Segmentation Beyond Defaults: Asymmetrical Byte Pair Encoding for Optimal Machine Translation Performance

Saumitra Yadav and **Manish Shrivastava**
Language Technologies Research Center, KCIS,
International Institute of Information Technology Hyderabad, India
saumitra.yadav@research.iiit.ac.in and m.shrivastava@iiit.ac.in

Abstract

Existing Machine Translation (MT) research often suggests a single, fixed set of hyperparameters for word segmentation models, **symmetric Byte Pair Encoding** (BPE), which applies the same number of merge operations (NMO) to train tokenizers for both source and target languages. However, we demonstrate that this uniform approach doesn't guarantee optimal MT performance across different language pairs and data sizes. This work investigates BPE segmentation recipes across various data volumes and language pairs to evaluate MT system performance. We find that utilizing **asymmetric BPE**—where the source and target languages have different NMOs—significantly improves results over the symmetric approach, especially in low-resource settings (50K, 100K, and 500K sentence pairs). Specifically, asymmetric BPE yield statistically significant ($p<0.05$) average gains of 5.32, 4.46, and 0.7 CHRF++ on English-Hindi in low-resource setups (50K, 100K, and 500K sentence pairs, respectively). We validated this trend across six additional language pairs (English↔Telugu, Shona, Norwegian, Kyrgyz, Hausa, and Inuktitut), observing statistically significant improvement in 10 out of 12 systems compared to symmetric BPE. Our findings indicate a high NMO for the source (4K to 32K) and a low NMO for the target (0.5K to 2K) provides optimal results, particularly benefiting low-resource MT.

1 Introduction

Efforts have been made to include low-resource language pairs in Neural Machine Translation (NMT), e.g. Workshop on Technologies for MT of Low Resource Languages. Often, successful past methodologies on high-resource language pairs, like hyperparameters for preprocessing, are used without considering their suitability for specific language pairs. For example, if we take a preprocessing step, such as word segmentation, a key preprocessing step, divides words into subwords to enhance learning and manage vocabulary size, handling rare and unknown words to boost MT performance. Notable Techniques include BPE (Sennrich et al., 2016), word piece (Devlin et al., 2019), sentence piece (Kudo and Richardson, 2018), and morfessor (Smit et al., 2014). BPE compresses data by merging frequent character pairs into symbols (Gage, 1994), with the *number of merge operations* (NMO) as a key parameter. A lower NMO (e.g., 500, Table 1) reduces vocabulary size with more segmentation, while a higher NMO (e.g., 32K) results in larger vocabularies and less segmentation. Typically, the same NMO is applied to both source and target languages. Recent work have shown that examining BPE parameters in low-resource MT is vital (Ding et al., 2019; Abid, 2020), but uniform NMOs for source and target (symmetrical BPE) (Huck et al., 2017; Ortega et al., 2020; Lankford et al., 2021; Domingo et al., 2023; Lee et al., 2024) prevail, with little exploration of asymmetrical BPE in MT. Earlier work Ngo Ho and Yvon (2021) looked at asymmetric BPE for language alignment, not for MT. Our work is a result of a multi-year exploration of the impact of asymmetrical subword segmentation in bilingual MT systems.

While we acknowledge the rise of multilingual and decoder-only models, our study focuses on the effect of asymmetric BPE in bilingual setups, particularly in low-resource conditions where pretrained tokenizers or joint vocabularies may be unavailable. Bilingual systems remain a research focus, with studies in Cantonese-Mandarin (Liu, 2022), English-Luganda (Kimera et al., 2025), Wolof-French (Dione et al., 2022), Bavarian-German (Her and Kruschwitz, 2024), and English-Manipuri (Singh et al., 2023; Singh and Singh, 2022) using bilingual data and transformer-based architectures with customized subword segmentation like BPE or morphology-aware tokenization. These efforts, along with Li et al. (2024),

Proceedings of the Twelfth Workshop on Asian Translation (WAT 2025), pages 35–53
December 24, 2025 ©2025 Association for Computational Linguistics

Sentence	bosusco , 54 , runs an adventure tourism bureau .
500 NMO	bo@@ su@@ sc@@ o , 5@@ 4 , r@@ un@@ s an ad@@ v@@ en@@ ture t@@ our@@ is@@ m bu@@ re@@ a@@ u .
32K NMO	bo@@ su@@ sco , 54 , runs an adventure tourism bureau .

Table 1: Effect of NMO variation: 500 NMO yields highly segmented tokens, while 32K retains most vocabulary

cover underrepresented languages and diverse writing systems, proving the continued relevance of bilingual systems. Our work investigates asymmetrical BPE's impact on bilingual MT systems, utilizing different merge operation counts for source and target languages across varied dataset sizes and resources. Extending these results to multilingual or decoder-only models is beyond this work's scope but represents an interesting future direction.

We define the "BPE configuration" as $m_1\_m_2$, with m_1 and m_2 representing the merge operations for source and target languages. Our study on symmetric and asymmetric BPE configurations for English–Hindi under varying data conditions shows asymmetric configurations performing best, especially in low-resource context. We extend these insights to six additional language pairs—English $\leftrightarrow$ Telugu, Shona, Norwegian, Kyrgyz, Hausa, Inuktitut—selected for diverse language families and morphological typologies. **Our findings consistently demonstrate that, in low-resource environments, the most effective BPE configuration for the majority of language translation directions tends to be asymmetric. Specifically, setups with *4K* to *32K* NMO for the source and *500* to *2K* for the target outperform symmetric BPE configurations.**

Section 2 summarizes previous efforts to use symmetric BPE merge operations to improve MT performance. Section 3 explains our motivation for finding optimal BPE configurations by exploring asymmetric BPE. Section 4 outlines our experimental setup and presents the performance of the English-Hindi MT system on FLORES and Domain testsets. Section 5 evaluates the setup for other language pairs in low resource context, concluding our observations in Section 6.

2 Related Work - Symmetrical BPE

Most bilingual MT systems—especially for low-resource pairs—use the same number of merge operations (NMO) for source and target languages. Studies show that smaller vocabularies (0–4K NMO) outperform the common 32K setting by up to 4 BLEU points in low-resource scenarios (Ding et al., 2019); similar patterns are reported for English–Egyptian, English–Levantine (Abid, 2020), and English–Irish (Lankford et al., 2021).

Other work adapts segmentation for polysynthetic languages (Ortega et al., 2020), rich morphology (Lee et al., 2024), or target-side variation (Domingo et al., 2023). Alternative strategies include cascading segmentations (Huck et al., 2017), vocabulary refinement (Xu et al., 2021), and multi-BPE–setting corpora (Poncelas et al., 2020). While (Ngo Ho and Yvon, 2021) varied NMOs for alignment, no prior study systematically evaluates asymmetric BPE—using different NMOs for source and target—across resource levels. This work addresses that gap.

Though multilingual MT research now dominates, bilingual MT remains vital for low-resource pairs, where symmetric BPE is still common (Liu, 2022; Kimera et al., 2025; Dione et al., 2022; Her and Kruschwitz, 2024; Singh et al., 2023; Singh and Singh, 2022). Recent work on Parity-Aware BPE (Foroutan et al., 2025) introduces fairness-oriented subword allocation, reducing disadvantages for low-resource languages in multilingual tokenization. Although our experiments are limited to bilingual MT, asymmetric BPE could complement such fairness-aware methods in multilingual systems; extending this remains outside our current scope.

3 Exploring Asymmetrical BPE

In practice, for a BPE configuration $m_1\_m_2$, the values of m_1 and m_2 are usually the same, with the number of merge operations (NMO) ranging from *8K* to *40K* (Wu, 2016; Denkowski and Neubig, 2017; Cherry et al., 2018; Renduchintala et al., 2019). However, Ding et al. (2019); Dewangan et al. (2021) found these settings suboptimal for low-resource language pairs. Ding et al. (2019) observed that $m_1 = m_2 \leq$ *4K* NMO outperforms *32K* in low-resource conditions, consistent with our experiments on 0.1 million sentence pairs (English $\leftrightarrow$ {Hindi, Telugu}) (Figure 1). Dewangan et al.

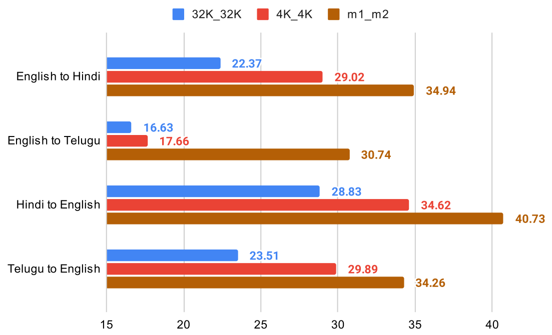


Figure 1: CHRF++ Scores for Symmetrical BPE (*32K*,*4K*) vs Asymmetrical BPE ($m_1 \neq m_2$)

(2021) further showed that identical BPE configurations yield differing performance across language pairs, exemplified by English-Hindi vs. English-Telugu comparisons at *4K* NMO (Figure 1).

