 12, 6] which struggle to capture the intricacies of a language. On the contrary, character-level methods require more computational power due to the resulting long sequence lengths and they also produce less meaningful individual tokens. Thus, *subword tokenization* [58] has been widely adopted as a middle-ground solution; if two characters often appear together, they are merged into a subword. This is achieved using the Byte-Pair Encoding algorithm [26] which iteratively merges the most frequent pair of characters into new subwords. A hurdle with such an approach on a multilingual level was representing languages that don't have clear spacings between words (e.g.

Chinese). Following this, success was found using bespoke tokenizers for each language as opposed to a multilingual tokenizer [55]. This is unsurprising given that monolingual tokenizers are typically trained by native-speaking experts who can identify phenomena present in their language. Alternatively, SentencePiece [37], an implementation of subword tokenization, is designed to work on languages with no segmentation and is one of the most popular algorithms used in state-of-the-art (SOTA) models [22, 44].

2.6 Unsupervised Learning

Whilst the fully supervised approach discussed so far dominated NMT literature, unsupervised approaches have also been explored [42, 4] which have the benefit of only requiring monolingual data (which is plentiful). When their pre-trained models are strong enough, they perform well on HRLs in the direction xx $\rightarrow$ yy, where yy is the language the model is trained on. However, they typically struggle with yy $\rightarrow$ xx [2]. To combat this, back translation [57], or more recently, iterative back translation [32] has been used to augment a training dataset with synthetic data created by translating monolingual sentences in the backward direction. This works well for MNMT in some situations as it supports the translation of each language in both directions and can be used to improve directions which have initially lower translation quality. However, doing this for all languages is intractable given its computation expense (the dataset grows quadratically with each language added). Additionally, the resulting examples often lack diversity [9]. Those who have attempted this strategy on a multilingual level have only performed it on a subset of LRLs [22].

2.7 Harnessing the Power of Language Models

Today's SOTA MNMT follows the emergent behaviours of LMs for downstream tasks, such as NMT. These are trained using self-supervised learning, borrowed from computer vision domains, where patterns and constructs of a language can be learnt from unlabeled data [5, 10, 31]. Instead of overfitting discriminative patterns between data during pre-training, this method focuses on capturing the intrinsic structure of the data so that we can generalise to unseen languages. To do this, we pre-train an LM with a self-supervised objective so that it can gain an understanding of language structure and nuances. This involves noising large amounts of monolingual data, for which there are several existing techniques, and feeding it as the source task with the original monolingual data as the target task. Below is a brief list of some self-supervised objectives and cited implementations, showing how LMs learn cross-lingual semantics.

Causal Language Modeling (CLM) involves a model learning to predict the next token in a sequence given the previous tokens.

Masked Language Modelling (MLM) / Denoising Autoencoder (DAE) introduced by BERT [20], sets the target as a sentence from a monolingual corpus. The source is a noised version of the target sentence. The aim is to maximize the likelihood of predicting the target given the noised source. Note that combining MLM and CLM objectives is not recommended to avoid interference [48].

Translation Language Modeling (TLM) [41] is an extension of the MLM approach where batches of parallel sentences are used instead of consecutive sentences. Words in both the source and target sentences are randomly masked.

Span corruption (variation of MLM) [52] presents a model with a source and target language sentence pair. A specific section of the text is masked in one of these sentences, and the model then predicts the missing words in the masked sentence. This strengthens the model's ability to represent and align semantic information between languages. This objective was also used by a closely related work [28].

The above list isn't exhaustive; whilst there are many other multilingual objectives and LMs with their own advantages and drawbacks [17, 11, 8], they all share a similar encoder-decoder structure (with the majority using some variation of the MLM objective) leaving little reason for further discussion. For our purposes, any pre-trained sequence-to-sequence LM can be used, rendering the differences trivial. Performance comparisons between different objectives are left for future work.

2.8 Semi-Supervised Learning

Semi-supervised learning proposes a blend of the best of both worlds: pretraining with a self-supervised objective on monolingual data to initialise NMT training, followed by fine-tuning on task-specific parallel data. It has been discovered that models trained in such a way reverse the curse, where adding more languages improves the model's ability to generalise [60, 39, 59], especially for LRLs [64, 29]. This is because they can learn an *interlingua* (shared semantic representation between languages). Furthermore, such models are magnitudes in size smaller than their fully supervised counterparts [28] as this transfer learning approach means that they don't require training on parallel corpus in every language and direction.

