Manipulating Multilingual Models for Neural Machine Translation

Fengjiao Sun*, Jordan Meyer*, Leon Rafael Gutierrez Angulo*
University of California, Berkeley – School of Information
{fengjiao.sun, jordan.meyer, leonrafael29}@berkeley.edu

Abstract

In recent years Neural Machine Translation (NMT) has made significant advances in fluency and faithfulness coming a long way from the early days of word-by-word translation lookup. In this paper we aim to further that exploration by building upon recent world class research.

In this paper we have fine-tuned two state of the art pre-trained multilingual translation models, mBART-50 and M2M-100 and created a pre-trained multilingual language model, BERT, in a new sequence-to-sequence architecture, BERT2BERT. Models were fine-tuned on the News Commentary dataset to significantly improve BLEURT performances with 1-3 training epochs. Our new BERT2BERT model outperforms the many-to-many mBART-50 state of the art model on single pair directional translations after 3 epochs by 4.7% with a 19.8% reduction in model size. The BERT2BERT model delivered a BLEURT performance with 77.9% of the M2M-100 single directionally fine-tuned model. We tried to reduce model size to deliver a model for app deployment using a many-to-many mBART-50 base but performance reduced linearly with the size of the model making it unusable in its current form. This new BERT2BERT architecture can be used to bridge the gap and drastically reduce model size, while still delivering a strong BLEURT performance in future applications.

1 Introduction

In an ever-increasingly connected world it is more important than ever to facilitate communication. It is currently estimated that there are over 7,100 unique languages spoken[1]. As a team of 3 members with different native languages we agreed that this topic is highly relevant and should be explored.

More commonly is the usage of Multilingual Machine Translation (MMT) where many new transformer models are being made available for many-to-many translations that were trained on some combination direct translation bitexts (ie. A ↔ B), many to one language (ie. M → B) translations and one to many translations (A → M). While these models have often achieved state of the art performance, practical use cases of these are often limited due to their large model sizes and language breadth.

Through our research we aimed to take pre-existing multilingual models and fine tune them on a subset of the News Commentary dataset using the top 3 languages of the United States: English, Spanish and Chinese, to deliver models that are smaller but remain performant in hopes of developing a compact application that can be used in a portable device rather than being served through a remote service API.

2 Background and Related Works

2.1 Overview

Our research builds off of the previous studies that introduced each of the pre-trained baseline models, namely Multilingual BART Model (mBART/mBART-50)(Tang et. al., 2020), Many to Many 100 Language Model (M2M-100)(Fan et. al., 2020), and a Transformer-based sequence-to-sequence BERT (Devlin et. al., 2018) model with the Multilingual BERT Base Uncased Checkpoint (mBERT) which were used in an Encoder-Decoder architecture. All of the models in scope have been pre-trained on the selected three languages in some methodology.

Additionally, the Google Research team introduced the idea of initialized encoder-decoder models using pre-trained encoder and/or decoder checkpoints for sequence generation resulting in state of the art

results for machine translation (Rothe et al. 2020).

2.1.1 Multilingual BART 50 (mBART-50)

Multilingual BART (mBART/mBART-50) is an evolution of the original large BART model (Bi-directional and Auto-Regressive Transformer)(Lewis, Liu, Goyal et. al., 2019) architecture released by the for general NLP applications such as sequence classification, token classification, generation and a rough version of translation.

The multilingual model uses an encoder that is similar to BERT with 12 encoder layers and a decoder that is similar to GPT with 12 decoder layers. The model checkpoint used for our study was mBART-50, which was pre-trained to translate between 50 languages in a one to many, many to one and many to many configurations. The paper demonstrated capability of learning new languages without hindering performance on previously learned languages which was seen in previous attempts for fine-tuning. While this model has good performance translating between many pairs of languages it is prohibitively too large for deploying in a compact application that can deliver quick translations and run on a mobile device.

2.1.2 M2M-100

M2M-100 (Fan et al., 2020) is a many-to-many multilingual translation model that can translate directly between any pair of 100 languages. It used the automatic construction of parallel corpora – CCMatrix (Schwenk et al., 2019) and CCAligned (El-Kishky et al., 2020), with a novel data mining strategy that exploits language similarity to avoid mining all directions, and leveraged back translation to improve the quality on zero-shot and low resource language pairs. The model is a large Transformer model with 12 Encoder and 12 Decoder layers, with 8192 feed-forward network (FFN) size and 1024 embedding dimensions. The model training dataset comprises 7.5B training sentences corresponding to 2200 directions for 100 languages. It brings gains of more than 10 BLEU when directly translating between non-English directions while performing competitively to the best single system of WMT.

