Evaluation of Pre-Trained Models for Translation in Different Languages

Stefan Liemawan Adji

July 29, 2024

1 Introduction

According to Ethnologue [4], 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 [28]). In modern times, the need for translation services has surged due to the growing exchange of information across different regions that speak various languages [14].

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 [26], 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 [7]. 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' [8]. 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) [5, 6]. Subsequently, statistics-based approaches for MT re-emerged, bolstered by the recent success of probabilistic techniques in speech recognition. Statistical machine translation [13] dominated the domain between late 1990s through the early 2010s, before largely being surpassed by neural machine translation (NMT) [2, 20]. Finally, since the introduction of Transformers in 2017 [25], 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 [3], GPT [15], and T5 [18], which revolutionised both MT and the whole field of NLP. Soon after, several pre-trained models that are specifically designed for machine translation emerged, namely mBART [12], mT5 [30], and NLLB [22].

Since LLMs require vast amounts of training data, the availability of parallel corpora is crucial for building effective models. The lack of extensive parallel corpora, especially for low-resource languages, leads to suboptimal performance in NMT techniques compared to their high-resource counterparts [19].

OPUS [16] 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 [21] 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.

pre-trained models...

Through simple experimentation and analysis, this paper aims to evaluate existing pre-trained models across different languages and contribute valuable insights into the current state of machine translation technology.

2 Literature Review

2.1 Neural Machine Translation (NMT)

The encoder-decoder approach [2] remains as the foundation architecture for many sequence-to-sequence models in natural language processing, especially machine translation.

2.2 One-to-one 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.

The OPUS-MT by Helsinki-NLP [24] is a good example of one-to-one translation project, they provide over 1,000 pre-trained translation models for many language pairs, allowing users to perform high-quality translations without needing to train models from scratch.

OpenNMT [10] is another framework that allows for one-to-one translation, although their pre-trained models only support English and German languages.

2.3 Many-to-many Models

With the rise of LLMs and pre-trained language models, multilingual machine translation models have received popularity, particularly many-to-many translation.

Multilingual neural machine translation (NMT) allows for the training of a single model capable of translating between multiple source and target languages [1].

Many-to-many translation refers to the capability of a model to translate text between multiple language pairs bidirectionally. This means the model can handle translations from any of the supported source languages to any of the supported target languages, not limited to a single source or target language.

mBart [12] (50 languages), NLLB-200 [22] (200 language), PolyLM [27] (32 languages), and mt5 [30] (101 languages) are some of the most popular many-to-many models.

2.4 Sequence-to-Sequence Models

GPT [15]

mBERT [3] is the multilingual version of BERT, introduced in the same paper, trained on 104 different languages. However, the model been shown to suffer on low-resource languages [29].

T5 [18] is a Text-to-Text Transfer Transformer model trained on English.

MistralAI [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 [21]. They currently have 12,132,349 sentences and 423 supported languages, with around one to two thousand new sentences added daily, on average

