BYTF: How Good Are Byte Level N-Gram F-Scores for Automatic Machine Translation Evaluation?

Raj Dabre **Hour Kaing** **Haiyue Song**
National Institute of Information and Communications Technology (NICT), Japan
{raj.dabre, hour_kaing, haiyue.song}@nict.go.jp

Abstract

CHRF and CHRF++ have become the preferred metrics over BLEU for automatic n-gram evaluation of machine translation, as they leverage character-level n-gram overlaps, which achieve better correlations with human judgments for translating into morphologically rich languages. Building on this insight, we observed that bytes capture finer, sub-character-level structures in non-Latin languages. To this end, we propose BYTF to capture sub-character-level information through byte-level n-gram overlaps. Furthermore, we augment it to BYTF+ and BYTF++ where we consider character and word n-gram backoffs. On machine translation metric meta-evaluation datasets from English into 5 Indian languages, Chinese and Japanese, we show that BYTF and its variants are comparable or significantly better compared to CHRF and CHRF++ with human judgments at the segment level. We often observe that backing off to characters and words for BYTF and to words for CHRF does not have the highest correlation with humans. Furthermore, we also observe that using fixed n-gram values often leads to scores having poorer correlations with humans, indicating the need for well-tuned n-gram metrics for efficacy.[1]

1 Introduction

Recently, CHRFand CHRF++ (Popović, 2015, 2017) have become the preferred metrics for automatic n-gram evaluation of machine translation (MT) (Robinson et al., 2024; J et al., 2024; Gala et al., 2023). Compared to BLEU (Papineni et al., 2002), they focus on fine-grained character-level n-grams. As a result, they appear to have better correlations with human judgments for translating into morphologically rich languages.

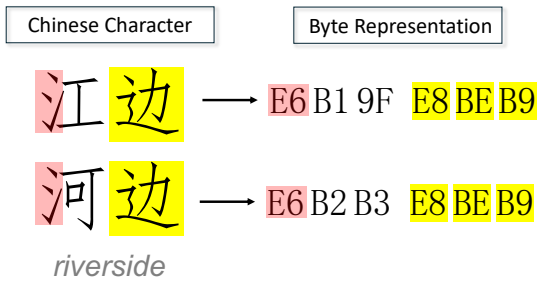

Figure 1: BYTF captures not only character-level similarity but also sub-character-level (named radical that usually conveys the meaning of a Chinese character) overlap.

However, for non-Latin languages with sub-character structures, as shown in Figure 1 for Chinese, we can go one step further to evaluate the sub-character-level structures, which are usually represented by bytes. This applies to a wide range of languages such as Japanese and Indian languages. To this end, we propose BYTF, in which we consider byte-level n-grams instead of character-level n-grams that can be implemented with a single line code change. Experimental results on WMT and Indian MT meta-evaluation datasets show that BYTF has a higher correlation (Pearson and Kendall Tau) with human judgments at the segment level compared to CHRF. We further extend BYTF to BYTF+/BYTF++ where we incorporate character- and word-level n-gram backoffs to show that this further enhances correlations.

Our contributions are as follows:

1. Novel metric: We propose BYTF a complete version of CHRF, to capture sub-character-level structural similarity for many non-Latin languages.

2. N-gram backoffs: We extend BYTF to BYTF+ and BYTF++ to incorporate character- and word-level n-gram backoffs.

3. Extensive meta-evaluation: Experimental results on 10 languages show comparable or higher

[1] https://github.com/shyyhs/bytF

Proceedings of Machine Translation Summit XX Volume 1, pages 378–387, June 23-27, 2025

Pearson and Kendall Tau correlations with human evaluations compared to BLEU, CHRF, and CHRF++.

4. Tuning is important: We show that the default choices of n-gram are not always optimal and should ideally be tuned based on the language pair.

2 Related Work

We introduce commonly used MT evaluation metrics in Section 2.1 and the recent trend of byte-level methods in Section 2.2.

2.1 Evaluation Metrics

BLEU (Papineni et al., 2002) is a long-standing, widely adopted word-level n-gram evaluation metric due to its simplicity:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \quad (1)$$

where p_n is the n-gram word-level precision, w_n is a weight smoothing factor and BP represents the brevity penalty (Post, 2018). There are two limitations of BLEU. First, it requires word boundary information, but many languages do not have it. For languages without explicit word boundaries—such as Japanese and Chinese, we have to apply an additional word segmenter, such as Juman++ (Tolmachev et al., 2018) or the Stanford Chinese word segmenter (Wang et al., 2014) to pre-process them. However, for low-resource languages such as Burmese, we do not even have high-quality word segmenters. Another limitation is that BLEU overlooks fine-grained character-level overlaps. As a result, it does not capture the difference between a critical translation error and a minor typographical or morphological variation.

