Evaluation of Machine Translation Pre-Trained Models for Many-to-English Translation

Stefan Liemawan Adji

August 22, 2024

Abstract

Machine translation (MT) has become an essential tool for overcoming language barriers in an increasingly globalised world. Today, there are pre-trained models specifically designed for automatic machine translation. This paper evaluates the performance of various machine translation pre-trained models (MT-PTMs) for many-to-English translation. The models assessed include OPUS-MT [52], mBART-50 [30], NLLB-200 [47], and M2M-100 [18]. Using a dataset curated from the Tatoeba repository, comprising 1,323 parallel sentence pairs across 14 source languages, translation quality is measured using BLEU, SacreBLEU, and METEOR scores. The results indicate that OPUS-MT, a one-to-one PTM, outperforms multilingual models in one-to-English translation tasks, with mBART-50 leading among the multilingual models. Chinese, Italian, Polish, and Turkish translations to English show the best performance on average across the four models used in this study, however, Ukrainian-to-English on the OPUS-MT model achieves the highest SacreBLEU and METEOR scores.

1 Introduction

According to Ethnologue [16], 7,164 languages currently exist and are in use today, with 40% of them considered endangered. With so many languages around us, the language barrier has become a known issue, especially in the current globalised world where it is essential for individuals who speak different languages to be able to communicate effectively [35]. Thus, the need for translation services to facilitate communication across different languages has also surged [36]. As of August 2024, only 243 languages are supported by Google Translate (according to Wikipedia [59]).

Machine translation (MT) refers to the task of automatically translating from one language to another, which typically can be done through text or audio. The research and practice can be traced back to 1949 [57], with the first public demonstration of an MT system on January 7, 1954, in collaboration with IBM, where 49 Russian sentences were translated into English using a limited vocabulary of 250 words and 6 grammar rules [23]. However, over the next several decades, growth was limited for machine translation, with 1956-1966 considered the decade of high expectation and disillusion, and 1967-1976 dubbed 'the quiet decade' [24]. Then in 1989, the dominance of the rule-based approach was challenged by the rise of new methods and strategies, collectively referred to as 'corpus-based' methods (data-driven) [21, 22]. Subsequently, statistics-based approaches for MT re-emerged, bolstered by the recent success of probabilistic techniques in speech recognition. Statistical machine translation [32] dominated the domain between the late 1990s through the early 2010s, before largely being surpassed by neural machine translation (NMT) [12, 45].

Since the introduction of Transformers in 2017 [56], Natural Language Processing (NLP) and machine translation in particular reached a giant milestone. The following years saw the birth of Large Language Models (LLMs) such as BERT [15], GPT [37], and T5 [42], which revolutionised both MT and the whole field of NLP. Then in the early 2020s, several pre-trained models that are specifically designed for machine translation emerged, these will be

referred to in this paper as Machine Translation Pre-Trained Models (MT-PTMs). OPUS-MT [52] produced numerous models designed for one-to-one translations. Additionally, multilingual models: mBART-50 [30], NLLB-200 [47], and M2M-100 [18] have also been published, growing popularity as multilingualism has been shown to allow for sharing information among languages [19]. Most of these models are trained on a diverse set of languages, allowing for many-to-many translation, which is able to translate between any of the supported pairs of languages. This allows the models to generalise over shared lexical and linguistic among languages, and have been shown to increase performance compared to one-to-one translation models [30].

Despite these advancements, pre-trained models are often evaluated using a different set of languages and benchmarks [30, 47, 18, 58], making it difficult to gauge their relative effectiveness across various languages. Multilingual datasets often do not contain parallel texts across different language pairs; instead, each language pair typically has its own unique set of texts that may not overlap with the texts available in other languages [51, 9].

Nevertheless, there does not seem to be much work on comparing or benchmarking different MT-PTMs in machine translation. Intento recently published 'The State of Machine Translation 2024' [49] providing an in-depth evaluation of 52 MT engines and LLMs across 11 language pairs. Figure 1 shows the reported translation quality across different domains and language pairs, showing that the results significantly vary based on these parameters. They also reported a lack of growth in the number of supported languages and concluded that translation quality significantly varies depending on languages [49]. However, the biggest drawback in this report is that the selected LLMs are general LLMs such as GPT-4 [37], LLaMa2 [54], and Mistral [25], instead of MT-specific models.

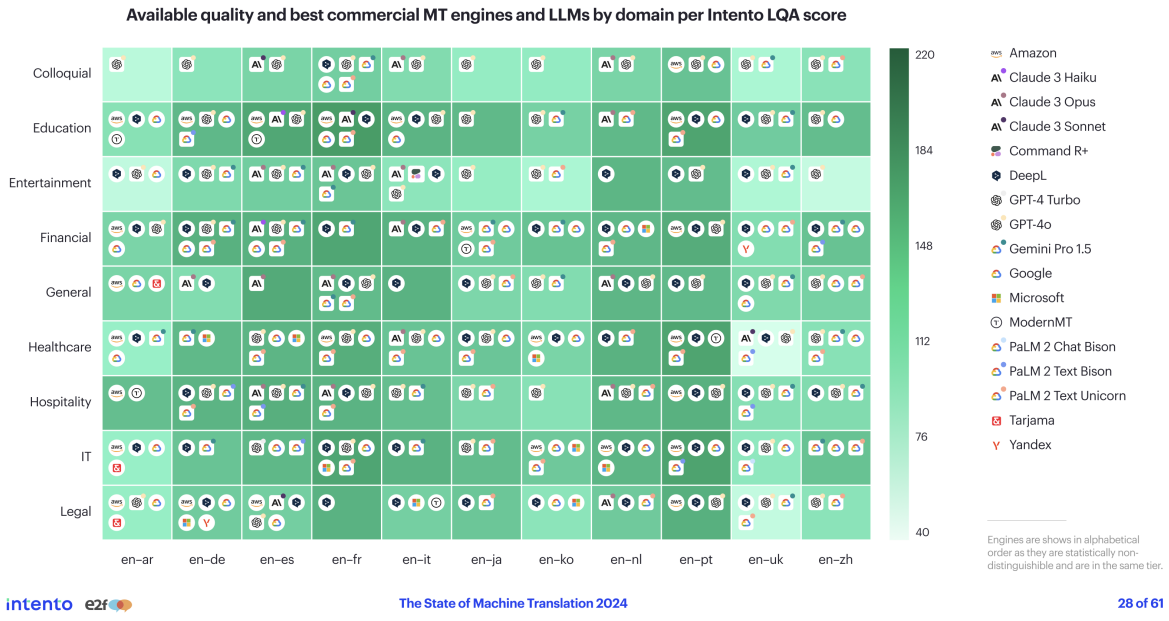

Figure 1: Quality of MT engines and LLMs by domains (y-axis) and language pairs [49]

Furthermore, BLEU [39] has been shown to be inconsistent depending on tokenisation and parameters that are used [41]. While most researchers do report on these specifications explicitly, different BLEU parametrisations between different papers still make a direct comparison challenging to do.

This study will lean towards evaluating MT-specific models rather than general LLMs. Through simple experimentations, this paper aims to evaluate the performance of existing and publicly available MT pre-trained models (MT-PTMs) on many-to-English translation across 14 source languages. Although fine-tuning multilingual MT-PTMs has been proven to increase

translation performance [13], no pre-training or fine-tuning is performed in this study due to simplicity reasons and resource limitations. A dataset is curated from the Tatoeba repository [46], containing 1,323 parallel sentence pairs across source and target languages. The models include a one-to-one PTMs: OPUS-MT [53, 52], and multilingual MT-PTMs: such as mBART-50 [30], NLLB-200 [47], and M2M-100 [18]. The performance of these models is evaluated using the BLEU [39], SacreBLEU [41], and METEOR [28].

The result indicates that OPUS-MT outperforms all other multilingual models in one-to-English translation on both SacreBLEU and METEOR scores, highlighting the superiority of specific languages MT-PTMs compared to unfine-tuned multilingual MT-PTMs. Among the multilingual models, mBART-50 shows the best performance, while NLLB-200 performs the worst, by a significant margin. Ukrainian-to-English translation by OPUS-MT receives the highest scores, however, Chinese, Italian, Polish, and Turkish are better on average across the four models.

2 Literature Review

2.1 Pre-Trained Models for Machine Translation

Attention mechanism has been used for Neural Machine Translation (NMT) [5, 33] even before the introduction of the transformer [56]. The Transformer, with its encoder-decoder architecture [56], has since become the foundational framework for many sequence-to-sequence models in machine translation [12]. By combining this architecture with training on large datasets, models were able to capture and retain vast amounts of linguistic information, leading to the development of pre-trained models.

Currently, pre-trained models have become a cornerstone in machine translation, achieving state-of-the-art performances [20]. While general Large Language Models (LLMs) such as BERT [15], GPT [37], or LLAMA [54] can be used for machine translation tasks, in this study the focus will be towards MT-specific PTMs (MT-PTMs) which are specifically designed, built, and trained for machine translation tasks. Most MT-PTMs typically support two kinds of translations: one-to-one translation and many-to-many translation (multilingual models).