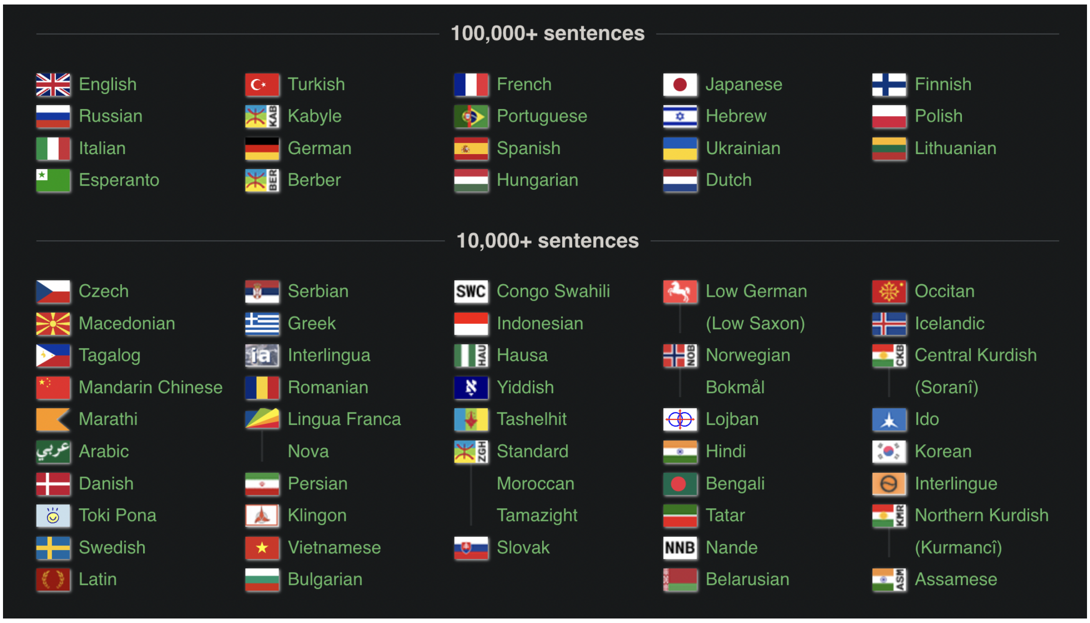
Tatoeba is chosen due to its sentence having more daily life sentences...?

To build the dataset,

Sentences in English are downloaded, 1,898,494 sentences (it is unclear why it is less than the number stated in the Tatoeba website). Then for each languages, download sentence pairs compared to English. Merge every language sentences into one big dataframe, only keep where sentences exist for every language

Tatoeba 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 of the languages supported in the website is shown in Figure 1 and Figure 2, sorted from the biggest corpus. Low-resource languages such as Rendille, Southern Haida, and Cuyonon can be seen at the bottom of the list, having only a single sentence example. Ancestor languages such as Old Saxon and Old Turkish can also be seen in the list, subsequently with low number of examples.

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.



