Benchmarking Low-Resource Machine Translation Systems

Ana Alexandra Morim da Silva[1], Nikit Srivastava[1], Tatiana Moteu Ngoli[1], Michael Röder[1], Diego Moussallem[1,2], and Axel-Cyrille Ngonga Ngomo[1]

{ana.silva | michael.roeder | axel.ngonga}@upb.de
[1]DICE group, Department of Computer Science, Paderborn University, Germany
[2]Jusbrasil, Data Science Team, Rio de Janeiro, Brazil

Abstract

Assessing the performance of machine translation systems is of critical value, especially to languages with lower resource availability. Due to the large evaluation effort required by the translation task, studies often compare new systems against single systems or commercial solutions. Consequently, determining the best-performing system for specific languages is often unclear. This work benchmarks publicly available translation systems across 4 datasets and 26 languages, including low-resource languages. We consider both effectiveness and efficiency in our evaluation. Our results are made public through BENG—a FAIR benchmarking platform for Natural Language Generation tasks.

1 Introduction

The Machine Translation (MT) task is increasingly relevant in today's connected world as accessibility enables knowledge transfer. Hence, MT systems are recognized as prime tools in the Natural Language Processing (NLP) domain (Goyal et al., 2022). In recent years, Neural Machine Translation (NMT) (Bahdanau et al., 2015) has led the field as it achieves state-of-the-art performance for many language pairs (Gulcehre et al., 2017). However, NMT systems can become computationally demanding and the abundance of new systems also complicates a cross-system comparison. As a result, newly-released systems often compare their performance against single systems (NLLB Team et al., 2022; Tang et al., 2020). Furthermore, recent system analyses also focus on assessing the capability of commercial translation solutions (Zhu et al., 2023). To the best of our knowledge, no work exclusively considers open-source translation systems. Thus, leading to a lack of clarity when determining the best-performing and when identifying shortcomings among existing translation systems, an especially critical task for Low-Resource Languages (LRLs). While the translation task is vital to progress in general, it is still largely unfeasible to the $7,000+$ languages in the world.[1] From these, only close to $2,500$ are represented in the NLP field, with 88% considered to be low-resource. LRLs have a minimal resource availability that causes them to be largely untouched by the benefits of language technology (Joshi et al., 2020). With our work, we aim to contribute to a more complete picture of the current state of the art of machine translation with a focus on LRLs.

We compare four open-source NMT systems—LibreTranslate[2], Opus MT (Tiedemann and Thottingal, 2020), NLLB (NLLB Team et al., 2022), and mBART50 (Tang et al., 2020)—on four parallel machine-translation benchmark datasets—OPUS100 (Zhang et al., 2020), Europarl (Koehn, 2005), IWSLT2017 (Cettolo et al., 2017), and FLORES-200 (NLLB Team et al., 2022). Our evaluation comprises data from 26 different languages. Our results suggest that using languages with lower resource availability does not necessarily translate to lower system performance. However, we did observe more substantial variations in the systems' performance for these languages. Our analysis also showed that LibreTranslate had the highest token throughput among the evaluated systems. Some systems showed proficiency in certain languages, while others performed better according to a certain dataset. Our experiments are shared via BENG (Moussallem et al., 2020), an open-source benchmarking platform that improves the accessibility of experiment results according to the FAIR data principles (Wilkinson et al., 2016).[3]

[1]https://www.ethnologue.com/
[2]https://libretranslate.com/
[3]https://beng.dice-research.org/gerbil

Proceedings of the The Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024), pages 175–185
August 15, 2024

2 Preliminaries and Related Work

Machine Translation (MT) is the process of translating from a source language into a target language autonomously, i.e., without human intervention (Kenny, 2018; Bhattacharyya, 2015). This can be achieved through different approaches. Wang et al. (2022) divide MT techniques into rule- and corpus-based approaches. Corpus-based approaches can be further divided into example-based, statistical, and, more recently, neural approaches. In this work, we evaluate approaches of the latter category with a focus on low-resource languages. We describe both further within this section, along with relevant MT tools and platforms.

2.1 Low-Resource Languages

There are more than 7,000 human languages, with the vast majority being classified as low-resource languages (LRLs) (Magueresse et al., 2020). In contrast to high-resource languages (HRLs), LRLs have a low density of computational corpora (Cieri et al., 2016). However, it is often challenging to identify languages as low- or high-resource as the distinction is often difficult to quantify.