One-to-one translation refers to a translation approach where a model is specifically trained to translate between one source language and one target language. This setup is characterised by having a dedicated model for each unique language pair. OPUS-MT [52] is a project dedicated to creating and providing free resources for machine translation, providing over a thousand MT-PTMs of direct one-to-one translations. Additionally, Facebook Fairseq [38] similarly offers one-to-one MT-PTMs, although they are only limited to translations between English, French, German, and Russian.

Many-to-many translation allows for a translation between multiple pairs of source and target languages. This is done by training a single multilingual model, typically trained on multilingual datasets. Multilingual models have recently gained significant prominence, shown to be simpler and efficient [3]. mBART [30], M2M [18], NLLB-200 [47], are examples of massively multilingual models that support many-to-many translation language.

2.2 Parallel Corpora

Corpora simply means a collection of texts, which in this case are used for the purpose of training models. Thus, parallel corpora are defined as sets of texts in a given source language along with their translations in another target language [29]. They can be bilingual or multilingual and this is crucial for training and evaluating machine translation. Languages without or with limited parallel corpora are referred to as low-resource languages.

Particularly for these low-resource languages, where there are not enough parallel corpora, English has often been used as an intermediary or 'pivot' language. This means texts originally

written in languages are first translated into English, and then from English into other languages, resulting in most target texts being translations of translations [29]. This is called indirect translation and poses many problems in translations as terms can be ambiguous with varying or similar meanings [4].

Since Neural Machine Translation (NMT) systems require vast amounts of training data, the availability of parallel corpora is crucial for building effective models [27]. The lack of extensive parallel corpora leads to suboptimal performance in NMT techniques [43]. Today, numerous websites, repositories, and datasets exist containing large amounts of parallel texts between specific languages.

OPUS [50] provides a comprehensive collection of open-source parallel corpora dedicated to machine translation (MT) tasks. It includes corpora for 744 languages and contains over 1,210 different datasets, amassing a total of 45,945,946,108 sentence pairs. Tatoeba [46] is another prominent resource in the field of MT and NLP, known for its extensive collection of translated sentences. As of July 2024, it contains 12,186,207 sentences over 423 supported languages, growing daily through volunteer contributions. There are currently 86,915 members contributed through the Tatoeba website.

A number of renowned datasets are commonly used for evaluation in machine translation. The Tatoeba Challenge [51] covers 487 languages in 4,024 language pairs, including 657 test sets sourced from the Tatoeba website, covering 138 languages. The TED Talks dataset [62] contains transcripts from TED talks for more than 50 languages. IWSLT [2] also contains TED talks data, but paired with English translations. Finally, WMT (Conference on Machine Translation) [6], previously named Workshop on Statistical Machine Translation, is an annual event that organises tasks for machine translation and provides a collection of datasets for benchmarking and evaluating translation systems.

While these datasets can be used to evaluate translation performance between any language pairs, they mostly do not contain parallel corpora across different languages. A corpus between English and French, for example, contains different sets of text than a corpus between English and Spanish. Therefore, they do not support easy evaluation for many-to-English translations. Furthermore, research papers often use different datasets and versions to evaluate their model performance, making direct comparisons even more challenging. mBART [30] uses WMT19 for English-German translation and TED15 for Chinese-English translation. M2M-100 [18] uses seven different datasets shared across languages. NLLB-200 [47] uses the FLORES dataset, designed for low-resources languages. OPUS-MT [53, 52] models are evaluated using the datasets from Tatoeba challenges.

3 Methodology

3.1 Dataset

The Tatoeba website [46] includes a vast, continuously expanding database consisting of sentences and their translations, built through the contributions of thousands of volunteers. They currently have 12,132,349 sentences of over 423 supported languages, with around one to two thousand new sentences added daily, on average. The English sentence dataset contains 1,905,089 sentences, the largest one in their repository, with Russian in second place with 1,066,633 sentences. Figure 2 shows the repository, sorted from the biggest corpus.

Table 1 shows the 14 language pairs selected for this study. Languages are chosen based on the resources' availability in Tatoeba and supported languages in most MT-PTMs. Thus, the languages chosen here can be considered moderate to high-resource languages.

To build the dataset, 1,898,494 English sentences were initially downloaded, though the number is less than what is listed on the Tatoeba website for reasons that remain unclear. For each of the 14 source languages, sentence pairs in English were individually downloaded and

Figure 2: Tatoeba's repository by languages, sorted by sentences count [46]

No.	Language Pairs	Sentence Pairs Count
1	Chinese-English	68,814
2	Dutch-English	155,856
3	Finnish-English	102,202
4	French-English	405,088
5	German-English	501,145
6	Hebrew-English	172,082
7	Hungarian-English	171,698
8	Italian-English	624,160
9	Japanese-English	270,116
10	Polish-English	77,345
11	Russian-English	722,837
12	Spanish-English	265,253
13	Turkish-English	710,279
14	Ukrainian-English	214,244
	Total	4,461,119

Table 1: List of chosen languages for evaluation

compiled. The English corpus was then merged with each compiled source language corpus to create parallel sentences across 14 languages. In each corpus, multiple translations for the same English phrase were dropped, retaining only the first translation to ensure that each English sentence had only a single translation reference. The final dataset contains 1,323 parallel sentences in English and all 14 source languages. This corpus will be used as a test set to evaluate the models' performance in each language.

Sentences typically consist of everyday phrases such as 'I have to go to sleep', 'That is intriguing', and 'Where do you live?'. They may also include single-word exclamations like 'Speak!', 'So what?', or 'Look!'. Additionally, multiple sentences such as 'You may write in any language you want. On Tatoeba, all languages are considered equal', and 'Guns don't kill people. People kill people' can be found inside the corpus. A few of them also include human names, 'Compare your answer with Tom's', and 'Muiriel is 20 now'. All the sentences seem to be straightforward and literal, without any use of linguistic features such as metaphors or sarcasm. A word cloud is illustrated in Figure 3, while Figure 2 shows a couple of examples of parallel sentences in the final dataset, across all 14 languages and the original English sentence.

Figure 3: Dataset word cloud illustrated.

Figure 4 shows a box plot of each language sentence's word count. Note that for Chinese and Japanese the word count is counted for each Chinese/Japanese letter instead, which explains their higher range and larger numbers of outliers.

Most of the languages' interquartile range (IQR) is quite low (around two) with a centre median, except for Chinese, Italian, and Japanese. Finnish and Hebrew have the lowest median, while Polish, Russian, Spanish, Turkish, and Ukrainian have a median closely similar to each other. English, Dutch, French, German, and Italian have a slightly higher median, with Italian also having a wider IQR. Lastly, Japanese and Chinese have the highest median, the widest range, and significantly more outliers.

Language	Sentence 1	Sentence 2
English	I have to go to sleep.	So what?
Chinese	我该去睡觉了。	那又怎?
Dutch	Ik moet gaan slapen.	Dus?
Finnish	Minun täytyy mennä nukkumaan.	Mitä sitten?
French	Je dois aller dormir.	Et alors?
German	Ich muss jetzt schlafen.	Na und?
Hebrew	<hidden-due-to-latex-incompatibility>	<hidden-due-to-latex-incompatibility>
Hungarian	Aludni kell mennem.	És akkor mi van?
Italian	Devo andare a dormire.	E allora?
Japanese	私は眠らなければなりませ	だから何?
Polish	Muszę iść spać.	No i co?
Russian	Мне пора идти	Так что?
Spanish	Tengo que irme a dormir.	¿Entonces qué?
Turkish	Yatmaya gitmek zorundayım.	Öyleyse ne yapmalı?
Ukrainian	Маю пти спати.	Ну то що?

Table 2: A snippet of the dataset

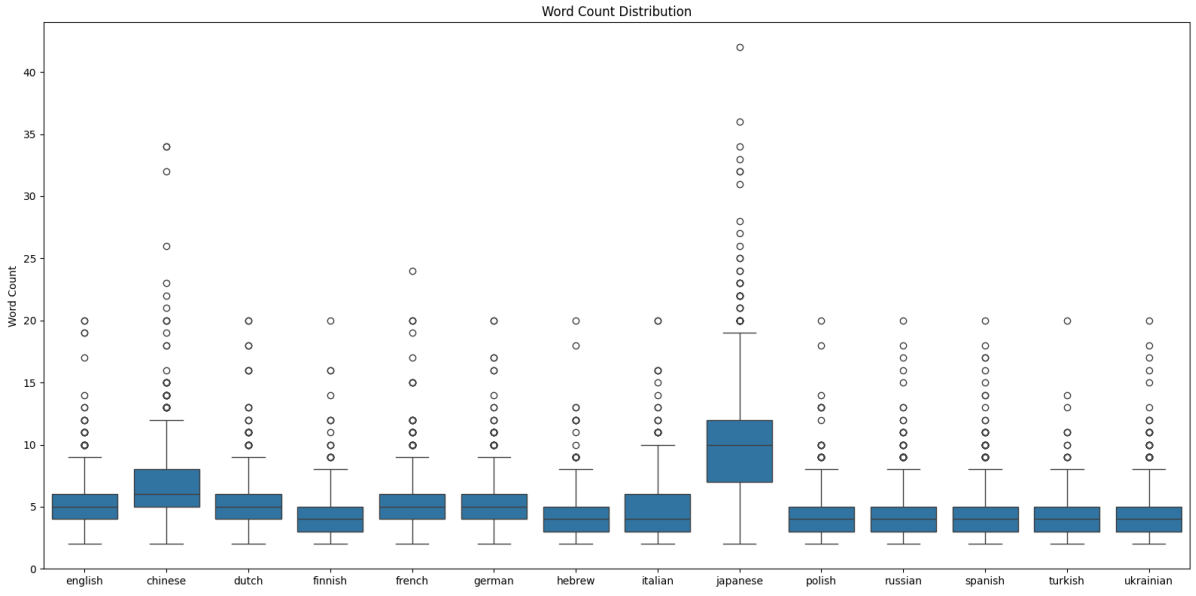

Figure 4: Dataset word count distribution per language, Chinese and Japanese sentences are counted per letter.