2.1.3 BERT2BERT

The BERT2BERT model is a new model created with a BERT encoder joined to a BERT decoder using a transformer based sequence-to-sequence model with the Multilingual BERT (mBERT) uncased checkpoint. mBERT's model architecture is similar to BERT. It has 12 layers, 12 attention heads and 768 hidden dimensions per token. (Devlin et al., 2019), the primary difference is that mBERT is trained on monolingual Wikipedia corpora of 104 languages instead of just English. Because mBERT is an encoder-only model, it can only map an input sequence to an output sequence of a priori known output length, which makes it unsuitable for sequence-to-sequence tasks, such as translation. Therefore, we warm-start a sequence-to-sequence model initializing both the encoder and decoder part from the encoder-only mBERT base uncased checkpoint for each directional pair. In order to use BERT as a decoder the forward facing attention mechanism is removed and cross-attention layers are added between the self-attention layer and the two feed-forward layers (Rothe, et al., 2020).

3 Methods

3.1 Objective

Our primary objective of our research was to deliver models which deliver high quality translations in portable sizes. We use BLEURT score as a method for scoring our models performance which we describe further in the following section.

To address model portability we also look at the number of parameters and the size of the binary Pytorch model in bytes.

3.2 Data

The dataset used for fine-tuning all models is the News Commentary (v11) Dataset from OPUS (Tiedemann, 2012). OPUS is a free language resource that provides available datasets and tools for various studies, including translation. This dataset contains news commentaries

provided by WMT. The sentences were collected from multiple news websites and translated by professional translators. The dataset comprises 12 languages in 63 bitext combinations. For our scope we have used the following 6 language pairs: English ↔ Spanish (En ↔ Es) – 239k, English ↔ Chinese(En↔ Zh) – 69k and Spanish ↔ Chinese (Es ↔ Zh) – 65k bitexts.

In order to train and fine-tune our models we have used the train evaluation and test split outlined in the table 1 below.

Bitext Pair	Training	Evaluation	Test
En ↔ Zh	70% (49k)	15% (10k)	5k
En ↔ Es	70% (167k)	15% (36k)	5k
Es ↔ Zh	70% (46k)	15% (10k)	5k

Table 1: Dataset Train, Evaluation and Test split for each bitext pair

3.3 Environment

For both training and evaluation purposes all of the models were run via Jupyter Notebooks using Google Colab Pro+ which includes access to an Nvidia Tesla T4 Tensore Core GPU with 16 GB of VRAM and a 52+ GB RAM runtime environment. As compute units were limited by budgetary constraints we opted to use standard GPUs rather than premium GPUs. This environment also enabled us to execute code for longer than 24 hours continuously in multiple notebooks.

3.4 Experiments

As our models had undergone very different pre-training we wanted to find common grounds for comparing our models to each other as well as testing different architectures as a type of Design of Experiments.

For model comparisons, BERT2BERT vs base mBART were scored against each other as both models used the same approximate size with 12 encoder and 12 decoder layers.

Another comparison was to view performance of M2M-100 vs BERT2BERT as both were fine-tuned on single directional bitext-pairs.

3.4.1 Model Specificity

We wanted to test two approaches of modeling to observe performance and sizes. The desired approach was to use a single multilingual pretrained model that is fine-tuned to translate between many-many languages (ie A↔ B, A → M, M →B) vs individual pre-trained models only fine-tuned on specific directional bitext pairs ie (A→B) for the sole purpose of that translation.

3.4.2 Model Size

We explored the possibility of creating a more "efficient" model in the number of parameters. Since most language translation models are computationally heavy, it is ideal if we could finetune a model into a smaller size yet achieve a competitive BLEURT score.

To evaluate size, the mBART model is modified following the initial thoughts outlined in the BART paper taking the 12 encoder layers, 12 hidden layer and 12 decoder layer (mBART Base 12/12/12) model and reducing the number of encoder, decoder and hidden layers based on the notion thought process covered in the BART paper of removing portions of the network that may not be needed for our task. We proceeded to split the model in half, keeping only the first pre-trained 6 encoder, 6 hidden and 6 decoder layers (mBART Mini, 6/6/6).

