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

Stefan Liemawan Adji

August 6, 2024

1 Introduction

According to Ethnologue [10], 7,164 languages currently exist and in use today, with 40% of them considered endangered. As of July 2024, 243 languages are supported by Google Translate (according to Wikipedia [39]). In modern times, the need for translation services has surged due to the growing exchange of information across different regions that speak various languages [21].

Machine translation (MT) is the task of automatically translating from one language to another. This can be done through text or audio. It can be traced back to 1949 [37], 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 [14]. However, over the next several decades, growth were limited for machine translation, with 1956-1966 considered the decade of high expectation and disillusion, and 1967-1976 dubbed 'the quiet decade' [15]. Then in 1989, the dominance of the rule-based approach has been challenged by the rise of new methods and strategies, collectively referred to as 'corpus-based' methods (data-driven) [12, 13]. Subsequently, statistics-based approaches for MT re-emerged, bolstered by the recent success of probabilistic techniques in speech recognition. Statistical machine translation [20] dominated the domain between late 1990s through the early 2010s, before largely being surpassed by neural machine translation (NMT) [7, 29].

Since the introduction of Transformers in 2017 [36], 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 [9], GPT [22], and T5 [26], which revolutionised both MT and the whole field of NLP. Then in early 2020s, several pre-trained models (PTM) that are specifically designed for machine translation emerged, namely mBART [19], mT5 [40], NLLB [31], M2M [11], and PolyLM [38]. Most of these models are multilingually trained, allowing for many-to-many translation: able to translate between any of the supported pair 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 [19].

Intento published 'The State of Machine Translation 2024' [32] providing an in-depth evaluation of popular MT engines and LLMs. However, the biggest drawback in this report is that the selected LLMs are general LLMs such as GPT, LLaMa, Mistral, instead of MT-specific LLMs.

Despite these advancements, pre-trained models are often evaluated using different set of benchmarks [19, 31, 11, 38], making it difficult to gauge their relative effectiveness across various languages. Nevertheless, there does not seem to be much work on comparing or benchmarking different pre-trained models in machine translation.

Through simple experimentations, this paper aims to evaluate the performance of existing pre-trained models (PTMs) on many-to-English translation across 14 source languages. Although fine-tuning multilingual PTMs has been proven to increase model performance [8], no pre-training or fine-tuning is performed in this study for simplicity reasons. A dataset is curated from the Tatoeba repository [30], containing 1,323 parallel sentence pairs across source and target languages. The models includes a one-to-one PTMs: OPUS-MT [35, 34], and multilingual PTMs: such as mBART-50 [19], NLLB-200 [31], and M2M-100 [11]. The performance of these models is evaluated using the BLEU score [24].

2 Literature Review

2.1 Pre-Trained Models for Machine Translation

The encoder-decoder approach [7] remains as the foundation architecture for many sequence-to-sequence models in machine translation.

In terms of pre-trained models (PTMs) for machine translation, it can be divided into two categories: one-to-one models and many-to-many models (multilingual).

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. An example of this setup is OPUS-MT by Helsinki-NLP [35, 34], which provides over 1,000 pre-trained models for translation between numerous language pairs.

With the advent of large language models and pre-trained language models, multilingual machine translation has gained prominence. This approach enables many-to-many translation, where a single model can translate between multiple source and target languages [3]. mBART [19] 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. M2M-100 [11] is designed to perform direct translation between 100 languages without relying on English as an intermediate language. NLLB-200: NLLB-200 [31] 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, encompassing numerous underrepresented languages. mBART, M2M-100, and NLLB-200 were all developed by Meta AI (formerly Facebook AI), showcasing the organisation's significant impact in the machine translation field through pre-trained models (PTMs). These models represent a substantial advancement in multilingual translation capabilities.

2.2 Parallel Corpora

Corpora are large and structured sets of texts used for linguistic research and analysis. Thus, parallel corpora are defined as sets of texts in a given source language along with their translations in another target language [18]. It can be bilingual or multilingual, and are crucial for training and evaluating machine translation. Languages without or with limited parallel corpora are referred as **low-resources language**.

