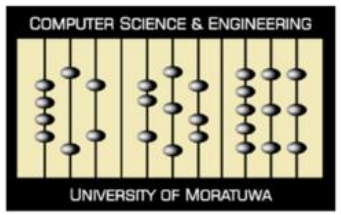
Quality Does Matter: A Detailed Look at the Quality and Utility of Web-Mined Parallel Corpora

Surangika Ranathunga[1], Nisansa de Silva[2], Menan Velayuthan[2], Aloka Fernando[2] and Charitha Rathnayake[2]

[1]Massey University, Palmerston North, New Zealand, 4443

[2]Dept. of Computer Science & Engineering, University of Moratuwa,10400, Sri Lanka

Presented by Menan Velayuthan



MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Background

- The continued significance of parallel corpora cannot be overstated in ensuring optimal performance of NMT models.
- Even Pretrained Language Models (PLMs) fail to close this gap in performance in data scarce settings (Lee et al., 2022).
- This situation becomes a curse for **low resource languages** (Ranathunga et al., 2023).
- **Web-mined parallel corpora** (bitext) s.a. CCMatrix, CCAlign, WikiMatrix, NLLB, and ParaCrawl represent a beacon of hope due to their vast quantities and coverage of hundreds of languages, including numerous low-resource languages.

Limitations of Web-mined Corpora

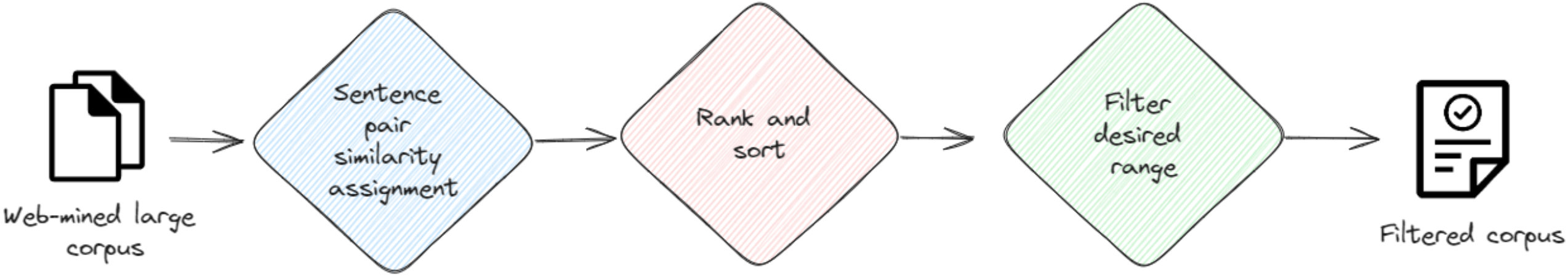
- Kreutzer et al. (2022) analyzed a sample of 100 sentence pairs from some of these corpora and showed that these web-mined corpora have serious quality issues, especially for low-resource languages.
- NMT performance exhibited a decrease when trained on 100k random samples sourced from CCAlign, as evidenced by Lee et al. (2022).
- Injecting common noises (identified through random samples from web-mined parallel corpora), into clean parallel corpora resulted in a decline in NMT performance, as demonstrated by Khayrallah and Koehn (2018).
- BUT, above studies simply utilized a very small random sample from web-mined corpora for their analysis, operating under the assumption of consistent quality throughout the corpus.
- In our work we challenge this assumption.

Outline of Our Work

We show that analyzing a random sample of such a large web-mined corpus can be misleading.

- We selected parallel corpora for two low-resource languages, Sinhala and Tamil, resulting in three language pairs: English-Sinhala (En-Si), English-Tamil (En-Ta), and Sinhala-Tamil (Si-Ta).
- Instead of evaluating the quality of a small random sample from a web-mined corpus, we employed a ranking mechanism based on a similarity measure. This allowed us to extract the top 25k, bottom 25k, and a random 25k subsets from each corpus.
- We conduct comprehensive evaluations on these datasets, which can be broadly categorized into two groups:
  - Intrinsic Evaluation: We improved the error taxonomy of Kreutzer et al. (2022) and carried out a human (intrinsic) evaluation on a random sample of 250 from each of these portions.
  - Extrinsic Evaluation: We separately trained NMT systems by using these top, bottom, and random 25k samples of the corpora as well as the full corpora, and tested them with two different evaluation sets.
- Furthermore, we manually cleaned the top 25k of the NLLB corpus in order to see whether there is any positive impact from human involvement.

