BPE

Sangwhan Moon
Google LLC
sangwhan@iki.fi

Tatsuya Hiraoka
MBZUAI
tatsuya.hiraoka@mbzuai.ac.ae

Naoaki Okazaki
Institute of Science Tokyo
okazaki@c.titech.ac.jp

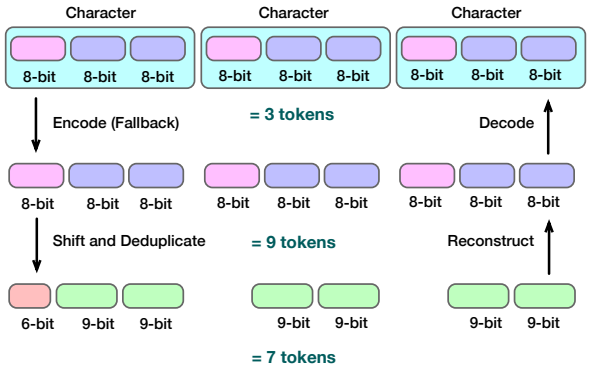

Figure 1:

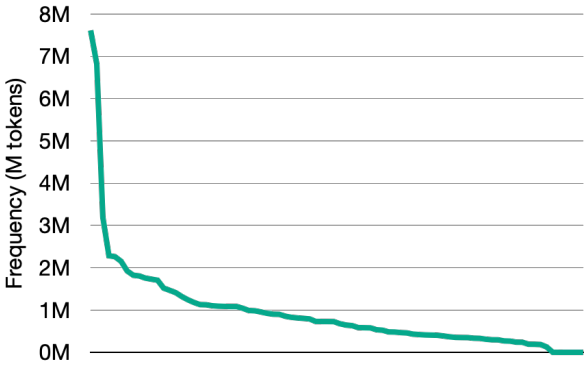

Figure 2: -

Abstract

OOVCJK

1

BPE (Sennrich et al., 2016) NLPBPEBPEOOVOOVUnicodeUTF-8OOV

Unicode

UTF-8

2

2.1 BPE

BPEBPEUnicodeBPE Radford et al. (2019) Wang et al. (2019) CJK

CJKCJK97,68011,172 [1] CJKOOVOOV

2.2

(Zhang and LeCun, 2017; El Boukkouri et al., 2020; Shaham and Levy, 2021; Xue et al., 2022) (Libovický and Fraser, 2020; Gowda and May, 2020; Goldman et al., 2024) (Mielke et al., 2021; Sreedhar et al., 2023) Rust et al. (2021)

LLMLLM (Ahia et al., 2023; Petrov et al., 2024)

[1] Unicode 15.1

2.3

OOV

UnicodeUnicode (Hoffmann et al., 2022)

[2] UTF-8 CJK OOV (Land and Bartolo, 2024)

Zouhar et al. (2023) Rényi V W_v Rényi $H_\alpha(W_v)$

$$H_\alpha(W_v) = \lim_{\alpha' \to \alpha} \frac{1}{1-\alpha'} \log \left(\sum_{w \in V} p_v(w)^{\alpha'} \right) \quad (1)$$

Rényi $E_\alpha(W_v)$ $|V|$

$$E_\alpha(W_v) \approxeq \frac{H_\alpha(W_v)}{\log |V|} \quad (2)$$

Rényi $|V|$ [3] Rényi

3

UTF-8 §3.1 §3.2 UTF-8 §3.3 1

3.1 UTF-8

2 UTF-8 1 UTF-8 E4 ED

UTF-8 2.3 1

[2]

[3] 255

zh-CN	UTF8gbsn	UTF8gbsn	UTF8gbsn
	E4 BC 97	E5 94 A4	E4 BC 97
ja-JP	UTF8min	UTF8min	UTF8min
	E6 A4 9C	E8 AA 8D	E8 A3 81
ko-KR			
	EC B2 A0	EC A0 80	ED 9E 88

Table 1:

		Length			Entropy	
		Byte	Ours	Diff	Byte	Ours
en-zh	T	131M	127M	3.13 %	0.764	0.634
	B	64M	60M	6.41 %	0.586	0.435
en-ja	T	176M	174M	0.83 %	0.818	0.763
	B	41M	40M	3.56 %	0.580	0.417
ja-ko	T	113M	111M	2.21 %	0.498	0.485
	B	59M	56M	4.25 %	0.485	0.446

Table 2: T B

	Decode error		Empty	
	Byte	Ours	Byte	Ours
en-zh	33	14	3156	2545
en-ja	136	0	218	87
ja-ko	1522	121	7	3