Since the early 2000s, English has often been used as an intermediary language, meaning 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 [18]. This is called indirect translation, and posed for many problems in translations as terms can be ambiguous with varying or similar meanings [4].

Neural Machine Translation (NMT) systems require vast amounts of training data, and thus the availability of parallel corpora is crucial for building effective models [16]. The lack of extensive parallel corpora, especially for low-resource languages, leads to suboptimal performance in NMT techniques compared to their high-resource counterparts [27]. OPUS [23] is a comprehensive collection of open-source parallel corpora used extensively in the field of machine translation (MT). It includes corpora for 744 languages and contains over 1,210 different datasets, amassing a total of 45,945,946,108 sentence pairs. Tatoeba [30] 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.

Several datasets are often used for evaluation in machine translation. The Tatoeba Challenge [33] covers 487 languages in 4,024 language pairs, including 657 test sets sourced from Tatoeba website, covering 138 languages. The TED Talks dataset [41] contains transcripts from TED talks for more than 50 languages. IWSLT [2] also contains TED talks data, but paired with English translations. Finally, WMT [5] 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 do not contain parallel corpora between different languages. Corpus between English and French for example, contains different text to the corpus between English and Spanish. Therefore, they do not support easy evaluation for many-to-English translations.

Furthermore, papers often use different sets and different versions of datasets to evaluate their model performance, incurring unstraightforward comparison.

mBart [19] uses WMT19 for English-German translation and TED15 for Chinese-English translation. M2M-100 [11] uses 7 different datasets shared across languages. NLLB-200 [31] uses FLORES dataset, designed for low-resources languages. OPUS-MT [35, 34] models use datasets from Tatoeba challenges.

2.3 Evaluation Metrics

Other metrics for machine translation exist, but not used nearly as widely as BLEU. Crosslingual Optimized Metric for Evaluation of Translation (COMET) [28] is a recent metric that leverages neural-network-based models for deeper understanding and evaluation. BERTScore.

In this project, a variant of BLEU called SacreBLEU and METEOR will be used due to their simplicity.

2.3.1 BLEU

Bilingual Evaluation Understudy (BLEU) [24] is the most commonly used metrics for machine translation (MT). It assesses how well a candidate translation matches the reference translation using precision metrics for n-grams and incorporates a brevity penalty to prevent overly short translations from achieving high scores.

The n-gram precision, as presented in the original BLEU paper [24], is calculated as:

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{\text{n-gram} \in C} \text{Count}_{\text{clip}}(\text{n-gram})}{\sum_{C \in \{Candidates\}} \sum_{\text{n-gram} \in C} \text{Count}(\text{n-gram})} \tag{1}$$

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.
- Count(n-gram) is the count of the n-gram in the candidate translation.

Thus, the BLEU score is calculated as:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{2}$$

Where:

- BP is the brevity penalty.
- p_n is the precision for n-grams.
- 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:

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-\frac{r}{c})} & \text{if } c \leq r \end{cases} \tag{3}$$

Where:

- c is the length of the candidate translation.
- r is the effective reference length.

The machine translation community's rely heavily BLEU score, however, it has several drawbacks. 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 [6].

Furthermore, It is challenging to directly compare BLEU scores between paper [25]. Thus, the author proposed a variation called SacreBLEU [25], which claimed to facilitate easy computation of BLEU scores that are shareable, comparable, and reproducible. This is done by including preprocessing steps and normalisation routines.

2.3.2 METEOR

Metric for Evaluation of Translation with Explicit ORdering (METEOR) [17] 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.

