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

**Anonymous ACL submission**

## Abstract

In Machine Translation (MT) research, we often come across recipes recommending a set of fixed hyperparameters to train segmentation models to segment words regardless of the amount of text or language pair involved. Although a fixed hyperparameter for the word segmentation model can reduce training resource overhead, we find that using the same number of merge operations (NMO) on both source and target languages - *symmetric* Byte Pair Encoding (BPE), for different language pairs and text sizes does not guarantee optimal Machine Translation system performance. In this work, we explore and identify BPE segmentation recipes across various data sizes and language pairs to obtain optimal performance. We find that using *asymmetric* BPE improves results compared to symmetric BPE, particularly in low-resource scenarios (50K, 100K, 500K) by (5.32,4.46,0.7) CHRF++ scores (with $p < 0.05$) on average for English-Hindi. We further validate our findings on the other six pairs, English↔Telugu, Shona, Norwegian, Kyrgyz, Hausa and Inuktitut, to show the consistency of this work. A statistically significant improvement is observed using asymmetric BPE configurations in 10 of 12 systems when comparing symmetric BPE configurations. Our findings indicate that using a high NMO for the source (*4K* to *32K*) and a low NMO (*0.5K* to *2K*) provides optimal results, particularly in low-resource contexts.

## 1 Introduction

There have been commendable and considerable efforts in the recent past to bring low resource language pairs into the fold of Neural Machine Translation paradigm (e.g. Workshop on Technologies for MT of Low Resource Languages). However, in this process, it is common to apply successful configurations or methodologies, such as hyperparameters for data preprocessing pipelines and neural networks from past work, without sufficient consideration of their suitability for the specific language pair in question. For example, if we take a preprocessing step, such as word segmentation, it plays a vital role by dividing words into "subwords" to improve learning. In machine translation, word segmentation/subword tokenisation reduces vocabulary size, manages rare and unknown words, and improves MT performance. Notable methods 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 the data by *merging* frequent character pairs into new symbols (Gage, 1994) and thus the resultant subword tokenisation model has *number of merge operations* (NMO) as a hyperparameter. The NMO determines the degree of word segmentation: a lower NMO, such as 500 (see Table 1), leads to smaller vocabularies and more segmentation, whereas a higher NMO, such as 32K, creates larger vocabularies with minimal segmentation (segmenting mostly rare or unknown words). More often than not, same number of merge operations are performed on both the source and the target languages. Most of the work using BPE as preprocessing step has taken the same NMO on source and target for various language pairs without taking the language pair into account or the amount of data involved. Recent research highlights the importance of examining BPE hyperparameters in MT systems, particularly in low-resource scenarios (Ding et al., 2019; Abid, 2020). However, there is still uniformity in terms of having the same NMO for both the source and the target (symmetrical BPE) (Huck et al., 2017; Ortega et al., 2020; Lankford et al., 2021; Domingo et al., 2023; Lee et al., 2024) and very little exploration has been done on asymmetrical BPE for MT. Work like Ngo Ho and Yvon (2021) explored asymmetric BPE for alignment between languages, but not for MT.

In the context of MT, we define the "BPE configuration" as $m_1\_m_2$ where $m_1$ and $m_2$ are the number of merge operations (NMO) for the source and target languages, respectively. Our work explores the effects of symmetric and asymmetric BPE configurations for English-Hindi language pair in different dataset availability scenarios and shows that we achieve best performance using asymmetric configuration. We further validate our findings on other six language pairs English $\leftrightarrow$\{Telugu, Shona, Norwegian, Kyrgyz, Hausa, and Inuktitut\} across different language families to show the consistency of this work. These languages were chosen to represent different typology and morphological structures. We find that, depending on data availability and the language pair involved, the optimal BPE configuration is likely to be asymmetric. Specifically, for low-resource scenarios, we find that, irrespective of the language pair, the optimal configuration should have *4K* to *32K* NMO on the source side and *500* to *2K* on the target.

Section 2 summarises 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. Section 5 evaluates the setup for other language pairs in low resource context, concluding our observations in Section 6.