As this was still large, we aim to improve our portability by further reducing size to a set of 3 layers (mBART Tiny, 3/3/3) and a set of single layers (mBART Single, 1/1/1).

3.4.3 Fine-Tuning Ordering

While training mBART, we iterated through bitext pairings sequentially by exhausting all of a single directional language pair before continuing onto training the next pairing. As we thought training in a new pairing may reduce performance on a previously trained pairing we wanted to verify if this notion was correct. Therefore, an additional test was added to compose a fully shuffled dataset where each subsequent training example had some random probability of being drawn from any of the 6 directional pairings.

3.4.4 Hyperparameter Optimization

As BERT2BERT was a unique challenge to build a sequence to sequence architecture we also tested some of the underpinning performance drivers that have been well established in the other existing models such as the maximum allowable repeated n-grams, beam width, minimum length and the maximum length.

While a higher maximum length for translation is desirable to avoid cutting off sentences and losing key sentiments, each additional token requires a significant increase in VRAM and RAM limiting potential for these tests on all models.

3.5 Evaluation Metrics

We opted to use Bilingual Evaluation Understudy with Representations from Transformers (BLEURT) as the metric to evaluate the translations. BLEURT is a trained metric, it is a regression model trained on a public collection of ratings data (the WMT Metrics Shared Task dataset). The model is based on BERT and RemBERT (Shellam, et al., 2020). BLEURT offers superior performance than previous approaches and higher correlation with human ratings. BLEURT builds upon recent advances in transfer learning and is able to capture extensive linguistic cases, such as paraphrasing, therefore, its level of quality is closer to human annotations and fluency.

We are using the distilled BLEURT which generates scores roughly between 0 and 1 where 1 indicates the output sentence conveys the meaning of the reference sentence (Shellam, et al., 2020).

3.6 Baseline

For mBART and M2M-100 models, we calculate the baseline by generating the target labels without fine-tuning the models. We don't have a baseline for our BERT2BERT model as it hasn't been trained on translation. Because our BERT2BERT model has just been initialized, we need to first warm-start the model to train the randomly initialized cross-attention weights for a translation sequence-to-sequence downstream task.

4 Results and Discussion

4.1 Model Results:

Table below highlights the average results for our model performance across 6 pairs of languages from the news commentary dataset. Refer to the appendix A for detailed language pair results.

Model	Stage	Parameters	BLEURT
M2M-100	Baseline	484M	0.4021
	Fine-tuned	484M	0.4820
mBART	Baseline	610M	0.3587
	Fine-tuned	610M	0.3045
BERT2BERT	Baseline	363M	N/A
	Fine-tuned	363M	0.2256

Table 2: Baseline and Fine-tuned Average BLEURT scores

4.2 Model Analysis:

We evaluated the three models on all 6 pairs of language translation by using the BLEURT score. After training the model with 70% from each language pair and tuning the models and hyperparameters, we observed different patterns of BLEURT score results.

4.2.1 Baseline and Model Specificity

(a) Baseline results: M2M-100 and mBART were both used to create a baseline that could be compared with the fine-tuned model and with the BERT2BERT model. The two baseline models demonstrated a similar BLEURT score in the Chinese→English pair. While M2M-100 outperformed mBART in Spanish→Chinese pair and English→Spanish translation, it delivered a worse score when translating Spanish→English.

(b) M2M-100: we found a progressive increase in BLEURT score across all 6 pairs, with the highest uplift in Spanish→English (0.4421 to 0.6775), which surpassed the mBART standard baseline model.

(c) mBART: the trained Standard model generally did not achieve a higher BLEURT score than the baseline model, which was probably because it was trained in a many-to-many approach. However, the trained Standard model generally performed the best among all mBART models except for English→Chinese and Spanish→Chinese translation, where the Mini model delivered a superior performance.
(d) BERT2BERT: The best fine tuned model (maximum length of 100, minimum length of 0, no_repeat_ngram_size of 3) not only achieved a distinct improvement in BLEURT score across the language pairs, but was able to achieve 77.9% of M2M-100 model's performance, and even delivered results superior to the many-to-many mBART model likely due to the benefit of specificity in this case. The results demonstrate a potential for the model to perform at parity with M2M-100 with further fine-tuning.