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

Where:

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

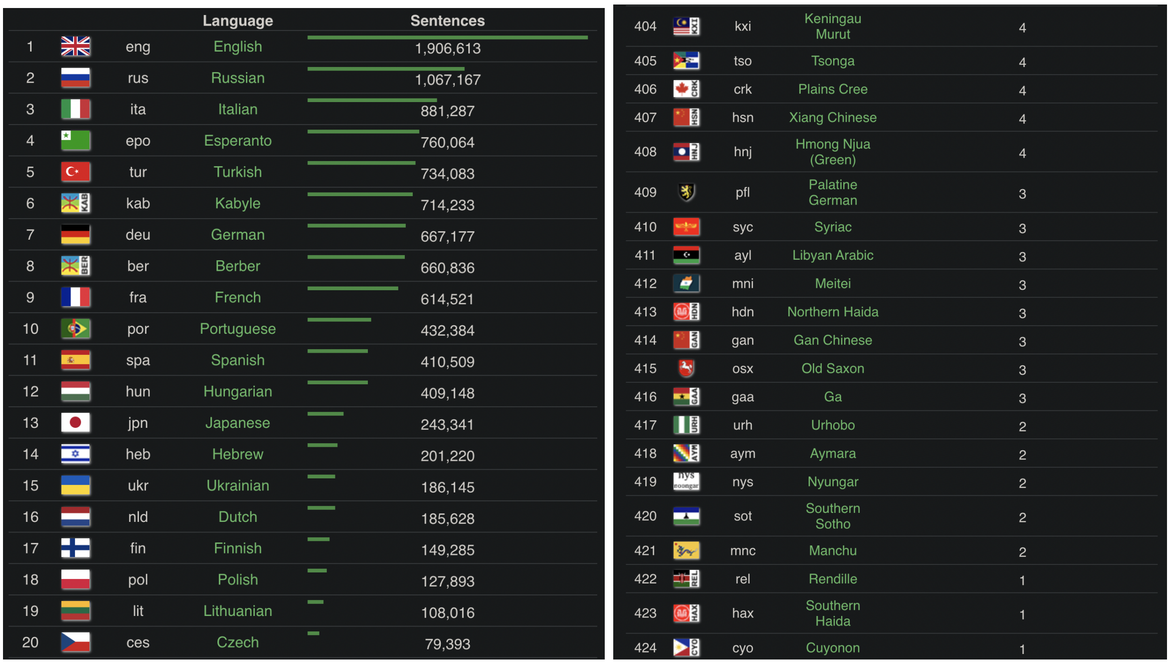
3 Experiments

3.1 Dataset

Tatoeba is a vast, continuously expanding database consisting sentences and their translations, built through the contributions of thousands of volunteers, offering a tool that allows users to see examples of how words are used in sentences [30]. They currently have 12,132,349 sentences and 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 the second place with 1,066,633 sentences. Some languages supported on the website is shown in Figure 1 and Figure 2, sorted from the biggest corpus.

Figure 1: Tatoeba's languages repository with 10,000+ sentences and 100,000+ sentences [30]

Table 1 show the 14 languages selected for this project. Languages are chosen based on its resources' availability in Tatoeba, as well as considering supported languages in most PTMs models. To build the dataset, sentences in English are first downloaded, containing 1,898,494 sentences (it is unclear why it is less than the number stated in the Tatoeba website). Then for each language, sentence pairs between English and source languages are downloaded individually and compiled. The result is a single Dataframe



		Language	Sentences
1	eng	English	1,906,613
2	rus	Russian	1,067,167
3	ita	Italian	881,287
4	epo	Esperanto	760,064
5	tur	Turkish	734,083
6	kab	Kabyle	714,233
7	deu	German	667,177
8	ber	Berber	660,836
9	fra	French	614,521
10	por	Portuguese	432,384
11	spa	Spanish	410,509
12	hun	Hungarian	409,148
13	jpn	Japanese	243,341
14	heb	Hebrew	201,220
15	ukr	Ukrainian	186,145
16	nld	Dutch	185,628
17	fin	Finnish	149,285
18	pol	Polish	127,893
19	lit	Lithuanian	108,016
20	ces	Czech	79,393