3.2 Proposed Pre-Trained Models

3.2.1 OPUS-MT

OPUS-MT [53, 52] provides over 1,000 pre-trained models for translation between numerous language pairs. The architecture is based on MARIAN-NMT [26], based on a standard transformer setup: 6 self-attentive layers in both the encoder and decoder networks, each with 8 attention heads per layer [52]. The models were trained for up to 72 hours on 1 or 4 GPUs, noting that not all of them fully converged within this time frame.

While the framework provides both monolingual and multilingual models, only the OPUS-MT monolingual models are used in this study. This is partly due to the lack of documentation on the multilingual models.

3.2.2 mBART-50

mBART [30] is a sequence-to-sequence denoising auto-encoder model specifically designed for multilingual tasks. The mBART-50 variant supports many-to-many translations for over 50 languages.

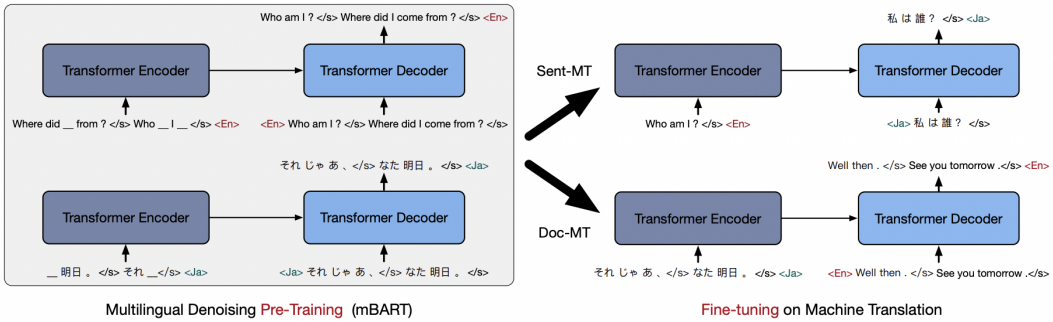

Figure 5: mBART framework for Multilingual Denoising Pre-training (left) and fine-tuning on downstream MT tasks (right) [30]

The mBART model is built as a typical sequence-to-sequence transformer architecture [56], using an embedding dimension of 1024 and 16 heads within 12 layers of encoder and 12 layers of decoder, totalling around 680M parameters [30].

Figure 5 shows mBART's key features: multilingual denoising pre-training and fine-tuning. The dataset is first corrupted by a noising function, which is applied to 35% of words of each span text [30]. During pre-training, the model will be made to denoise the texts, improving the model's ability to generate coherent and accurate translations. After the pre-training step, the model is then able to be fine-tuned for downstream tasks. It is shown that the model achieved consistent performance gains through pre-training in low-to-medium resource sentence-level translation [30]

3.2.3 M2M-100

M2M-100 [18] is designed to perform direct translation between any pair of 100 languages without relying on English as an intermediate language. Similarly, the architecture is of a transformer sequence-to-sequence with a typical encoder that outputs embeddings of the same length, while the decoder is an autoregressive component that sequentially produces the target sentence one token at a time. The parameters include 12 encoder layers and 12 decoder layers, 8192 hidden layer sizes, and 1024 embedding dimensions, totalling a 1.2B parameter count [18].

Figure 6 shows the inner workings of the M2M model. The English-centric dataset (top left) includes training data exclusively involving translations to and from English, while the

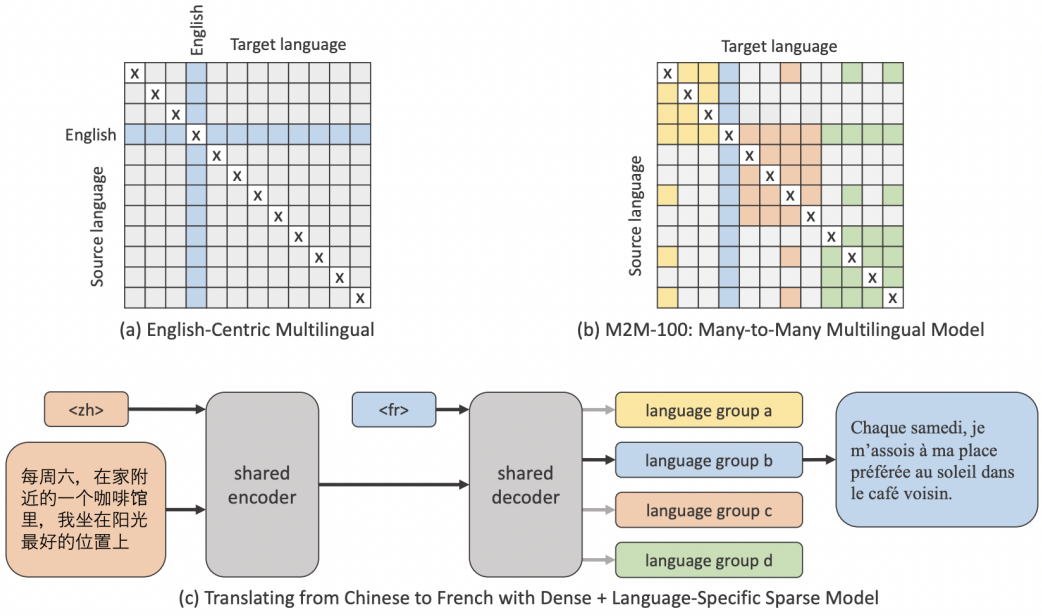

Figure 6: Summary of M2M dataset and multilingual model.

many-to-many multilingual setting (top right) involves direct translation data among multiple language pairs [18]. Finally, the model combines sparse language-specific parameters with dense parameters (bottom part of Figure 6),

3.2.4 NLLB-200

NLLB-200 [47] is built to handle translation tasks across a broad spectrum of languages, including many that are low-resource or underrepresented in existing datasets. It supports translations for 200 languages, including numerous underrepresented languages, and is currently one of the most extensive multilingual machine translation models. The model employs a Mixture of Experts (MoE) architecture and achieves state-of-the-art (SoTA) results across many language pairs, even surpassing Meta's previous model, M2M-100 [47].

The 'nllb-200-distilled-600M' variant has a total of 600M parameters, distilled from originally 54B parameters [47]

Figure 7 shows the Dense Transformer and MoE Transformer layers implemented within the model. MoE [34] utilises a gating mechanism to route different inputs to different subsets of experts (sub-networks), allowing the model to handle various linguistic phenomena efficiently. Additionally, the model is then trained on a diverse set of languages, complemented with back-translation and data augmentation to generate additional data for low-resource languages [47].

3.3 Evaluation Metrics

In this study, BLEU [39], SacreBLEU [41], and METEOR [28] are applied as evaluation metrics due to their popularity, simplicity, and ease-of-use. While other metrics such as Crosslingual Optimized Metric for Evaluation of Translation (COMET) [44] and BERTSCore [63] exist, they involve using a deep neural network or a transformer to evaluate, which naturally increase computational cost and slow down calculation greatly.

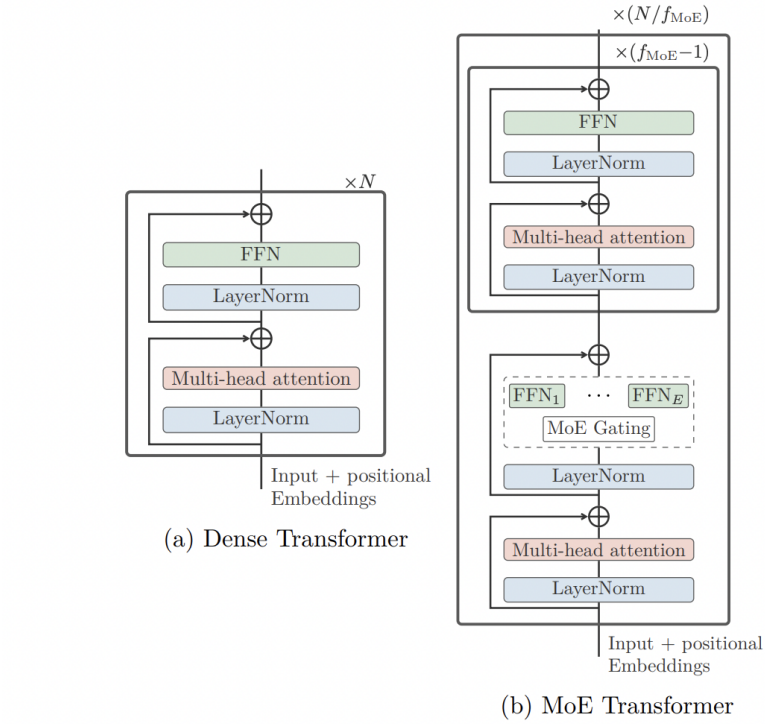

Figure 7: NLLB pipeline [47]

3.3.1 BLEU

Bilingual Evaluation Understudy (BLEU) [39] is the most commonly used metric for machine translation (MT). It assesses how well a candidate translation matches the reference translation by using precision metrics for n-grams and a brevity penalty to ensure a good match in regard to length, word choice, and word order.