Data Filtering Pipeline



Intrinsic Evaluation Results

Error Taxonomy

Error (E) Codes	
NL	at least one of source and target are not linguistic content
WL	Source OR target in some other language, but both still linguistic content
UN	Most part of the source/target has been copied to target/source
X	Correct source and target language, but the translation is completely wrong.

Correct (C) Codes	
CS	Correct translation but very short sentences
CB	Correct translation but boilerplate or low-quality. Requires considerable effort to derive the correct translation.
CN	Near-perfect translation (minor grammar or spelling mistakes). Requires minor effort to derive the correct translation
CC	Perfect translation (no modification by the human is needed)

Human Evaluation Results

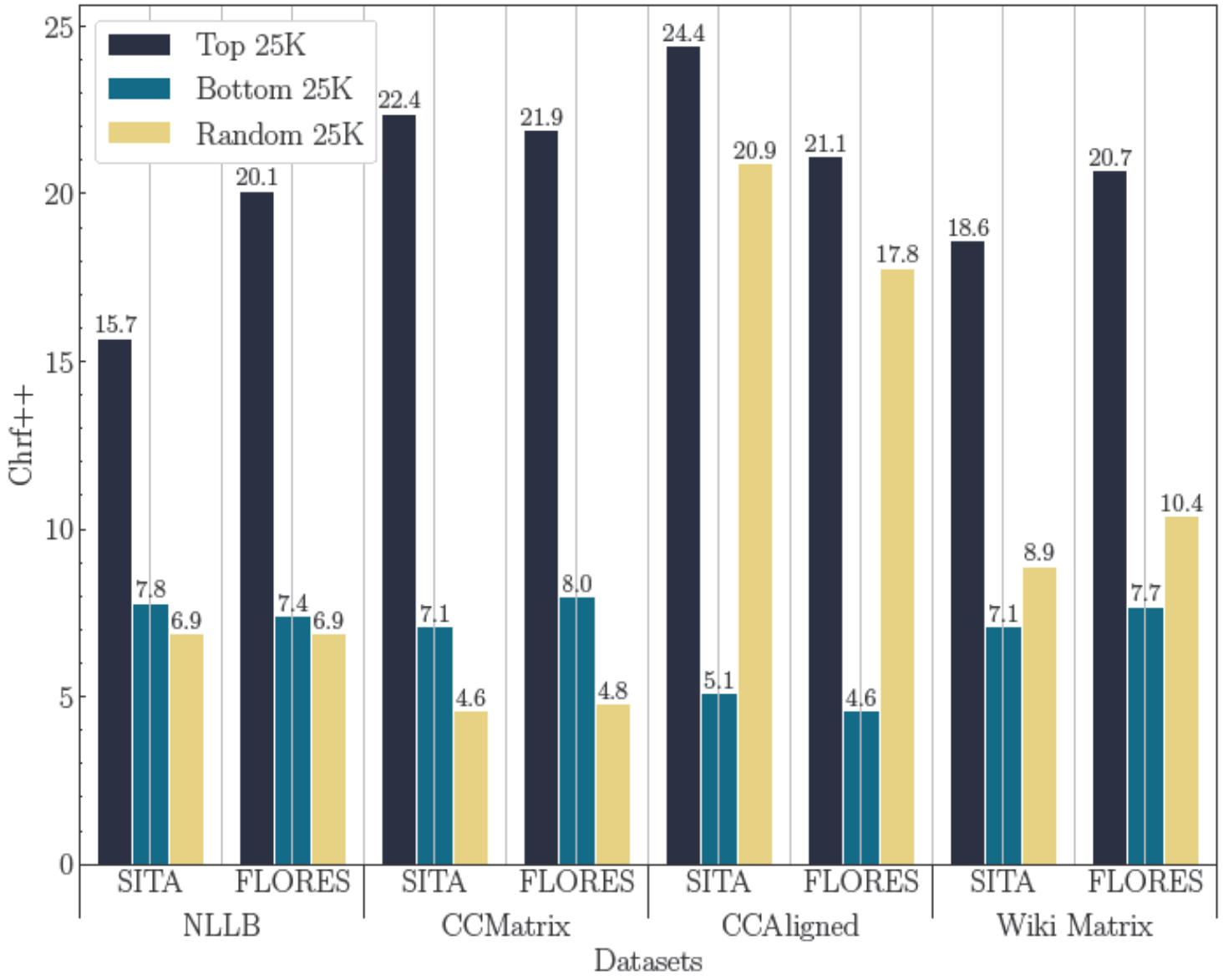
Dataset		En-Si										En-Ta										Si-Ta									
		NL	WL	UN	X	E	CS	CB	CN	CC	C	NL	WL	UN	X	E	CS	CB	CN	CC	C	NL	WL	UN	X	E	CS	CB	CN	CC	C
CCAligned	Top	0.0	0.0	1.9	0.3	2.2	13.2	59.7	10.5	14.4	97.8	0.1	0.1	5.5	0.4	6.1	18.7	33.6	25.3	16.3	93.9	–	–	–	–	–	–	–	–	–	–
	Random	2.0	0.1	5.9	8.9	16.9	17.9	36.1	13.2	15.9	83.1	0.4	0.0	0.9	25.9	27.2	9.1	28.1	19.9	15.7	72.8	–	–	–	–	–	–	–	–	–	–
	Bottom	0.5	0.0	0.1	60.4	61.0	4.3	17.5	11.3	5.9	39.0	1.9	0.1	1.1	45.5	48.6	8.8	10.0	16.1	16.5	54.1	–	–	–	–	–	–	–	–	–	–
WikiMatrix	Top	0.3	0.1	2.1	16.3	18.8	6.1	40.8	12.0	22.3	81.2	0.9	6.3	15.9	46.1	69.2	1.6	10.3	7.5	11.5	30.9	–	–	–	–	–	–	–	–	–	–
