Evaluation of Machine Translation in Languages

Stefan Liemawan Adji

July 13, 2024

1 Introduction

Machine translation (MT) is the task of automatically translating from one language to another. This can be done through text or audio. Ideally, machine translation should be able to conserve the contextual, idiomatic and pragmatic nuances of both languages.

Machine translation can be traced down to as early as 1949, [19], with the first public demonstration of an MT system happened on 7th January 1954 through a collaboration with IBM [6]. A meticulously chosen set of 49 Russian sentences was translated into English, utilising a highly limited vocabulary of 250 words and only 6 grammar rules. While this had minimal scientific impact, it was impressive enough to encourage substantial funding for MT research in the USA.

The decade of high expectation and disillusion, 1956-1966 [7].

The 1960s, particularly 1967-1976 is dubbed as 'the quiet decade' for machine translation by Hutchins [7]

Rule-based MT relied on extensive linguistic rules and bilingual dictionaries. They were based on syntactic and grammatical rules to parse and generate translations.

Since 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 [4, 5]. Subsequently, statistics-based approaches for MT re-emerged, bolstered by the recent success of probabilistic techniques in speech recognition.

1980s: Statistical Machine Translation (SMT) Emerges, use statistical models, probability

Late 2010s: Rise of Neural Networks Neural Machine Translation (NMT): Starting around 2014, neural networks began to dominate MT. NMT systems, like those using the sequence-to-sequence (Seq2Seq) model with attention mechanisms (e.g., the Transformer model), demonstrated significant improvements in translation quality. Advantages Over SMT: NMT models often produce more fluent and coherent translations as they consider the entire sentence context, rather than breaking it down into smaller units like phrases.

...recently growing as a popular research field within Natural Language Processing (NLP). Statistical machine translation [11], which previously dominated for many years in the early 2000s with its reliance on various count-based models, has largely been surpassed by neural machine translation (NMT) in (year). Since the introduction of Transformers in 2017 [18], 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 [2] and GPT [12].

Difference between LLMs and MT engines? MT engines are specifically designed to translate text from one language to another while preserving the meaning and context as accurately as possible. MT engines can also be based on transformer architectures, but they are usually fine-tuned and optimized for the task of translation. They often incorporate additional components like alignment models and post-processing steps to ensure translation quality.

Primary Function: LLMs focus on generating and understanding natural language text, whereas MT engines focus exclusively on translating text between languages. Training Data: LLMs are trained on diverse text data from various sources, while MT engines are trained specifically on parallel corpora (texts in two or more languages that are translations of each other). Output: LLMs generate coherent text in the same language as the input, often in a more creative or conversational style. MT engines produce translated text in a different language, aiming for accuracy and fidelity to the original meaning.

According to Ethnologue [3], 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 [21])

Many language models are trained on predominantly English NLP tasks, before fine-tuned for downstream tasks in other languages.

There is little insights on how the current state of machine translation is across different language. Therefore...

There does not seem to be a lot of pre-trained models that focuses on a single language. This could be because multilingual LLMs perform better? find proof

Focus on single-language models?

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 Literature Review

DeepL [1] is a machine translation service that has gained widespread acclaim, currently proclaimed itself as the world's most accurate translator. There is not much information on how the model or techniques as it is essentially a private business. (irrelevant?)

Instruction Learning and decoder-based models...

GPT [12]

Multilingual Language Models are on the rise for the task of machine translation that involves several languages

A major drawback within most multilingual LMs nowadays is that they are either trained, tested, prioritised on English first.

mBERT [2] 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 [22].

M-BERT does create multilingual representations,

T5 [14] is a Text-to-Text Transfer Transformer model trained on English.

mt5 [23] is a multilingual variant of T5 [14], covering a total of 101 languages.

PolyLM [20]

mBart [10]

NLLB [16]

MistralAI [8]

Massive Parallel Corpora: The availability of large parallel corpora (datasets containing aligned text pairs in different languages) has been crucial for training effective translation models. Datasets like Europarl, OpenSubtitles, and Common Crawl provide a wealth of parallel text data. Unsupervised Translation: Techniques have been developed to perform translation without the need for large parallel corpora, using monolingual data to train models that can translate between languages. This is especially useful for low-resource languages.

Multilingual and Zero-Shot Translation:

Multilingual NMT Models: Models like Google's multilingual NMT and Facebook's M2M-100 can translate directly between many language pairs without relying on an intermediate language (like English). These models are trained on datasets containing multiple languages simultaneously. Zero-Shot Translation: These models can sometimes translate between language pairs they were not explicitly trained on, leveraging knowledge from related languages or shared linguistic features.

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 [15]. 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

Figure 1: Tatoeba languages repository with 10,000+ sentences and 100,000+ sentences [15]

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.

		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 [15]

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.

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) [13], METEOR (Metric for Evaluation of Translation with Explicit ORdering) [9] 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 [12] try GPT and see

READ THIS [17] (exactly what this paper should do, maybe compare results and find insights)

4 Evaluation

5 Conclusion

References

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

[2] 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`.

[3] 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`.

[4] 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`.

[5] 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`.

[1] `https://tatoeba.org/`

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

[6] 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.

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

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

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

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

[16] 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.

[17] *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.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.