The BLEU score is calculated as:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{1}$$

where BP is the brevity penalty, p_n is the precision for n-grams, and w_n is the weight for each n-gram (often uniformly distributed, so $w_n = \frac{1}{N}$).

The brevity penalty (BP) is calculated as:

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-\frac{r}{c})} & \text{if } c \leq r \end{cases} \tag{2}$$

where c is the length of the candidate translation and r is the effective reference length.

The n-gram precision, as presented in the original BLEU paper [39], is calculated as:

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{\text{n-gram} \in C} \text{Count}_{\text{clip}}(\text{n-gram})}{\sum_{C \in \{Candidates\}} \sum_{\text{n-gram} \in C} \text{Count}(\text{n-gram})} \tag{3}$$

where p_n is the precision for n-grams, $\sum_{C \in \{Candidates\}}$ denotes the summation over all candidate translations, $\sum_{\text{n-gram} \in C}$ denotes the summation over all n-grams in a candidate translation C. $\text{Count}_{\text{clip}}(\text{n-gram})$ is the clipped count of the n-gram, which is the count of the n-gram in the candidate translation limited by the maximum count of that n-gram in any reference translation. $\text{Count}(\text{n-gram})$ is the count of the n-gram in the candidate translation.

The machine translation community relies heavily upon the BLEU score, despite its several drawbacks [31]. The metric has been reported to not correlate strongly with human judgement, showing variations in translation that could mean that a higher BLEU score does not necessarily indicate a true enhancement in translation quality [8]. Furthermore, the metric is tokenisation-dependant, making it challenging to directly compare BLEU scores between paper [41]. Thus, the author proposed a standardised variant called SacreBLEU [41], which is essentially a Python script that applies its own metric-internal preprocessing and produces the same values as WMT [6], utilising the WMT (2008–2018) and IWSLT 2017 [10] test sets.

3.3.2 METEOR

Metric for Evaluation of Translation with Explicit ORdering (METEOR) [28] assesses a translation by calculating a score that reflects explicit word-to-word matches between the reference and a candidate translation [1]. It is designed to address some limitations of the BLEU score, allowing matches between simple morphological variants and synonyms. The formula is defined as:

$$\text{METEOR} = (1 - \gamma \cdot \text{frag}) \cdot \frac{P \cdot R}{\alpha \cdot P + (1 - \alpha) \cdot R} \tag{4}$$

where P is the precision, R is the recall, and frag is the fragmentation penalty. γ is a parameter that controls the weight of the fragmentation penalty, commonly set to 0.5, and α is a parameter that controls the balance between precision and recall, commonly set to 0.9.

Although the latest official version of METEOR (METEOR-1.5) was released in 2014 [14], the NLTK library [7] still implements the older version based on the 2007 paper by Lavie and Agarwal [28]. Therefore, this study will use the older version of METEOR, as the calculations will rely on the NLTK library.

3.4 Inference Details

OPUS-MT models are taken from the Helsinki-NLP repository through the Hugging Face website [17, 64] with the exception of Hebrew-to-English models taken directly from 'tiedeman/opus-mt-he-en'. The 'mbart-large-50-many-to-many-mmt', the 'm2m100_418M', and the 'nllb-200-distilled-600M' model variants are used for mBART-50 [30], M2M-100 [18], and NLLB-200 [47], respectively. Similarly, all multilingual models are downloaded using the transformers [60] library from Hugging Face [17]. A batch size of 4 is used in all models. Translations from each source language to English are first generated through inference and saved into CSV files, before calculating the metrics score.

For BLEU and METEOR, tokenisation and metrics calculation use the NLTK [7] package, specifically:

- `nltk.word_tokenize`
- `nltk.translate.meteor_score`
- `nltk.translate.bleu_score.corpus_bleu`

Additionally, the smoothing method 7 is used for the BLEU calculation, which interpolates two other smoothing functions: smoothing method 4, which divides by $\frac{1}{\ln(\text{len}(T))}$, where T is the length of the translation; and smoothing method 5, which averages the $n-1$, n, and $n+1$ gram matched counts [48]. The `sacrebleu` [41] package from pip is used to calculate SacreBLEU [41].

All codes are written in Python 3.12.2 and run locally on MacOS Sonoma 14.4.1 with an M2 chip and 16 GB memory. Datasets and scripts will be available on GitHub: `https://github.com/stefanliemawan/eval-machine-translation`

4 Evaluation

Language	BLEU	SacreBLEU	METEOR
Chinese	0.0971	*59.5098*	0.7953
Dutch	0.0974	69.8803	0.8471
Finnish	0.0965	66.6267	0.8296
French	0.0967	69.8185	0.8357
German	0.0969	69.7422	0.8419
Hebrew	0.0977	66.5149	0.8229
Italian	0.0979	74.1298	0.8584
Japanese	0.0960	63.1435	*0.7893*
Polish	*0.0952*	61.9026	0.8425
Russian	0.0974	66.7046	0.8179
Spanish	0.0979	71.4174	0.8463
Turkish	**0.0980**	72.6551	0.8460
Ukrainian	0.0978	**75.4447**	**0.8667**

Table 3: OPUS-MT result.

Language	BLEU	SacreBLEU	METEOR
Chinese	0.0962	54.7322	0.7600
Dutch	0.0970	63.6482	0.8007
Finnish	0.0963	47.2194	0.7067
French	0.0968	57.2482	0.7598
German	0.0967	63.1666	0.8022
Hebrew	0.0972	58.0846	0.7590
Italian	0.0972	**65.9415**	**0.8068**
Japanese	0.0943	43.9547	0.7151
Polish	**0.0978**	62.7550	0.7923
Russian	0.0974	58.8820	0.7686
Spanish	*0.0883*	*35.1593*	0.7306
Turkish	0.0975	50.9377	*0.6982*
Ukrainian	0.0964	55.8637	0.7496

Table 4: mBART50 result.

Language	BLEU	SacreBLEU	METEOR
Chinese	0.0975	43.3490	0.6848
Dutch	0.0976	52.6920	**0.7448**
Finnish	0.0975	49.9149	0.7145
French	0.0969	51.4090	0.7288
German	0.0969	52.7877	0.7423
Hebrew	0.0981	50.3287	0.7198
Italian	0.0975	**53.1644**	0.7341
Japanese	*0.0959*	*41.6610*	*0.6511*
Polish	**0.0984**	52.5151	0.7301
Russian	0.0976	48.2725	0.7027
Spanish	0.0977	52.9323	0.7371
Turkish	0.0978	51.3009	0.7262
Ukrainian	0.0978	46.5555	0.6890

Table 5: M2M-100 result.

Language	BLEU	SacreBLEU	METEOR
Chinese	**0.0949**	**50.4454**	**0.7180**
Dutch	0.0915	20.8988	0.2899
Finnish	0.0921	32.3926	0.4006
French	0.0950	32.1389	0.4338
German	0.0903	7.1436	0.1528
Hebrew	0.0897	2.3455	0.1412
Italian	0.0926	24.2702	0.3059
Japanese	*0.0842*	*1.9569*	*0.0919*
Polish	0.0937	38.4395	0.4739
Russian	0.0918	20.6734	0.2977
Spanish	0.0929	26.3313	0.3578
Turkish	0.0945	45.2721	0.6187
Ukrainian	0.0914	9.0340	0.1613

Table 6: NLLB-200 result.

Table 3, 4, 5, and 6, and shows the result from all four PTMs. All models perform similarly low on BLEU score (below 10) across all languages, with NLLB-200 reaching 0.01 below on Japanese translation. As BLEU is notoriously unreliable for very short texts, these results are not surprising. The BLEU score can be disproportionately affected by small differences in shorter texts due to the limited number of n-grams available for comparison, as well as the brevity penalty [11, 39].

OPUS-MT model shows the best performance in all languages, consistently reaching mostly above 60 SacreBLEU and 0.8 METEOR. The highest-rated languages are Ukrainian-to-English translation with 75.44 SacreBLEU and 0.86 METEOR, while the lowest scores are achieved on Chinese-to-English translation: 59.50 SacreBLEU and Japanese-to-English: 0.78 METEOR.

The mBART-50 model is the second-best-performing PTM and the best-performing multilingual model, achieving the best SacreBLEU and METEOR score in Italian-to-English translation while suffering the most in Spanish-to-English (SacreBLEU) and Turkish-to-English (METEOR) translation. Interestingly, Spanish receives the worst score in SacreBLEU while it performs decently in METEOR, and Turkish performs well in SacreBLEU score but has the worst in METEOR. Additionally, the model slightly outperforms OPUS-MT in Polish-to-English translation on SacreBLEU score (0.85 higher).

Similarly, SacreBLEU for Italian-to-English translation outperforms other languages with the M2M-100 model, although Dutch-to-English surpasses Italian-to-English by 0.01 in METEOR score. Japanese received the worst score in both SacreBLEU and METEOR for this model.