	Random	0.3	0.1	0.0	86.1	86.5	1.2	7.9	2.9	1.5	13.5	0.7	0.9	1.2	91.9	94.7	0.3	4.1	0.7	0.3	5.4	–	–	–	–	–	–	–	–	–	–
	Bottom	0.0	2.7	0.3	88.5	91.5	1.2	6.9	0.4	0.0	8.5	1.3	5.2	0.8	88.7	96.0	0.3	2.1	1.2	0.4	4.0	–	–	–	–	–	–	–	–	–	–
CCMatrix	Top	0.0	0.0	7.1	0.1	7.2	8.7	37.5	14.3	32.4	92.9	0.0	0.0	51.5	3.6	55.1	2.7	27.5	8.5	6.3	45.0	0.1	5.5	0.5	2.1	8.2	9.3	26.4	34.3	21.7	91.7
	Random	0.0	0.0	1.6	31.3	32.9	6.1	27.6	22.5	10.8	67.0	0.1	0.0	2.9	83.5	86.5	0.3	8.5	3.1	1.6	13.5	0.0	2.1	0.8	31.3	34.2	0.9	34.7	23.2	6.9	65.7
	Bottom	0.0	1.3	0.8	27.2	29.3	7.3	47.3	8.5	7.5	70.7	0.0	0.1	0.0	83.1	83.2	0.0	10.1	4.3	2.4	16.8	0.0	1.2	0.1	50.1	51.4	1.1	31.7	11.3	4.4	48.6
NLLB	Top	0.0	0.5	0.4	19.1	20.0	6.5	36.0	18.8	18.7	80.0	0.0	0.4	0.3	11.1	11.8	0.4	21.6	25.2	41.1	88.3	0.0	0.0	0.3	1.9	2.2	0.3	22.3	40.5	34.8	97.9
	Random	0.1	0.4	0.7	54.5	55.7	1.3	27.5	10.5	4.9	44.2	0.1	0.0	0.5	43.3	1.9	43.9	31.6	10.9	11.6	98.0	56.0	0.3	0.0	20.0	20.3	1.2	44.5	22.3	11.7	79.7
	Bottom	0.0	0.0	1.9	56.9	58.8	4.7	27.1	8.1	1.3	41.2	0.0	0.0	0.0	51.9	51.9	1.6	28.9	11.1	6.5	48.1	0.0	0.0	0.1	34.7	34.8	0.0	42.0	20.3	2.9	65.2
NLLB (cleaned)	Translator 1	0.0	0.0	0.0	1.9	1.9	10.7	24.0	14.3	49.1	98.1	0.1	0.0	0.0	0.0	0.1	0.5	16.2	12.6	70.6	99.9	0.0	0.0	0.0	0.3	0.3	0.3	1.9	24.0	73.6	99.7
	Translator 2	0.0	0.1	0.0	1.9	2.0	7.2	21.3	13.0	55.6	98.0	0.0	0.0	0.0	0.1	0.1	0.8	16.9	10.1	72.1	99.9	0.0	0.0	0.0	0.4	0.4	0.3	4.3	38.0	57.0	99.6
	Translator 3	0.0	0.4	0.0	1.8	2.1	9.1	22.5	7.0	59.3	97.9	0.0	0.0	0.1	0.4	0.5	0.6	8.1	15.8	75.0	99.5	0.0	0.0	0.0	0.0	0.0	0.1	1.9	29.5	68.5	100.0