4.2.2 Model Size

To evaluate the model size, mBART is tuned in a number of parameters to be comparable with BERT2BERT at approximately 343M parameters (mBART Tiny model), and with M2M-100 at approximately 435M parameters (mBART Mini model).

None of the reduced size mBART models were able to deliver a superior or on par performance, which could be due to mBART being trained as a many-to-many model. It has been reported in the past that many-to-many translation systems did not perform as well as bilingual models (Johnson et al. 2017).

Further experiments can be conducted to fine-tune mBART for single-pair performance,or to train M2M-100 and BERT2BERT into a many-to-many language translation model.

An important feature about the mBART50 model is the inclusion of the beginning of speech token for the target language which acts as a flag to jump start the conditional probability calculations to generate new text following the starting point. If this is not provided the model makes a (poor) guess as to what you want to do with it.

For the mini model (6/6/6) we evaluate the baseline prior to any fine-tuning. As expected, performance dropped significantly, however the model seems to still understand basic parts of speech, such as spelling and word flow, only in Enlgi. When we attempt to translate to Chinese we are only provided the Chinese beginning of speech token and when we attempt a translation to Spanish it is only in English. This may be due to the extensive training of the English language being encoded in the lower layers and the fine tuned language levers sitting in the layers that were removed. This impact was even more pronounced when cutting the model size down further into a size of 3 and 1 layers of pairwise encoder and decoders with hidden layers.

Example:
Input (English): "<s>en_XX ['Companies tend not to invest in talent or develop new skills, and the quality of existing skills can suffer. In the longer run, flexible labor markets also increase structural unemployment and fuel the informal economy.</s>"

Prediction (Undirected): "</s><unk> ['One cannot afford to take the table over human rights in one country, but to keep one's views in one another – a wholesome occurrence. In other words, one's views can be held back from</s>"

Predicted (Chinese):
"</s>zh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CNzh_CN</s>"

Predicted (Spanish):
</s>es_XX ['In some cases, they involve border disputes (Bolivia and Venezuela, Bolivia, Bolivia, and Nicaragua, etc.). In others, they concern the armed factions (Colombia and Venezuela, Colombia and</s>

The final mBART model sizes and average BLEURT scores for all pair combinations are shown in the following table:

Model	Parameters	Model Size	BLEURT
Baseline	610M	2.28 GB	0.35874
Mini	435M	1.74 GB	0.19306
Tiny	343M	1.39 GB	0.16354
Single	288M	1.07 GB	0.17383

Table 3: Baseline and Fine-tuned Average BLEURT scores for mBART's models: baseline vs. three models with different sizes.

Although we were able to successfully reduce the model size from 2.28 GB to 1.07 GB, we suffered from an almost 50% reduction in performance as well. While using mBART, all language pairs are available in a single model whereas with individual pairwise models a user may need multiple different models requiring an even larger footprint as a tradeoff for improved performance. However, a single reduced model size of 1.07 GB is still unwieldy to be deployed for execution on an application that is to be run on a low resource device.

4.2.3 Fine-Tuning Ordering

While we did notice a reduced performance in some pairings when other language pairings were trained, our BLEURT score improved slightly via shuffling the training pairings, seen in table 4 below.

One challenge was that only one large epoch was able to be conducted due to the overhead requirements of individual tokenization on a sentence to sentence basis due to the randomness of the source and target language. With appropriate system optimization and further tuning, this test can be repeated with better results.

Model	Epoch	BLEURT
Ordered	1	0.1039
Shuffled	1	0.1179

Table 4: Average BLEURT performance for dataset ordering.

4.2.4 Hyperparameter Optimization

Maximum sentence length (`max_length`), minimum sentence length (`min_length`) and number of repetitions of each n-gram (`no_repeat_ngram_size`) were optimized with BERT2BERT. The default parameters of BERT are `max_length = 20`, `min_length = 0`, and `no_repeat_ngram_size = 0`.

4.2.4.1 Maximum Length Optimization (`max_length`)

The default maximum sentence length of BERT is 20, so long sentences were truncated.