CHRF (Popović, 2015) relies on character-level n-gram precision and recall, whereas CHRF++ (Popović, 2017) uses word-level m-gram backoffs and fine-tunes the hyperparameter n (from 1 to 4) and m (from 1 to 2) to achieve the optimal correlations with human judgments. However, they ignore sub-character-level structures, which are important for non-Latin languages, a gap that we explore.

In contrast to the simplicity of statistical metrics, neural metrics leverage neural models trained to minimize the difference between predicted evaluations and human judgments. BERTScore (Zhang et al., 2020), BLEURT (Sellam et al., 2020), and COMET (Rei et al., 2020) are based on pre-trained models such as BERT (Devlin et al., 2019) or XLM (Conneau et al., 2020). They are then fine-tuned on annotated MT quality evaluation datasets including *Direct Assessments (DA)* (Graham et al., 2013) and *Multidimensional Quality Metrics (MQM)* (Lommel et al., 2014). However, they rely on at least hundreds of annotated samples (Rei et al., 2022), which are hard to obtain for low-resource languages, making them language-specific. We do not compare with them as our goal is not to beat them but to complete CHRF.

2.2 Byte-Level Methods

The byte-level method is a path to language-agnostic NLP. For pre-processing, byte-level BPE (BBPE) (Wang et al., 2019) handles unseen characters in Chinese and Japanese by segmenting them into seen byte-subwords. The ByT5 model (Xue et al., 2021) processes input text as raw UTF-8 bytes, thereby enabling it to handle any language, increasing its robustness to noise, and simplifying the pre-processing pipelines. The byte latent transformer (Pagnoni et al., 2024) is a purely tokenizer-free model that learns from raw byte data. This paper aims to find the missing piece: a byte-level evaluation method.

3 Proposed Methods

This section introduces our proposed BYTF metric and the extended BYTF+ and BYTF++ variants.

3.1 BYTF

We compute the byte-level F-score, BYTF_β, similarly as CHRF, as

$$\text{BYTF}_\beta = (1+\beta^2)\,\frac{\text{BYTP}\cdot\text{BYTR}}{\beta^2\,\text{BYTP}+\text{BYTR}}, \quad (2)$$

where BYTP and BYTR denote the overall byte-level n-gram precision and recall, respectively, which are obtained by averaging the scores over all n-gram orders. For each n (with $n = 1, \ldots, N$), let $\mathcal{G}_n$ be the multiset of all byte n-grams in the candidate text, and let $\text{Count}(g, \cdot)$ denote the number of occurrences of an n-gram g in the candidate or reference text. For each n, we define the n-gram precision and recall as

$$P_n = \frac{\sum_{g\in\mathcal{G}_n} \min\left\{\text{Count}(g,\text{cand}), \text{Count}(g,\text{ref})\right\}}{\sum_{g\in\mathcal{G}_n} \text{Count}(g,\text{cand})}, \quad (3)$$

$$R_n = \frac{\sum_{g\in\mathcal{G}_n} \min\left\{\text{Count}(g,\text{cand}), \text{Count}(g,\text{ref})\right\}}{\sum_{g\in\mathcal{G}_n} \text{Count}(g,\text{ref})}. \quad (4)$$

The overall byte-level precision and recall are computed as the arithmetic mean over all n-gram orders:

$$\text{BYTP} = \frac{1}{N}\sum_{n=1}^{N} P_n, \quad \text{BYTR} = \frac{1}{N}\sum_{n=1}^{N} R_n. \quad (5)$$

The parameter β assigns β times more importance to recall than to precision. In our experiments, we set $\beta = 1$ so that they are equally weighted. To capture more input details while tolerating some redundancy, one can consider using $\beta > 1$ to favor recall over precision.

Note that for languages using the Roman alphabet such as English, BYTF reduces to CHRF, with only minor differences (e.g., accent decomposition in languages like Finnish).

3.2 BYTF+ and BYTF++

BYTF does not leverage character or word-level information. Inspired by CHRF++ (Popović, 2017), we propose BYTF+, which integrates byte-level n-grams and character-level m-grams, and BYTF++, which further integrates word-level l-grams, within the same F-score framework.

We define the extended metrics as

$$\text{BYTF+/++}_\beta = (1+\beta^2)\frac{\text{BYTP+/++} \cdot \text{BYTR+/++}}{\beta^2\,\text{BYTP+/++} \;+\; \text{BYTR+/++}}. \quad (6)$$

where BYTP+/++ and BYTR+/++ denote the overall precision and recall computed by averaging the n-gram byte-level scores, m-gram character-level scores (and, l-gram word-level scores for BYTF++) statistics.