## 2 Related Work - Symmetrical BPE

In quite a lot of work on BPE to find the optimal segmentation of sub-words for low to high resource settings, the recipe/BPE configuration of keeping NMO the same for both source and target is prevalent. Ding et al. (2019) showed that in low-resource settings with transformer architectures, using 0-4000 (4K) merge operations is optimal, as traditional 32K operations can reduce BLEU scores by up to 4 points. Abid (2020) found lower merge operations are better for English, Egyptian, and Levantine in low-resource scenarios. Domingo et al. (2023) revealed that the performance of the system varies with tokenizers in the target language. There is a class of works which modify the segmentation strategy based on the typological and morphological properties of one or both of the languages in the pair. Ortega et al. (2020) improved BPE by restricting segmentation based on token suffixes in polysynthetic languages. Lankford et al. (2021) highlighted the importance of NMO in low-resource settings for MT performance affecting English-to-Irish systems. Lee et al. (2024) addressed BPE's over-segmentation issue in morphologically rich languages like Korean by incorporating longer words from monolingual corpora, linking difficulties to morphological complexity. However, there also has been works which have shown potential directions to find different subword tokenising strategies. Work like Huck et al. (2017) and Poncelas et al. (2020) used cascading segmentation strategies. Xu et al. (2021) applied Marginal Utility to select appropriate vocabularies (VOLT), which refined segmentation methods and further improved results. Ngo Ho and Yvon (2021) used various BPE NMO in source and target languages independently to improve word alignment between language pairs, but this was not extended to check the performance in the MT system. Unfortunately, we did not find any study that explored asymmetric BPE configurations in diverse resource availability scenario as presented here.

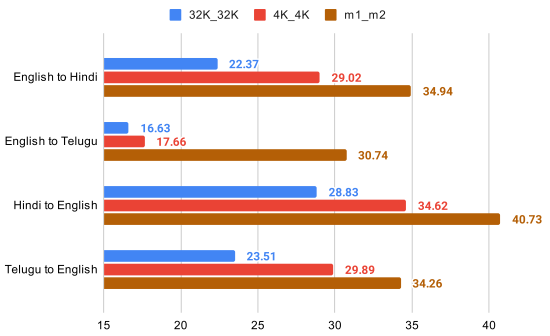



Figure 1: CHRF++ Scores for Symmetrical BPE (32K,4K) vs Asymmetrical BPE (m1 $\neq$ m2)

## 3 Exploring Asymmetrical BPE

In practice, for a BPE configuration $m_1\_m_2$, the values of $m_1$ and $m_2$ are the same and the NMO for both the source and target range from *8K* to *40K* (Wu, 2016; Denkowski and Neubig, 2017; Cherry et al., 2018; Renduchintala et al., 2019), but Ding et al. (2019); Dewangan et al. (2021) found that these are not ideal for low-resource language pairs. Ding et al. (2019) observed $m_1 = m_2$ <=*4K* NMO is better for low-resource settings than *32K*. This finding aligns with our experiments on 0.1 million sentence pairs (English $\leftrightarrow$ {Hindi, Telugu}), showing better performance with *4K* NMO (Fig-

| | |
|---|---|
| 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: Different NMO Effect

ure 1). Dewangan et al. (2021) found that identical BPE configurations do not guarantee similar performance across different language pairs, as demonstrated by the English-Hindi vs. English-Telugu comparisons (Figure 1) for *4K* NMO.

To choose NMO, one should consider datasize and language pair as nuanced BPE strategies have benefitted morphologically complex languages (Ortega et al., 2020; Mujadia and Sharma, 2021). We examine symmetrical BPE configurations that employ the same NMO for both source and target languages, and investigate alternative BPE configurations by independently altering the NMO $m_1$ and $m_2$ for the source and target languages in English-Hindi, examining datasets ranging from low resource (50K) to high resource (8 million). We find that this approach delivered better results (Figure 1) in low-resource contexts. Our extensive experiments with English-Hindi, evaluated on FLORES (Goyal et al., 2022), show that this kind of atypical BPE configuration yields better performance. We further validate the exploration of BPE configurations for optimal results by extending these experiments to English $\leftrightarrow$\{Telugu, Shona, Norwegian, Kyrgyz, Hausa, and Inuktitut\}.