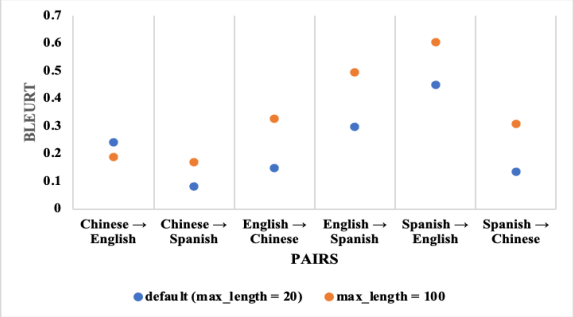

Figure 1. BLEURT scores BERT2BERT models, each fine-tuned for a respective language pair with `max_length = 100` vs. the default `max_length = 20`, keeping all the other hyperparameters the same.

In Figure 1, with the exception of Chinese → English, there is an improvement of 92% in BLEURT score at a maximum length of 100.

4.2.4.2 Minimum Length Optimization (`min_length`)

Since most models performed better with a maximum length of 100 than the default of 20, we continued testing with only a maximum length of 100 and a minimum length of 50. We compared these BLEURT scores against those from the previous test.

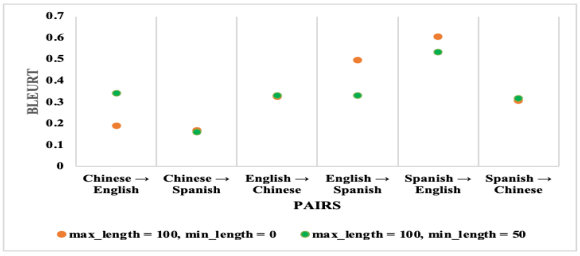

Figure 2. BLEURT scores for BERT2BERT models, each fine-tuned for a respective language pair with `max_length = 100` and `min_length = 0` vs. `max_length = 100` and `min_length = 50`.

Figure 2 illustrates how for 3 out of the 6 pairs the models with a maximum length of 100 and a minimum length of 0 outperforms the models tested with a minimum length of 50. From the three that had a higher score, only the Chinese→English pair had a substantial improvement of 81.47%.

By setting minimum length at 50, the model needs to output at least 50 tokens, which results in the model repeating or generating random tokens if the target sentence is smaller than 50 tokens which degrades our results for the outperformed models by 0.08 BLEURT points.

In the example below, we can see how for the English→Spanish fine-tuned model, there are some sequences of tokens that are repeated multiple times (highlighted words).

Example:
Input (English): The process resembles the selection of a standard in, say, consumer electronics.

Label (Spanish): El proceso se parece a la selección de una norma, digamos, entre productos electrónicos.

Predicted (Spanish):el proceso se parece a la seleccion de un estandar en el que se asemeja a la seleccion de un estandar en la industria de consumo, por ejemplo, la seleccion de un estandar en la industria de consumo de consumo de consumo, digamos, por ejemplo, la electronica de consumo de los consumidores.

4.2.4.3 N-Gram Repetition Optimization (`no_repeat_ngram_size`)

To avoid repetition of sequences of n consecutive words, we tested the `no_repeat_ngram_size` hyperparameter of BERT2BERT with a maximum length of 100.

For clarity, the default hyperparameter for the number of repetitions of each n-gram at a value of 0 indicates no restriction on maximum number of repetitions rather than no repetitions allowed.

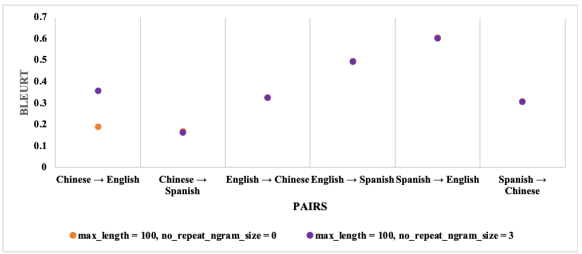

Figure 3. BLEURT scores for BERT2BERT models, each fine-tuned for a respective language pair with `max_length = 100` and `no_repeat_ngram_size = 0` vs `max_length = 100` and `no_repeat_ngram_size = 3`

Figure 3 shows how increasing the allowable repetitions of n-grams has almost no effect on the BLEURT score for 5 of the 6 pairs.

Only the BERT2BERT model fine-tuned with the Chinese→English pair saw considerable improvement in BLEURT of 0.168, which is a 89.26% increase from the model with no restriction on repetition.

4.3 Challenges and Future Work

First, all three models were not able to perform particularly well when translation pairs include Chinese. This may be due to the large imbalances between the number of Chinese tokens in the models' vocabularies and those of the Latin alphabet (Fan et al., 2020) or due to the imbalance of pre-training data with Chinese pairs.