Joshi et al. (2020) propose a language taxonomy based on the quantities of labeled and unlabeled data available in each language. The labeled data is measured through the LDC catalog and the ELRA Map repositories, and the unlabeled data is based on Wikipedia articles.[4] The taxonomy separates languages into six types of languages: *The Left-Behinds* (0), *The Scraping-Bys* (1), *The Hopefuls* (2), *The Rising Stars* (3), *The Underdogs* (4), and *The Winners* (5). Simplified, class 0 languages have neither labeled nor unlabeled data; class 1-4 languages have unlabeled data available, but whose labeled data amount ranges from virtually non-existent to high; and class 5 languages have both high volumes of labeled and unlabeled data.

Hedderich et al. (2021) classify low-resource based on the availability of three data types: 1) task-specific labeled data that supports supervised NLP approaches, 2) unlabeled data that supports unsupervised learning, and 3) auxiliary data that supports learning by proxy. When both labeled and unlabeled data are insufficient in either quantity or quality, other methods can be used to bridge the gap, e.g., transfer learning, data augmentation techniques, distant supervision, and others (Burlot and Yvon, 2018; Gibadullin et al., 2019). Similar statistical studies revealed that more languages should benefit from the availability of NLP tools.

Simons et al. (2022) introduce an automatic approach to measure Digital Language Support for every language by measuring a language's presence across 143 digital tools. Digital support is measured by analyzing different categories of a language's digital presence, such as the level of content provision in a language, system encodings, surface-level tools for text processing, localized user interfaces, text meaning processing, speech processing, and the existence of virtual assistants. The languages are then classified as either still, emerging, ascending, vital, or thriving according to their level of digital support.

2.2 Neural Machine Translation Systems

In recent years, Neural Machine Translation (NMT) has transformed the MT task. By leveraging the currently available large parallel corpora, the MT task has been able to improve translation quality significantly thanks to recent developments in language models. However, large parallel corpora are not available for LRLs, making it difficult to tailor classic NMT models towards LRLs. Open-source translation toolkits like OpenNMT (Klein et al., 2017) and Marian NMT (Junczys-Dowmunt et al., 2018) also provide different neural architecture implementations, forming the backbone of many open-source systems. Below are some examples of open-source NMT systems that cater to LRLs.

LibreTranslate is an open-source NMT service that supports the translation across 46 languages including LRLs.[5] The tool relies on the open-source Argos Translate library to train a transformer-based model from OpenNMT (Klein et al., 2017).[6]

Fairseq (Ott et al., 2019) provides pre-trained convolutional and transformer-based MT models for the English, French, German, and Russian languages with English as source or target language. It is also a development toolkit for NMT tools.

Opus MT (Tiedemann and Thottingal, 2020) is an MT tool trained on the OPUS data (Zhang et al., 2020) based on Marian NMT (Junczys-Dowmunt et al., 2018). Opus MT is a transformer-based NMT system with 6 self-attention layers in the encoder

[4] LDC catalog: https://catalog.ldc.upenn.edu/; ELRA Map: https://catalog.elra.info/en-us/.

[5] https://libretranslate.com/

[6] Argos Translate: https://github.com/argosopentech/argos-translate

and the decoder network, with 8 attention heads in each layer.

mBART50 (Tang et al., 2020) is an extension of mBART (Liu et al., 2020) to demonstrate that multilingual translation models can be created through multilingual fine-tuning. mBART is a sequence-to-sequence generative pretraining model that incorporates languages by concatenating data. While mBART was trained on 25 mainly high-resource languages, Tang et al. (2020) enlarge the embedding layers and combine the monolingual data of the original 25 languages with additional languages to extend the model to more than 50 languages—including LRLs—without requiring to retrain from scratch.

NLLB (No Language Left Behind) (NLLB Team et al., 2022) is a collection of language models created to fill the void left in MT for LRLs. NLLB aims to narrow the performance gap between low and high-resource languages. The model is developed based on a sparsely gated mixture of experts trained on data obtained with novel data mining techniques tailored for LRLs. The model's performance was evaluated across $40,000$ translation directions on the human-translated benchmark dataset FLORES-200.