M2M-100 performs worse in SacreBLEU compared to mBART-50, but comparably similar in METEOR scores.

NLLB-200 scores much lower compared to other models in both SacreBLEU and METEOR in almost all languages. It scores highest in Chinese-to-English translation in all metrics, while showing significant variability among other languages. However, Japanese-to-English translation suffers with extremely low scores of 1.95 SacreBLEU and 0.09 METEOR. Additionally, German, Hebrew, and Ukrainian also have low SacreBLEU and METEOR scores, among the other languages.

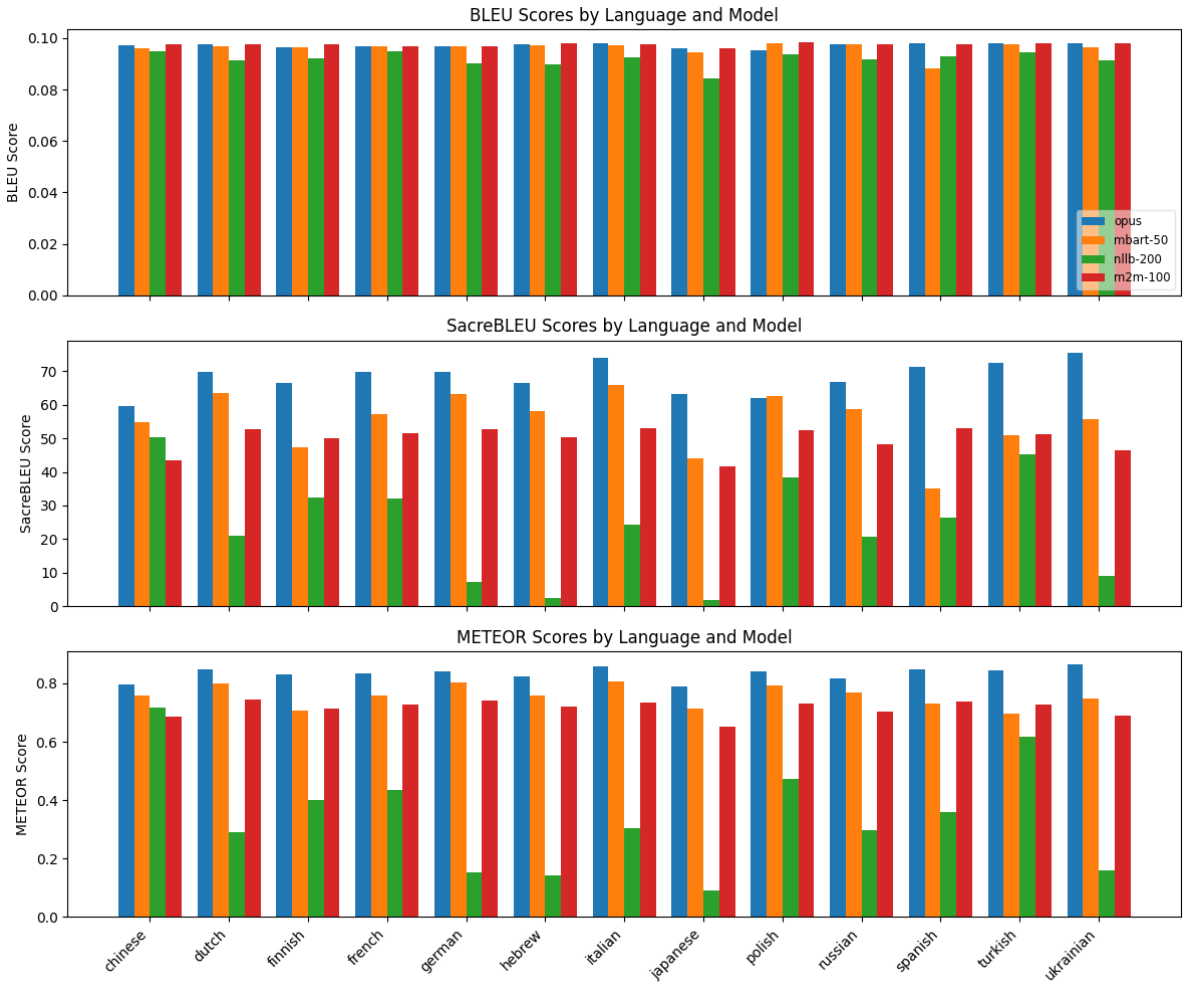

Figure 8: BLEU, SacreBLEU, and METEOR scores of each model on each language pair.

The BLEU, SacreBLEU, and METEOR scores gained by each model, categorised by every language pair are illustrated in Figure 8. Dutch, Italian, and Ukrainian translations generally score higher across all metrics, suggesting that these languages are well-supported by the pre-trained models. Japanese translations, however, pose significant challenges for all models, as indicated by the consistently low scores across all metrics. Japanese-to-English translation seems to have the worst score on average across the four models, but particularly bad on multilingual models. In contrast, Dutch and German have consistently high results on all models except for NLLB-200, often reaching close to the highest language SacreBLEU and METEOR scores.

Figure 9 shows the SacreBLEU and METEOR scores for each language, categorised by model. The OPUS-MT and mBART-50 models are again shown as the clear winners in almost all languages. M2M-100 also performs competitively with mBART-50 in most languages, surpassing mBART-50 in Spanish-to-English translation. NLLB-200 is clearly shown to suffer greatly across

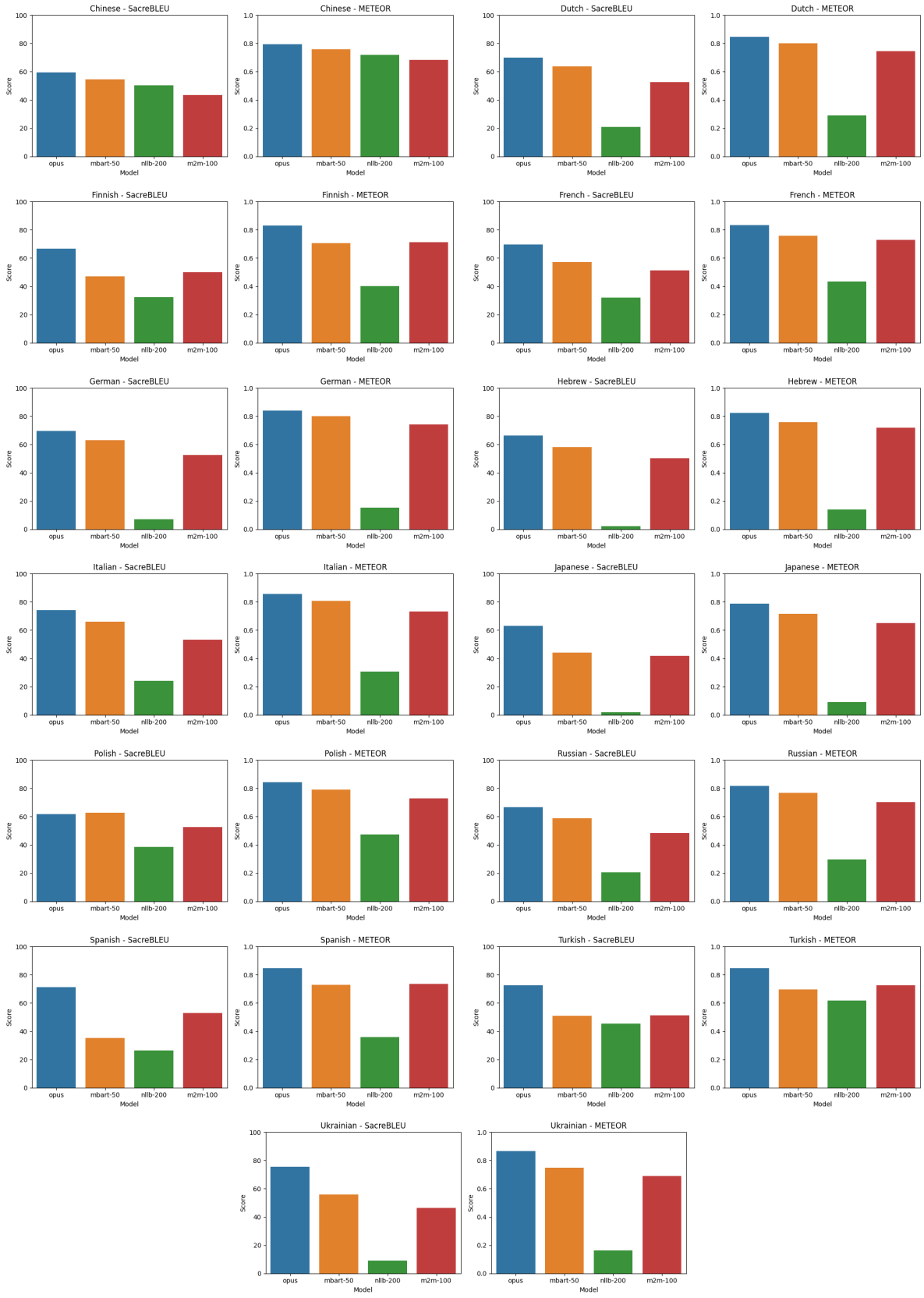

Figure 9: Performance of every language.

all languages besides Chinese. All four models seem to perform consistently against each other in Chinese-to-English and Turkish-to-English translations.