2.9 Few-Shot Learning

FSL at inference allows a model to generalise to an unseen language using just a few examples. For example, say a model can translate between English, French and German in all directions. If you wanted to be able to translate from Spanish to English, for instance, then you would provide a few high-quality Spanish parallel examples. FSL has had notable triumph on Transformer-based LMs in recent years [8] following Lu et al. [45] who state that, unlike supervised pre-training which focuses on the distinct features of the seen classes, a self-supervised model displays less bias towards such classes and is instead able to generalise to unseen classes by providing just a few examples at inference. At the same time, Garcia et al. [28] highlight that this approach is highly dependent on the quality of the FSL demonstrations. To find these, beam search is popularly used which, in an NMT context, finds the sentence with the highest estimated probability as the prime example. However, this assumption has been questioned by some [25, 21], who propose minimum Bayes risk (MBR) decoding that additionally uses a risk factor in the decision-making process. FSL also proposes some controllability on the style/formality of the translation [28]. Whilst this is largely subjective in English, many other languages have separate vocabularies for representing different levels of formality. For example, the use of *vuvoiement* in French.

The closest related attempt only operates on a bi and trilingual level [28]. It also uses a decoder-only architecture where the source and target sentences are concatenated during training. We would instead use the widely adopted (for NMT) encoder-decoder structure.

3 Requirements

3.1 Measuring success

BLEU [49] is the most commonly used metric for evaluating the quality of machine-generated translations by comparing how similar a machine-translated text is to a human-generated reference translation. It calculates a similarity score based on n-gram precision where n refers to the number of words in a sequence. Whilst useful, it is imperfect [47, 35, 24] as a good translation can get a poor BLEU score because it has a low n-gram overlap with human translation. Additionally, the score can also vary based on the domain. For example, Arivazhagan et al. [3] show a 3-5 BLEU difference between the WMT dev and test sets for the same language pair.

Additionally, COMET [53] would also be used as it considers semantic quality meaning that fluency, adequacy and preservation of meaning are also taken into account. The model scores between 0 and 1 for low and high quality translations respectively.

Translation Error Rate (TER) and chrF could also be included for a more diverse evaluation. TER [61] measures the number of editing operations required to transform the system's output into the reference translation. chrF [50] is not tied to a specific language and has been shown to capture intricate structures more accurately.

In addition to this, human evaluation would also be used to understand the quality of the machine translations. Each evaluator would be able to score a translation on a scale from 1-10 on this model and a SOTA multilingual competitor as done by [22]. This is particularly crucial in this environment where FSL has been found to produce a different style of translations that are still semantically correct but appear much lower on metrics like BLEU [28].

Note that translations would be measured in both directions. For example, when measuring English → French, French → English would also be measured.

4 Solution Sketch

To summarise, the proposed model would be able to learn strong linguistic representations following pre-training on monolingual data which is available for most languages. The model would then be fine-tuned for the downstream task (i.e. Machine Translation) using parallel data, where available. In the case where parallel data isn't available for a language, it is usually available for a neighbouring language (i.e. one that possesses similar linguistic properties). Then, at inference, the model would be shown five examples of parallel data for the source-target language, plus the source sentence. It would then use both of these to produce a high-quality translation.

4.1 Architecture

In our example, mBART [44], a sequence-to-sequence DAE pre-trained on large-scale monolingual corpora in many languages, is adopted in a similar setup to Thillainathan et al. [65]. This architecture has been chosen arbitrarily; any pre-trained multilingual encoder-decoder Transformer could be used to initialise NMT training. However, one should be careful choosing an LM as the training times would vary significantly with parameters ranging from 130 million to >13B for some LMs (e.g. mBART is a sensible choice with 611M parameters, though any smaller could lead to underfitting).