ALMA (Advanced Language Model-based Translator) (Xu et al., 2024) is a language model based on LLaMA-2 (Touvron et al., 2023) built specifically for machine translation. ALMA introduces a new fine-tuning scheme to improve translation in a zero-shot scenario. It first fine-tunes the model on monolingual data and then fine-tunes it on a parallel corpus. It currently supports 10 language pairs.

With the recent drive of using language models for machine translation, studies such as Zhu et al.'s have emerged to assess the machine translation quality of language models. Zhu et al. (2023) compared 10 different language models across 102 languages, with three languages, English, French, and Chinese, as either source or target language translations. The study provides a good reference point for translation for commercial solutions, as gate-kept models often performed better than open-source solutions. However, due to the large evaluation effort, and the cost of using commercial APIs, the study was only conducted on the first 100 sentences of one dataset: Flores-101 (Goyal et al., 2022). Furthermore, the language models are assessed in an in-context learning setting, where instructions are provided in addition to the translation as context. The authors also observed the influence of different instructions in 6 language pairs.

2.3 Translation Evaluation

The increasing demand for more and better MT tools led to the development of frameworks to simplify their usage. Multiple frameworks streamline the building and training process of language models for translation and offer efficiency. These tools standardize evaluation procedures and enable the user to either tune the models per their requirements or use them as-is. The user trades off fine-grained control over the models for simplicity of use.

2.3.1 Metrics

BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) is an n-gram-based metric used to evaluate text generation systems, mostly chosen due to its low computational cost. In MT, BLEU correlates to human evaluation—the current gold standard—over the entire output. BLEU focuses on the precision between the n-grams in the generated text against those in a reference text. **BLEU NLTK** is an implementation of BLEU from the NLTK library[7] with smoothing applied to sentence-level BLEU scores.

METEOR (Banerjee and Lavie, 2005) is an MT metric that measures the harmonic mean between precision and recall of unigram matches, assigning a higher weight to recall. The word-to-word matching also considers synonyms via the WordNet synset. METEOR scores correlate to human evaluation at the sentence level, in contrast to BLEU.

chrF++ (Popović, 2015) is a variant of the chrF score where the F-score is calculated for both the character n-grams and the word n-grams with the default order being 6 and 2, respectively. chrF is a character-based n-gram F-score metric for MT. It also shows sentence and document-level correlation with human evaluation.

TER (Translation Edit Rate) (Snover et al., 2006) measures the minimum number of edits required to make an output match the corresponding reference. The edits include insertions, deletions, substitutions, word reordering, capitalization, and punctuation. Thus, making the method computationally expensive. The TER score is calculated by computing the number of edits divided by the average referenced words.

[7] https://www.nltk.org/

2.3.2 Benchmarking Frameworks

Systematic evaluations can be a key factor in a research field as they allow a clean comparison between the performance of different approaches over a set of tasks. Benchmarking frameworks support such evaluations and aim to standardize the evaluation for a specific task, including a common task definition, implementation of metrics, and the set of data that is used throughout the evaluation. In the past, different benchmarking frameworks have been proposed for the MT task. The majority of them are local frameworks, i.e., these frameworks compute a set of metrics over the system's output locally. sacreBLEU (Post, 2018) is such a framework and calls for reproducible BLEU scores in the community. Despite its name, it not only supports the BLEU metric, but also chrF, chrF++, and TER. COMET (Rei et al., 2020) trains multilingual MT evaluation models. It allows the user to either train a metric or use the available default models to score the translation output with its COMET-score. Appraise (Federmann, 2018) and HOPE (Gladkoff and Han, 2021) are local human-centric evaluation frameworks. They rely on human intervention due to the low agreement between human quality evaluation and automatic evaluation metrics for MT. Moussallem et al. (2020) propose BENG, an online benchmarking platform for natural language generation that abides by the FAIR data principles (Wilkinson et al., 2016).[8] BENG allows for the submission of multiple systems to be checked against a reference dataset and returns a unique experiment URI with the results. It computes the BLEU, METEOR, chrF++, and TER scores.