Our work yields compelling results in favour of optimising NMO based on training data size and language pair. Figure 2 offers a conceptual overview of the **optimal ranges** of the **BPE configuration** derived from our experiments in **English-Hindi**, as we move from low to high-resource settings. In this context, "ranges" denote the spectrum of NMO values that serve as hyperparameters for the source and target subword tokenization models in word segmentation. The performance gap between the best and symmetrical BPE systems is depicted by varying shades of green, with the most significant improvement seen in low-resource settings (darker shade of green). As the dataset size grows (for medium to high resource contexts), the performance variance between different configurations decreases (lighter shade of green).

## 4 Evaluation on English $\iff$ Hindi

We explore BPE configurations with the Samanantar dataset (Ramesh et al., 2022) for English-Hindi[1], with 8 million parallel sentences. English text is tokenised, normalised and lowercase using scripts from Moses, while for the preprocessing of the Hindi text we use the Indic NLP library (Kunchukuttan, 2020). We simulate various training set size scenarios by grouping sentences based on English sentence length (Table 2) and randomly sample datasets of sizes 0.05 million (M), 0.1M, 0.5M, 1M, 4M and 8M, maintaining sentence length propotion. For each language and dataset size, the BPE tokenizer is trained on eight different NMOs - *0.5K*, *1K*, *2K*, *4K*, *8K*, *16K*, *25K* and *32K*. After subword segmentation, all possible BPE configurations ($src_{500}$_$tgt_{500}$, $src_{500}$_$tgt_{1000}$, $src_{500}$_$tgt_{2000}$, etc.) are trained with Transformer architecture (Vaswani et al., 2017) using Fairseq (Ott et al., 2019)[2]. Training a single BPE configuration $m_1$_$m_2$ for all dataset sizes required an average of 1040 GPU hours using a 1080 Ti. This results in 64 system configurations trained per language direction for 6 datasets, totalling 768 systems (64×6×2), and resultant systems are evaluated on FLORES (Goyal et al., 2022) Dataset using CHRF++ (Popović, 2015) to assess the impacts of the BPE configurations.

Figures 3 present the performance of all configurations for English $\iff$ Hindi systems in a low resource scenario (for data set sizes of 0.05M, 0.1M and 0.5M). And Figures 4 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 3 for low-resource environments (0.05M, 0.1M and 0.5M) systems, as noted by (Ding et al., 2019), the use of symmetric BPE

[1]English Hindi bitext under the created directory of version 2 of Samanantar is used.

[2]Transformer Base Architecture

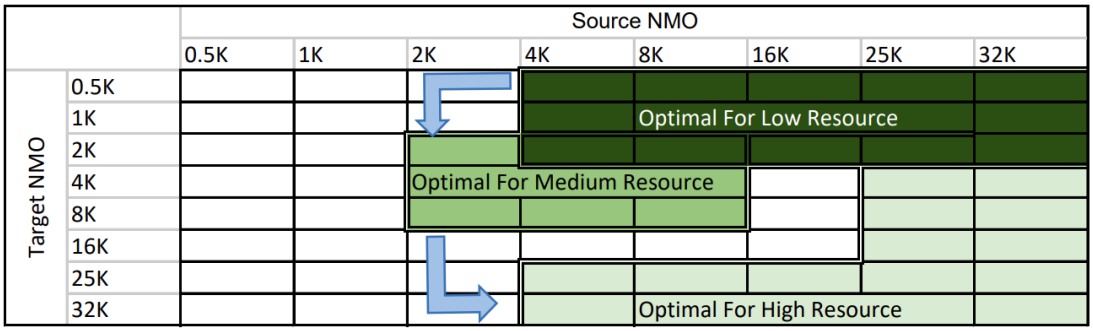



Figure 2: Changes in Optimal BPE Configuration from Low- to High-Resource Settings

| 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

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.