This may also be because Chinese sentence structure does not contain any whitespaces or letter casing which makes the languages English and Spanish increasingly different from Chinese.

These differences together may also help to offer explanation as to why English and Spanish perform better together, despite coming from Germanic and Romance roots.

The below example shows how M2M-100, the overall best performer, only reached 0.2957 Bleurt score when translating from English to Chinese.

Example:
Input (English): With the stakes so high, the Russian-US initiative to organize a peace conference has not only been welcomed, but is seen as an indispensable initiative that must be planned with excruciating care. We are likely to hear of delay after delay in the coming days, as various sides try to choose their representatives.

Label (Chinese): 因为利害攸关，俄美组织和平会议的计划不仅颇受欢迎，而且被视为不可或缺的行动而受到慎重对待。因为各方都力争选定代表，和平会议的日程可能一推再推。但我们必须保持耐心，因为会议一旦开始除成功外就没有别的选择。没有结果的努力不会有任何用处。

Predicted (Chinese):
因此,俄罗斯-美国组织和平会议的意图不但得到了欢迎,也被视为一个不可或缺的计划,必须以谨慎的态度进行规划。我们很可能会在未来几天中听到延迟之后的延迟,因为各方试图选择代表。

Second, due to the time and computing resource limit, we are using only 6 pairs of languages from the news commentary dataset and trained by only 1 to 3 epochs. Though the trained models have produced a satisfying BLEURT score, we do hope to apply the same approach on broader pairs of languages and train in more epochs to see how better the result could be.

Third, since we only trained mBART into a many-to-many language translation model while both BERT2BERT and M2M-100 remained to be one-pair language only, further work can be performed to finetune M2M-100 and BERT2BERT model into many-to-many models.

5 Conclusion

In this paper, we fine tuned two different pre-trained translation models, mBART and M2M-100, and created a pre-trained BERT2BERT model on six bilingual translation tasks (English ↔ Spanish ↔ Chinese) with a new dataset, News Commentary. We observed that the BLEURT score generally improved across all models after fine-tuning and hyperparameter tuning. The new BERT2BERT architecture demonstrated the potential to deliver a strong BLEURT performance with a reduced model size.

Since many-to-many language models delivered modest results compared to bilingual models, and models with more parameters outperformed small-size ones, our future research will explore expanding models to broader language pairs and fine-tune bilingual models to many-to-many ones.

References:

Ahmed El-Kishky et al. “CCAligned: A Massive Collection of Cross-Lingual Web-Document Pairs.” In *Proc. of EMNLP, 2020.* https://arxiv.org/pdf/1911.06154

Alexis Conneau et al., “XNLI: Evaluating Cross-lingual Sentence Representations.” arXiv, Sep. 13, 2018 http://arxiv.org/abs/1809.05053.

Angela Fan et al., “Beyond English-Centric Multilingual Machine Translation.” arXiv, Oct. 21, 2020 http://arxiv.org/abs/2010.11125
M2M-100 Model API
https://huggingface.co/docs/transformers/model_doc/m2m_100

Colin Raffel, Noam Shazeer, Adam Robert, Katherine Lee et al., “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.” arXiv, Jul. 28, 2020 http://arxiv.org/abs/1910.10683.

Dzmitry Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate.” arXiv, May 19, 2016 http://arxiv.org/abs/1.0473.

Ethnologue “How many languages are there in the world?,” Ethnologue, May 03, 2016. https://www.ethnologue.com/guides/how-many-languages

Holger Schwenk et al., “CCMatrix: Mining Billions of High-Quality Parallel Sentences on the WEB.” arXiv, May 01, 2020 http://arxiv.org/abs/1911.04944.

Jacob Devlin et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” arXiv, May 24, 2019 http://arxiv.org/abs/1810.04805,
BERT/mBERT Model Checkpoint
https://huggingface.co/bert-base-multilingual-uncased.

Jörg Tiedemann, 2012, Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation, (LREC 2012)* http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf,
News Commentary Dataset API:
https://huggingface.co/datasets/news_commentary

Melvin Johnson, Mike Schuster, et al. Google’s multilingual neural machine translation system: Enabling zero-shot translation. In *Transactions of the Association for Computational Linguistics, 5:339–351, 2017.* Available: https://aclanthology.org/Q17-1024/

OpenAI “Better Language Models and Their Implications,” OpenAI, Feb. 14, 2019. https://openai.com/blog/better-language-models/.