3 Evaluation

3.1 Experimental setup

We evaluated the performance of four NMT tools—LibreTranslate[9], Opus MT (Tiedemann and Thottingal, 2020), NLLB (NLLB Team et al., 2022), and mBART50 (Tang et al., 2020). We chose NMT approaches that are open-source, locally deployable, and support several languages, including LRLs. We executed our experiments using the Naïve Entity Aware Machine Translation (NEAMT) tool introduced by Srivastava et al. (2023). This framework was originally implemented as a step in a multilingual knowledge graph question-answering pipeline. It supports a combination of named entity recognition, entity linking, and MT systems. We've used NEAMT for the standard MT pipelines without any of the entity-awareness features as it allows modular and local deployments of new components and serves them through an API[10].

We measured both the quality of the systems' translation and the inherent time cost. Our first experiment compared the system performance across multiple languages. However, some datasets were small and offered limited support for LRLs. So in our second experiment, we compared the performance in languages across the largest datasets and considered 26 languages from all language classes of the taxonomy proposed by Joshi et al. (2020). All of our experiments consider the target language to be English.

3.2 Datasets

We considered four parallel machine-translation benchmark datasets OPUS100 (Zhang et al., 2020), Europarl (Koehn, 2005), IWSLT2017 (Cettolo et al., 2017), and FLORES-200 (NLLB Team et al., 2022). The statistics of the datasets are in Table 1 in the form of token and parallel pair counts. All the datasets have the same number of parallel pairs across languages, except for IWSLT2017. In this case, we averaged the number of pairs for the languages considered in this experiment.

OPUS100 (Zhang et al., 2020) is a parallel translation dataset randomly sampled from the OPUS corpus (Tiedemann, 2012) that covers 100 languages, focused on English. The represented domains in the dataset were not balanced, but sampling filters were applied to ensure no cross-lingual data leakage. This also means that the dataset is not sentence-aligned across languages, i.e., the test sets have different content w.r.t. the language, despite having the same document size.

Europarl (Koehn, 2005) is a parallel translation dataset from the Proceedings of the European Parliament that covers 11 languages. We used the *common-test-set*, a cross-lingual sentence-aligned split, as presented by Koehn (2005) in our experiments.

IWSLT2017 (Cettolo et al., 2017) is a parallel dataset based on TED talks introduced for the IWSLT 2017 multilingual translation task evaluation with language pairs from 5 languages. IWSLT

[8] https://beng.dice-research.org/gerbil/

[9] https://libretranslate.com/

[10] The MT models were deployed on a system with Intel(R) Xeon(R) CPU E5-2695 v3 @ 2.30GHz, 128 GB RAM, and Debian GNU/Linux 11.

		Datasets			
		OPUS100	Europarl	IWSLT2017	FLORES-200
LC	Language \ Parallel pairs	2 000	11 369	4 835	1 012
5	French (FR)	60 497	470 159	233 492	38 842
	German (DE)	43 834	482 529	198 713	36 321
	Japanese (JA)	24 617	–	244 772	44 660
4	Dutch (NL)	37 636	479 949	40 413	36 769
	Finnish (FI)	34 806	540 970	–	41 844
	Hindi (HI)	61 235	–	–	51 218
	Italian (IT)	39 612	444 961	38 468	37 577
	Korean (KO)	26 310	–	261 553	40 255
	Russian (RU)	53 537	–	–	41 700
3	Bengali (BN)	63 760	–	–	56 407
	Bulgarian (BG)	30 210	–	–	44 817
	Estonian (ET)	44 883	–	–	41 940
	Hebrew (HE)	28 239	–	–	40 810
	Indonesian (ID)	23 755	–	–	33 015
	Lithuanian (LT)	76 771	–	–	43 636
	Romanian (RO)	31 144	–	47 187	43 676
	Thai (TH)	48 232	–	–	78 226
	Ukrainian (UK)	31 266	–	–	44 289
2	Irish (GA)	92 241	–	–	54 910
	Xhosa (XH)	62 678	–	–	53 541
1	Macedonian (MK)	37 718	–	–	45 400
	Malayalam (ML)	47 946	–	–	75 526
	Nepali (NE)	25 228	–	–	54 488
	Norwegian Bokmål (NB)	46 924	–	–	36 110
	Telugu (TE)	26 491	–	–	61 108
0	Sinhala (SI)	15 369	–	–	23 886

(Token count)

Table 1: Dataset statistics of the test corpora. The token counts were measured with the cased BERT multilingual base model tokenizer (Devlin et al., 2019).

also introduced an unofficial bilingual task to follow previous editions of the venue that extended the English-centric dataset to 4 other languages. The content and the document size of each test set differ for each language.