Work by Ortega et al. (2020); Mujadia and Sharma (2021) suggests that selecting NMO should be done while considering dataset size and language pair, as nuanced BPE strategies benefit morphologically complex languages. We study symmetrical BPE configurations with identical NMOs for source and target, and investigate alternatives by varying m_1 and m_2 independently in English-Hindi across datasets from 50K to 8M sentences. This approach improves results in low-resource settings (Figure 1). Extensive experiments on English-Hindi, evaluated on FLORES (Goyal et al., 2022), confirm better performance of atypical BPE for tokenization. We further validate these findings by extending experiments to English $\leftrightarrow$ {Telugu, Shona, Norwegian, Kyrgyz, Hausa, Inuktitut}. Our results strongly support optimizing NMO based on training data size and language pair. Figure 2 presents a conceptual overview of the **optimal ranges** for **BPE configurations** found in English-Hindi across resource settings. Here, "ranges" indicate the spectrum of NMO values used as hyperparameters for source and target subword tokenization in word segmentation. The performance gap between the best and symmetrical BPE systems is shown by shades of green, with the largest gains in low-resource scenarios (darker green). As dataset size increases, performance differences among configurations diminish (lighter green).

4 Evaluation on English $\leftrightarrow$ Hindi

We explore BPE configurations with the Samanantar dataset (Ramesh et al., 2022) for English-Hindi containing 8 million parallel sentences. English text is tokenized, normalized, and lowercased using Moses scripts[1], while preprocessing of Hindi utilizes the Indic NLP library (Kunchukuttan, 2020). We simulate various training set sizes by grouping sentences based on English sentence length (Table 2) and randomly sample datasets of sizes 0.05M, 0.1M, 0.5M, 1M, 4M, and 8M, maintaining sentence length proportions (see Appendix A.1 for details). The BPE tokenizer is trained per language and dataset size with eight NMOs: *0.5K*, *1K*, *2K*, *4K*, *8K*, *16K*, *25K*, and *32K*.

All possible BPE configurations (e.g., src_{500}-tgt_{500}, src_{500}-tgt_{1000}) are trained using the Transformer architecture (Vaswani et al., 2017) with hyperparameters detailed in Appendix A.2. Training a single BPE configuration $m_1\_m_2$ across all dataset sizes averages 1040 GPU hours on a 1080TI, resulting in 64 configurations per language direction and 768 total systems (64 configurations $\times$ 6 dataset sizes $\times$ 2 directions). For evaluation, we use the FLORES dataset (Goyal et al., 2022) and report CHRF++ scores (Popović, 2015) to analyze the impact of different BPE configurations. We adopt CHRF++ rather than embedding-based metrics such as COMET (Rei et al., 2022), as not all language pairs have COMET support and we aim to compare performance using a consistent metric across all pairs. Validation and test set statistics are provided in Appendix A.8.

4.1 Best and Worst Configurations

To maintain clarity and brevity in our observations, Tables 3 and Table 4 show the performance of five selected configurations out of 64. For each dataset size, the systems represented are:

- High A and B: The two systems with the highest performance across all asymmetric configurations for each dataset size.
- Low A and B: The two systems with the lowest performance across all asymmetric configurations for each dataset size.
- Baseline: The best system among all symmetric BPE configurations (*m_m*, where $m \subset \{500, 1K, 2K, 4K, 8K, 16K, 25K, 32K\}$).

Performance of all configurations for all systems is provided in the Appendix A.3.

[1] https://github.com/moses-smt/mosesdecoder/

Length bin	1 to 10	11 to 15	16 to 20	21 to 25	26 to 30	31 to 35	35 to 40	>=41	Total
No. of sentences	2792334	1655162	1150396	854091	617318	420583	275774	414926	8180584
Percentage	34.13	20.23	14.06	10.44	7.55	5.14	3.37	5.07	100

Table 2: Distribution of sentences in groups based on token length for full data

Dataset Size	0.05 M				0.1 M				0.5 M			
Performance Tier	src	tgt	CHRF++	δ	src	tgt	CHRF++	δ	src	tgt	CHRF++	δ
Low A	500	1K	19.56	-3.93	500	25K	23.36	-15.92	2K	32K	48.92	-3.53
Low B	500	2K	19.58	-3.91	1K	32K	24.2	-15.08	25K	32K	49.62	-2.83
Baseline	4K	4K	23.49	0	500	500	39.28	0	4K	4K	52.45	0
High B	25K	500	**28.47***	4.98	16K	500	**40.66***	1.38	8K	2K	**53.19***	0.74
High A	16K	500	**29.33***	5.84	8K	500	**40.75***	1.47	4K	500	**53.37***	0.92
Dataset Size	1 M				4 M				8 M			
Performance Tier	src	tgt	CHRF++	δ	src	tgt	CHRF++	δ	src	tgt	CHRF++	δ
Low A	500	32K	53.27	-1.77	500	1K	56.1	-1.73	500	2K	56.26	-2.45
Low B	1K	32K	53.58	-1.46	1K	2K	56.3	-1.53	500	500	56.43	-2.28
Baseline	8K	8K	55.04	0	32K	32K	57.83	0	32K	32K	58.71	0
High B	16K	8K	55.19	0.15	32K	16K	58.06	0.23	16K	25K	58.74	0.03
High A	16K	4K	55.39	0.35	25K	16K	58.18	0.35	4K	32K	58.75	0.04

Table 3: Performance of the top 2 (High A, High B) and bottom 2 (Low A, Low B) tokenization configurations compared to the symmetric baseline for Hindi-to-English across dataset sizes. Bold indicates statistically significant improvement over baseline ($p < 0.05$); bold with * denotes high significance ($p < 0.01$). δ shows CHRF++ difference from best baseline. **src** and **tgt** are source and target merge operations (NMO).

Dataset Size	0.05 M				0.1 M				0.5 M			
Performance Tier	src	tgt	CHRF++	δ	src	tgt	CHRF++	δ	src	tgt	CHRF++	δ
Low A	1K	25K	13	-5.39	500	32K	16.49	-12.55	500	32K	43.57	-3.5
Low B	500	4K	13.55	-4.84	500	25K	16.74	-12.3	1K	32K	43.88	-3.19
Baseline	8K	8K	18.39	0	4K	4K	29.04	0	4K	4K	47.07	0
High B	16K	500	**23.19***	4.8	16K	500	**34.73***	5.69	8K	500	47.12	0.05
High A	8K	500	**23.83***	5.44	8K	500	**35***	5.96	4K	500	**47.55**	0.48
Dataset Size	1 M				4 M				8 M			
Performance Tier	src	tgt	CHRF++	δ	src	tgt	CHRF++	δ	src	tgt	CHRF++	δ
Low A	1K	32K	47.23	-1.93	8K	2K	50.64	-1.12	500	1K	50.79	-1.84
Low B	2K	32K	47.83	-1.33	500	2K	50.73	-1.03	32K	2K	51.29	-1.34
Baseline	8K	8K	49.16	0	16K	16K	51.76	0	25K	25K	52.63	0
High B	4K	2K	**49.74**	0.58	16K	32K	51.95	0.19	25K	32K	52.63	0
High A	8K	2K	**49.75**	0.59	32K	25K	52	0.24	16K	25K	**53**	0.37

Table 4: Performance of the top 2 (High A, High B) and bottom 2 (Low A, Low B) tokenization configurations compared to the symmetric baseline for English-to-Hindi across dataset sizes. Bold indicates statistically significant improvement over baseline ($p < 0.05$); bold with * denotes high significance ($p < 0.01$). δ shows CHRF++ difference from best baseline. **src** and **tgt** are source and target merge operations (NMO).

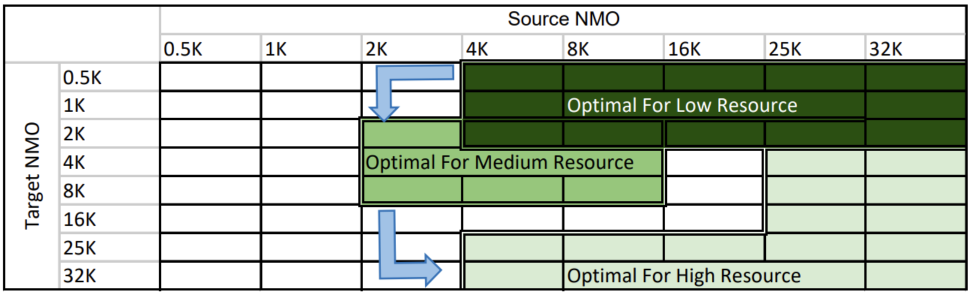


Figure 2: Changes in Optimal BPE Configuration from Low- to High-Resource Settings

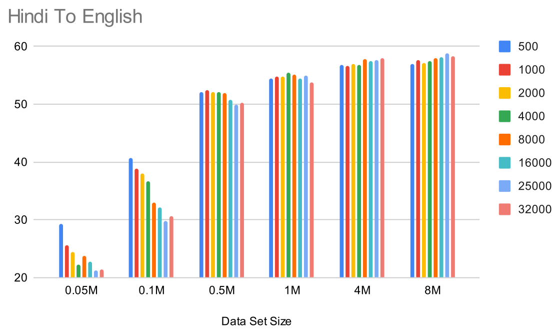


Figure 3: CHRF++ scores for 0.1M sentence pairs for *Hindi-to-English* MT systems using configurations of the form *16K_x*, where $x \in \{$*500*, *1K*, *2K*, *4K*, *8K*, *16K*, *25K*, *32K*$\}$.