4 Experimental Setup

We describe our datasets, language pairs and meta-evaluation setup.

4.1 Datasets and Language Pairs

We evaluate our n-gram metrics on the IndicMT Eval (Sai B et al., 2023) and WMT2017-2022 (Bojar et al., 2017, 2018; Barrault et al., 2019, 2020; Akhbardeh et al., 2021; Kocmi et al., 2022) datasets. The IndicMT Eval dataset contains MQM scores, and the WMT dataset contains DA scores, both of which are annotated by professional translators or raters. The languages included in this study comprise six Indian languages—Hindi (Hin), Gujarati (Guj), Malayalam (Mal), Tamil (Tam), Marathi (Mar), and Bengali (Ben)—as well as two East Asian target languages, Japanese (Jpn) and Chinese (Zho). Their source language is primarily English, except for Ben↔Hin. The WMT datasets we used primarily belong to the news domain (News*), except for Ben↔Hin, which is sourced from Wikimedia (Wiki21).

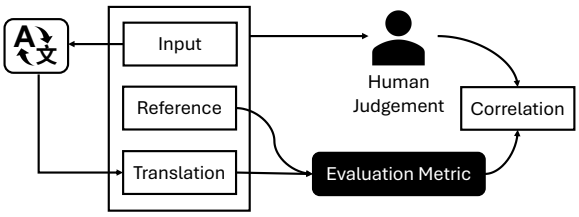

Figure 2: The flowchart of meta evaluation. We calculate the correlation between human judgment and our evaluation metrics.

4.2 Meta Evaluation

To assess the reliability of evaluation metrics, meta evaluation is commonly used to measure the correlation between an evaluation metric and human judgment, as illustrated in Figure 2. There are two levels of meta evaluation: segment-level and system-level. Segment-level correlation evaluates how well a metric aligns with human scores on individual translations, while system-level correlation assesses its effectiveness in ranking entire systems based on their aggregated performance. In this work, we evaluate correlation only at the segment level.

For correlation measurement, we employ Pearson correlation and Kendall's Tau just as previous works (Sai B et al., 2023; Singh et al., 2024). Pearson correlation measures the linear relationship between two sets of numerical values, making it useful for evaluating metrics that predict absolute human scores. In contrast, Kendall's Tau measures ordinal association, which is particularly valuable in ranking-based evaluations where the relative ordering of scores is more important than their exact values.

5 Results

We now describe our results to determine whether byte-based metrics can be used to replace character-based metrics. Tables 1 and 2 provide the Pearson and Kendall Tau correlations with human scores, along with the winning metric and the optimal configuration. For BYTF and its variants, the configuration is given as a tuple ***a, b, c*** for byte, character and word n-gram values, respectively. Similar for

Direction	Pearson Correlation Coefficient			Kendall's Tau		
	BLEU	CHRF	BYTF	BLEU	CHRF	BYTF
Eng-Hin (IndicMT)	0.2600	$0.2918_{12,0}$	$\mathbf{0.3462}_{20,0,0\dagger}$	0.1725	$0.2012_{9,0}$	$\mathbf{0.2631}_{20,0,0\dagger}$
Eng-Guj (IndicMT)	0.2978	$0.4269_{2,0}$	$\mathbf{0.4725}_{6,0,0\dagger}$	0.2472	$0.2857_{6,0\ddagger}$	$\mathbf{0.3284}_{13,0,0\dagger}$
Eng-Mal (IndicMT)	0.2793	$0.4175_{6,0}$	$\mathbf{0.4426}_{20,0,0\dagger}$	0.3076	$0.3463_{6,2\ddagger}$	$\mathbf{0.3746}_{20,0,0\dagger}$
Eng-Tam (IndicMT)	0.2647	$0.3668_{6,0}$	$\mathbf{0.4043}_{20,0,0\dagger}$	0.2069	$0.2579_{6,0}$	$\mathbf{0.2896}_{20,0,0\dagger}$
Eng-Mar (IndicMT)	0.1954	$0.2656_{4,2\ddagger}$	$\mathbf{0.3327}_{13,0,0\dagger}$	0.1468	$0.1709_{4,2\ddagger}$	$\mathbf{0.2268}_{13,0,0\dagger}$
Ben-Hin (Wiki21)	0.0901	$0.1156_{2,0}$	$\mathbf{0.1165}_{6,4,0\dagger}$	0.0563	$0.0669_{6,0}$	$\mathbf{0.0673}_{16,9,0\dagger}$
Hin-Ben (Wiki21)	0.1116	$0.1915_{2,2}$	$\mathbf{0.1974}_{2,2,0\dagger}$	0.0956	$0.1144_{6,4}$	$\mathbf{0.1162}_{16,0,0\dagger}$
Eng-Guj (News19)	0.3992	$0.4760_{6,2\ddagger}$	$\mathbf{0.4774}_{16,6,2\ddagger}$	0.2845	$0.3366_{4,0}$	$\mathbf{0.3377}_{6,6,2\dagger}$

Table 1: Translation Performance Metrics for Indian languages. † underneath BYTF denotes BYTF+. ‡ underneath CHRF and BYTF denotes CHRF++ and BYTF++ respectively.