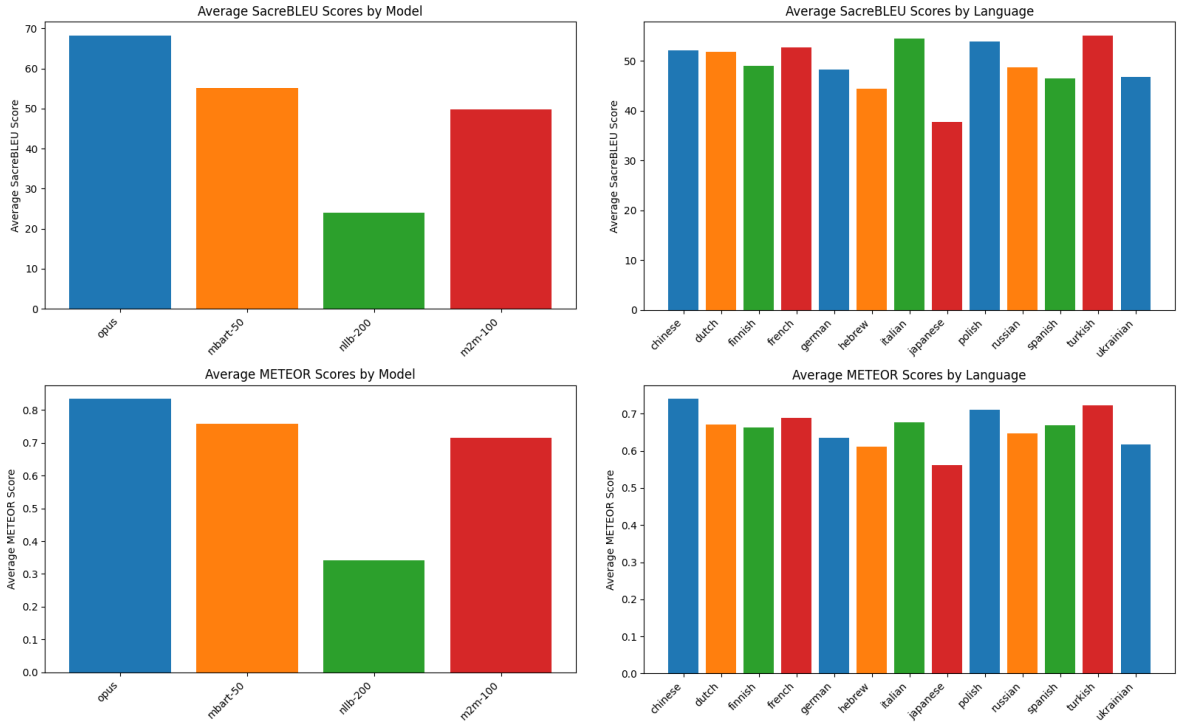

Figure 10: The average BLEU, SacreBLEU, and METEOR scores by model (left) and by language (right)

Visualisations of the average SacreBLEU and METEOR, categorised by the model (left figures) and language (right figures) are shown in Figure 10, providing a closer insight into how each model and language perform independently. It can be seen that when averaged across all languages, mBART-50 performs more closely to OPUS-MT on the METEOR score than the SacreBLEU score. Language-wise, Italian, Polish, and Turkish achieve the best average SacreBLEU scores while Chinese, Polish, and Turkish are the top three languages with the highest average METEOR scores.

Figure 11 highlights the difference between each OPUS-MT model's reported BLEU scores compared to the SacreBLEU result of this study. The reported BLEU is evaluated on the Tatoeba dataset [52]. It can be seen that the scores are quite aligned, with Ukrainian and Italian reaching the two highest scores while Chinese, Polish, and Japanese suffer.

While this study's experiments resulted in a moderately higher overall SacreBLEU, the dataset used here is also significantly smaller than the full Tatoeba challenge dataset [51] that is used in the OPUS-MT reported BLEU. In contrast, this study employs the same set of parallel sentences for all languages, which should lead to more consistent results. Nevertheless, a direct comparison between these results is not fully pertinent.

Additionally, Table 7 shows the full scores of both results SacreBLEU and reported BLEU from each OPUS-MT model [52]. The reported BLEU scores are taken from each model card in the HuggingFace repository [17]. Unfortunately, other models are either not benchmarked on the Tatoeba dataset, not evaluated on the source languages used in this study, or use different metrics [30, 18, 47], and therefore cannot be directly compared to the results of this study.

Lastly, Figure 12 shows the time taken to translate all 1,323 sentences for each language across experimented models. The OPUS-MT model is the fastest among the other three multilingual models. M2M-100 and NLLB-200 achieve similar timings, while mBART-50 has slightly longer

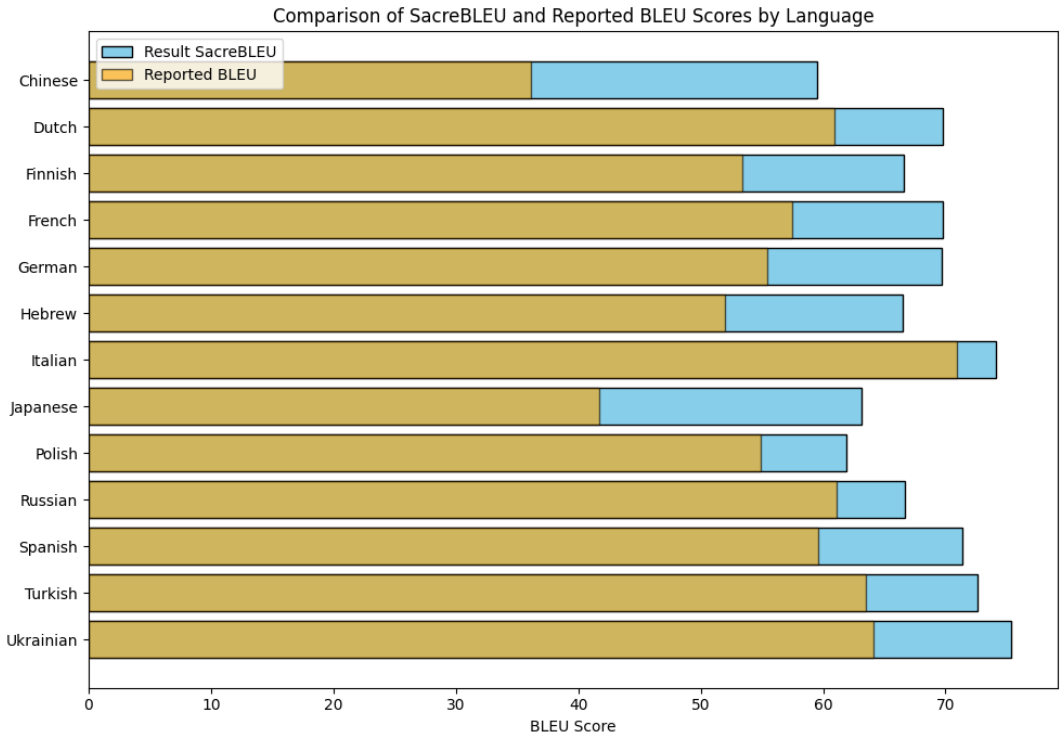

Figure 11: Comparison between reported OPUS-MT models performance and performance in this experiment.

Language	**SacreBLEU**	**Reported BLEU**
Chinese	*59.5*	*36.1*
Dutch	69.8	60.9
Finnish	66.6	53.4
French	69.8	57.5
German	69.7	55.4
Hebrew	66.5	52.0
Italian	74.1	70.9
Japanese	63.1	41.7
Polish	61.9	54.9
Russian	66.7	61.1
Spanish	71.4	59.6
Turkish	72.6	63.5
Ukrainian	**75.4**	**64.1**

Table 7: OPUS-MT result compared with reported BLEU result

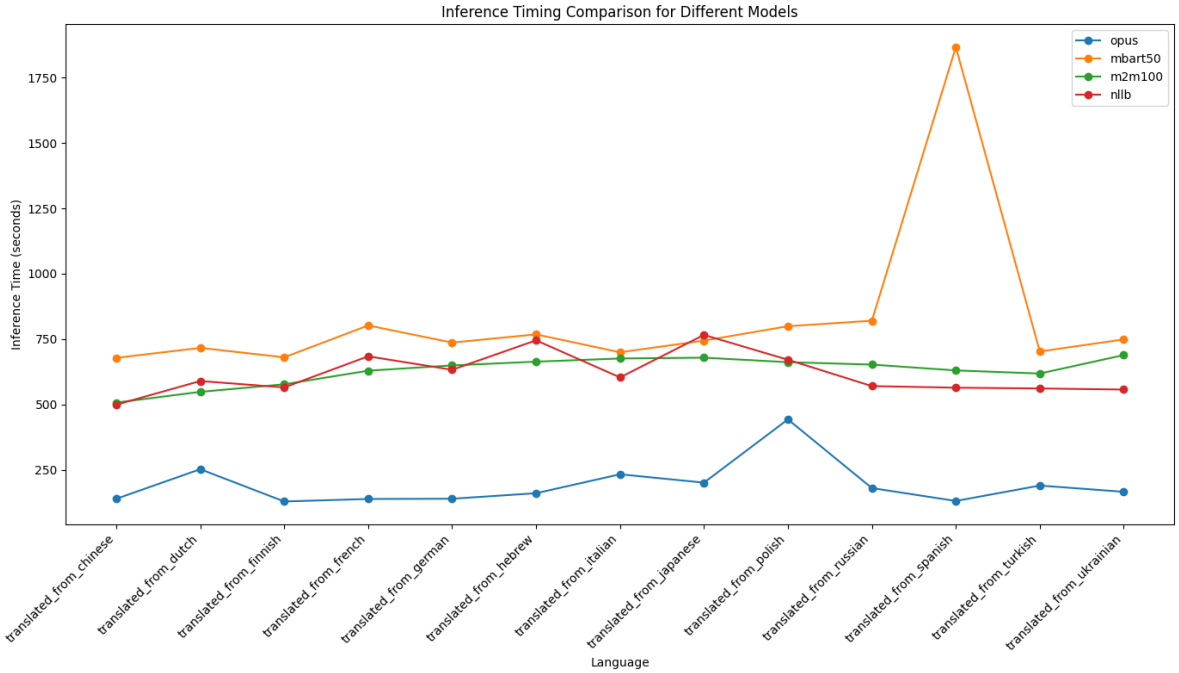

Figure 12: Inference timings for every model.