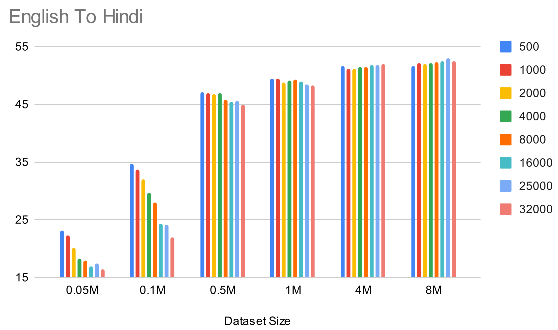


Figure 4: CHRF++ scores for 0.1M sentence pairs for *English-to-Hindi* MT systems using configurations of the form *16K_x*, where $x \in \{$*500*, *1K*, *2K*, *4K*, *8K*, *16K*, *25K*, *32K*$\}$.

As shown in Tables 3 and 4, for low-resource settings (<1M), the best system outperforms the weakest by ≈15 CHRF++ scores and the best symmetric BPE by ≈5. In medium-resource scenarios (1M), the optimal source and target NMO shift to the medium range (2K–8K), with smaller performance variation (≈3 CHRF++). For high-resource settings, the difference between best and worst configurations is minimal (< 2 CHRF++), with the best system using 32K NMO on the target. This highlights the advantage of asymmetric BPE in low-resource contexts. This trend of shifting optimal BPE values with dataset size also appears when varying target NMO while keeping source NMO fixed. For example, English↔Hindi systems with source NMO fixed at 16K on 0.1M data (Figures 3 and 4) show gradual performance changes as target NMO varies from 500 to 32K. Similar patterns with other fixed source or target values are detailed in Appendix A.3. This highlights that modifying the NMO on the target side, especially in a low-resource scenario, plays a vital role in determining the optimal BPE configuration.

We conclusively find that symmetric BPE configurations underperform compared to asymmetric ones in low-resource MT systems. As dataset size grows, symmetric configurations perform comparably to asymmetric. Nonetheless, asymmetric BPE yields statistically significant improvements in low-resource settings.

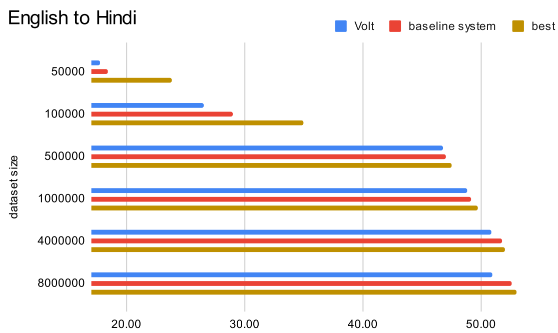


Figure 5: CHRF++ score comparison of Asymmetric BPE with VOLT for English to Hindi

We compare our systems with optimal BPE configurations against VOLT (Xu et al., 2021)[2]. Figures 5 and 6 show CHRF++ comparisons between VOLT tokenization, optimal BPE, and "best" baseline symmetric BPE (source NMO = target NMO)

[2]Using hyperparameters specified in the original paper.

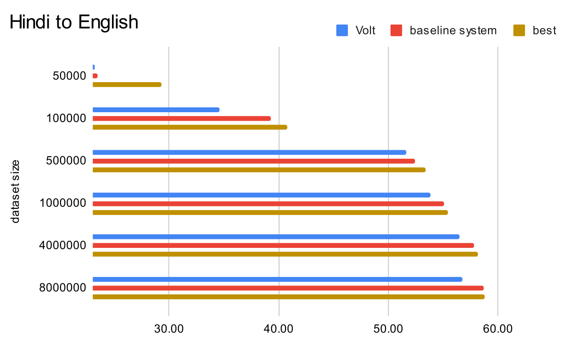


Figure 6: CHRF++ score comparison of Asymmetric BPE with VOLT for Hindi to English

Domain	# of Sentences	English Tokens	Hindi Tokens
Artificial Intelligence	389	6965	8441
Chemistry	392	7761	9368

Table 5: Statistics of ICON 2020 Domain Adaptation Testset

configuration. Systems using asymmetric BPE outperform VOLT across all dataset sizes, with statistically significant improvements ($p < 0.05$) especially in low-resource settings.

4.2 Performance on Domain Test

Subword models must handle rare or unseen words, making domain-specific datasets effective for evaluating asymmetric BPE in MT systems. Thus, to demonstrate the impact of segmentation strategies, we evaluate all systems on Artificial Intelligence (AI) and Chemistry (CH) domain test sets from the ICON 2020 Domain Adaptation Task[3]. Table 5[4] presents domain test data statistics. Table 6 show the performance of configurations from Table 4 on domain datasets for English-to-Hindi systems. Performance of Hindi-to-English systems is given in Appendix A.4.

For English↔Hindi domain test set translation, we observe:

- In low- to medium-resource settings, asymmetric BPE systems outperform baselines significantly when source NMO is much higher than target NMO. This aligns with FLORES results (Tables 3 and 4) and highlights asymmetric BPE benefits for domain translation with limited data.
- In high-resource settings, symmetric and asymmetric systems perform similarly.

[3]We thank task organizers for access.

[4]After removing 12 and 5 lines from AI and CH test sets respectively, that overlapped with the 8M training set.

These results demonstrate the potential translation improvements from asymmetric BPE in new domains under limited-resource conditions. Performances of all systems on AI and CH test sets is in Appendices A.5 and A.6, respectively.

Figure 7 illustrates, with an example on AI domain, the advantage of asymmetric BPE over symmetric BPE for 0.1M parallel sentences. Configurations like *16K_500* or *8K_500* produce more natural, semantically faithful Hindi translations than symmetric *32K_32K* or *4K_4K* setups. Translation improves as we move from symmetric high NMO (*32K_32K*), to symmetric low NMO (*4K_4K*), to asymmetric (*16K_500* or *8K_500*).

- ***32K_32K*** – In the output with delimiters, most of the tokens are already fully merged into complete words. While this segmentation yields a large vocabulary, in low-resource conditions, it results in sparsity: many source and target tokens appear too infrequently for effective parameter learning. Consequently, the network fails to learn robust mappings, leading to incomplete or inaccurate translations despite having fully merged tokens.
- ***4K_4K*** – The glossary shows an improvement in overall translation fluency, but important content words such as system, commonly and click are missing, both explicitly and implicitly (meaning that they cannot be inferred from context). The improvement is due to the increased recurrence of subword units in the training data from the reduced vocabulary size, which strengthens learned associations, but at the cost of certain semantic details.
- **Asymmetric (*16K_500*, *8K_500*)**: Better meaning preservation than symmetric. Whereas *16K_500* omits "post" and drops final language reference, *8K_500* conveys almost full meaning but mistranslates "post" as a job title. From a learning perspective, the smaller decoder vocabulary improves the alignment and connection learning between the source and target segments (similar to Ngo Ho and Yvon (2021)), aligning with previous findings (Domingo et al., 2023) that the target side vocabulary influences NMT performance. Although overly constrained vocabularies can still introduce semantic drift in rare or domain-specific terms, overall transla-

Dataset Size	0.05M				0.1M				0.5M			
Performance Tier	src	tgt	AI	CH	src	tgt	AI	CH	src	tgt	AI	CH
Low A	1K	25K	15.98	14.13	500	32K	18.46	16.67	500	32K	53.32	47.44
Low B	500	4K	15.97	15.03	500	25K	18.80	16.86	1K	32K	53.99	47
Baseline	8K	8K	20.76	19.34	4K	4K	35.79	32.19	4K	4K	58.63	50.64
High B	16K	500	**26.76***	**24.03***	16K	500	**42.97***	**37.94***	8K	500	58.91	50.94
High A	8K	500	**28.28***	**25.14***	8K	500	**44.05***	**38.57***	4K	500	58.70	**51.53**
Dataset Size	1M				4M				8M			
Performance Tier	src	tgt	AI	CH	src	tgt	AI	CH	src	tgt	AI	CH
Low A	1K	32K	58.58	51.78	8K	2K	62.23	54.55	500	1K	61.91	54.78
Low B	2K	32K	58.88	51.65	500	2K	61.51	54.01	32K	2K	62.52	54.63
Baseline	8K	8K	61.22	53.6	16K	16K	63.12	55.14	25K	25K	63.95	55.65
High B	4K	2K	60.39	53.55	16K	32K	63.21	**55.84**	25K	32K	63.9	55.92
High A	8K	2K	60.01	53.27	32K	25K	63.6	55.74	16K	25K	63.53	55.69

Table 6: Performance of the top 2 (High A and High B) and bottom 2 (Low A and Low B) systems with respective tokenisation configurations compared to the symmetric baseline for *English-to-Hindi* systems across dataset sizes for **AI** and **CH Domains**. Bold scores indicate statistically significant improvements over the baseline ($p < 0.05$); bold scores with an asterisk ($*$) indicate high significance ($p < 0.01$)