In medium resource settings (1M), optimal configuration for both the source and target NMO *moves* towards *medium* (*2K* - *8K*) range of values, with relatively small score variations across different configurations ≈ 3 CHRF++ between extremes. In contrast, among configurations with low resources, the best system outperforms the weakest system by a margin of ≈ 15 CHRF++ scores and the best symmetric BPE configuration by ≈ 5 CHRF++ scores. For the high-resource scenario, the score disparity between the most effective and least effective configuration is minimal (< 2 CHRF++), with the most effective system employing a 32K NMO for the target. This pattern of shifting optimal BPE configurations across the size of the dataset is also evident if we consider only the last point (*32K* on x-axis) in all subplots in Figures 3 and 4. This highlights that modifying the NMO on the target side, especially in a low resource scenario, plays a more vital role in determining the optimal BPE configuration. For example, for English to Hindi MT systems considering all the configurations *32K_x* for all dataset sizes, we see a gradual change in the configurations of the best performing systems starting with *500* NMO for 0.05M to *32K* NMO for 8M.

We also see that as far as symmetric BPE configurations are concerned, for low resource MT systems it under performs compared to the asymmetric configuration. But as we move to larger dataset size, symmetrical configurations start to perform comparable to asymmetrical configurations. However, using BPE configurations in an asymmetric fashion has yielded a statistically significant improvement in a low-resource environment.

We also compare our systems using optimal BPE configurations with VOLT (Xu et al., 2021)[3]. Figure 5 shows the comparison of the CHRF++ scores between the systems using VOLT tokenisation, optimal BPE, and baseline configurations[4]. Systems using optimal BPE settings surpass those tokenised with VOLT in all dataset sizes, especially for low resource setting with statistical significant ($p<0.05$).

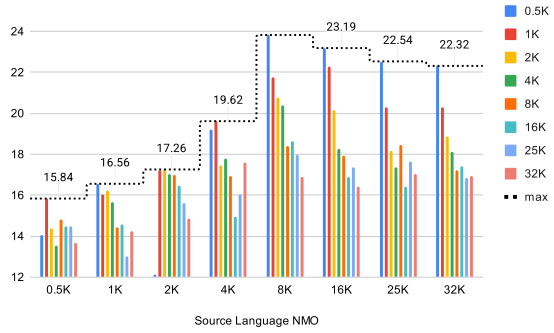
## 5 Exploring Asymmetrical BPE Configurations for other language pairs

To confirm the effectiveness of asymmetric BPE configurations to improve machine translation (MT), we extended our work with atypical BPE configurations using 0.1 million sentence pairs in English ↔ {Telugu, Shona, Norwegian, Kyrgyz, Hausa and Inuktitut}[5]. The English-Telugu

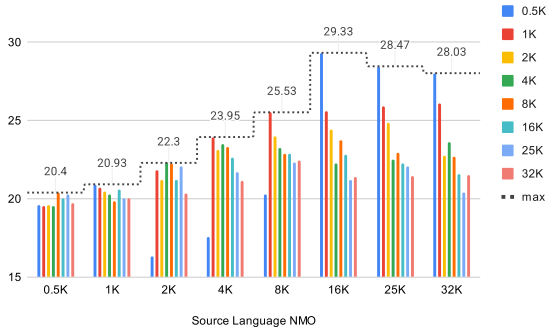
[3]using hyperparameters specified in their work

[4]We selected the top performing BPE configuration as baseline configuration where source NMO = target NMO

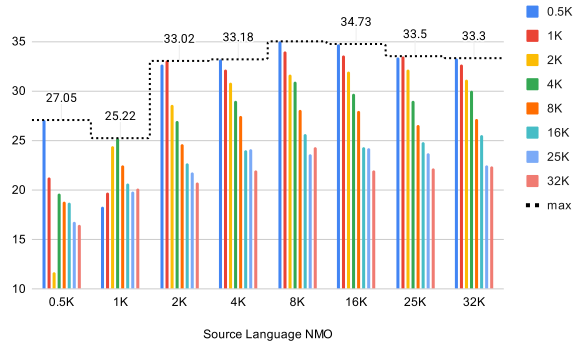
[5]Sampled in a manner similar to English-Hindi sampling