Extrinsic Evaluation Results

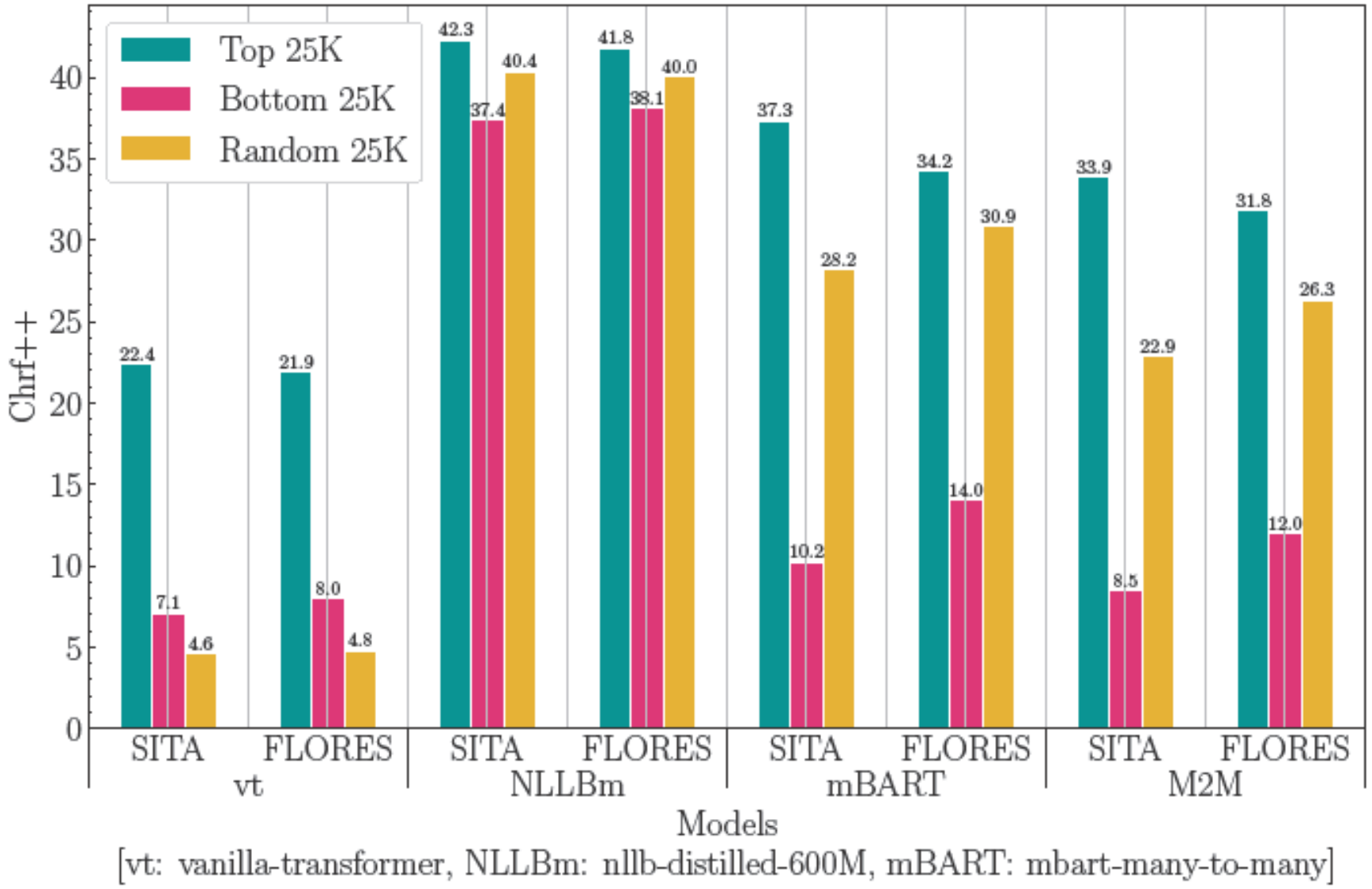
NMT Model Evaluation Experimental Setup

- Dataset
  - Web-mined corpora: NLLB, CCMatrix, CCAligned and WikiMatrix.
  - For each corpus, we trained separate NMT models using the top, bottom, and random 25k portions of each of the web-mined En-Si & En-Ta corpora.
  - We used two separate datasets for testing: FLORES-101 (Goyal et al., 2022), and the test set of the SITA parallel corpus (Fernando et al., 2020). FLORES was created from Wikipedia articles, and SITA from government documents of Sri Lanka.
- Models
  - Vanilla transformers.
  - Ablation study with: NLLB (Team et al., 2022) (henceforth referred to as NLLBm, to distinguish from the NLLB dataset), mBART (Tang et al., 2021) and M2M (Fan et al., 2021) were selected as our base models.