Direction	Pearson Correlation Coefficient			Kendall's Tau		
	BLEU	CHRF	BYTF	BLEU	CHRF	BYTF
Eng-Jpn (News20)	0.3615	$0.4144_{2,2\ddagger}$	$\mathbf{0.4213}_{2,2,2\ddagger}$	0.2509	$\mathbf{0.2769}_{2,2\ddagger}$	$0.2576_{6,2,0\dagger}$
Eng-Jpn (News21)	0.2645	$0.3157_{2,2\ddagger}$	$\mathbf{0.3189}_{2,2,2\ddagger}$	0.1740	$\mathbf{0.1953}_{2,2\ddagger}$	$0.1895_{2,2,2\ddagger}$
Eng-Zho (News17)	0.4197	$\mathbf{0.4717}_{2,2\ddagger}$	$0.4708_{6,2,2\ddagger}$	0.2951	$\mathbf{0.3203}_{2,2\ddagger}$	$0.3196_{6,2,2\ddagger}$
Eng-Zho (News18)	0.3101	$0.3492_{2,2\ddagger}$	$\mathbf{0.3545}_{2,2,2\ddagger}$	0.2209	$0.2424_{2,2\ddagger}$	$\mathbf{0.2444}_{2,2,2\ddagger}$
Eng-Zho (News19)	0.2262	$0.2481_{2,2\ddagger}$	$\mathbf{0.2503}_{2,2,2\ddagger}$	0.1350	$\mathbf{0.1491}_{2,2\ddagger}$	$0.1489_{6,2,2\ddagger}$
Eng-Zho (News20)	0.2672	$0.3097_{2,2\ddagger}$	$\mathbf{0.3147}_{2,2,2\ddagger}$	0.1720	$0.1954_{2,2\ddagger}$	$\mathbf{0.1962}_{2,2,2\ddagger}$
Eng-Zho (News21)	0.1703	$\mathbf{0.1834}_{2,2\ddagger}$	$0.1820_{6,2,2\ddagger}$	0.1050	$\mathbf{0.1149}_{2,2\ddagger}$	$0.1137_{6,2,2\ddagger}$

Table 2: Translation Performance Metrics for Eng-Jpn and Eng-Zho. † underneath BYTF denotes BYTF+. ‡ underneath CHRF and BYTF denotes CHRF++ and BYTF++ respectively.

CHRF and CHRF++, the configuration is ***a, b*** for character and word n-gram values respectively.

5.1 Byte Based Metrics Are Competitive

As shown in Tables 1 and 2, BLEU consistently has the lowest correlation. This aligns with previous findings that BLEU struggles to capture translation quality in non-Latin and low-resource languages (Kocmi et al., 2021). Its reliance on exact word matching makes it less effective for languages with flexible word order and rich inflections, such as Indian and East Asian languages. Kocmi et al. (2021) suggest using CHRF among string-based metrics for non-Latin languages.

The byte-based metric, BYTF, achieves the highest correlation with human judgments across various language pairs, suggesting that byte-level representations effectively capture essential aspects of translation quality. While CHRF remains competitive in some cases, BYTF operates at a more granular level than characters and words, making it more language-agnostic and a potentially superior alternative to traditional string-based metrics.

5.2 Correlation Improvements Are Domain And Language Pair Specific

The effectiveness of BYTF varies depending on the language pair and domain of the dataset, showing a strong advantage in Indian languages but a competitive performance with CHRF in Japanese and Chinese. Correlation patterns differ depending on the dataset, reinforcing that a single metric may not perform best across all domains (e.g., News vs. IndicMT). This suggests that while byte-level evaluation is effective, its application needs to be carefully adapted to language- and domain-specific characteristics. Future research should explore adaptive evaluation strategies based on the specific characteristics of the dataset.