404	kxi	Keningau Murut	4
405	tso	Tsonga	4
406	crk	Plains Cree	4
407	hsn	Xiang Chinese	4
408	hnj	Hmong Njua (Green)	4
409	pfl	Palatine German	3
410	syc	Syriac	3
411	ayl	Libyan Arabic	3
412	mni	Meitei	3
413	hdn	Northern Haida	3
414	gan	Gan Chinese	3
415	osx	Old Saxon	3
416	gaa	Ga	3
417	urh	Urhobo	2
418	aym	Aymara	2
419	nys	Nyungar	2
420	sot	Southern Sotho	2
421	mnc	Manchu	2
422	rel	Rendille	1
423	hax	Southern Haida	1
424	cyo	Cuyonon	1

Figure 2: Tatoeba top 20 and bottom 20 languages based on sentences count [30]

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

Table 1: List of chosen languages for evaluation

containing 1,323 parallel sentences in all 14 languages, this will be treated as a test set to evaluate the models performance on 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!' 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', 'Muiriel is 20 now'. All of the sentences are straightforward and literal, without the use of linguistic devices such as metaphors or sarcasm. Therefore, machine translation process should be straightforward on this level.

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 2 shows a few examples of parallel sentences in the final dataset.

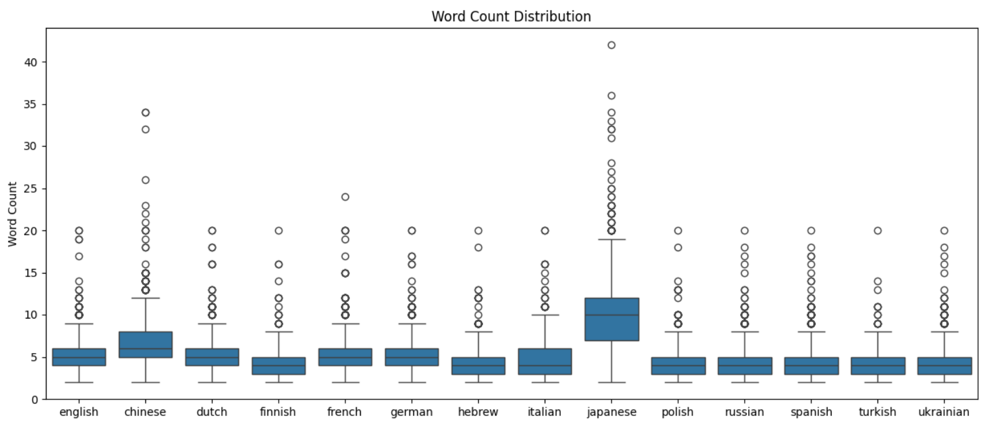


Figure 3: Dataset word count distribution, consisting of mostly short sentences.

Figure 3 shows a box plot of sentences word count. It can be seen that the majority of sentences are short.

Chinese and Japanese are counted per letter.

While most languages cluster around 4-6 words per sentence, there are notable exceptions like Japanese, which exhibits much less variability. This analysis can be useful for understanding language-specific characteristics in sentence structure, which could inform tasks like translation, text processing, or linguistic studies.

4 Evaluation

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 3: M2M-100

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

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 5: NLLB-200

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 6: OPUS-MT

Figure 4 shows the visualisation of all scores.

5 Conclusion

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] 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`.