FLORES-200 (NLLB Team et al., 2022) is a manually curated dataset that covers 204 languages, based on Wikinews, Wikijunior, and Wikivoyage. The translations were done by professional translators and followed a series of automatic and manual quality review processes. All documents have the same content. As the test set of the dataset is kept blind, in our experiments we evaluated the performance on the *devtest* split.

3.3 Results

The results of the FLORES-200 and OPUS100 are listed in Table 2. NLLB performed better in the FLORES-200 dataset for 20 of the 26 languages with a statistically significant difference to the second-best system.[11] Likewise, Opus MT performed better in the OPUS100 for 19 of the 26 tested languages. The results of the Europarl and IWSLT2017 are in Table 3. LibreTranslate performed best in the Europarl dataset, while mBART50 performed better in IWSLT2017. Language-wise, LibreTranslate performed well in Russian and Estonian, mBART50 in Japanese,

[11]The significance tests were performed with paired bootstrap resampling (Post, 2018) with a 95% confidence interval.

LC	Language	FLORES-200				OPUS100			
		Libre	OPUS	NLLB	mBART	Libre	OPUS	NLLB	mBART
5	FR	42.10	41.93	42.42	39.60	34.45	**38.94**	32.84	36.04
	DE	36.22	40.73	**41.49**	40.48	33.63	**36.55**	27.01	35.22
	JA	13.48	10.67	22.91	**23.93**	03.93	**16.00**	13.33	10.72
4	NL	29.51	29.67	**31.04**	25.89	23.78	**34.92**	30.80	27.29
	FI	24.71	29.55	**30.41**	26.04	18.29	**28.58**	24.70	22.74
	HI	26.97	09.90	**38.37**	32.46	12.30	**33.78**	25.44	25.46
	IT	28.70	29.94	**33.36**	27.35	34.37	**38.20**	33.55	30.12
	KO	14.31	15.80	**25.33**	20.70	05.60	**21.12**	14.59	12.91
	RU	**36.88**	30.15	33.29	31.78	37.28	36.84	31.13	34.18
3	BN	16.03	16.16	**32.85**	09.25	22.42	**28.58**	20.96	07.33
	BG	35.28	34.35	**38.11**	–	34.25	34.52	32.03	–
	ET	**38.83**	32.03	32.71	31.08	**42.14**	39.83	28.63	33.80
	HE	32.53	34.02	**38.19**	30.41	26.69	**39.74**	35.74	29.70
	ID	28.44	33.44	**40.56**	30.36	21.26	**41.33**	34.59	26.99
	LT	26.63	26.58	29.13	28.49	49.43	**50.06**	37.83	37.74
	RO	39.77	39.96	**42.39**	36.85	39.11	**40.24**	36.51	30.65
	TH	15.28	01.06	**25.69**	09.25	**20.48**	08.55	20.35	07.44
	UK	27.98	24.26	**36.79**	27.57	11.11	**33.37**	26.31	21.73
2	GA	30.52	12.11	**34.74**	–	57.98	58.35	46.46	–
	XH	–	02.28	**32.78**	12.21	–	**25.41**	23.48	08.47
1	MK	–	33.75	**39.49**	28.02	–	**42.37**	30.55	24.62
	ML	–	00.38	**32.87**	23.98	–	02.86	18.21	**19.90**
	NE	–	00.99	**37.32**	29.66	–	**63.92**	15.20	49.14
	NB	38.25	24.27	38.35	–	35.36	**45.15**	35.37	–
	TE	–	00.54	**36.40**	15.39	–	59.13	25.88	60.98
0	SI	–	06.52	**30.15**	23.50	–	**33.89**	21.68	23.31

Table 2: BLEU scores of the evaluation for the 17 LRLs and 9 HRLs of the FLORES-200 and OPUS100 datasets. The corresponding URIs are linked with the experiment's BLEU, METEOR, chrF++, and TER scores. The results in bold mark the system with the best BLEU value on a dataset and a statistically significant difference to the second-placed system. The underlined values are the best BLEU values without a significant difference to the next highest value on that dataset.