5.3 The Optimal Metric And Configuration Needs Tuning

The above results present the optimal configuration for BYTF and CHRF, determined based on their correlation with human scores. One key takeaway is that BYTF and CHRF require tuning to achieve their best performance. The n-gram order of bytes,

characters, and words plays a significant role in influencing these correlations. The optimal configuration is language-specific, where the best settings for Indian languages differ from those for Japanese or Chinese, which use distinct scripts or writing systems. Therefore, rather than viewing tuning as a limitation, it should be seen as a necessary step in improving the reliability of automatic metrics.

5.4 Backing Off To Larger Granularities Is Not Always Reliable

The BYTF metric follows a common strategy in evaluation metrics, which involves backing off to larger linguistic units (e.g., moving from byte-level to character-level, and then to word-level evaluation). However, our results suggest that this strategy is not always effective. Specifically, we found that for Indian languages, particularly those in the IndicMT Eval dataset, the backing-off strategy is often unnecessary, as byte-level evaluation alone provides adequate alignment with human judgment. This suggests that, for these languages, smaller linguistic units may be more appropriate or sufficient for capturing translation quality. On the other hand, for languages like Japanese and Chinese, the backing-off strategy remains consistently effective, highlighting the varying effectiveness of this approach depending on the linguistic characteristics of the language in question.

5.5 N-gram Metrics Appear To Have Decreasing Correlation With Humans Over The Years

Our results in Table 2 show that the correlation of n-gram metrics with human judgments decreases over time. This phenomenon can be explained by several key factors: (1) modern neural machine translation systems tend to generate more fluent or natural-sounding translations rather than n-gram matches with a reference translation, and (2) as NMT becomes more fluent and context-aware, human evaluation criteria focus more on overall meaning rather than literal word choices (Barrault et al., 2019), making n-gram metrics less aligned with human judgments. This suggests that while n-gram metrics remain useful for basic assessments, they should be supplemented with more sophisticated semantic-based metrics like COMET (Falcão et al., 2024) to provide a comprehensive evaluation of translation quality.

5.6 Visualizing Impact Of Configuration On Correlations

Figure 3 highlights that the choice of configuration plays a crucial role in the n-gram metrics. BYTF could be highly sensitive to its configuration especially on Hindi, Malayalam, and Tamil, but the variation is more stable on Gujarati and Marathi. A similar tendency can be observed for CHRF but its sensitivity is lower compared to BYTF. These findings further emphasize the importance of per-language tuning to align with human judgment.

We further observe the overall tendency of the optimal configuration for Indic languages in Figure 4. The results show that the configuration is optimal when the orders of character and word are smaller and when the byte order is larger. This suggests that the configuration for the Indic languages should have a larger byte order and smaller character and word order. For example, most Indic languages in Table 1 have an optimal configuration with a byte order of 20 and character and word order of zero. A similar analysis for Japanese and Chinese is provided in Appendix A.

5.7 Recommendations

Based on our findings, we provide the following recommendations for future evaluation:

- **Byte-Based Metrics as Preferred Choice:** Given their strong performance, BYTF should be prioritized over BLEU and CHRF, especially for Indian languages.
- **Configuration Tuning:** Metric configurations should be fine-tuned per language and domain, as the optimal settings vary across Indic, Japanese, and Chinese translations. Backing off to larger granularity is not always reliable.
- **Complementing N-Gram Metrics:** As modern NMT evolves, we recommend supplementing n-gram metrics with semantic-based metrics like COMET.

6 Conclusion

We proposed BYTF, a byte-level n-gram evaluation metric that captures sub-character-level similarities for machine translation. We further augment BYTF with character- and word-level backoffs as BYTF+ and BYTF++. Our experiments show that they achieve higher correlations with human

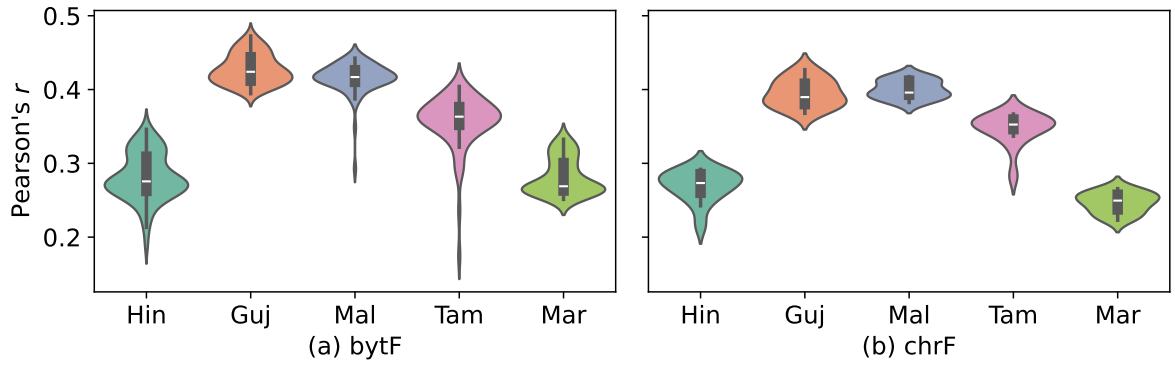

Figure 3: Correlation of various configurations on Indian languages in IndicMT Eval.