timings on average. It is unknown why mBART50 suffers from a significantly higher timing on the Spanish-to-English translation. Re-running the experiment shows that this timing is consistent and not an anomaly.

5 Conclusion

This study created a simple dataset containing 1,323 parallel sentence pairs in 14 moderate-to-high-resource languages and evaluated several monolingual and multilingual Machine Translation Pre-Trained Models (MT-PTMs). The findings suggest that while multilingual models have been shown to outperform one-to-one translation models [30], this premise does not necessarily hold in our experiments. Specifically, with a small dataset consisting of short sentences, many-to-English translations across 14 languages with monolingual MT-PTMs (OPUS-MT) outperform unfine-tuned multilingual MT-PTMs (mBART-50, M2M-100, NLLB-200).

OPUS-MT achieves the best performance across all 14 languages. mBART-50 is the second-best-performing model, beating OPUS-MT only in Polish SacreBLEU. Nevertheless, both the mBART-50 and the M2M-100 models perform generally well in all languages. NLLB-200, however, achieves much lower scores in both SacreBLEU and METEOR for all languages except for Chinese, slightly outperforming M2M-100 but still slightly below mBART-50. NLLB-200 should then be avoided for use without any further fine-tuning on these languages. Note that the distilled version is used in this study for NLLB-200 instead of the larger 3.3 billion parameters model variant.

Chinese, Italian, Polish, and Turkish have the best average scores over all four models. Ukrainian by OPUS-MT has the highest SacreBLEU and METEOR scores across all the other models and languages. These results show that these languages are well-supported within the four (MT-PTMs) evaluated in this study, achieving decent results without any fine-tuning.

6 Discussion

In curating the dataset, only the first translation of the same English phrases was taken, while additional translations were discarded. This ignores several possible translation references which might include synonyms or different phrasings, thus potentially affecting the evaluation score, especially for BLEU and SacreBLEU metrics. Moreover, the dataset mainly consists of short sentences with limited linguistic depth. Whether the result of this study would stay consistent with longer sentences or more complex documents should be further explored. The Tatoeba dataset [51] consisting of short sentences has been reported to be too optimistic when evaluated with other datasets that are more realistic [52].

The METEOR implementation in this study through the NLTK library is based on the original METEOR paper [28], which is described as "relatively simple and naive." A newer version, METEOR-1.5 [14] offers improved features and has been reported to produce significantly different results [55]. Therefore, re-evaluating translation models using the updated metric might be beneficial. Experimenting with other metrics such as chrF [40], which is widely used by Meta AI [47] could also provide a more comprehensive evaluation.

As Tatoeba [46] is run mostly by volunteers, there are also several limitations with its corpus. The quality of sentences can vary significantly depending on the contributors' language proficiency and the level of community engagement for specific languages. Moreover, not all sentences are reviewed or verified by native speakers or trusted users.

The languages chosen are mostly high-resource languages, while models such as NLLB-200 [47] are specifically designed to address the problem with low-resource languages. Data leakage is also a possibility, as sentences included in this dataset are fairly common phrases, it is highly likely that the model has seen these during training.

Future works should incorporate more Pre-Trained Models (PTMs) and implement fine-tuning multilingual models for specific languages on languages-to-English translation as it can significantly improve performance [64]. Furthermore, it may be beneficial to consider more general multilingual PTMs such as mBERT [61] and PolyLM [58], which are trained for a wide range of NLP tasks instead of just machine translation.

References

[1] Abhaya Agarwal and Alon Lavie. "METEOR, M-BLEU and M-TER: evaluation metrics for high-correlation with human rankings of machine translation output". In: *Proceedings of the Third Workshop on Statistical Machine Translation*. StatMT '08. Columbus, Ohio: Association for Computational Linguistics, 2008, 115–118. ISBN: 9781932432091.

[2] Milind Agarwal et al. "Findings of the IWSLT 2023 Evaluation Campaign". In: *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*. 20th International Conference on Spoken Language Translation. IWSLT 2023 (Toronto, Kanada, July 13–14, 2023). Association for Computational Linguistics (ACL), 2023, 1–61.

[3] Roee Aharoni, Melvin Johnson, and Orhan Firat. *Massively Multilingual Neural Machine Translation*. 2019. arXiv: 1903.00089 [cs.CL]. URL: https://arxiv.org/abs/1903.00089.

[4] Hanna Pięta Alexandra Assis Rosa and Rita Bueno Maia. "Theoretical, methodological and terminological issues regarding indirect translation: An overview". In: *Translation Studies* 10.2 (2017), pp. 113–132. DOI: 10.1080/14781700.2017.1285247. eprint: https://doi.org/10.1080/14781700.2017.1285247. URL: https://doi.org/10.1080/14781700.2017.1285247.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. 2016. arXiv: 1409.0473 [cs.CL]. URL: https://arxiv.org/abs/1409.0473.

[6] Loïc Barrault et al. "Findings of the 2020 Conference on Machine Translation (WMT20)". In: *Proceedings of the Fifth Conference on Machine Translation*. Ed. by Loïc Barrault et al. Online: Association for Computational Linguistics, Nov. 2020, pp. 1–55. URL: https://aclanthology.org/2020.wmt-1.1.

[7] Steven Bird, Edward Loper, and Ewan Klein. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

[8] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. "Re-evaluating the Role of Bleu in Machine Translation Research". In: *11th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by Diana McCarthy and Shuly Wintner. Trento, Italy: Association for Computational Linguistics, Apr. 2006, pp. 249–256. URL: https://aclanthology.org/E06-1032.

[9] Mauro Cettolo, Christian Girardi, and Marcello Federico. "WIT3: Web Inventory of Transcribed and Translated Talks". In: *Proceedings of the 16th Annual conference of the European Association for Machine Translation*. Trento, Italy: European Association for Machine Translation, 2012, pp. 261–268. URL: https://www.aclweb.org/anthology/2012.eamt-1.60.

[10] Mauro Cettolo et al. "Overview of the IWSLT 2017 Evaluation Campaign". In: *Proceedings of the 14th International Conference on Spoken Language Translation*. Ed. by Sakriani Sakti and Masao Utiyama. Tokyo, Japan: International Workshop on Spoken Language Translation, 2017, pp. 2–14. URL: https://aclanthology.org/2017.iwslt-1.1.

[11] ChatGPT. *Explanation on BLEU's Limitations on Short Sentences*. OpenAI. 2024. URL: https://www.openai.com/chatgpt.

[12] Kyunghyun Cho et al. "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches". In: *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Ed. by Dekai Wu et al. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 103–111. DOI: 10.3115/v1/W14-4012. URL: https://aclanthology.org/W14-4012.

[13] Asa Cooper Stickland, Xian Li, and Marjan 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. DOI: 10.18653/v1/2021.eacl-main.301. URL: http://dx.doi.org/10.18653/v1/2021.eacl-main.301.

[14] Michael Denkowski and Alon Lavie. "Meteor Universal: Language Specific Translation Evaluation for Any Target Language". In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar et al. Baltimore, Maryland, USA: Association for Computational Linguistics, June 2014, pp. 376–380. DOI: 10.3115/v1/W14-3348. URL: https://aclanthology.org/W14-3348.

[15] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL]. URL: https://arxiv.org/abs/1810.04805.

[16] David M. Eberhard, Gary F. Simons, and Charles D. Fennig, eds. *Ethnologue: Languages of the World*. 27th. Dallas, Texas: SIL International, 2024. URL: http://www.ethnologue.com.

[17] Hugging Face. *Hugging Face: Natural Language Processing Made Easy*. https://huggingface.co. Accessed: 2024-08-07. 2024.

[18] Angela Fan et al. *Beyond English-Centric Multilingual Machine Translation*. 2020. arXiv: 2010.11125 [cs.CL]. URL: https://arxiv.org/abs/2010.11125.

[19] Xavier Garcia et al. *A Multilingual View of Unsupervised Machine Translation*. 2020. arXiv: 2002.02955 [cs.CL]. URL: https://arxiv.org/abs/2002.02955.