BPE configuration	32K_32K	4K_4K	16K_500	8K_500
Segmented Source	integrated machine translation systems are commonly used in social media services where the posts of users can be translated into the reader 's own language with a single click .	integr@@ ated mach@@ ine trans@@ l@@ ation systems are comm@@ only used in social media services where the po@@ sts of us@@ ers can be trans@@ l@@ ated into the read@@ er 's own language with a single cl@@ ick .	integrated machine translation systems are commonly used in social media services where the posts of users can be translated into the reader 's own language with a single click .	integrated machine transl@@ ation systems are comm@@ only used in social media services where the posts of users can be transl@@ ated into the read@@ er 's own language with a single click .
Output with Delimiters	एकीकृत मशीन प्रणाली में प्रयोग किया जाता है जहां सोशल मीडिया साइटों का प्रयोग अंग्रेजी में किया जाता है , जहां वे भाषा का प्रयोग कर सकते हैं ।	ए@@ की@@ कृत मशी@@ न अनु@@ वाद सोशल मीडिया में ही किया जाता है , जहां उप@@ योग@@ कर्ताओं के साथ एक ही भाषा के साथ अनु@@ वाद किया जा सकता है ।	ए@@ की@@ कृ@@ त म@@ श@@ ी @@ न अनु@@ वा@@ द सा@@ मा@@ न्य रूप से सो@@ श@@ ल मी@@ डिया से@@ वा@@ ओं में प्र@@ योग किया जाता है ज@@ हां उप@@ यो@@ ग@@ कर्@@ ता@@ ओं को एक ही क्@@ लि@@ क के साथ पा@@ ठ@@ क में अनु@@ वा@@ द किया जा सकता है ।	ए@@ की@@ कृ@@ त म@@ श@@ ी @@ न अनु@@ वा@@ द प्र@@ ण@@ ा @@ लि@@ यों का इ@@ स्@@ ते@@ मा@@ ल सो@@ श@@ ल मी@@ डिया से@@ वा@@ ओं में किया जाता है ज@@ हां उप@@ यो@@ ग@@ कर्@@ ता@@ ओं के प@@ द@@ ों को एक ही क्@@ लि@@ क के साथ अनु@@ वा@@ द किया जा सकता है ।
Output with Glossary	एकीकृत मशीन प्रणाली → Integrated machine system में → in प्रयोग किया जाता है → is being used जहां सोशल मीडिया साइटों का प्रयोग अंग्रेजी में किया जाता है , → where social media sites are used in English जहां वे भाषा का प्रयोग कर सकते हैं । → where they can use the language	एकीकृत मशीन अनुवाद सोशल मीडिया में ही किया जाता है → Integrated machine translation is only used in social media जहां उपयोगकर्ताओं के साथ → where with users एक ही भाषा के साथ अनुवाद किया जा सकता है । → translation can be done in only one language	एकीकृत मशीन अनुवाद सामान्य रूप → Integrated machine translation commonly से सोशल मीडिया सेवाओं में प्रयोग किया जाता है → used in social media services जहां उपयोगकर्ताओं को एक ही क्लिक के साथ → where users, with one click पाठक में अनुवाद किया जा सकता है । → can translate into the reader	एकीकृत मशीन अनुवाद प्रणालियों का इस्तेमाल → Integrated machine translation systems are used सोशल मीडिया सेवाओं में किया जाता है → in social media services जहां उपयोगकर्ताओं के पदों को → where users' posts (here "posts" refers to positions in an organisation, not social media posts) एक ही क्लिक के साथ अनुवाद किया जा सकता है । → can be translated with a single click

Figure 7: Examples of English-to-Hindi translations across different BPE configurations, showing segmented source text, outputs with delimiters '@@', and output without delimiters with corresponding English glossaries for each segment.

tion remains improved compared to symmetric configurations.

5 Exploring Asymmetrical BPE Configurations for other language pairs

To evaluate the transferability of optimal subword segmentation from English–Hindi to typologically diverse languages, we extend experiments to English↔{Telugu, Shona, Norwegian, Kyrgyz, Hausa, Inuktitut}. Corpora sources are:

- **English–{Hausa, Shona, Norwegian, Kyrgyz}**: Gowda et al. (2021)
- **English–Telugu**: Ramesh et al. (2022)
- **English–Inuktitut**: Joanis et al. (2020)

To simulate low-resource settings, we sampled 0.1M sentence pairs per language via sentence-length binning, analogous to English–Hindi, statistics are in Appendix A.7.

These language pairs were chosen to assess the impact of symmetric and asymmetric BPE configurations in low-resource scenarios across diverse language families with varying morphological and typological complexity. Baselines used symmetric BPE (*4K_4K*, *32K_32K*), while asymmetric settings (*8K_500*, *16K_500*) derive from English-Hindi optimal configurations at 0.1M sentence pairs. For evaluating we use the FLORES test set, except English↔Inuktitut tested on Joanis et al. (2020) (Appendix A.8).

Experiments are repeated three times for reproducibility (sampling, BPE training, model training). Figures 8 and 9 compare average asymmetric and symmetric BPE results for translations to and from English. Asymmetric BPE significantly improves four of six *L-to-English* systems and all *English-to-L* systems ($p < 0.05$, indicated by *), underscoring the benefits of asymmetric BPE and the need to explore beyond conventional settings for low-resource pairs.

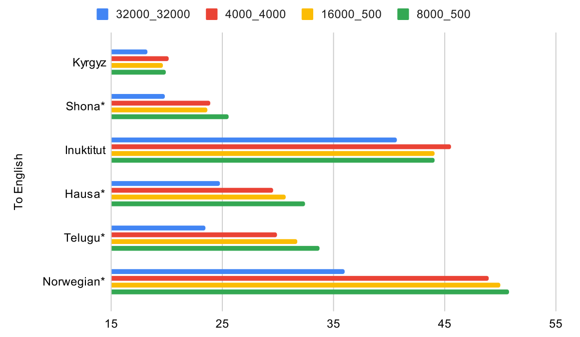


Figure 8: CHRF++ scores improvement with asymmetrical over symmetrical BPE for English to L Languages

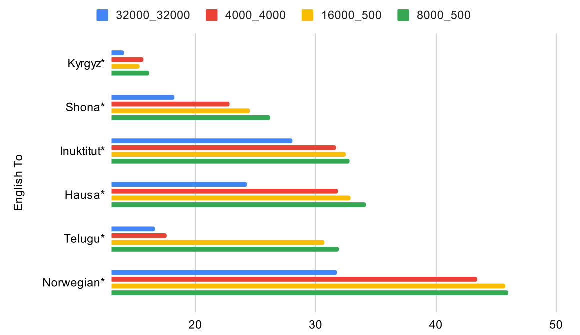


Figure 9: CHRF++ scores improvement with asymmetrical over symmetrical BPE from L Languages to English

6 Conclusion

In-depth examination of BPE configurations across diverse language pairs and differing dataset sizes reveals that typical configurations (*n_n*) do not always produce optimal results. As referenced in Section 2, in low-resource settings, systems benefit from using symmetric n NMO configurations when n is significantly smaller than *32K*; our experiments with asymmetric BPE *n_m* show that further improvement in translation performance is possible, under low-resource conditions, when $n \gg m$ where n, m represent NMOs for source and target respectively. This study highlights the need to go beyond default segmentation in machine translation, especially for low-resource languages. While symmetric BPE configurations may suffice with medium to large datasets, their effectiveness drops in low-resource settings. Using asymmetric BPE—with a higher number of merge operations for the source language and fewer for the target—yields significant translation quality gains. These configurations consistently outperform across varied language families and morphological complexities, underscoring the importance of tailored segmentation for optimizing low-resource translation.

Limitation

This study is limited by the computational cost of exhaustively analysing all BPE configurations for each language pair and by its focus only on bilingual encoder–decoder NMT. However, the results show that certain configuration ranges consistently improve translation quality in low-resource settings, substantially reducing the search space. These findings suggest promising extensions to multilingual models, potentially combined with fairness-aware tokenisation such as Parity-Aware BPE (Foroutan et al., 2025) to deliver both performance gains and balanced vocabulary distribution.

References

Wael Abid. 2020. The SADID evaluation datasets for low-resource spoken language machine translation of Arabic dialects. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6030–6043, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Colin Cherry, George Foster, Ankur Bapna, Orhan Firat, and Wolfgang Macherey. 2018. Revisiting character-based neural machine translation with capacity and compression. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4295–4305, Brussels, Belgium. Association for Computational Linguistics.

Michael Denkowski and Graham Neubig. 2017. Stronger baselines for trustable results in neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 18–27, Vancouver. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Shubham Dewangan, Shreya Alva, Nitish Joshi, and Pushpak Bhattacharyya. 2021. Experience of neural machine translation between indian languages. *Machine Translation*, 35(1):71–99.

Shuoyang Ding, Adithya Renduchintala, and Kevin Duh. 2019. A call for prudent choice of subword merge operations in neural machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 204–213, Dublin, Ireland. European Association for Machine Translation.

Cheikh M. Bamba Dione, Alla Lo, Elhadji Mamadou Nguer, and Sileye Ba. 2022. Low-resource neural machine translation: Benchmarking state-of-the-art transformer for Wolof<->French. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6654–6661, Marseille, France. European Language Resources Association.

Miguel Domingo, Mercedes García-Martínez, Alexandre Helle, Francisco Casacuberta, and Manuel Herranz. 2023. How much does tokenization affect neural machine translation? In *Computational Linguistics and Intelligent Text Processing*, pages 545–554, Cham. Springer Nature Switzerland.

Negar Foroutan, Clara Meister, Debjit Paul, Joel Niklaus, Sina Ahmadi, Antoine Bosselut, and Rico Sennrich. 2025. Parity-aware byte-pair encoding: Improving cross-lingual fairness in tokenization. *Preprint*, arXiv:2508.04796.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316, Online. Association for Computational Linguistics.

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.

Wan-hua Her and Udo Kruschwitz. 2024. Investigating neural machine translation for low-resource languages: Using Bavarian as a case study. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 155–167, Torino, Italia. ELRA and ICCL.