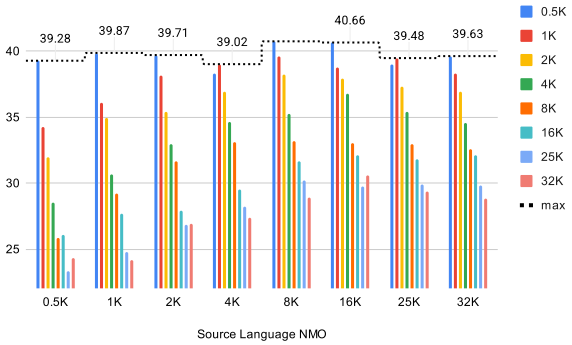

(a) 0.05 Million English to Hindi



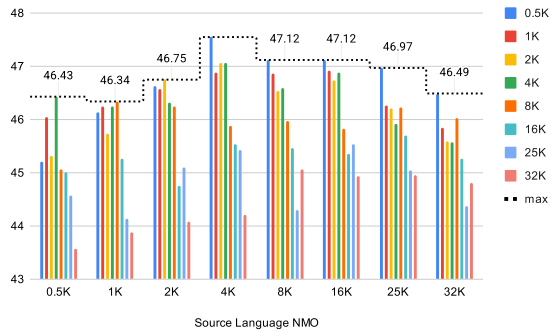

(b) 0.05 Million Hindi to English



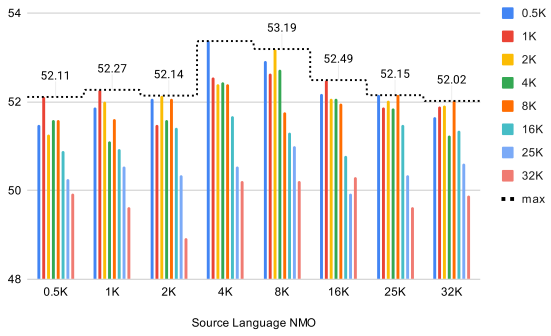

(c) 0.1 Million English to Hindi




(d) 0.1 Million Hindi to English

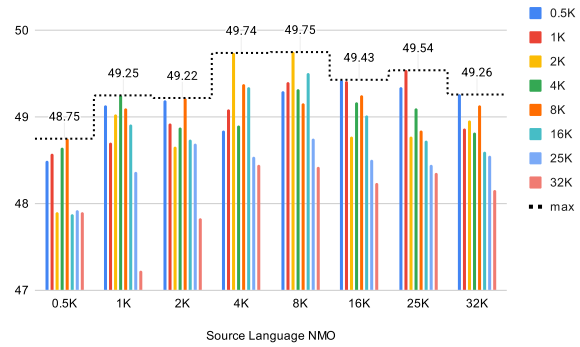



(e) 0.5 Million English to Hindi

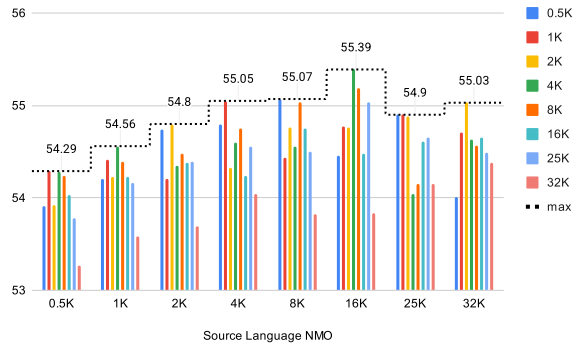



(f) 0.5 Million Hindi to English

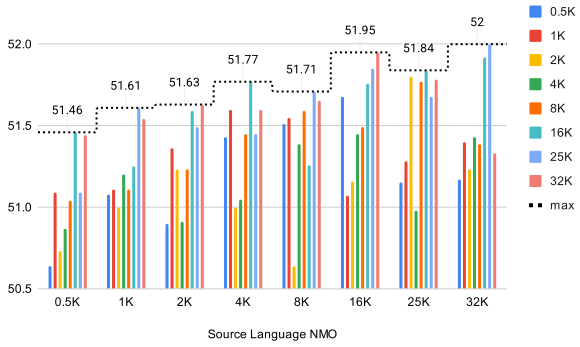
Figure 3: Evaluation of English ↔ Hindi MT Systems for 0.05M, 0.1M and 0.5M dataset sizes on FLORES, x-axis is source NMO and y-axis is CHRF++ scores