Opus MT in Romanian, and NLLB in French and German.

3.4 Discussion

We observe a tendency of NLLB and Opus MT towards achieving a better performance on the evaluation part of the dataset on which they have been trained on in comparison to their overall performance. Especially Opus MT seems to be overfitting to its training data, which is reflected by its performance on the FLORES-200 dataset. Opus MT achieves high BLEU scores for the languages Hindi, Irish, Xhosa, Nepali, Telugu, and Sinhala in the OPUS100, but very low scores for the same languages in the FLORES-200 dataset. For the NLLB system, this phenomenon was only observed for the Nepali language.

As expected, the results indicate that some languages are supported better than others. This is underlined by Figure 1, which summarizes the BLEU scores of all four systems on the FLORES-200 dataset. However, the diagram also shows that the evaluated systems do not always perform better on class 5 languages when compared to languages in lower classes. All four systems perform well when translating French and German to English. However, the translation of Japanese is not well supported by all four of them. Instead, all four

LC	Language	Europarl Libre	Europarl OPUS	Europarl NLLB	Europarl mBART		IWSLT2017 Libre	IWSLT2017 OPUS	IWSLT2017 NLLB	IWSLT2017 mBART	
5	FR	**28.38**	25.95	23.43	25.97		39.95	42.34	**43.06**	42.38	
	DE	**25.19**	22.18	20.33	22.15		33.76	37.17	**38.49**	38.08	
	JA	-	-	-	-	-	6.39	8.50	15.60	**17.03**	
4	NL	14.16	**21.54**	19.14	18.93		35.01	40.15	39.96	**43.13**	
	FI	19.78	**22.17**	18.48	21.89		-	-	-	-	
	IT	**26.71**	24.52	21.47	20.93		33.19	36.14	36.84	**39.48**	
	KO	-	-	-	-	-	9.38	23.44	20.91	**23.95**	
3	RO	-	-	-	-	-	37.98	**38.94**	37.87	34.59	

Table 3: BLEU scores of the evaluation of the Europarl and IWSLT2017 datasets. The experiment URIs are linked with the corresponding BLEU, METEOR, chrF++, and TER scores. The results in bold mark the system with the best BLEU value on a dataset and a statistically significant difference to the second-placed system. The underlined values are the best BLEU values without a significant difference to the next highest value on that dataset.

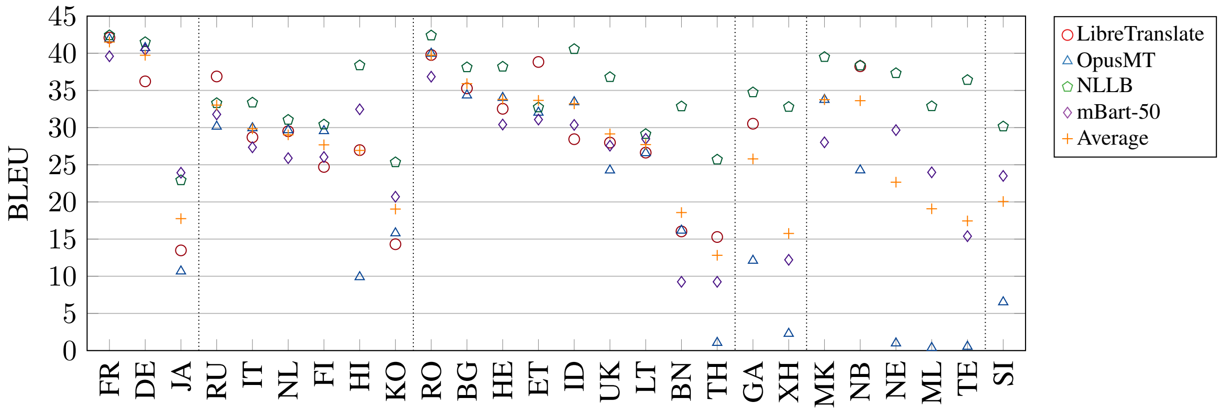

Figure 1: BLEU scores of all four systems and their average on the FLORES-200 dataset for 9 HRLs and 17 LRLs. The languages are sorted by their class from class 5 on the left to class 0 on the right. Within their class, the languages are sorted by the average system performance (orange).