Matthias Huck, Simon Riess, and Alexander Fraser. 2017. Target-side word segmentation strategies for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 56–67, Copenhagen, Denmark. Association for Computational Linguistics.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart, and Jeffrey Micher. 2020. The Nunavut Hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2562–2572, Marseille, France. European Language Resources Association.

Richard Kimera, DongNyeong Heo, Daniela N. Rim, and Heeyoul Choi. 2025. Data augmentation with back translation for low resource languages: A case of english and luganda. In *Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '24, page 142–148, New York, NY, USA. Association for Computing Machinery.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Anoop Kunchukuttan. 2020. The IndicNLP Library. https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf.

Seamus Lankford, Haithem Alfi, and Andy Way. 2021. Transformers for low-resource languages: Is féidir linn! In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 48–60, Virtual. Association for Machine Translation in the Americas.

Jungseob Lee, Hyeonseok Moon, Seungjun Lee, Chanjun Park, Sugyeong Eo, Hyunwoong Ko, Jaehyung Seo, Seungyoon Lee, and Heuiseok Lim. 2024. Length-aware byte pair encoding for mitigating over-segmentation in Korean machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2287–2303, Bangkok, Thailand. Association for Computational Linguistics.

Fuxue Li, Beibei Liu, Hong Yan, Mingzhi Shao, Peijun Xie, Jiarui Li, and Chuncheng Chi. 2024. A bilingual templates data augmentation method for low-resource neural machine translation. In *Advanced Intelligent Computing Technology and Applications: 20th International Conference, ICIC 2024, Tianjin, China, August 5–8, 2024, Proceedings, Part III*, page 40–51, Berlin, Heidelberg. Springer-Verlag.

Evelyn Kai-Yan Liu. 2022. Low-resource neural machine translation: A case study of Cantonese. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 28–40, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Vandan Mujadia and Dipti Misra Sharma. 2021. English-Marathi neural machine translation for LoResMT 2021. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 151–157, Virtual. Association for Machine Translation in the Americas.

Anh Khoa Ngo Ho and François Yvon. 2021. Optimizing word alignments with better subword tokenization. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 256–269, Virtual. Association for Machine Translation in the Americas.

John E Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

Alberto Poncelas, Jan Buts, James Hadley, and Andy Way. 2020. Using multiple subwords to improve English-Esperanto automated literary translation quality. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 108–117, Suzhou, China. Association for Computational Linguistics.

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.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Adithya Renduchintala, Pamela Shapiro, Kevin Duh, and Philipp Koehn. 2019. Character-aware decoder for translation into morphologically rich languages. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 244–255, Dublin, Ireland. European Association for Machine Translation.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Loitongbam Sanayai Meetei, Sivaji Bandyopadhyay, and Thoudam Doren Singh. 2023. NITS-CNLP low-resource neural machine translation systems of English-Manipuri language pair. In *Proceedings of the Eighth Conference on Machine Translation*, pages 967–971, Singapore. Association for Computational Linguistics.

Salam Michael Singh and Thoudam Doren Singh. 2022. Low resource machine translation of english–manipuri: A semi-supervised approach. *Expert Syst. Appl.*, 209(C).

Peter Smit, Sami Virpioja, Stig-Arne Grönroos, and Mikko Kurimo. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 21–24, Gothenburg, Sweden. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yonghui Wu. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Jingjing Xu, Hao Zhou, Chun Gan, Zaixiang Zheng, and Lei Li. 2021. Vocabulary learning via optimal transport for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7361–7373, Online. Association for Computational Linguistics.

A Appendix

A.1 English–Hindi Training Data Statistics

We use an 8-million-sentence English–Hindi corpus from the Samanantar dataset and execute stratified random sampling across sentence length bins to simulate different resource availability levels. Table 7 summarises the statistics for sentence pairs corresponding to each level of resource availability.

A.2 Hyperparameters for Training Transformer Model

We followed the official Fairseq tutorial instructions for preprocessing, training, and translation[5], and customised the parameters given in Table 8 with respective values for all experiments.

A.3 Performance of all systems for English ↔ Hindi for all dataset scenarios

Figures 10 present the performance of all configurations for English ↔ Hindi systems in a low resource scenario (for data set sizes of 0.05M, 0.1M and 0.5M). And Figures 11 show the performance of all configurations on 1M, 4M and 8M dataset sizes. Each subgraph represents performance on a particular dataset size, with the x-axis being the source NMO. The black stepped dotted lines indicate the maximum CHRF++ score for each dataset size considering for each source NMOs. In figure 10 for low-resource environments (0.05M, 0.1M and 0.5M) systems, as noted by (Ding et al., 2019), the use of symmetric BPE configuration with lower NMOs improves performance over high NMOs. However, the best results are achieved using asymmetric BPE configurations when the source has a higher NMO than the target. We see a maximum performance gain when the source NMO is very high and the target NMO very low (we see consistent performance with the target NMO = *500*). Conversely, when the target's NMO is greater than that of the source, performance declines, like for the Hindi to English 0.1M dataset, performance of *500_25K* and *500_32K* was worse than symmetric BPE configurations.

A.4 Performance of Hindi-To-English Selected Configurations on Domain Test set

Table 9 shows the performance of the Highest and Lowest performing asymmetric BPE systems with baseline systems for Hindi-To-English systems. Like in English to Hindi systems, we see significant improvement when using asymmetric BPE configurations in low-resource settings.

A.5 Evaluation of English ↔ Hindi systems on AI for all BPE Configurations

Figures 12 and 13 depict the performance of all configurations for English ↔ Hindi systems during translations in the **AI** domain. A similar performance pattern appears across configurations here, as observed with the FLORES test set (see Appendix A.3).

A.6 Evaluation of English ↔ Hindi systems on Chemistry for all BPE Configurations

Figures 14 and 15 depict the performance of all configurations for English ↔ Hindi systems during translations in the **Chemistry** domain. A similar performance pattern appears across configurations here, as observed with the FLORES test set (see Appendix A.3).

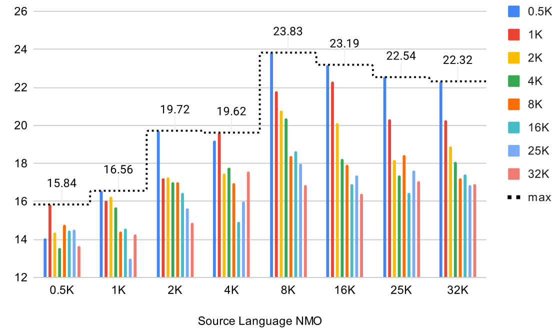
A.7 Statistics of Bitext for secondary set of experiments

Table 10 gives the statistics of the original bitext that we obtained for the secondary set of experiments, to see the transferability of asymmetric BPE configurations. And to simulate low-resource settings, we sampled 0.1M sentence pairs per language using sentence-length binning, as done for English–Hindi; statistics are shown in Table 11.

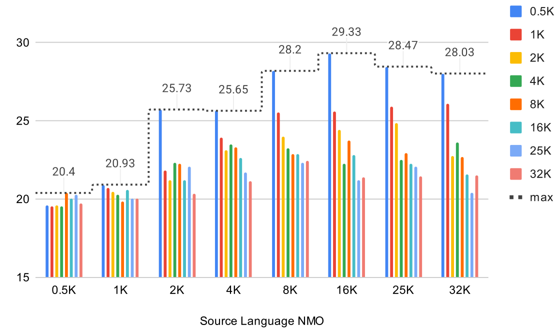
A.8 Validation and Test Set Statistics

As noted, for English–Inuktitut validation and test sets, we use Joanis et al. (2020). For all other language pairs, the FLORES dataset was used. Table 12 shows token-level statistics for validation and test sets across all language pairs.

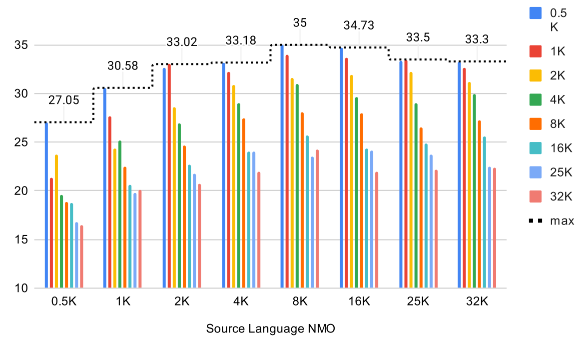
[5] https://fairseq.readthedocs.io/en/latest/getting_started.html