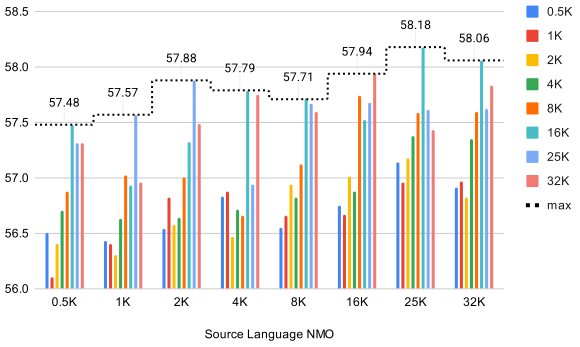

(a) 1 Million English to Hindi



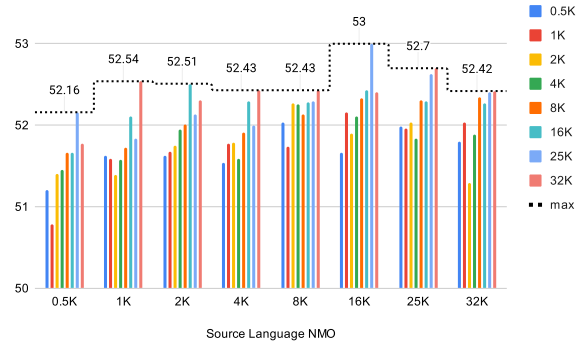

(b) 1 Million Hindi to English



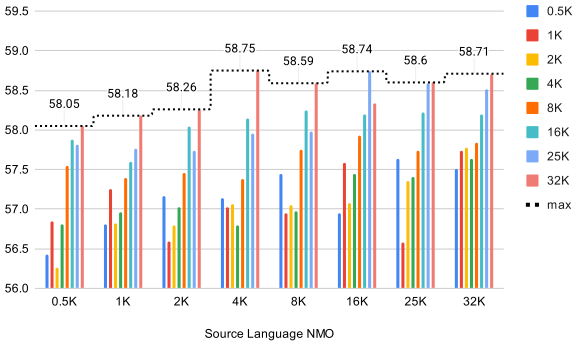

(c) 4 Million English to Hindi




(d) 4 Million Hindi to English




(e) 8 Million English to Hindi




(f) 8 Million Hindi to English

Figure 4: 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

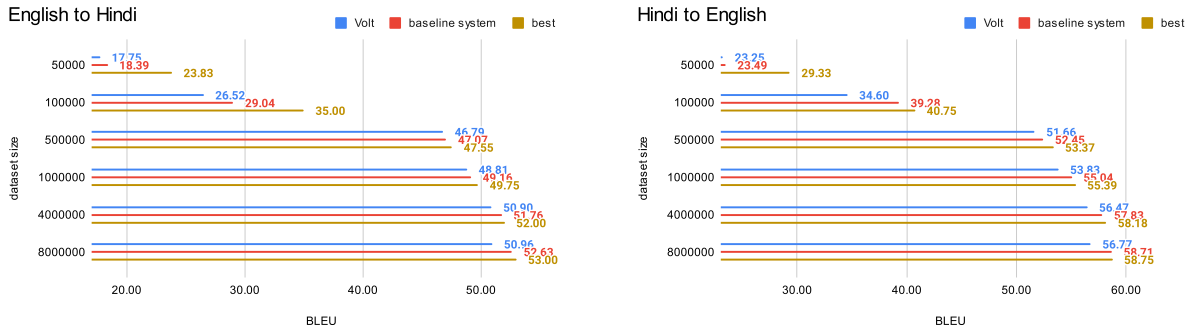



Figure 5: CHRF++ score comparison of Asymmetric BPE with VOLT

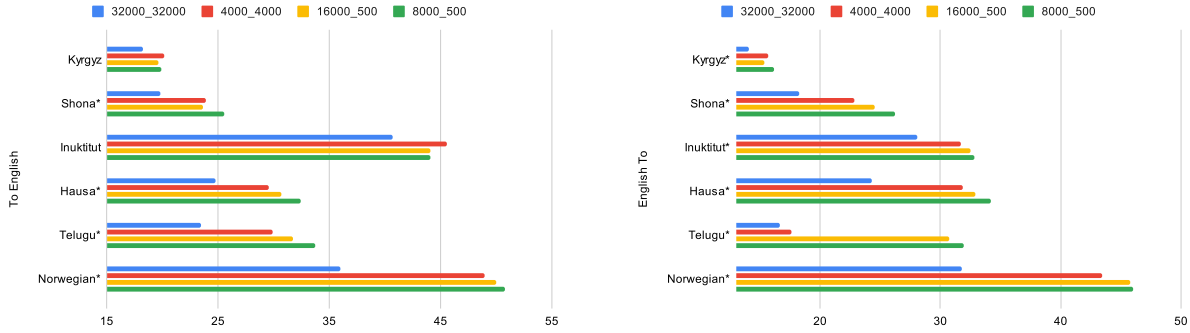



Figure 6: CHRF++ scores improvement with asymmetrical over symmetrical BPE

pair is taken from Ramesh et al. (2022), English-{Shona, Norwegian, Kyrgyz, Hausa} from Gowda et al. (2021) and English-Inuktitut from Joanis et al. (2020). These language pairs are intended to examine the impact of both symmetric and asymmetric BPE configurations in a low resource scenario across languages belonging to various language families and exhibiting diverse levels of morphological complexity. Using symmetric configurations (*4K_4K*, *32K_32K*) as baselines and asymmetric configurations (*8K_500*, *16K_500*) based on insights gained from optimal BPE settings in English-Hindi. We perform evaluations on the FLORES test set, with the exception of Inuktitut, which is evaluated using the test set from Joanis et al. (2020). Figure 6 compares the performance of asymmetric and symmetric BPE. Asymmetric BPE improves translation in four of six *X to English* translation systems and for all systems while translating from *English to X* languages. The improvements are statistically significant ($p<0.05$, languages indicated by * in the figure) compared to symmetric BPE, underscoring the importance of investigating BPE configurations beyond standard choices, particularly for low-resource language pairs.

## 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 $n \gg m$ and *n*, *m* represent NMOs for source and target respectively.

In conclusion, this study underscores the importance of moving beyond the default segmentation options typically used in machine translation, especially when working with languages that have limited data available. In cases with medium- to high-availability of training data, symmetric Byte Pair Encoding (BPE) configurations might be adequate. However, their effectiveness tends to diminish when applied to low-resource scenarios. By utilising configurations with larger number of merger operations (NMO) for the source language and a

smaller NMO for the target language, significant improvements in translation quality can be realised. These asymmetric BPE configurations prove to be consistently effective across various language families, regardless of their differing morphological complexities, thereby emphasising the necessity for specifically tailored configurations to optimise translation results in a low resource scenario.

## Limitation

This study is restricted by the computationally expensive resources required to analyse all BPE configurations for each language pair. However, the findings demonstrate that particular configuration ranges can consistently elevate translation quality, especially in low-resource setups, thus reducing the search grid by significant margin.

## 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.

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.

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.

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.

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.

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.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

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.

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.