ID	Objective	MOSCOW	Justification
1	Improved accuracy over MNMT using back translation.	MUST	Back translation is used in nearly every NMT system to reach optimal performance through the computationally expensive augmentation of parallel data.
2	Match accuracy of bi+tri-lingual FSL models [28] on WMT.	MUST	This is the best-performing FSL NMT model which only supports a maximum of three languages. It is hoped that our attempt would scale multilingually without a drop in performance.
3	Match performance of SOTA models on >100 languages.	MUST	Whilst most MNMT models stop at around 100 languages [3, 22, 69], the FSL capabilities should allow for the support of any language, provided that it or a neighbouring language has been seen during training. For extreme LRLs, this could only be measured using human judgement.
4	Have fewer parameters and reduced training time over SOTA models.	SHOULD	Whilst the overall goal is LRL support, it is important that the model isn't unreasonably large to support its real-world deployment and distribution.
5	Matched performance on LRLs vs HRLs.	SHOULD	Using a similar approach to Arivazhagen et al. [3], performance would be compared across languages of varying parallel data (e.g. group into the top 25 HRLs and the bottom 25 LRLs) to verify a balanced model.
6	Superior LRL performance against fine-tuned multilingual sequence-sequence models.	SHOULD	Diverse, high-quality, balanced data collection and FSL should mean that translation accuracy is improved for LRLs.
7	Formality and stylistic control over the target translation.	SHOULD	An emergent behaviour of using FSL at inference for NMT is that the model can match the style and formality of the examples provided [28].
8	ZSL capabilities.	COULD	The model may be able to generalise to new languages without providing any examples if the pre-training and fine-tuning stages are strong enough, though this isn't the aim of this paper.

Table 1. Requirements

4.2 Training Data

The model must be trained in a data-centric manner to ensure high-quality representations of languages. To achieve this, data would be collected from a diverse range of sources and domains. Both monolingual and parallel data would be required for pre-training and fine-tuning respectively. Following success in related literature [59], we use our self-supervised objective approach but also append a target language label (e.g. <2es> for Spanish) [33] as opposed to separate source and target language embeddings, so that the model can distinguish between source and target sentences. This labelling method has been shown to work on multilingual pre-training [7] and is supported by mBART [2].

4.2.1 Monolingual data for pre-training

This dataset aims to provide a vast amount of text data in various languages. This broad exposure would help the model learn the fundamental structures and vocabulary of each language. Aimlessly crawling the internet would be avoided as a lot of it is machine-translated so may not be accurate [66]. To ensure that LRLs aren't underrepresented, sentences would be sampled according to a multinomial distribution as previously mentioned [41].

- Common Crawl: covers 100+ languages as opposed to Wikipedia articles which cover fewer languages and aren't long enough for our purposes [16].
- Open-source dataset from M2M-100: the largest non English-centric LM [22].
- WMT monolingual data (e.g. news-crawl, news-commentary, common-crawl, europarl-v9, news-

[2] In the format [lang-code] X [eos], where [lang-code] is source language id for source text and target language id for target text, with X being the source or target text respectively.

discussions etc.) [70].

The monolingual data would be cleaned by removing duplicates and discarding sentences with more than 50 tokens to improve pattern recognition and reduce training times [54].

4.2.2 Parallel data for fine-tuning

- WMT: the highest quality parallel dataset available. However, it only covers a limited number of languages. WMT2023 wouldn't be used to prevent test/train overlap.
- TED Talks dataset: covers 59 languages and isn't English-centric [51].
- OPUS-100: a dataset of 55M translations covering 100 languages as used by Zhang et al. [69].
- The Bible would be used to fill any gaps as it covers 30k sentences in over 1000 different languages [47].

Parallel data would also be cleaned using Bifixer and/or Bicleaner [46].

4.3 Tokenization

Similar languages would be grouped as previously discussed. This would be done in a similar fashion to the proposed method by Chung et al. [15] where languages are clustered based on the distributional similarities of their individual subword vocabularies and then tokenized using the MBartTokenizer (a variation of the SentencePiece algorithm [22]).

4.4 Methodology

Pre-training For each monolingual sentence, feed a noised version of it into the encoder and train the model to generate the original sentence using the self-supervised objective. In our case, this could be token masking, deletion, infilling etc.

Fine-tuning Fine-tuning on parallel data is crucial for supporting LRLs [65]. 1k examples for each available language and direction should be sufficient [71]. It doesn't matter if the model isn't fine-tuned on all languages as the model should have covered at least one neighbouring language during pre-training and would be able to generalise from this.

Few-shot generation Five FSL examples could be produced either by manual selection or they can be automatically generated using MBR Decoding as previously discussed. These examples should be similar in style and formality to the desired output.

