Evaluation of Machine Translation in Languages

Stefan Liemawan Adji

July 6, 2024

1 Introduction

According to Ethnologue [2], 7,164 languages currently exist and in use today, with 40% of them considered endangered.

As of July 2024, 243 languages and languages varities are supported by Google Translate (according to Wikipedia [13])

Machine translation (MT) can be traced down to as early as 1949 [12], recently growing as a popular research field within Natural Language Processing (NLP).

Statistical machine translation, which previously dominated MT research for many years with its reliance on various count-based models, has largely been surpassed by neural machine translation (NMT) [9]. Nowadays, Large Language Models (LLMs) such as mBart [5], M2M-100 [3], NLLB-200 [11], T5 [8], and GPT-4 [6] dominated the field of machine translation. DeepL [1] currently proclaimed itself as 'The world most accurate translator'.

Through simple experimentation and analysis, this paper aims to evaluate existing techniques across different languages and contribute valuable insights into the current state of machine translation technology. By comparing metrics among languages, I hope to inform future researchers of the effectiveness of machine translation on each individual languages that are included in this paper.

2 Related Work

As is the case of most fields within Natural Language Processing (NLP), the current state-of-the-art (SOTA) in machine translation (MT) is characterised by the significant advancements brought by large language models (LLMs) like OpenAI's GPT-4. These models have surpassed traditional neural machine translation (NMT) systems in several areas, offering more nuanced and accurate translations, especially in handling context and idiomatic expressions.

3 Dataset

3.1 Overview

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

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 langauges, 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.

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

		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 [10]

3.2 Analysis

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.

4 Methodology

.

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) [7], METEOR (Metric for Evaluation of Translation with Explicit ORdering) [4] 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.

5 Evaluation

6 Conclusion

References

[1] DeepL GmbH. *DeepL Translator*. https://www.deepl.com/en/translator. Accessed: 2024-07-06.

[2] 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.

[1] https://tatoeba.org/

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

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

[4] 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. Prague, Czech Republic: Association for Computational Linguistics, 2007, 228–231.

[5] 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.

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

[7] 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.

[8] 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.

[9] Felix Stahlberg. "Neural Machine Translation: A Review". In: *Journal of Artificial Intelligence Research* 69 (Oct. 2020), 343–418. ISSN: 1076-9757. DOI: 10.1613/jair.1.12007. URL: http://dx.doi.org/10.1613/jair.1.12007.

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

[11] 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.

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

[13] 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.