systems perform better when translating the class 3 language Romanian than on Japanese or any class 4 language we look at in our evaluation. Similarly, LibreTranslate performs better on Estonian, Opus MT and NLLB better on Indonesian, Hebrew, and Ukrainian, when compared to Italian or Dutch. This even includes class 1 languages like Macedonian or Norwegian Bokmål for which the four systems achieve better performance than for most class 4 languages. As counter-examples, Thai, and Xhosa are not well supported by the majority of translation systems. Hence, our results suggest that freely available NMT systems can show a high BLEU score even on LRLs. At the same time, this result raises the question, which features of languages influence the performance of the NMT systems. It seems reasonable that an NMT system achieves a similar performance for similar languages, e.g., languages that originate from the same language family. However, although Romanian, French, and Italian belong to the group of Romance languages and the two latter even to the smaller group of Italo-Western languages, the performance of all four systems was significantly lower on Italian than on French or Romanian data. Similarly, German and Dutch belong to the group of languages but lead to quite different BLEU scores. Other language families like West Germanic (Dutch, German), Midlands Indo Aryan (Hindi, Nepali), and Neva (Estonian, Finnish) show similar results in our evaluation, while the languages of the families East Slavic (Russian, Ukrainian) and Macedo-Bulgarian (Bulgarian, Macedonian) let to similar BLEU scores within the families. Although our results point into this direction, the set of languages in our evaluation is too small to refute the hypothesis that families or groups of languages influence the performance of NMT systems. Hence, answering these questions remains future work.

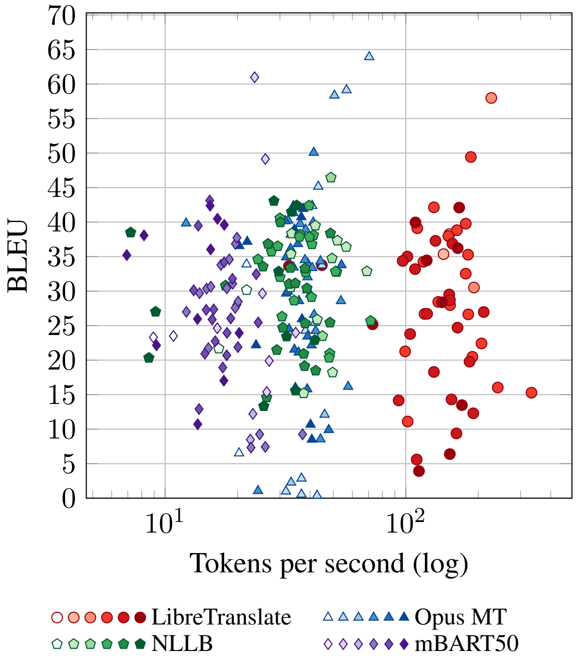

Figure 2: Comparison of the effectiveness (BLEU scores) and the efficiency (throughput). The latter is calculated as tokens per second. The filling of the marks represents the language class, i.e., unfilled marks represent a class 0 language while fully filled marks represent a class 5 language. Up and to the right is better.

Figure 2 shows a comparison of the effectiveness and efficiency of the single systems during all experiments that have been carried out within our evaluation. LibreTranslate shows the highest throughput in most experiments measured in tokens per second. Opus MT and NLLB achieve similar runtimes while mBART50 had the lowest throughput in most experiments. At the same time, we couldn't find a big difference between LRLs and HRLs concerning efficiency.

Figure 3 shows the average standard deviation per language sorted by language class. We observe increased deviations for LRLs when compared to HRLs. Despite the models being trained on LRLs-based data and the systems' language support for LRLs, the performance on these languages is still inconsistent. The Telugu, Malayalam, and Nepali languages are class 1 languages and show the highest deviation. While Bulgarian, a class 3 language, shows the lowest, followed by French and German, two class 5 languages. Hindi, a class 4 language, also shares an increased deviation following other Middle-Modern Indo-Aryan languages like Bengali and Nepali. Malayalam and Telugu are two South Dravidian languages with higher variations as well. This hints at systems having difficulties