Inference These few-shot examples could then be passed in at inference along with the source text. Thanks to the monolingual data from pre-training, it is likely that data already exists for the target language. This could be used in combination with the few-shot examples provided and the existing cross-lingual capabilities of the model to produce a translation.

4.5 Validation and test data

WMT is the main dataset that MNMT is tested on. Previously it focused on news data, but from 2023 onwards it diversified the domains and styles of content covered making it more rigorous [34]. Whilst a valuable benchmark, it only covers 8 language pairs. In our use case, it would therefore only be useful to verify that performance is maintained on HRLs. As a result, for languages that aren't covered by WMT2023 and to explore the translation accuracy of LRLs, FLORES-200 [48] would be used as it supports evaluation across 200 languages (though the test data is lower quality). Papers With Code provides leaderboards for all the WMT benchmarks making it easy to compare with the SOTA [1].

4.6 Potential weaknesses and limitations

English-centrism Whilst the mentioned efforts would be taken to ensure prioritisation for LRLs, the model may still possess English-centric tendencies given that the majority of the internet is in English. For example, when mBART was originally trained, 55,608M English tokens were used versus <600M for most LRLs [65].

Computational requirements Despite the proposed model being smaller than its fully supervised counterpart, there is still a significant computational expense for pre-training an LM. For example, mBART trained for 2.5 weeks on 256 Nvidia V100 GPUs [44].

Lack of FSL examples Although five parallel sentences is a reasonable ask in most cases, this still wouldn't be available in some situations (e.g. extremely distant languages). Back translation [57] could be a possible solution to this.

4.7 Future Work

Architecture variations Whilst this proposal assumes an encoder-decoder architecture, neighbouring works have found success using encoder/decoder-only architectures [27, 28] which generally have even fewer parameters.

Experimentation with k-FSL Different values of k may produce significantly different results. It would be interesting to explore this to find the optimal ROI. [44].

Sharing or freezing parameters during pre-training and fine-tuning There are many different ways of experimenting with this to best leverage the pre-trained model [18]. Language-Family Adapters [14] could be created from the existing language groups where the core pre-trained model remains largely frozen and the adapters adjust the model's behaviour for specific languages without disrupting the knowledge in the main model.

5 Conclusion

This paper proposes an MNMT model that would be able to leverage the cross-lingual capabilities of LMs and FSL at inference, offering scalability and stable performance on a universal scale. Following joint training of a single translation model that is comparatively smaller than the SOTA and trained on a holistic dataset, it is hoped that the representations of HRLs could be shared to improve the translation performance on LRLs as opposed to existing NMT models that see diminishing returns when increasing their language coverage. FSL also offers a level of control over the style of translation produced, something which could be particularly beneficial in certain environments.

References

[1] Papers with code : Machine translation, 2019.

[2] R. Aharoni, M. Johnson, and O. Firat. Massively multilingual neural machine translation, 2019.

[3] N. Arivazhagan, A. Bapna, O. Firat, D. Lepikhin, M. Johnson, M. Krikun, M. X. Chen, Y. Cao, G. Foster, C. Cherry, W. Macherey, Z. Chen, and Y. Wu. Massively multilingual neural machine translation in the wild: Findings and challenges, 2019.

[4] M. Artetxe, G. Labaka, E. Agirre, and K. Cho. Unsupervised neural machine translation, 2018.

[5] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas. Self-supervised learning from images with a joint-embedding predictive architecture, 2023.

[6] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate, 2016.

[7] A. Bapna, I. Caswell, J. Kreutzer, O. Firat, D. van Esch, A. Siddhant, M. Niu, P. Baljekar, X. Garcia, W. Macherey, T. Breiner, V. Axelrod, J. Riesa, Y. Cao, M. X. Chen, K. Macherey, M. Krikun, P. Wang, A. Gutkin, A. Shah, Y. Huang, Z. Chen, Y. Wu, and M. Hughes. Building machine translation systems for the next thousand languages, 2022.

[8] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020.

[9] L. Burchell, Birch, and A. Kenneth Heafield. Exploring diversity in back translation for low-resource machine translation. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*. Association for Computational Linguistics, 2022.

[10] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations, 2020.

[11] Z. Chi, L. Dong, S. Ma, S. H. X.-L. Mao, H. Huang, and F. Wei. Mt6: Multilingual pretrained text-to-text transformer with translation pairs, 2021.