Sascha Rothe, et al., “Leveraging Pre-trained Checkpoints for Sequence Generation Tasks.” arXiv, Apr. 16, 2020 http://arxiv.org/abs/1901.07291
BERT2BERT Model API:
https://huggingface.co/docs/transformers/v4.24.0/en/model_doc/encoder-decoder#transformers.EncoderDecoderModel

School of Information, w266 Course Material
https://github.com/datasci-w266/2022-fall-main/tree/master/materials/

Thibault Shellam, “BLEURT: Learning Robust Metrics for Text Generation.” arXiv, May. 21, 2020 https://arxiv.org/pdf/2004.04696

Yinhan Liu, Jiatao Gu, Naman Goyal et al., “Multilingual Denoising Pre-training for Neural Machine Translation.” arXiv, Jan. 23, 2020 http://arxiv.org/abs/2001.08210.
mBART Model API:
https://huggingface.co/docs/transformers/main/model_doc/mbart

Yonghui Wu et al., “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.” arXiv, Oct. 08, 2016 http://arxiv.org/abs/1609.08144

Yuqing Tang et al., “Multilingual Translation with Extensible Multilingual Pretraining and Finetuning.” arXiv, Aug. 02, 2020 http://arxiv.org/abs/2008.00401
mBART-50 Model API:
https://huggingface.co/docs/transformers/main/model_doc/mbart

Appendix A. Full results table

Model	Name / Modifier	Parameters	Encoder/ Decoder/ Hidden Layers	Status	en-zh	zh-en	en-es	es-en	es-zh	zh-es	Average
M2M-100	**Baseline**	484M	12 / 12 / 12	Baseline	0.3785	0.4018	0.6232	0.4421	0.2651	0.3021	**0.4021**
	Trained	484M	12 / 12 / 12	Fine-tuned	0.4154	0.4288	0.6482	0.6775	0.3906	0.3313	**0.4820**
mBART	**Standard**	610M	12 / 12 / 12	Baseline	0.3259	0.4138	0.5017	0.6128	0.1417	0.1566	**0.3587**
	Standard	610M	12 / 12 / 12	Fine-tuned	0.2174	0.3257	0.5861	0.3764	0.0965	0.2249	**0.3045**
	Mini	435M	6 / 6 / 6	Baseline	0.0714	0.1031	0.1244	0.1348	0.0948	0.0865	**0.1025**
	Mini	435M	6 / 6 / 6	Fine-tuned	0.2762	0.1220	0.2158	0.1833	0.2764	0.0846	**0.1931**
	Tiny	343M	3 / 3 / 3	Baseline	0.0712	0.0510	0.0304	0.0538	0.0720	0.0319	**0.0517**
	Tiny	343M	3 / 3 / 3	Fine-tuned	0.2337	0.1261	0.1213	0.1742	0.2454	0.0806	**0.1635**
	Single	288M	1 / 1 / 1	Baseline	0.0712	0.0422	0.0304	0.0356	0.0719	0.0319	**0.0472**
	Single	288M	1 / 1 / 1	Fine-tuned	0.0491	0.1470	0.4301	0.2448	0.0487	0.1233	**0.1738**
	Shuffled	288M	1 / 1 / 1	Fine-tuned	0.0713	0.1419	0.0758	0.3001	0.0645	0.0538	**0.1179**
BERT2BERT	**Default**[1]	363M	12 / 12 / 768	Baseline	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**
		363M	12 / 12 / 768	Fine-tuned	0.1479	0.2425	0.2963	0.4502	0.1340	0.0827	**0.2256**
	Max Length 100	363M	12 / 12 / 768	Fine-tuned	0.3267	0.1889	0.4951	0.6044	0.3076	0.1686	**0.3485**
	Max Length 100 Min Length 50	363M	12 / 12 / 768	Fine-tuned	0.3320	0.3428	0.3304	0.5336	0.3185	0.1600	**0.3362**
	Max Length 100 N-Gram Repetition 3	363M	12 / 12 / 768	Fine-tuned	0.3269	0.3575	0.4937	0.6036	0.3077	0.1635	**0.3755**

[1] Default parameters for BERT2BERT are a Maximum length of 20, a Minimum length of 0 and an Maximum n-gram repeat of 0.