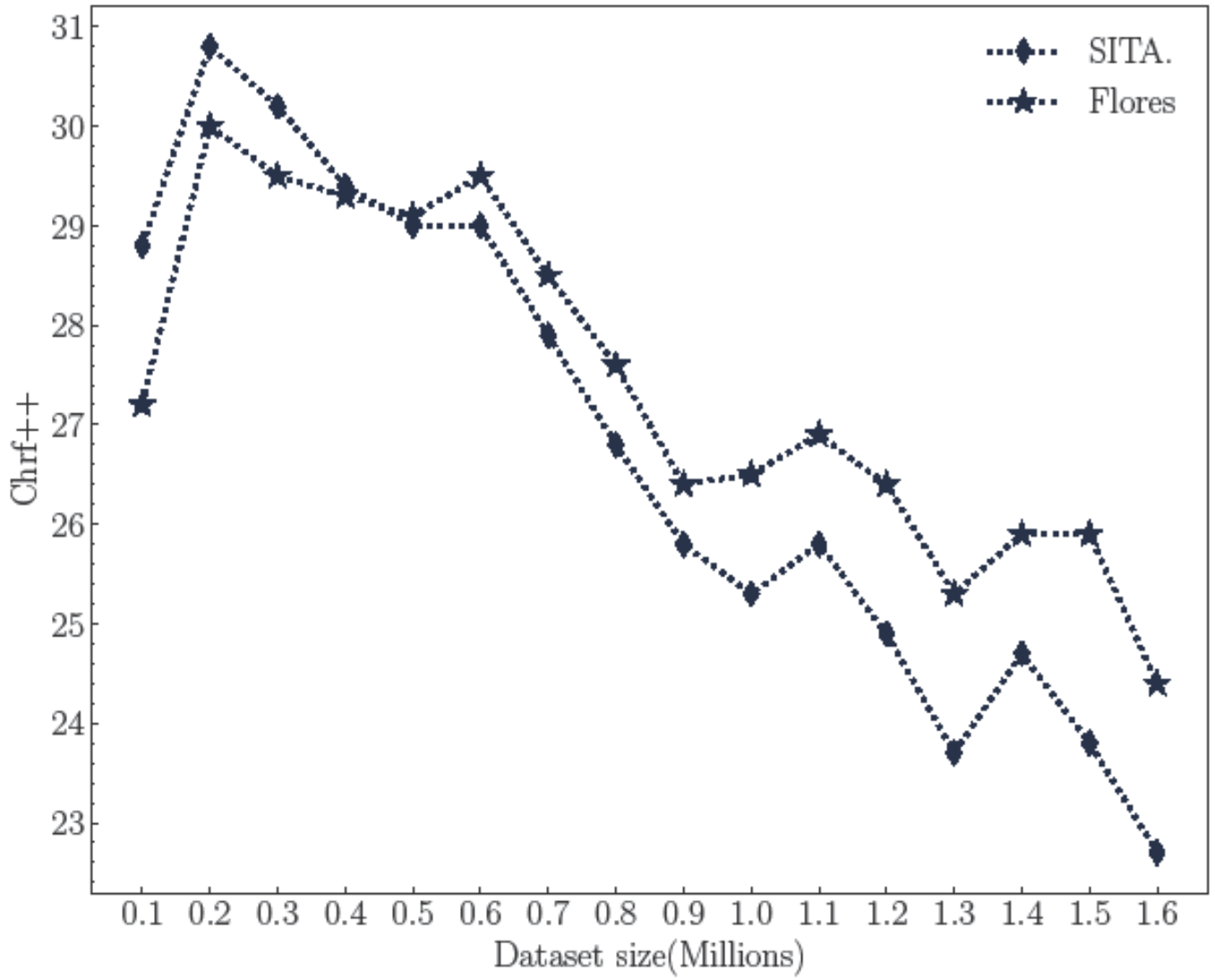
Evaluation of Vanilla Transformer Model Trained on Top, Bottom and Random 25k of Four Web-Mined Corpora (En-Si)