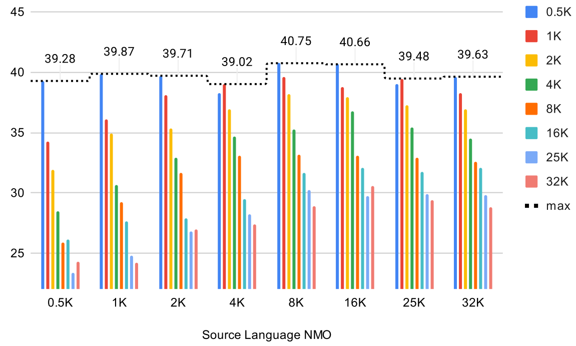
(a) 0.05 Million English to Hindi



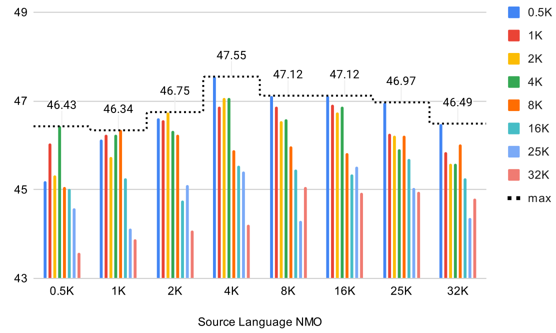
(b) 0.05 Million Hindi to English



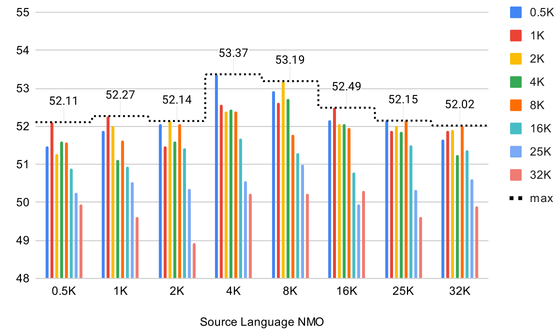
(c) 0.1 Million English to Hindi



(d) 0.1 Million Hindi to English

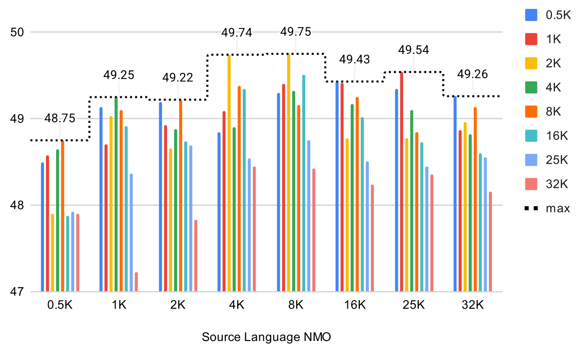


(e) 0.5 Million English to Hindi

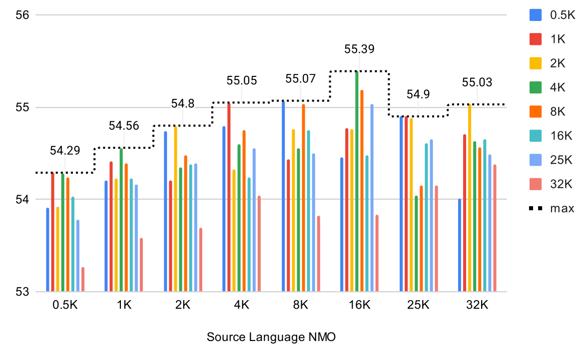


(f) 0.5 Million Hindi to English

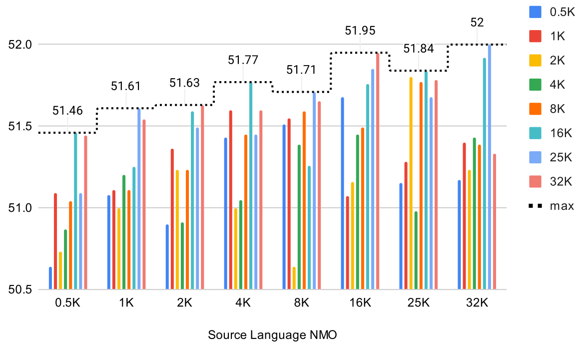
Figure 10: Evaluation of English ↔ Hindi MT Systems for 0.05M, 0.1M and 0.5M dataset sizes on FLORES, x-axis is source NMO and y-axis is CHRF++ scores



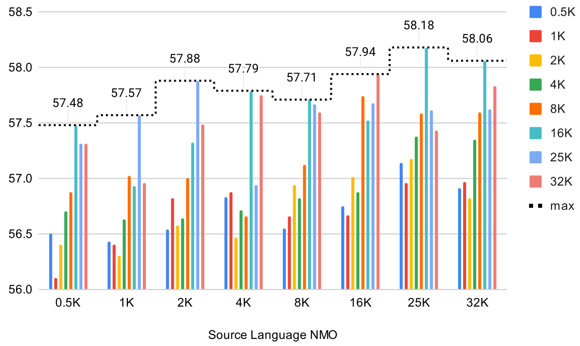
(a) 1 Million English to Hindi



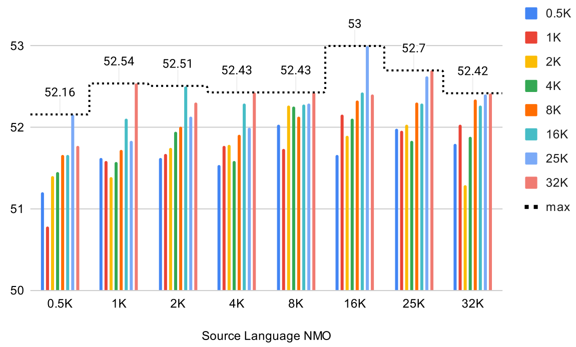
(b) 1 Million Hindi to English



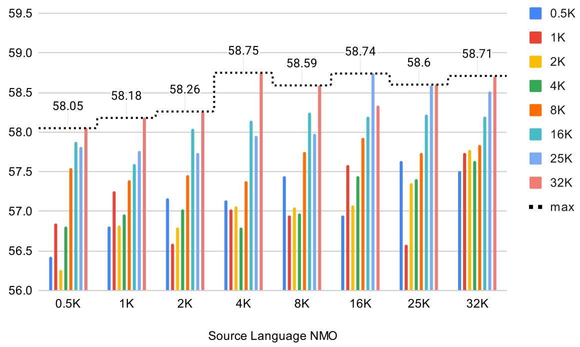
(c) 4 Million English to Hindi



(d) 4 Million Hindi to English

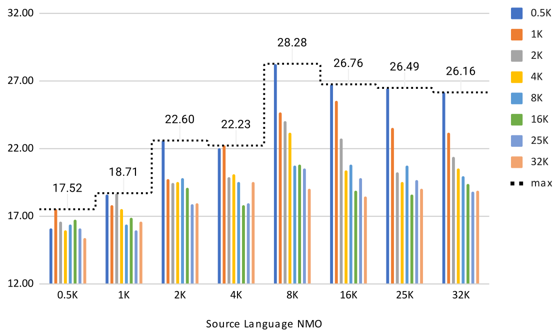


(e) 8 Million English to Hindi

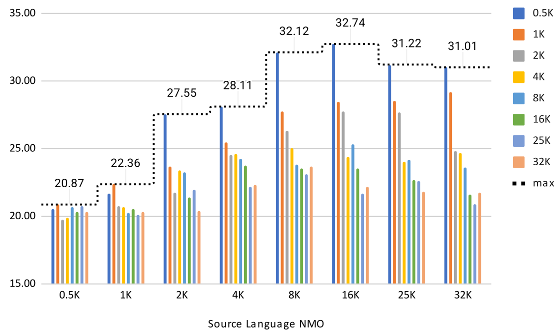


(f) 8 Million Hindi to English

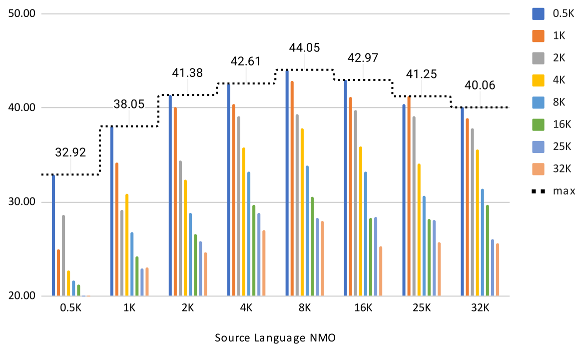
Figure 11: Evaluation of English ↔ Hindi MT Systems for 1M, 4M and 8M dataset sizes on FLORES, x-axis is source NMO and y-axis is CHRF++ scores