Table 3: 5000

3.2

1 UTF-8 [4] **??**

?? jako - `0xEC-0xEF` zh `0xE4-0xE7`

UTF-8

3.3 8

3.3.1

1 logits

`0b111001` `0x39` E5 01 2^8 768 "" $2^{10} - 2^8$ "" 9 256

1 UTF8gbsn

```
E4 BC 97 E5 94 A4 E4 BC 97
```

```
E4 BC 97
```

```
111001 00 10111100 10010111
111001 00 10111100 10010111
111001 00 1011110 0 10010111
```

`p1=0b111001` 6 [5] b_1 b_2 b_3 $\hat{b_1}$ 9 $\hat{b_2}$ $\hat{b_3}$ [6]

$$\hat{b_1} = (b_1 \wedge 127) \gg 2 \quad (3)$$

$$\hat{b_2} = (((b_1 \wedge 3) \ll 7) \vee (((b_2 \wedge 254) \gg 1) \quad (4)$$

$$\hat{b_3} = ((b_2 \wedge 1) \ll 8) \vee b_3 \quad (5)$$

```
p1 5E 97 p1 CA A4 p1 5E 97
```

[7] `p1` 6 0x39

```
p1 5E 97 CA A4 5E 97
```

22.22 %66 p $_n$ UTF8min `E8 AA 8D` UTF8gbsn UTF8min

```
p1 5E 97 CA A4 5E 97 p2 55 8D
```

- 16.66 %

Unicode9

```
p1 5E 97 p1 CA A4 p1 5E 97 p2 55 8D
```

UTF-8

$$b_1 = (\hat{b_1} \ll 2) \vee (\hat{b_2} \gg 7) \quad (6)$$

$$b_2 = ((\hat{b_2} \wedge 127) << 1) \vee (\hat{b_3} >> 8) \quad (7)$$

$$b_3 = \hat{b_3} \wedge 255 \quad (8)$$

```
E4 BB BD E5 81 87 E7 AE 80 E8 94 B5
```

CJKUnicode

4

4.1

---WMT20 (Barrault et al., 2020) AI Hub [8] WMT20WikiMatrix230350300 [9] 50004

4.2

MTLlama2CJK 2 §4.3.1 Llama2BPEByteLlama2Ours256 `0x100-0x1FF` `p1, p2, p3`

BPE65MTransformer (Vaswani et al., 2017) [10] CJKCJKUTF-8

sacreBLEU (Post, 2018) BLEUchrFsacreBLEU 5 1000 [11] 5000

Marian (Junczys-Dowmunt et al., 2018) Huggingface Transformers (Wolf et al., 2020)

4.3

4.3.1

2 Llama2 Llama2 en-zh131M 64M CJK 23.4 % 48.8 % 52 %

[4] Fizz buzz

[5] `p1=0x39` `0x39` 9

[6] $\ll$ $\gg$

[7]

[8] `https://aihub.or.kr`

[9]

[10] 10epoch

[11] 1000

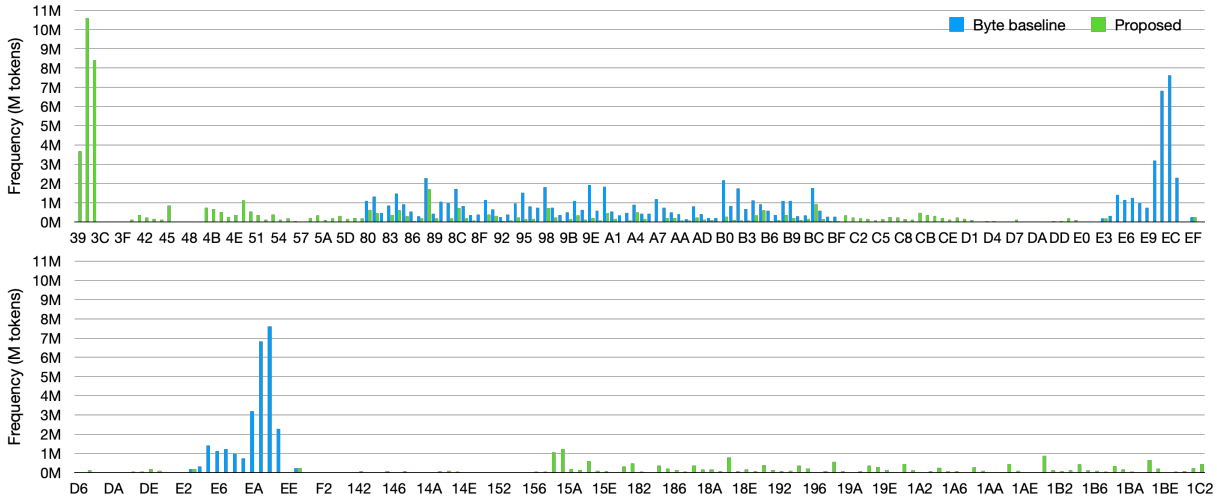

Figure 3: - 0x39-0x3C

	Size	BLEU		chrF		TER		Byte BLEU		Byte chrF	
		Byte	Ours	Byte	Ours	Byte	Ours	Byte	Ours	Byte	Ours
en-zh	2.23M	0.5	1.7	1.9	3.2	127.3	100.4	3	7.3	3	5.5
en-ja	3.47M	1.6	3.2	6.5	8.5	467.2	100.3	15.2	20	19.4	23.5
ja-ko	2.99M	19	24.6	35.8	35.5	114.4	100.2	49	53.3	54.7	53.5