[12] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014.

[13] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel. Palm: Scaling language modeling with pathways, 2022.

[14] A. Chronopoulou, D. Stojanovski, and A. Fraser. Language-family adapters for low-resource multilingual neural machine translation, 2023.

[15] H. W. Chung, D. Garrette, K. C. Tan, and J. Riesa. Improving multilingual models with language-clustered vocabularies, 2020.

[16] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019.

[17] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale, 2020.

[18] A. Cooper Stickland, X. Li, and M. Ghazvininejad. Recipes for adapting pre-trained monolingual and multilingual models to machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, 2021.

[19] A. Currey and K. Heafield. Zero-resource neural machine translation with monolingual pivot data. In A. Birch, A. Finch, H. Hayashi, I. Konstas, T. Luong, G. Neubig, Y. Oda, and K. Sudoh, editors, *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 99–107, Hong Kong, Nov. 2019. Association for Computational Linguistics.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[21] B. Eikema and W. Aziz. Is map decoding all you need? the inadequacy of the mode in neural machine translation, 2020.

[22] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, and A. Joulin. Beyond english-centric multilingual machine translation, 2020.

[23] O. Firat, K. Cho, and Y. Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism, 2016.

[24] M. Freitag, G. Foster, D. Grangier, V. Ratnakar, Q. Tan, and W. Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474, 2021.

[25] M. Freitag, D. Grangier, Q. Tan, and B. Liang. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825, 2022.

[26] P. Gage. A new algorithm for data compression. *The C Users Journal archive*, 12:23–38, 1994.

[27] Y. Gao, C. Herold, Z. Yang, and H. Ney. Is encoder-decoder redundant for neural machine translation?, 2022.

[28] X. Garcia, Y. Bansal, C. Cherry, G. Foster, M. Krikun, F. Feng, M. Johnson, and O. Firat. The unreasonable effectiveness of few-shot learning for machine translation, 2023.

[29] V. Goyal, S. Kumar, and D. M. Sharma. Efficient neural machine translation for low-resource languages via exploiting related languages. In S. Rijhwani, J. Liu, Y. Wang, and R. Dror, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 162–168, Online, July 2020. Association for Computational Linguistics.

[30] T.-L. Ha, J. Niehues, and A. Waibel. Toward multilingual neural machine translation with universal encoder and decoder, 2016.

[31] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners, 2021.

[32] V. C. D. Hoang, P. Koehn, G. Haffari, and T. Cohn. Iterative back-translation for neural machine translation. In A. Birch, A. Finch, T. Luong, G. Neubig, and Y. Oda, editors, *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[33] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. B. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558, 2016.

[34] T. Kocmi, E. Avramidis, R. Bawden, O. Bojar, A. Dvorkovich, C. Federmann, M. Fishel, M. Freitag, T. Gowda, R. Grundkiewicz, B. Haddow, P. Koehn, B. Marie, C. Monz, M. Morishita, K. Murray, M. Nagata, T. Nakazawa, M. Popel, M. Popović, and M. Shmatova. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not

quite there yet. In P. Koehn, B. Haddow, T. Kocmi, and C. Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore, Dec. 2023. Association for Computational Linguistics.

[35] T. Kocmi, C. Federmann, R. Grundkiewicz, M. Junczys-Dowmunt, H. Matsushita, and A. Menezes. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussa, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, T. Kocmi, A. Martins, M. Morishita, and C. Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online, Nov. 2021. Association for Computational Linguistics.

[36] P. Koehn. *Statistical machine translation*. Cambridge University Press, 2009.

[37] T. Kudo and J. Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, 2018.

[38] S. M. Lakew, M. Cettolo, and M. Federico. A comparison of transformer and recurrent neural networks on multilingual neural machine translation, 2018.

[39] S. M. Lakew, M. Federico, M. Negri, and M. Turchi. Multilingual neural machine translation for low-resource languages. *Italian Journal of Computational Linguistics*, 4:11–25, 06 2018.

[40] S. M. Lakew, M. Federico, M. Negri, and M. Turchi. Multilingual neural machine translation for zero-resource languages, 2019.

[41] G. Lample and A. Conneau. Cross-lingual language model pretraining, 2019.