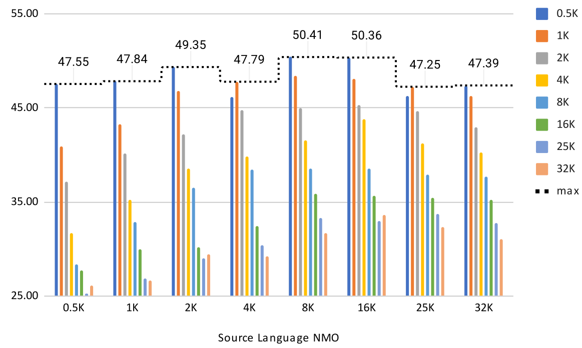
(a) 0.05 Million English to Hindi



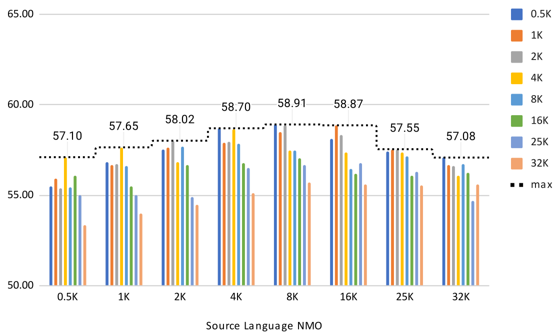
(b) 0.05 Million Hindi to English



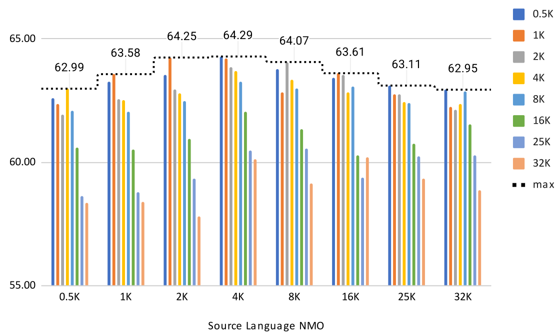
(c) 0.1 Million English to Hindi



(d) 0.1 Million Hindi to English

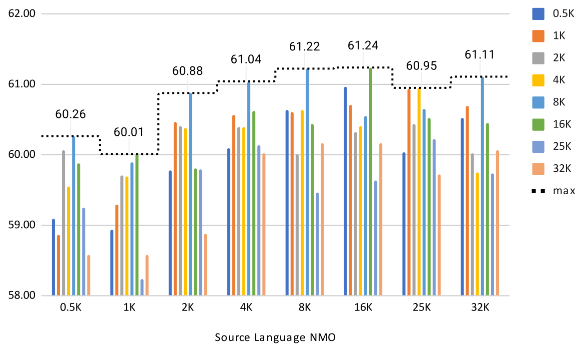


(e) 0.5 Million English to Hindi

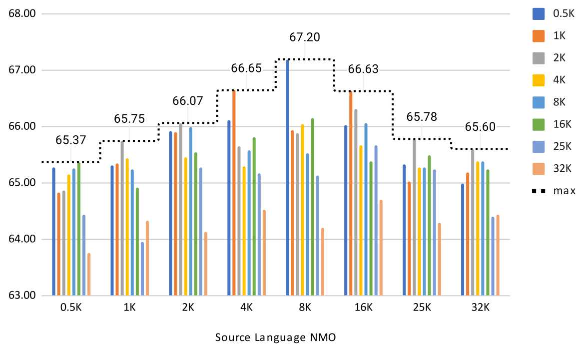


(f) 0.5 Million Hindi to English

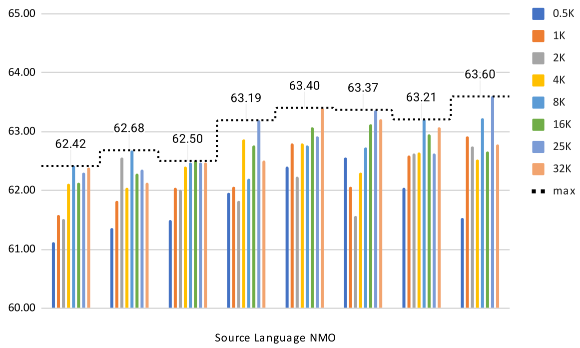
Figure 12: Evaluation of English ↔ Hindi MT Systems for 0.05M, 0.1M and 0.5M dataset sizes on **AI**, x-axis is source NMO and y-axis is CHRF++ scores



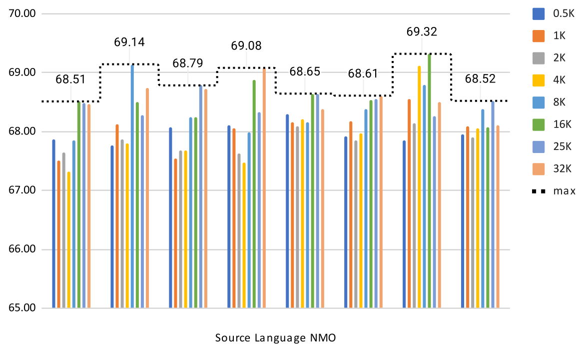
(a) 1 Million English to Hindi



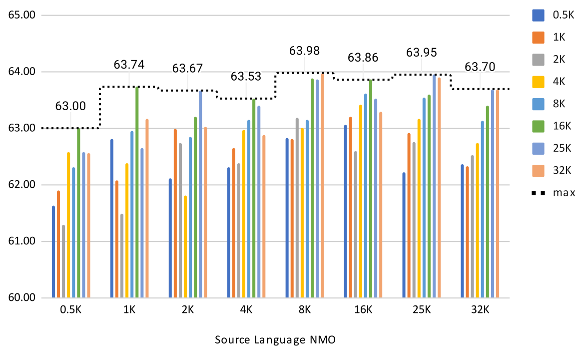
(b) 1 Million Hindi to English



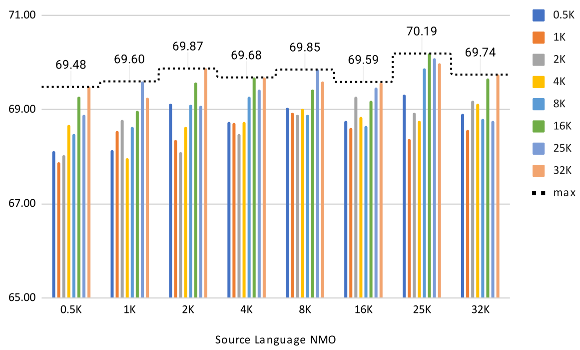
(c) 4 Million English to Hindi



(d) 4 Million Hindi to English

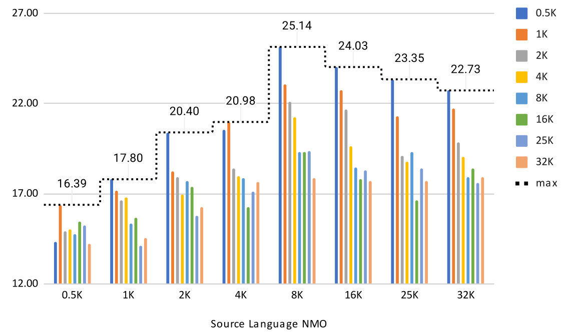


(e) 8 Million English to Hindi

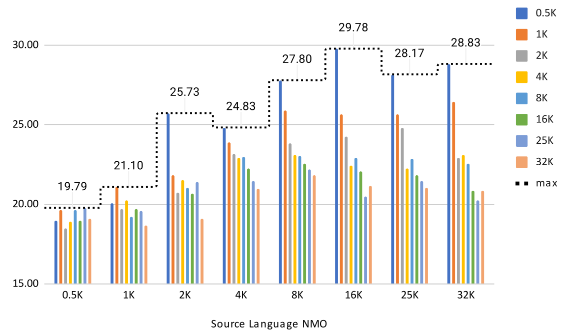


(f) 8 Million Hindi to English

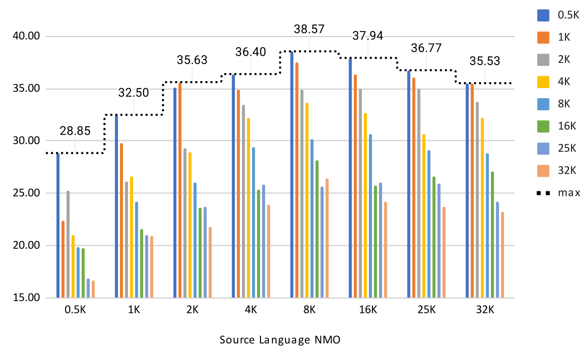
Figure 13: Evaluation of English ↔ Hindi MT Systems for 1M, 4M and 8M dataset sizes on **AI**, x-axis is source NMO and y-axis is CHRF++ scores