Table 4: 5000Byte BLEUByte chrF

2 (2) Renyi Zouhar et al. (2023) 2 3
3 [12] 100 2

4.3.2

4 BLEU
TransformerBLEU

5 TPS

5 logitsTPS
TPS 5 TPSTPS
TPSTPSTPSTPS $\frac{|T_c|}{|T_e|}$ T_c T_e
[13]

1.03343.34 %TPSTPS 5 TPSTPS

6

UnicodeUTF-8
TPSTPSUTF-8UTF-8

7

CJK -

TPSLlama2 (32K)Llama3 (128K)

[12] 2005K95 %
[13]

8

7 CJK
6,9041,220CJK

9

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. 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

	en-zh		en-ja		ja-ko	
	Byte	Ours	Byte	Ours	Byte	Ours
Tokens Out	188,361	194,891	578,446	465,401	168,233	302,867
AvgTok Out	33.64	60.57	115.69	93.08	37.67	38.98
Total time (s)	72.41	201.65	529.05	271.60	41.10	41.59
Tokens per Second (TPS)	464.69	300.39	183.91	342.71	916.43	937.21
Tokens in test reference	291,857	282,423	253,540	251,261	189,800	185,628
Relative Gain	1	1.0334	1	1.00091	1	1.0225
Perceived TPS	464.70	310.43	183.91	345.81	916.43	958.28

Table 5: TPSTokens outtokenAvgTok5KTokens per Secondtoken5TPS

machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Jun'ichi Tsujii. 2020. CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6903–6915, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Omer Goldman, Avi Caciularu, Matan Eyal, Kris Cao, Idan Szpektor, and Reut Tsarfaty. 2024. Unpacking tokenization: Evaluating text compression and its correlation with model performance. *arXiv preprint arXiv:2403.06265*.

Thamme Gowda and Jonathan May. 2020. Finding the optimal vocabulary size for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3955–3964, Online. Association for Computational Linguistics.

Masanori Hirano, Masahiro Suzuki, and Hiroki Sakaji. 2023. llm-japanese-dataset v0: Construction of Japanese Chat Dataset for Large Language Models and its Methodology.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *Preprint*, arXiv:2203.15556.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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

Sander Land and Max Bartolo. 2024. Fishing for magikarp: Automatically detecting under-trained tokens in large language models. *Preprint*, arXiv:2405.05417.

Jindřich Libovický and Alexander Fraser. 2020. Towards reasonably-sized character-level transformer NMT by finetuning subword systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2572–2579, Online. Association for Computational Linguistics.

Sabrina J Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y Lee, Benoît Sagot, et al. 2021. Between words and characters: A brief history of open-vocabulary modeling and tokenization in nlp. *arXiv preprint arXiv:2112.10508*.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2024. Language model tokenizers introduce unfairness between languages. *Advances in Neural Information Processing Systems*, 36.

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.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. 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 3118–3135, Online. Association for Computational Linguistics.

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.

Uri Shaham and Omer Levy. 2021. Neural machine translation without embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 181–186, Online. Association for Computational Linguistics.

Makesh Narsimhan Sreedhar, Xiangpeng Wan, Yu Cheng, and Junjie Hu. 2023. Local byte fusion for neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7199–7214, Toronto, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2019. Neural machine translation with byte-level subwords. *Preprint*, arXiv:1909.03341.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Xiang Zhang and Yann LeCun. 2017. Which encoding is the best for text classification in chinese, english, japanese and korean? *arXiv preprint arXiv:1708.02657*.

Vilém Zouhar, Clara Meister, Juan Gastaldi, Li Du, Mrinmaya Sachan, and Ryan Cotterell. 2023. Tokenization and the noiseless channel. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5184–5207, Toronto, Canada. Association for Computational Linguistics.