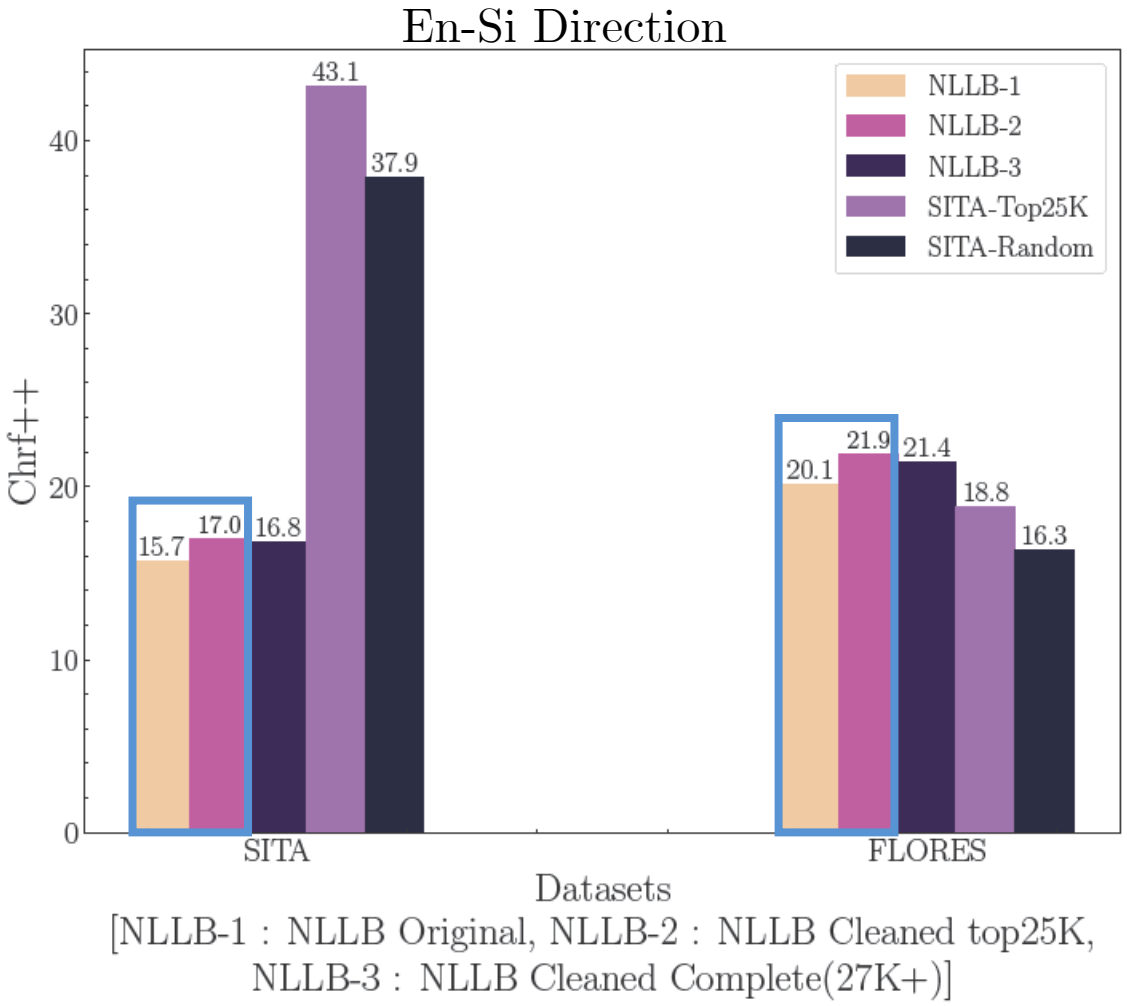
Evaluation of NMT Models Trained on Top, Bottom and Random 25k Portions of CCMatrix (En-Si)