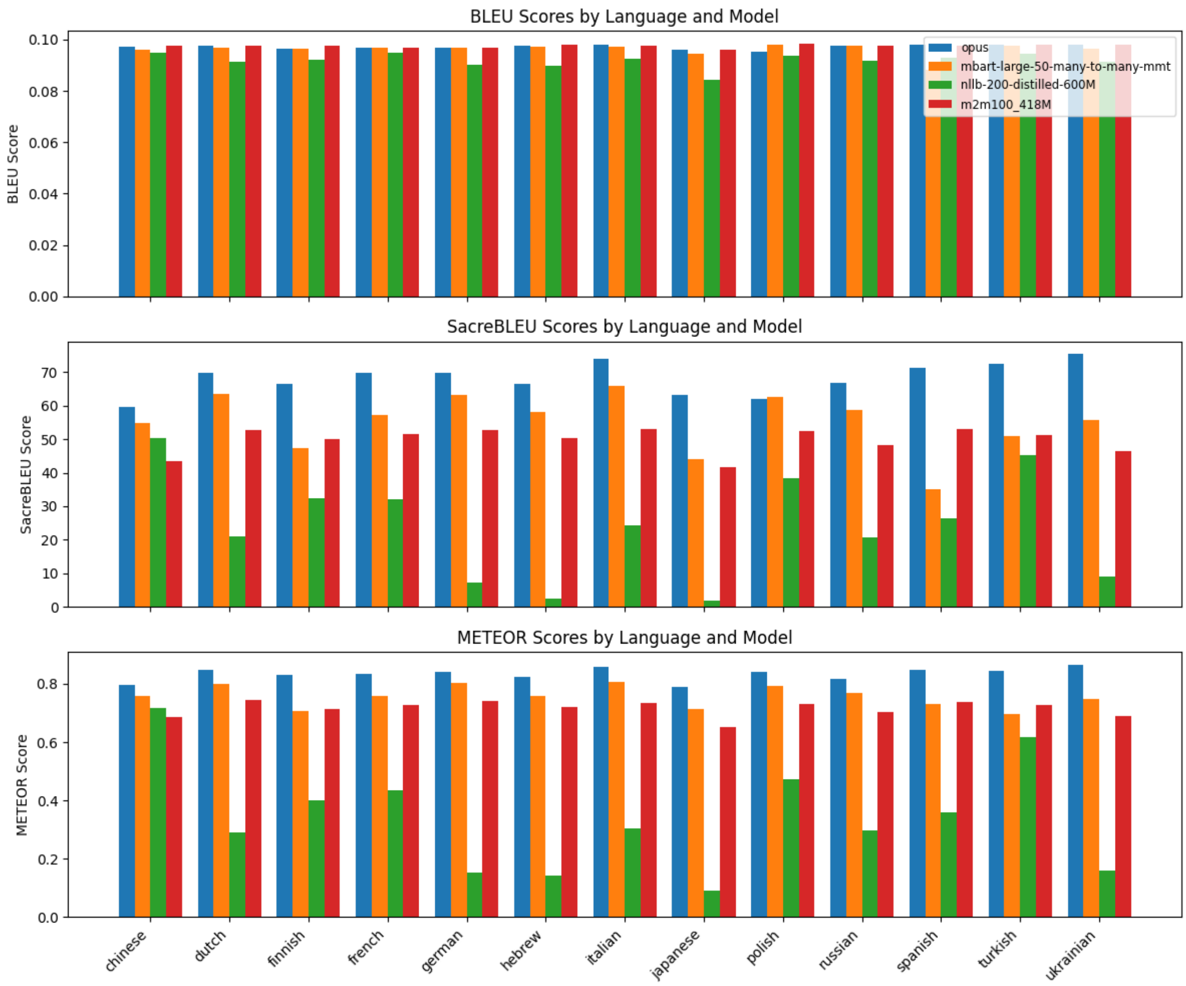


Figure 4: BLEU, SacreBLEU, and METEOR scores of each model on each language.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

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

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

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

[23] OPUS. *OPUS: The Open Parallel Corpus*. https://opus.nlpl.eu/. Accessed: 2024-07-29. 2024.

[24] 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.

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

[26] 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.

[27] 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.

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

[29] 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.

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

[31] 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.

[32] *The State of Machine Translation 2020.* Independent multi-domain evaluation of commercial Machine Translation engines. Intento, 2020. URL: https://try.inten.to/mt_report_2020.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

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

[38] 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.

[39] 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.

[40] Linting Xue et al. *mT5: A massively multilingual pre-trained text-to-text transformer.* 2021. arXiv: 2010.11934 [cs.CL]. URL: https://arxiv.org/abs/2010.11934.

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