A

A.1

Llama2 7B [14] (Hu et al., 2022)
Llama2 7B9.07M (Hirano et al., 2023)
`0x1AF 0xAF p1, p2, p3 p1 0xE4-0xE7`
20K 4
9 `0x100-0x1FF`

A.2

Nvidia H100 HBM294GBLoRANvidia A600048GB
247296H100x2LoRAH100x4120A6000x160
200 [15] 200(95 % +)100

- - byte: 296
- ours: 296
- -200
- en-ja 201
- ja-ko 93
- ja-ko ours: 93

A.3

MIT[CR]Llama2Llama2
-ID

[14]

[15] 10

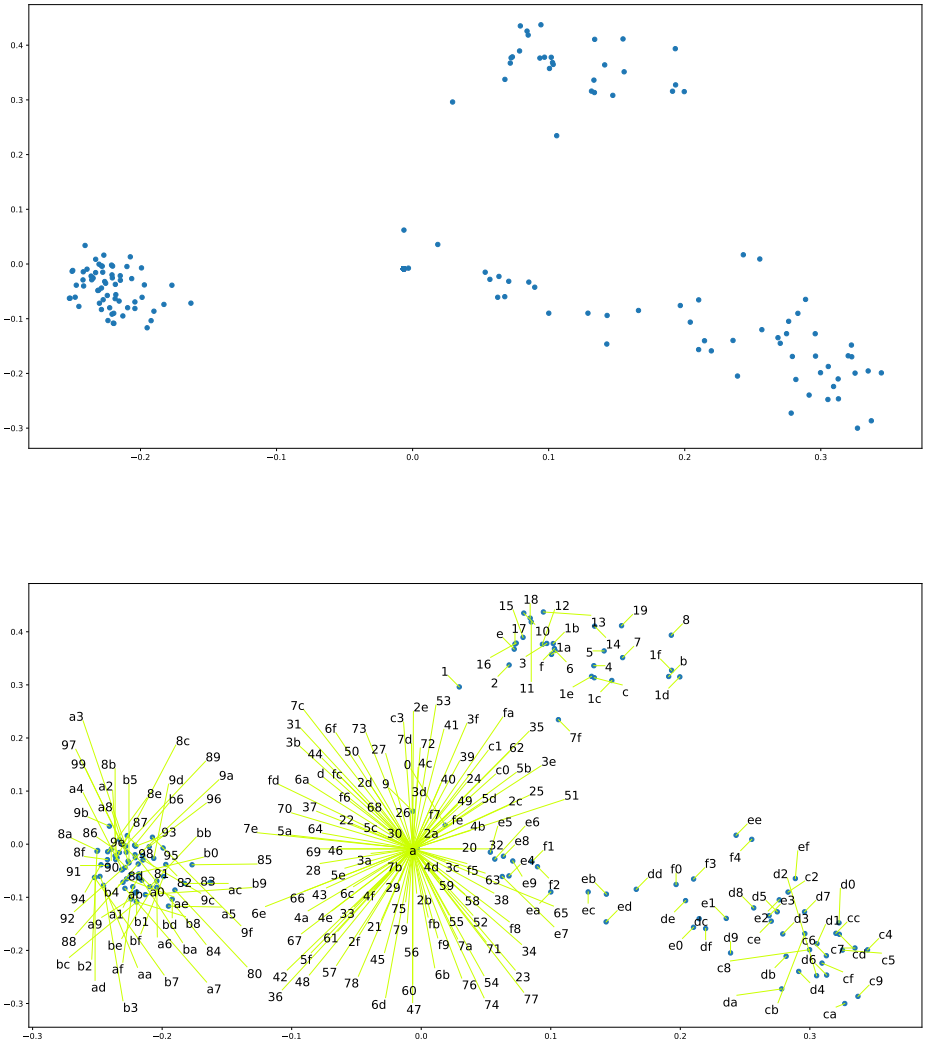

Figure 4: Llama2 $d = 2$

```
# Chinese (bleu, chrf, ter)
nrefs:1|case:mixed|eff:no|tok:zh|smooth:exp|version:2.4.2
nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.4.2
nrefs:1|case:lc|tok:tercom|norm:no|punct:no|asian:yes|version:2.4.2

# Japanese (bleu, chrf, ter)
nrefs:1|case:mixed|eff:no|tok:ja-mecab-0.996-IPA|smooth:exp|version:2.4.2
nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.4.2
nrefs:1|case:lc|tok:tercom|norm:no|punct:no|asian:yes|version:2.4.2

# Korean (bleu, chrf, ter)
nrefs:1|case:mixed|eff:no|tok:ko-mecab-0.996/ko-0.9.2-KO|smooth:exp|version:2.4.2
nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.4.2
nrefs:1|case:lc|tok:tercom|norm:no|punct:no|asian:yes|version:2.4.2

# Bytes (bleu, chrf)
nrefs:1|case:mixed|eff:no|tok:none|smooth:exp|version:2.4.2
nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.4.2
```

Table 6: sacreBLEU