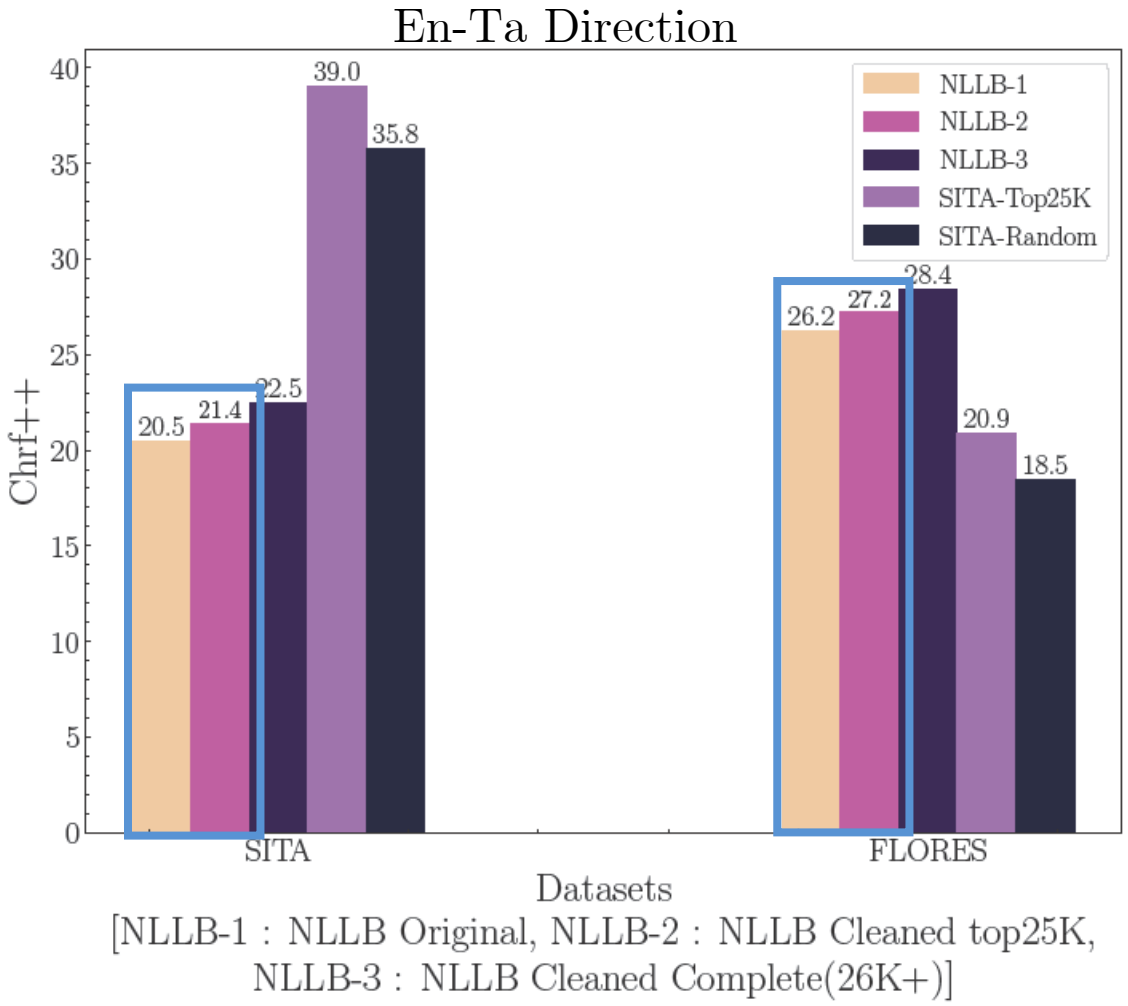


NMT Results of Vanilla Transformer Model Trained on CCMatrix En-Si in Jumps of 100k