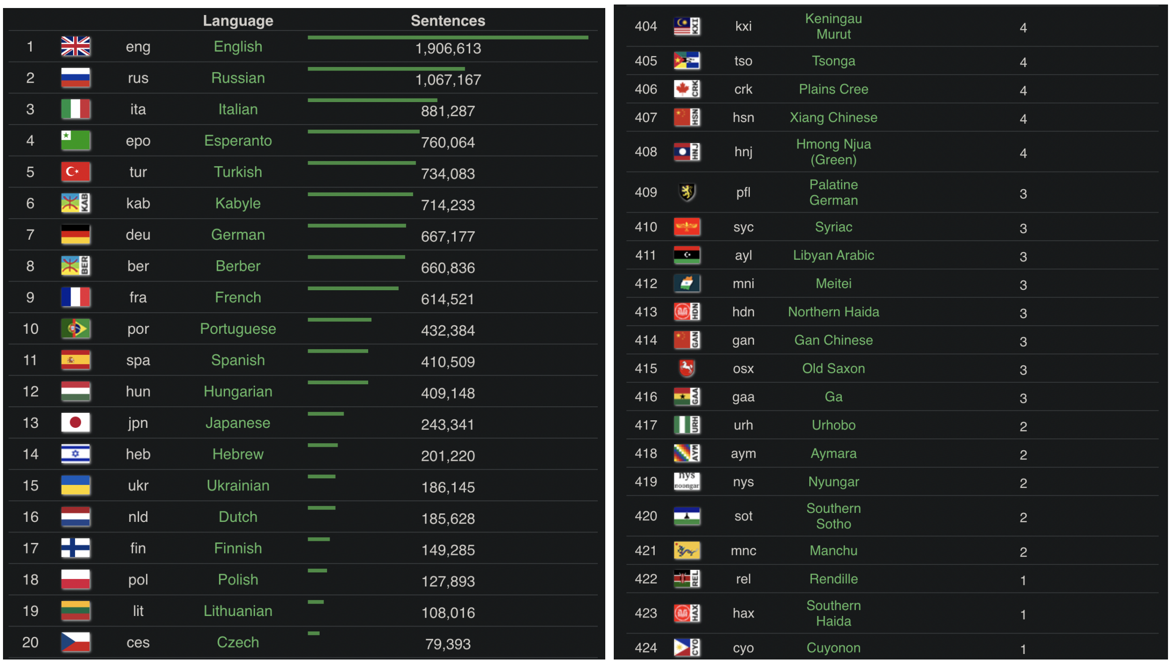
100,000+ sentences

English
Russian
Italian
Esperanto
Turkish
Kabyle
German
Berber
French
Portuguese
Spanish
Hungarian
Japanese
Hebrew
Ukrainian
Dutch
Finnish
Polish
Lithuanian

10,000+ sentences

Czech
Macedonian
Tagalog
Mandarin Chinese
Marathi
Arabic
Danish
Toki Pona
Swedish
Latin
Serbian
Greek
Interlingua
Romanian
Lingua Franca Nova
Persian
Klingon
Vietnamese
Bulgarian
Congo Swahili
Indonesian
Hausa
Yiddish
Tashelhit
Standard Moroccan Tamazight
Slovak
Low German (Low Saxon)
Norwegian Bokmål
Lojban
Hindi
Bengali
Tatar
Nande
Belarusian
Occitan
Icelandic
Central Kurdish (Soranî)
Ido
Korean
Interlingue
Northern Kurdish (Kurmancî)
Assamese

Figure 1: Tatoeba languages repository with 10,000+ sentences and 100,000+ sentences [21]



		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 [21]

The languages chosen for evaluation in this study represent a wide overview, striking a balance between resources and diversity. Each language will then be translated into English (Many-to-English) and evaluated. The evaluation criteria encompass standard metrics such as BLEU (Bilingual Evaluation Understudy) [17], METEOR (Metric for Evaluation of Translation with Explicit ORdering) [11] to capture nuances in translation quality. These metrics not only quantify the fidelity of translations but also offer insights into the models' adaptability and robustness across different linguistic pairs.

No.	Language
1	Dutch
2	Finnish
3	French
4	German
5	Hebrew
6	Hungarian
7	Italian
8	Japanese
9	Mandarin Chinese
10	Polish
11	Portuguese
12	Russian
13	Spanish
14	Turkish
15	Ukrainian

Table 1: List of chosen languages for evaluation

Sentences dataset from Tatoeba is used [1]. Languages that has more than fifty thousand sentences are selected. Accordingly, languages that are available for mbart is also selected based on the list here [2]

For translation, all languages are translated into English as the target language. Then compare the true English sentence and the predicted one, calculated BLEU.

GPT-4 [15] try GPT and see

READ THIS [23] (exactly what this paper should do, maybe compare results and find insights)

4 Evaluation

5 Conclusion

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[1] https://tatoeba.org/

[2] https://dl-translate.readthedocs.io/en/latest/available_languages/

[6] 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`.

[7] 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`.

[8] William J. Hutchins. “Machine translation over fifty years”. In: 2001. URL: `https://api.semanticscholar.org/CorpusID:6196527`.

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

[10] Guillaume Klein et al. “OpenNMT: Open-Source Toolkit for Neural Machine Translation”. In: *Proceedings of ACL 2017, System Demonstrations.* Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 67–72. URL: `https://www.aclweb.org/anthology/P17-4012`.

[11] 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.

[12] 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`.

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

[14] Margaret Dumebi Okpor. “Machine Translation Approaches: Issues and Challenges”. In: 2014. URL: `https://api.semanticscholar.org/CorpusID:11483090`.

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

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

[17] 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. Philadelphia, Pennsylvania: Association for Computational Linguistics, 2002, 311–318. DOI: `10.3115/1073083.1073135`. URL: `https://doi.org/10.3115/1073083.1073135`.

[18] 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`.

[19] 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`.

[20] 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`.

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

[22] 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`.

[23] *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`.

[24] 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.

[25] 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`.

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

[27] 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`.

[28] 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.

[29] 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.

[30] 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.