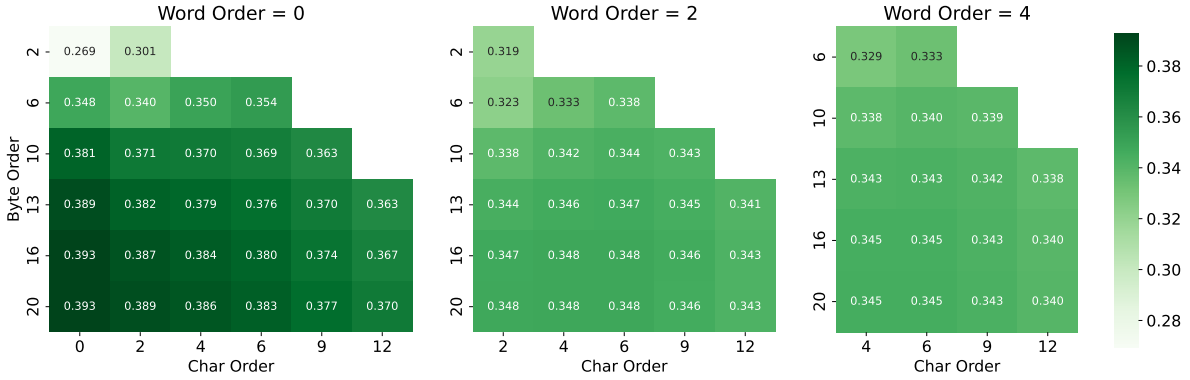

Figure 4: Pearson Correlation in relation between n-gram order of byte, character, and word on IndicMT Eval.

judgments than BLEU and CHRF, though language-specific hyper-parameter tuning is applied. Finally, we recommend (1) avoiding excessive reliance on backing off to larger granularities, as it weakens correlation with human judgment; and (2) complementing n-gram metrics with semantic-based metrics like COMET, as exact n-gram matching may fail to capture high-level semantics.

7 Sustainability Statement

In this work, we are using existing translations, therefore, there is no need to train NMT models or perform any inference. All results are based purely on numerical correlations and were computed using only CPUs, leading to significantly lower energy consumption. This approach is both efficient and environmentally friendly. We believe that our experimental setup used in this study is highly sustainable.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussa, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. ACL.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. 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. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Júlia Falcão, Claudia Borg, Nora Aranberri, and Kurt Abela. 2024. COMET for low-resource machine translation evaluation: A case study of English-Maltese and Spanish-Basque. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3553–3565, Torino, Italia. ELRA and ICCL.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Jaavid J, Raj Dabre, Aswanth M, Jay Gala, Thanmay Jayakumar, Ratish Puduppully, and Anoop Kunchukuttan. 2024. RomanSetu: Efficiently unlocking multilingual capabilities of large language models via Romanization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15593–15615, Bangkok, Thailand. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494.

Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463.

Artidoro Pagnoni, Ram Pasunuru, Pedro Rodriguez, John Nguyen, Benjamin Muller, Margaret Li, Chunting Zhou, Lili Yu, Jason Weston, Luke Zettlemoyer, Gargi Ghosh, Mike Lewis, Ari Holtzman, and Srinivasan Iyer. 2024. Byte latent transformer: Patches scale better than tokens.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. 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.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

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

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

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.

Nathaniel Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, Naome Etori, Vijay Murari Tiyyala, Olanrewaju Samuel, Matthew Stutzman, Bismarck Odoom, Sanjeev Khudanpur, Stephen Richardson, and Kenton Murray. 2024. Kreyòl-MT: Building MT for Latin American, Caribbean and colonial African creole languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3083–3110, Mexico City, Mexico. Association for Computational Linguistics.

Ananya Sai B, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra, and Raj Dabre. 2023. IndicMT eval: A dataset to meta-evaluate machine translation metrics for Indian languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228, Toronto, Canada. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Anushka Singh, Ananya Sai, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, and Mitesh Khapra. 2024. How good is zero-shot MT evaluation for low resource Indian languages? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 640–649, Bangkok, Thailand. Association for Computational Linguistics.

Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. 2018. Juman++: A morphological analysis toolkit for scriptio continua. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 54–59, Brussels, Belgium. Association for Computational Linguistics.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2019. Neural machine translation with byte-level subwords.

Mengqiu Wang, Rob Voigt, and Christopher D. Manning. 2014. Two knives cut better than one: Chinese word segmentation with dual decomposition. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 193–198, Baltimore, Maryland. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2021. Byt5: Towards a token-free future with pre-trained byte-to-byte models.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Visualizing Impact Of Configuration On Correlations On Japanese and Chinese

Figure 5 illustrates how performance varies across different configurations for both Japanese and Chinese. Additionally, we observe that sensitivity is influenced not only by the language but also by the domain, with some domains being more sensitive than others. This reinforces our conclusion about the significance of configuration tuning. Moreover, Figures 6 and 7 demonstrate the general trends of optimal configurations for Japanese and Chinese, respectively.

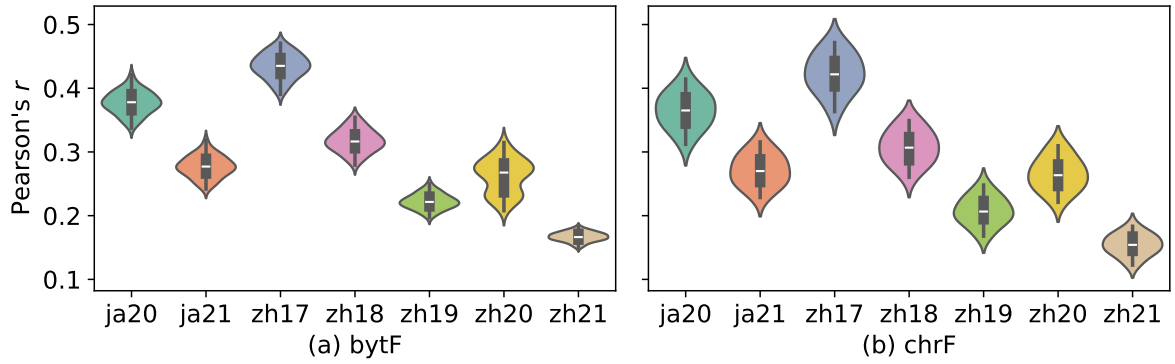

Figure 5: Correlation of various configurations on Japanese and Chinese.

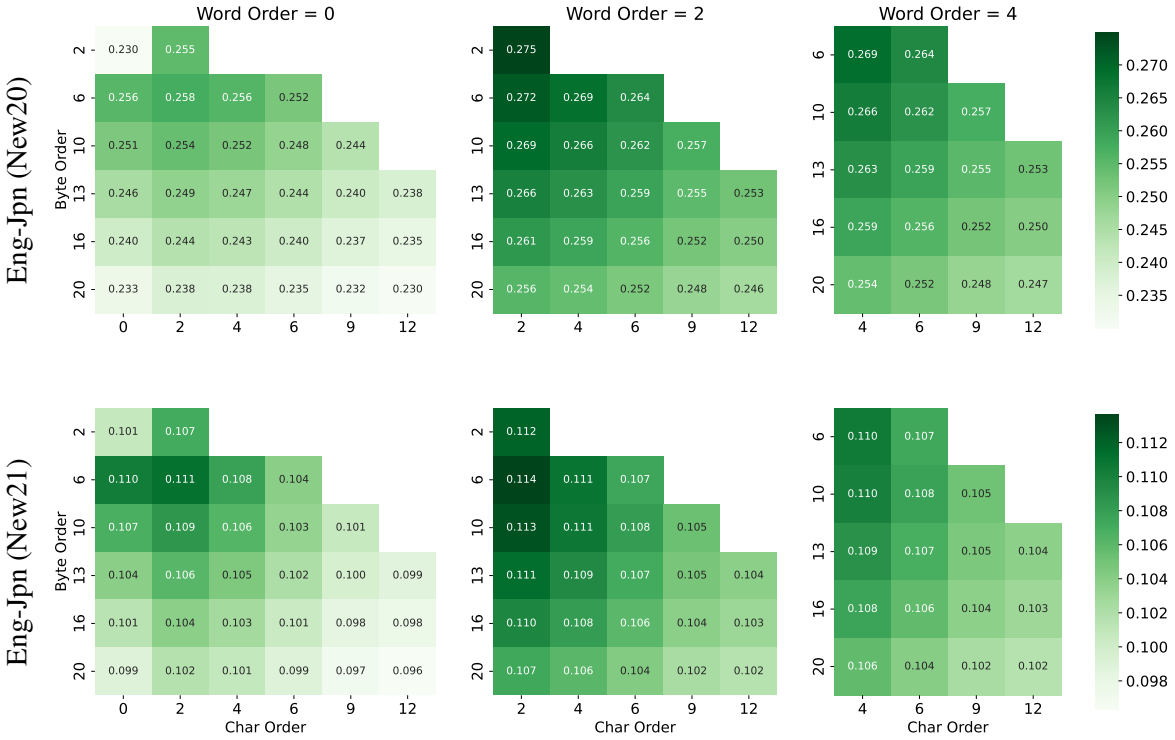

Figure 6: Pearson Correlation in relation between n-gram order of byte, character, and word on Japanese.