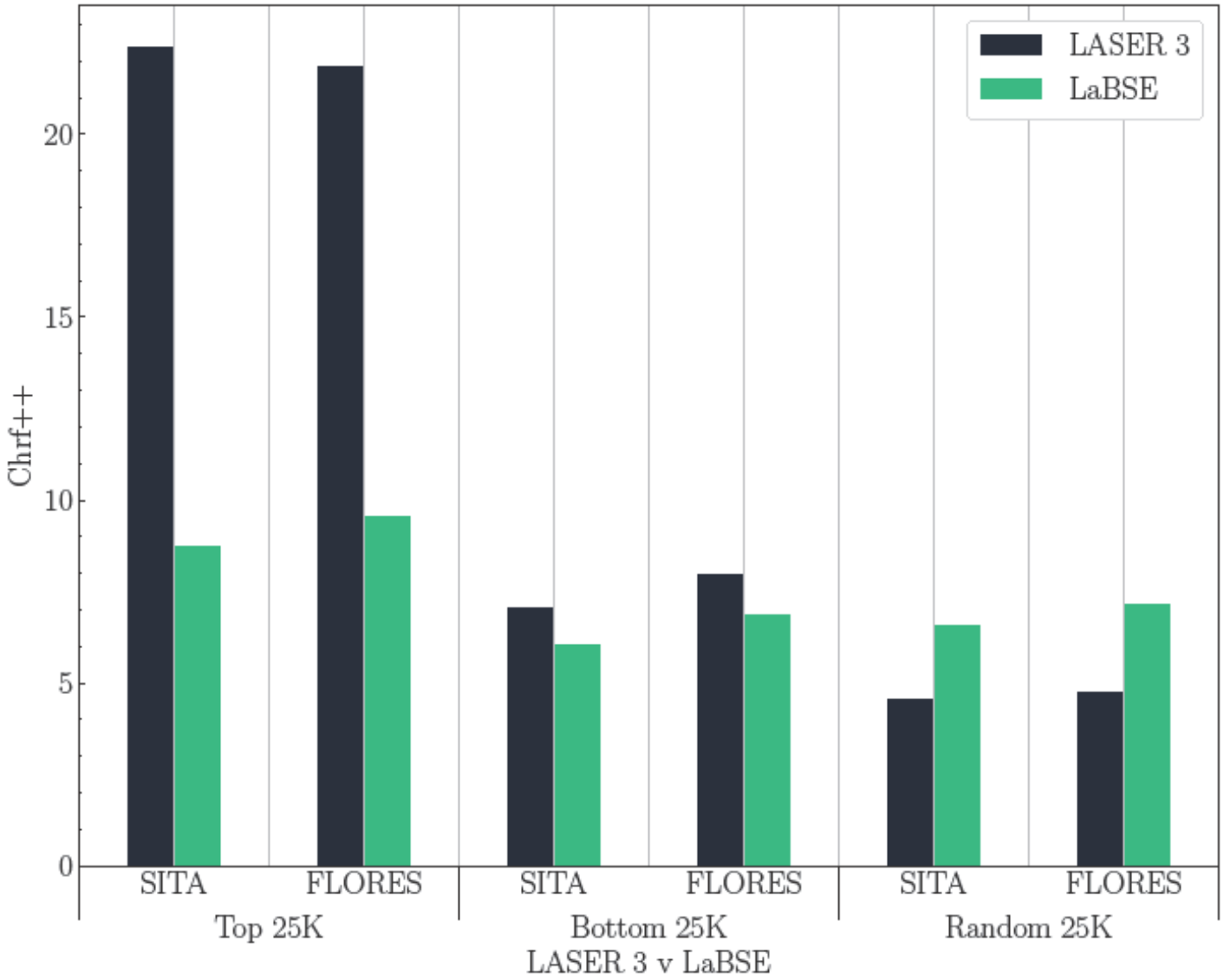


Vanilla Transformer Results for Original NLLB Top 25k, NLLB Cleaned Top 25k, NLLB Cleaned Full (27k+), SITA Top 25k, and SITA Random 25k on Language Directions En-Si and En-Ta





Ablation Study on Embedding Models for Measuring Similarity Between Parallel Sentences



Our Findings

Our findings can be summarized as follows,

- Intrinsic Evaluation
  - There are significant quality differences between the three portions.
  - Quality of the top 25k portion is much better than the other portions.
  - Noted major variations of quality across web-mined corpora belonging to different language pairs.
- Extrinsic Evaluation
  - NMT models trained with the top 25k portion are significantly better.
  - NMT models trained with the full version of some of these corpora were even lagging behind models trained with their top 25k portion.
  - The NMT model trained with the top 25k portion of the En-Si and En-Ta parts of the NLLB corpus performed even better than a model trained with a human-curated corpus.
- We show that addressing translation issues in the top 25k of the NLLB corpus using human translators resulted in a slightly cleaner corpus, achieved in slightly less time than translating from scratch. Although the NMT model trained with this cleaned corpus outperformed the uncleaned corpus, the resultant gains were meager and cannot justify the time and financial investment in the translators.

Key Take Aways

- We showed that the quality of such web-mined corpora significantly varies across different portions.
- Simply ranking a web-mined corpus beforehand and subsequently utilizing only the high-quality portion can lead to improved accuracy within significantly less training time.
- Our findings also suggest that utilizing only the highest quality portion of a web-mined corpus can lead to NMT results comparable to those obtained from human-curated corpora in certain cases.
- Our results serve as a cautionary note to researchers against indiscriminate use of web-mined corpora through random sampling alone.
- Project artefacts are released and the details are shared in the project Github (https://github.com/nlpcuom/quality-matters)

References

En-Shiun Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Adelani, Ruisi Su, and Arya McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low resource language translation? In Findings of the Association for Computational Linguistics: ACL 2022, pages 58–67, Dublin, Ireland. Association for Computational Linguistics.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. ACM Computing Surveys, 55(11):1–37.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. Transactions of the Association for Computational Linguistics, 10:50–72.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, PengJen 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.

Aloka Fernando, Surangika Ranathunga, and Gihan Dias. 2020. Data augmentation and terminology integration for domain-specific sinhala-englishtamil statistical machine translation. arXiv preprint arXiv:2011.02821.

NLLB Team, Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. pages 3450–3466.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. The Journal of Machine Learning Research, 22(1):4839–4886.