[20] Xu Han et al. "Pre-trained models: Past, present and future". In: *AI Open* 2 (2021), pp. 225–250. ISSN: 2666-6510. DOI: https://doi.org/10.1016/j.aiopen.2021.08.002. URL: https://www.sciencedirect.com/science/article/pii/S2666651021000231.

[21] John Hutchins. "Research methods and system designs in machine translation: a ten-year review, 1984-1994". In: *BCS International Academic Conference*. 1994. URL: https://api.semanticscholar.org/CorpusID:15952756.

[22] John Hutchins. "The development and use of machine translation systems and computer-based translation tools in Europe, Asia, and North America". In: 1998. URL: https://api.semanticscholar.org/CorpusID:18918684.

[23] John Hutchins. "The first public demonstration of machine translation : the Georgetown-IBM system, 7th January 1954". In: 2006. URL: https://api.semanticscholar.org/CorpusID:132677.

[24] William J. Hutchins. "Machine translation over fifty years". In: 2001. URL: https://api.semanticscholar.org/CorpusID:6196527.

[25] Albert Q. Jiang et al. *Mistral 7B*. 2023. arXiv: 2310.06825 [cs.CL]. URL: https://arxiv.org/abs/2310.06825.

[26] Marcin Junczys-Dowmunt et al. "Marian: Fast Neural Machine Translation in C++". In: *Proceedings of ACL 2018, System Demonstrations*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 116–121. URL: http://www.aclweb.org/anthology/P18-4020.

[27] Philipp Koehn and Rebecca Knowles. *Six Challenges for Neural Machine Translation*. 2017. arXiv: 1706.03872 [cs.CL]. URL: https://arxiv.org/abs/1706.03872.

[28] Alon Lavie and Abhaya Agarwal. "Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments". In: *Proceedings of the Second Workshop on Statistical Machine Translation*. StatMT '07. Association for Computational Linguistics, 2007, 228–231.

[29] Marie-Aude Lefer. "Parallel Corpora". In: *A Practical Handbook of Corpus Linguistics*. Ed. by Magali Paquot and Stefan Th. Gries. Cham: Springer International Publishing, 2020, pp. 257–282. ISBN: 978-3-030-46216-1. DOI: 10.1007/978-3-030-46216-1_12. URL: https://doi.org/10.1007/978-3-030-46216-1_12.

[30] Yinhan Liu et al. *Multilingual Denoising Pre-training for Neural Machine Translation*. 2020. arXiv: 2001.08210 [cs.CL]. URL: https://arxiv.org/abs/2001.08210.

[31] Arle Lommel. "Blues for BLEU: Reconsidering the Validity of Reference-Based MT Evaluation". In: *Proceedings of the LREC 2016 Workshop "Translation Evaluation – From Fragmented Tools and Data Sets to an Integrated Ecosystem"*. Ed. by Georg Rehm, Aljoscha Burchardt, et al. Portorož, Slovenia, 2016.

[32] Adam Lopez. "Statistical machine translation". In: *ACM Comput. Surv.* 40.3 (Aug. 2008). ISSN: 0360-0300. DOI: 10.1145/1380584.1380586. URL: https://doi.org/10.1145/1380584.1380586.

[33] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. *Effective Approaches to Attention-based Neural Machine Translation*. 2015. arXiv: 1508.04025 [cs.CL]. URL: https://arxiv.org/abs/1508.04025.

[34] Saeed Masoudnia and Reza Ebrahimpour. "Mixture of experts: a literature survey". In: *Artificial Intelligence Review* 42.2 (May 2012), 275–293. ISSN: 1573-7462. DOI: 10.1007/s10462-012-9338-y. URL: http://dx.doi.org/10.1007/s10462-012-9338-y.

[35] Satoshi Nakamura. "Overcoming the Language Barrier with Speech Translation Technology". In: *Science & Technology Foresight Center (NISTEP)* (2009).

[36] Margaret Dumebi Okpor. "Machine Translation Approaches: Issues and Challenges". In: 2014. URL: https://api.semanticscholar.org/CorpusID:11483090.

[37] OpenAI et al. *GPT-4 Technical Report*. 2024. arXiv: 2303.08774 [cs.CL]. URL: https://arxiv.org/abs/2303.08774.

[38] Myle Ott et al. "fairseq: A Fast, Extensible Toolkit for Sequence Modeling". In: *Proceedings of NAACL-HLT 2019: Demonstrations*. 2019.

[39] Kishore Papineni et al. "BLEU: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL '02. Association for Computational Linguistics, 2002, 311–318. DOI: 10.3115/1073083.1073135. URL: https://doi.org/10.3115/1073083.1073135.

[40] Maja Popović. "chrF: character n-gram F-score for automatic MT evaluation". In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar et al. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 392–395. DOI: 10.18653/v1/W15-3049. URL: https://aclanthology.org/W15-3049.

[41] Matt Post. *A Call for Clarity in Reporting BLEU Scores*. 2018. arXiv: 1804.08771 [cs.CL]. URL: https://arxiv.org/abs/1804.08771.

[42] Colin Raffel et al. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. 2023. arXiv: 1910.10683 [cs.LG]. URL: https://arxiv.org/abs/1910.10683.

[43] Surangika Ranathunga et al. "Neural Machine Translation for Low-resource Languages: A Survey". In: *ACM Comput. Surv.* 55.11 (2023). ISSN: 0360-0300. DOI: 10.1145/3567592. URL: https://doi.org/10.1145/3567592.

[44] Ricardo Rei et al. *COMET: A Neural Framework for MT Evaluation*. 2020. arXiv: 2009.09025 [cs.CL]. URL: https://arxiv.org/abs/2009.09025.

[45] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. *Sequence to Sequence Learning with Neural Networks*. 2014. arXiv: 1409.3215 [cs.CL]. URL: https://arxiv.org/abs/1409.3215.

[46] Tatoeba Community. *About Tatoeba*. Accessed: 2024-07-06. 2024. URL: https://tatoeba.org/en.

[47] NLLB Team et al. *No Language Left Behind: Scaling Human-Centered Machine Translation*. 2022. arXiv: 2207.04672 [cs.CL]. URL: https://arxiv.org/abs/2207.04672.

[48] NLTK Team. *NLTK Documentation*. 2023. URL: https://www.nltk.org/api/nltk.translate.bleu_score.html.

[49] *The State of Machine Translation 2024*. Intento, 2024. URL: https://inten.to/machine-translation-report-2024/.

[50] Jörg Tiedemann. "Parallel Data, Tools and Interfaces in OPUS". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Ed. by Nicoletta Calzolari et al. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 2214–2218. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

[51] Jörg Tiedemann. "The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT". In: *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1174–1182. URL: https://www.aclweb.org/anthology/2020.wmt-1.139.

[52] Jörg Tiedemann and Santhosh Thottingal. "OPUS-MT — Building open translation services for the World". In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT)*. Lisbon, Portugal, 2020.

[53] Jörg Tiedemann et al. "Democratizing neural machine translation with OPUS-MT". In: *Language Resources and Evaluation* 58.2 (Dec. 2023), 713–755. ISSN: 1574-0218. DOI: 10.1007/s10579-023-09704-w. URL: http://dx.doi.org/10.1007/s10579-023-09704-w.

[54] Hugo Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. 2023. arXiv: 2307.09288 [cs.CL]. URL: https://arxiv.org/abs/2307.09288.

[55] tuetschek. *METEOR scores very different from Meteor-1.5 #2655*. https://github.com/nltk/nltk/issues/2655. Accessed: 2024-08-22. 2021.

[56] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[57] Warren Weaver. "Memorandum on Translation". In: *MT News International* 22 (1999), pp. 5–6, 15.

[58] Xiangpeng Wei et al. *PolyLM: An Open Source Polyglot Large Language Model*. 2023. arXiv: 2307.06018 [cs.CL]. URL: https://arxiv.org/abs/2307.06018.

[59] Wikipedia contributors. *Google Translate — Wikipedia, The Free Encyclopedia*. [Online; accessed 6-July-2024]. 2024. URL: https://en.wikipedia.org/w/index.php?title=Google_Translate&oldid=1232822378.

[60] Thomas Wolf et al. "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[61] Shijie Wu and Mark Dredze. "Are All Languages Created Equal in Multilingual BERT?" In: *Proceedings of the 5th Workshop on Representation Learning for NLP*. Ed. by Spandana Gella et al. Online: Association for Computational Linguistics, July 2020, pp. 120–130. DOI: 10.18653/v1/2020.repl4nlp-1.16. URL: https://aclanthology.org/2020.repl4nlp-1.16.

[62] Qi Ye et al. "When and Why are pre-trained word embeddings useful for Neural Machine Translation". In: *HLT-NAACL*. 2018.

[63] Tianyi Zhang et al. *BERTScore: Evaluating Text Generation with BERT*. 2020. arXiv: 1904.09675 [cs.CL]. URL: https://arxiv.org/abs/1904.09675.

[64] Xuan Zhang et al. "Machine Translation with Large Language Models: Prompting, Few-shot Learning, and Fine-tuning with QLoRA". In: *Proceedings of the Eighth Conference on Machine Translation*. Ed. by Philipp Koehn et al. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 468–481. DOI: 10.18653/v1/2023.wmt-1.43. URL: https://aclanthology.org/2023.wmt-1.43.