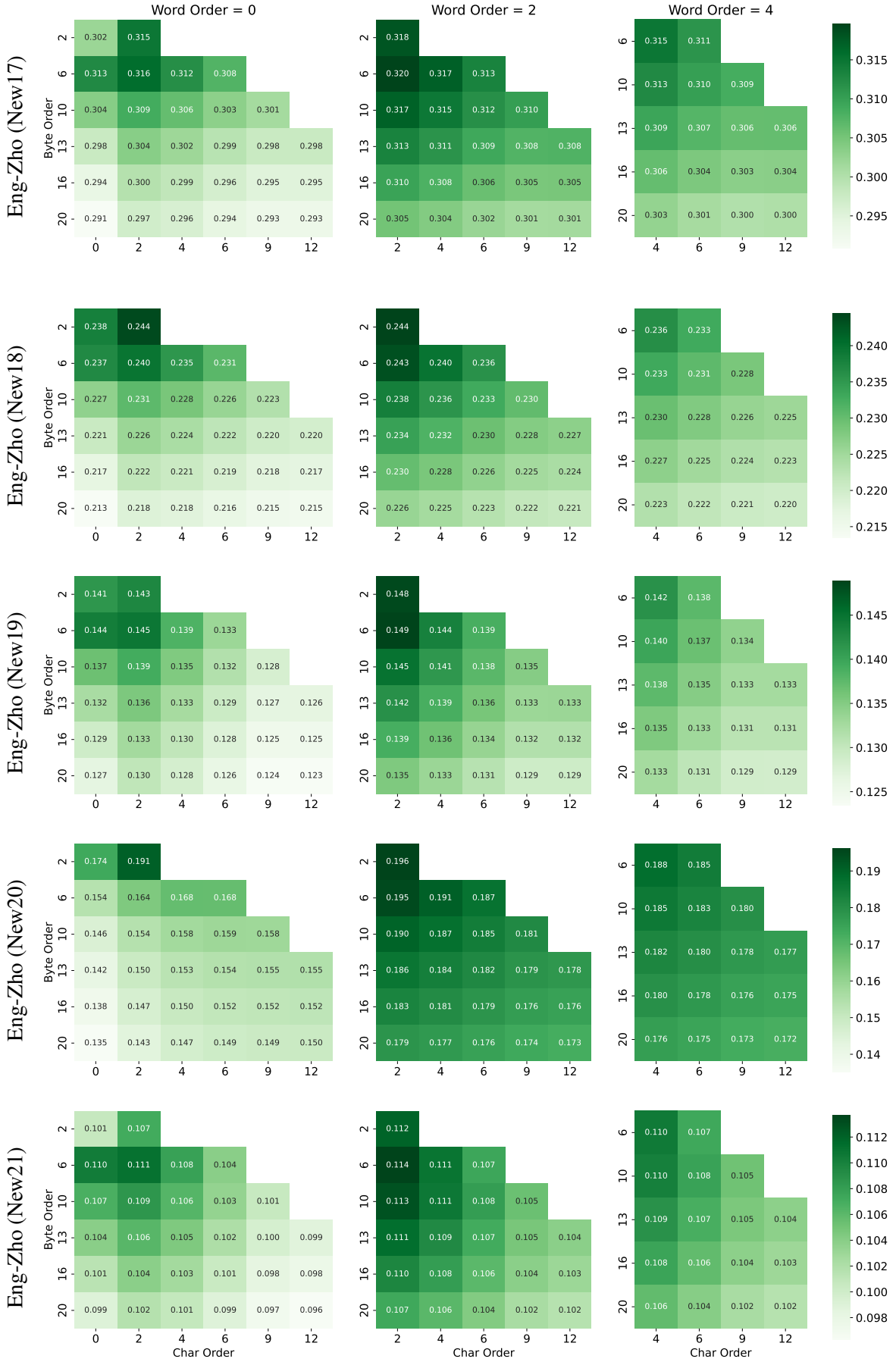

Figure 7: Pearson Correlation in relation between n-gram order of byte, character, and word on Chinese.