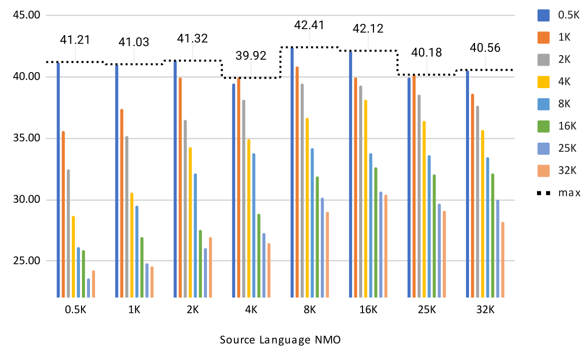
(a) 0.05 Million English to Hindi



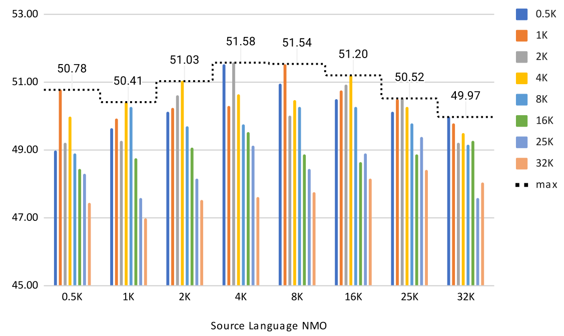
(b) 0.05 Million Hindi to English



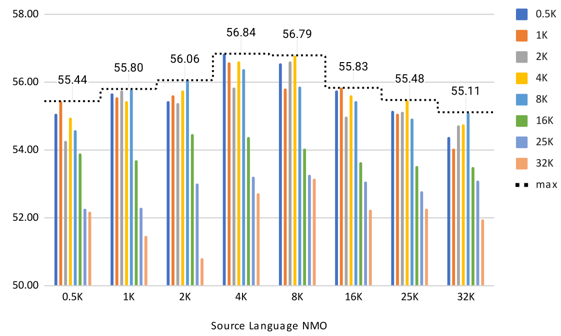
(c) 0.1 Million English to Hindi



(d) 0.1 Million Hindi to English

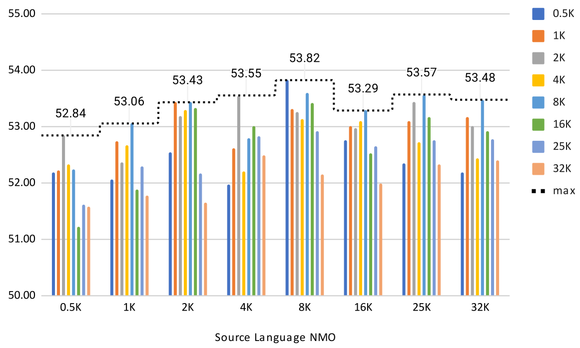


(e) 0.5 Million English to Hindi

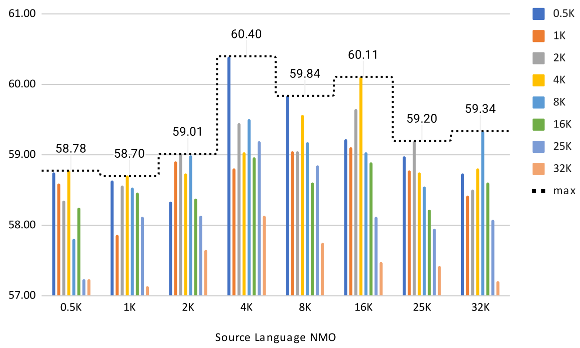


(f) 0.5 Million Hindi to English

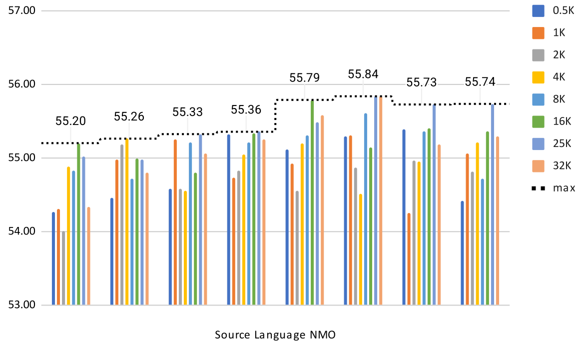
Figure 14: Evaluation of English ↔ Hindi MT Systems for 0.05M, 0.1M and 0.5M dataset sizes on **CH**, x-axis is source NMO and y-axis is CHRF++ scores



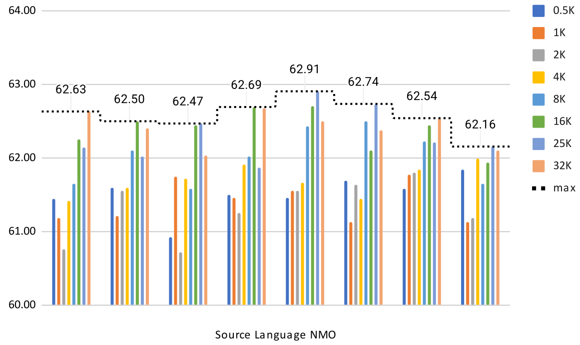
(a) 1 Million English to Hindi



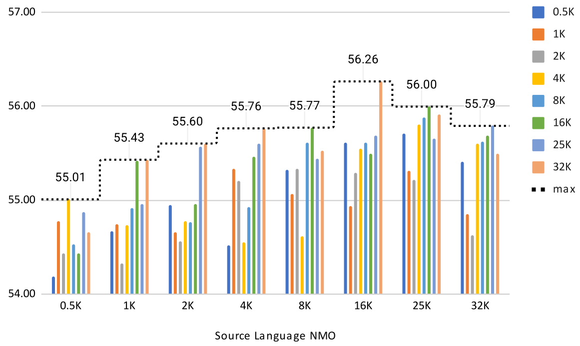
(b) 1 Million Hindi to English



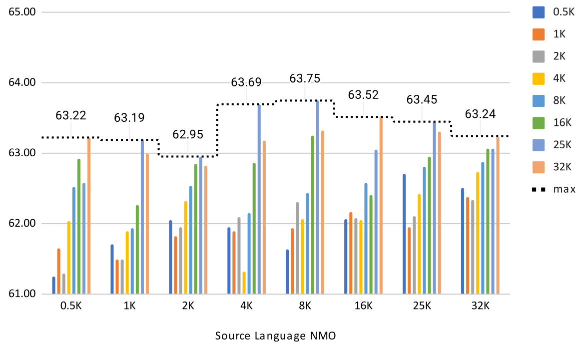
(c) 4 Million English to Hindi



(d) 4 Million Hindi to English



(e) 8 Million English to Hindi



(f) 8 Million Hindi to English

Figure 15: Evaluation of English ↔ Hindi MT Systems for 1M, 4M and 8M dataset sizes on **CH**, x-axis is source NMO and y-axis is CHRF++ scores

Length Range	**# of Lines**	**% of Total**	**4M**	**1M**	**0.5M**	**0.1M**
1 to 10	2,792,334	34.13%	1,365,200	341,300	170,650	34,130
11 to 15	1,655,162	20.23%	809,200	202,300	101,150	20,230
16 to 20	1,150,396	14.06%	562,400	140,600	70,300	14,060
21 to 25	854,091	10.44%	417,600	104,400	52,200	10,440
31 to 35	420,583	5.14%	205,600	51,400	25,700	5,140
36 to 40	275,774	3.37%	134,800	33,700	16,850	3,370
≥ 41	414,926	5.07%	202,800	50,700	25,350	5,070
Total	**8,180,584**		**3,999,600**	**999,900**	**499,950**	**99,990**

Table 7: Distribution of English–Hindi sentence pairs sampled from Samanantar across sentence length bins and different dataset sizes.

Parameter	**Value**
`arch`	transformer
`optimizer`	adam
`adam-betas`	(0.9, 0.98)
`clip-norm`	0.0
`lr`	5e-4
`lr-scheduler`	inverse_sqrt
`warmup-updates`	4000
`warmup-init-lr`	1e-07
`dropout`	0.3
`attention-dropout`	0.1
`activation-dropout`	0.1
`weight-decay`	0.0001
`criterion`	label_smoothed_cross_entropy
`label-smoothing`	0.1
`max-tokens`	6000
`max-update`	300000
`patience`	20
`update-freq`	10

Table 8: Training hyperparameters used across all experiments.

Dataset Size	0.05M				0.1M				0.5M			
Performance Tier	src	tgt	AI	CH	src	tgt	AI	CH	src	tgt	AI	CH
Low A	500	1K	20.87	19.64	500	25K	25.22	23.56	2K	32K	57.8	50.82
Low B	500	2K	19.71	18.46	1K	32K	26.65	24.61	25K	32K	59.35	52.27
Baseline	4K	4K	24.61	22.92	500	500	47.55	41.21	4K	4K	63.7	56.61
High B	25K	500	**31.22***	**28.17***	16K	500	**50.36***	**42.12***	8K	2K	64.07	56.61
High A	16K	500	**32.74***	**29.78***	8K	500	**50.41***	**42.41***	4K	500	**64.29***	56.84
Dataset Size	1M				4M				8M			
Performance Tier	src	tgt	AI	CH	src	tgt	AI	CH	src	tgt	AI	CH
Low A	500	32K	63.75	57.23	500	1K	67.51	61.19	500	2K	68.02	61.3
Low B	1K	32K	64.33	57.13	1K	2K	67.86	61.55	500	500	68.12	61.24
Baseline	8K	8K	65.52	59.18	32K	32K	68.1	62.1	32K	32K	69.74	63.24
High B	16K	8K	**66.07***	59.03	32K	16K	68.08	61.94	16K	25K	69.47	63.05
High A	16K	4K	65.68	60.11	25K	16K	**69.32**	62.45	4K	32K	69.68	63.18

Table 9: Performance of the top 2 (High A and High B) and bottom 2 (Low A and Low B) systems with respective tokenisation configurations compared to the symmetric baseline for *Hindi-to-English* systems across dataset sizes for **AI** and **CH Domains**. Bold scores indicate statistically significant improvements over the baseline ($p < 0.05$); bold scores with an asterisk ($*$) indicate high significance ($p < 0.01$)

Language	# Sentence Pairs	English Tokens	L Tokens
Telugu	508,557	9,277,916	6,861,361
Shona	9,463,612	98,089,812	76,046,554
Norwegian	1,454,765	22,223,984	20,541,537
Kyrgyz	21,603,490	251,345,836	168,333,543
Hausa	4,452,045	57,987,583	64,016,592
Inuktitut	733,624	15,751,147	7,991,818

Table 10: Original corpus statistics English - L Language for secondary language pair.

Language	English Tokens	L Tokens
Telugu	2,471,877	1,919,321
Shona	1,228,485	965,502
Norwegian	1,791,571	1,641,309
Kyrgyz	1,385,891	936,543
Hausa	1,531,132	1,679,785
Inuktitut	2,148,188	1,089,834

Table 11: Token statistics after sampling 0.1 million training sentence pairs per language pair (English - L).

Language	Split	# Sentences	English Tokens	L Tokens
Hindi	validation	997	23,586	27,325
	test	1,012	24,722	28,534
Telugu	validation	997	23,586	19,443
	test	1,012	24,722	20,213
Shona	validation	997	23,586	19,116
	test	1,012	24,722	19,958
Norwegian	validation	997	23,586	23,472
	test	1,012	24,722	24,213
Kyrgyz	validation	997	23,586	18,935
	test	1,012	24,722	20,022
Hausa	validation	997	23,586	27,031
	test	1,012	24,722	28,018
Inuktitut	validation	5,433	66,431	37,321
	test	6,139	86,661	47,813

Table 12: Validation and test set statistics for all language pairs.