[42] G. Lample, A. Conneau, L. Denoyer, and M. Ranzato. Unsupervised machine translation using monolingual corpora only, 2018.

[43] B. Liu and I. R. Lane. Attention-based recurrent neural network models for joint intent detection and slot filling. *CoRR*, abs/1609.01454, 2016.

[44] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer. Multilingual denoising pre-training for neural machine translation, 2020.

[45] Y. Lu, L. Wen, J. Liu, Y. Liu, and X. Tian. Self-supervision can be a good few-shot learner, 2022.

[46] M. Malli and G. Tambouratzis. Evaluating corpus cleanup methods in the WMT'22 news translation task. In P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno Yepes, T. Kocmi, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, M. Negri, A. Névéol, M. Neves, M. Popel, M. Turchi, and M. Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 335–341, Abu Dhabi, United Arab Emirates (Hybrid), Dec. 2022. Association for Computational Linguistics.

[47] A. Mueller, G. Nicolai, A. D. McCarthy, D. Lewis, W. Wu, and D. Yarowsky. An analysis of massively multilingual neural machine translation for low-resource languages. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3710–3718, Marseille, France, May 2020. European Language Resources Association.

[48] NLLB Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. Mejia-Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang. No language left behind: Scaling human-centered machine translation. 2022.

[49] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, and D. Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[50] M. Popović. chrF: character n-gram F-score for automatic MT evaluation. In O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, and P. Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics.

[51] Y. Qi, D. S. Sachan, M. Felix, S. J. Padmanabhan, and G. Neubig. When and why are pre-trained word embeddings useful for neural machine translation?, 2018.

[52] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[53] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie. COMET: A neural framework for MT evaluation. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, Nov. 2020. Association for Computational Linguistics.

[54] D. Ruiter, C. España-Bonet, and J. van Genabith. Self-supervised neural machine translation. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1828–1834, Florence, Italy, July 2019. Association for Computational Linguistics.

[55] P. Rust, J. Pfeiffer, I. Vulić, S. Ruder, and I. Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models, 2021.

[56] D. S. Sachan and G. Neubig. Parameter sharing methods for multilingual self-attentional translation models, 2018.

[57] R. Sennrich, B. Haddow, and A. Birch. Improving neural machine translation models with monolingual data, 2016.

[58] R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units, 2016.

[59] A. Siddhant, A. Bapna, Y. Cao, O. Firat, M. Chen, S. Kudugunta, N. Arivazhagan, and Y. Wu. Leveraging monolingual data with self-supervision for multilingual neural machine translation, 2020.

[60] A. Siddhant, A. Bapna, O. Firat, Y. Cao, M. X. Chen, I. Caswell, and X. Garcia. Towards the next 1000 languages in multilingual machine translation: Exploring the synergy between supervised and self-supervised learning, 2022.

[61] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA, Aug. 8-12 2006. Association for Machine Translation in the Americas.

[62] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks, 2014.

[63] X. Tan, J. Chen, D. He, Y. Xia, T. Qin, and T.-Y. Liu. Multilingual neural machine translation with language clustering, 2019.

[64] M. Tars, A. Tättar, and M. Fišel. Extremely low-resource machine translation for closely related languages, 2021.

[65] S. Thillainathan, S. Ranathunga, and S. Jayasena. Fine-tuning self-supervised multilingual sequence-to-sequence models for extremely low-resource nmt. In *2021 Moratuwa Engineering Research Conference (MERCon)*, pages 432–437, 2021.

[66] B. Thompson, M. P. Dhaliwal, P. Frisch, T. Domhan, and M. Federico. A shocking amount of the web is machine translated: Insights from multi-way parallelism, 2024.

[67] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023.

[68] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Łukasz Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. Google's neural machine translation system: Bridging the gap between human and machine translation, 2016.

[69] B. Zhang, P. Williams, I. Titov, and R. Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online, July 2020. Association for Computational Linguistics.

[70] W. Zhang. IOL research machine translation systems for WMT23 general machine translation shared task. In P. Koehn, B. Haddow, T. Kocmi, and C. Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 187–191, Singapore, Dec. 2023. Association for Computational Linguistics.

[71] D. Zhu, P. Chen, M. Zhang, B. Haddow, X. Shen, and D. Klakow. Fine-tuning large language models to translate: Will a touch of noisy data in misaligned languages suffice?, 2024.