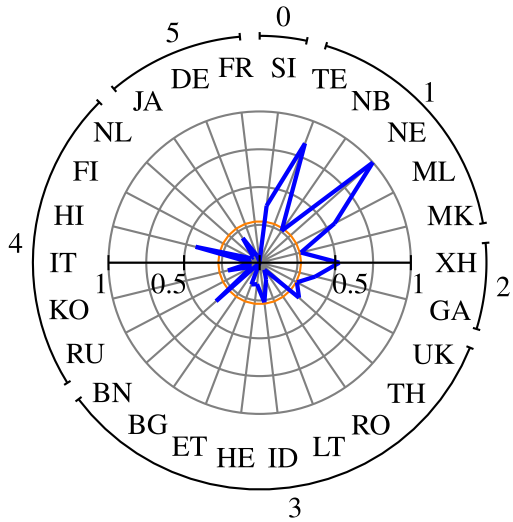

Figure 3: Average standard deviation of BLEU scores per language over all datasets sorted by language class. The values have been normalized using the highest standard deviation (22.06). The orange ring marks the average value over all languages.

processing languages from these families. No other family tree in this experiment presented higher deviations, e.g., Romance, Germanic, Slavic, or Finnic families.

4 Conclusion

We compared four open-source NMT systems on high and low-resource languages regarding their effectiveness and efficiency, filling a gap in the literature that focused on the evaluation of single systems or the comparison of commercial solutions. Our experiments show that open-source systems can perform well on LRLs, showcasing the NLP community's efforts in bridging the gap. However, the performance of the systems in these languages remains variable. Assessing the impact of the domain and genre of the training datasets on the translation quality remains a question for future work. Despite the existence of numerous evaluation frameworks for MT, we used BENG to share the evaluation data via a common space and hope that it boosts comparability across systems and datasets. The influence of language families and writing systems on the translation consistency of these systems requires further investigation.

Acknowledgements

This work has been supported by the Ministry of Culture and Science of North Rhine-Westphalia (MKW NRW) within the project SAIL under the grant no NW21-059D.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Pushpak Bhattacharyya. 2015. *Machine Translation*, 1st edition. Chapman and Hall/CRC.

Franck Burlot and François Yvon. 2018. Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155, Brussels, Belgium. Association for Computational Linguistics.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14, Tokyo, Japan. International Workshop on Spoken Language Translation.

Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. 2016. Selection criteria for low resource language programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549, Portorož, Slovenia. European Language Resources Association (ELRA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Ilshat Gibadullin, Aidar Valeev, Albina Khusainova, and Adil Khan. 2019. A survey of methods to leverage monolingual data in low-resource neural machine translation. *CoRR*, abs/1910.00373.

Serge Gladkoff and Lifeng Han. 2021. Hope: A task-oriented and human-centric evaluation framework using professional post-editing towards more effective MT evaluation.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, and Yoshua Bengio. 2017. On integrating a language model into neural machine translation. *Computer Speech & Language*, 45:137–148.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Dorothy Kenny. 2018. Machine translation. In J.P. Rawling and P. Wilson, editors, *The Routledge Handbook of Translation and Philosophy*, 1st edition. Routledge.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *CoRR*, abs/2006.07264.

Diego Moussallem, Paramjot Kaur, Thiago Ferreira, Chris van der Lee, Anastasia Shimorina, Felix Conrads, Michael Röder, René Speck, Claire Gardent, Simon Mille, Nikolai Ilinykh, and Axel-Cyrille Ngonga Ngomo. 2020. A general benchmarking framework for text generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 27–33, Dublin, Ireland (Virtual). Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Gary F. Simons, Abbey L. L. Thomas, and Chad K. K. White. 2022. Assessing digital language support on a global scale. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4299–4305, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Nikit Srivastava, Aleksandr Perevalov, Denis Kuchelev, Diego Moussallem, Axel-Cyrille Ngonga Ngomo, and Andreas Both. 2023. Lingua franca – entity-aware machine translation approach for question answering over knowledge graphs. In *Proceedings of the 12th Knowledge Capture Conference 2023*, K-CAP '23, page 122–130, New York, NY, USA. Association for Computing Machinery.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2022. Progress in machine translation. *Engineering*, 18:143–153.

Mark D. Wilkinson, Michel Dumontier, Jan I. Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C. Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis.