HFT: High Frequency Tokenizer for Better Low-Resource NMT

Edoardo Signoroni, Pavel Rychlý
e.signoroni@mail.muni.cz, pary@fi.muni.cz

NLP Centre, Masaryk University

Oct 16, 2022

Outline

1. Introduction
2. High Frequency Tokenizer
3. Metrics Evaluation
4. Downstream NMT evaluation
5. Limitations and Future Work
6. Summary

Introduction

- Tokenization impacts downstream Neural Machine Translation (NMT) [1, 2, 7]
- A more meaningful subword vocabulary will improve NMT quality, (e.g. for low-frequency words) [8, 3]

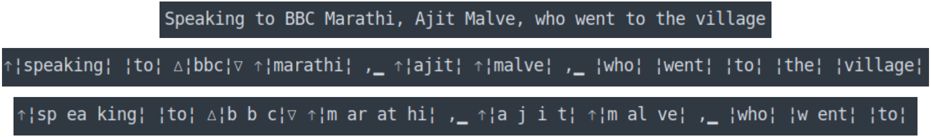
High Frequency Tokenizer

- High Frequency Tokenizer (HFT) is a new language-independent subword tokenization algorithm aimed at improving the frequency of the tokens in the vocabulary.
- HFT uses the advantage of pretokenization, using the regular expression $\backslash b$ of the Unix `sed`[1] command and a set of meta-characters.

[1] https://www.gnu.org/software/sed/manual/sed.html

¦	<token-delimiter>
↑	<single-uppercase>
▁	<explicit-whitespace>
∇	<all-uppercase>
Δ	<end-of-uppercase>

Figure 1: Special characters in the pretokenization and tokenization.



```
Speaking to BBC Marathi, Ajit Malve, who went to the village
↑¦speaking¦ ¦to¦ ∆¦bbc¦∇ ↑¦marathi¦ ,▁ ↑¦ajit¦ ↑¦malve¦ ,▁ ¦who¦ ¦went¦ ¦to¦ ¦the¦ ¦village¦
↑¦sp ea king¦ ¦to¦ ∆¦b b c¦∇ ↑¦m ar at hi¦ ,▁ ↑¦a j i t¦ ↑¦m al ve¦ ,▁ ¦who¦ ¦w ent¦ ¦to¦
```

Figure 2: Sample of text in the raw, pretokenized, and tokenized stages with HFT.

How HFT works

1. it processes pretokenized text to find the best subword segmentation using only subwords from the current vocabulary;
2. counts the frequencies of each subword and of all possible pairs of succeeding subwords;
3. selects the top K candidates with the highest frequency and adds them as new subwords;
4. removes from the vocabulary all non-single-character subwords with frequency lower than the last added candidate;
5. repeat from 1. until the requested vocabulary size is reached

Metrics evaluation

We test HFT against BPE [7] and Unigram [4] on:

- **Frequency at 95%**:
  Minimum frequency at the 95th percentile of the vocabulary (higher>lower)
- **Average Sentence Length**:
  Average length of the tokenized output sentence in number of tokens (lower>higher)
- **Frequency-Rank Weighted Average**

$$\nu = \frac{\sum_{i=1}^{n}(i \cdot f_{x_i})}{\sum_{i=1}^{n} i} \tag{1}$$

where i is the frequency rank of token x

Datasets (1)

Language		Dataset		Sent.
Amharic	Afro-Asiatic	am	Bible	30.580
Arabic	Afro-Asiatic	ar	Bible	31.102
Cherokee	Iroquian	chr	Bible-NT	7.957
Czech	Indo-Eur.	cs	Bible	38.116
English	Indo-Eur.	en	LoResMT	20.933
Finnish	Ugro-Finnic	fi	Bible	38.613
Irish	Indo-Eur.	ga	LoResMT	8.112
Hindi	Indo-Eur.	hi	IITB	20.000
Italian	Indo-Eur.	it	Bible	38.536
Japanese	Japonic	ja	Bible	31.087
Jakaltek	Mayan	jak	Bible-NT	12.509
Lithuanian	Indo-Eur.	lt	Europarl	20.000
Marathi	Indo-Eur.	mr	LoResMT	20.933
Burmese	Sino-Tibetan	my	Bible	30.928
Ojibwe	Algic	ojb	Bible-NT	7.945
Swedish	Indo-Eur.	sv	Bible	38.879
Syriac	Afro-Asiatic	syr	Bible-NT	7.954
isiZulu	Niger-Congo	zu	Bible-NT	9.095

Datasets (2)

Language	Script	Sample
Amharic	Ge’ez	በመጀመሪያ እግዚአብሔር ሰማይንና
Arabic	Arabic	في البدء خلق الله السموات والارض
Cherokee	Cherokee	ᎯᎠ ᎪᏪᎵ ᎧᏃᎮᏛ ᏧᏁᏢᏔᏅᏒ ᏥᏌ ᎦᎶᏁᏛ,
Hindi	Devanagari	यह कार्य दोनों प्रकार के हैं - ऑनलाइन और बाहरी।
Japanese	Kana/Kanji	はじめに神は天と地とを創造された。
Marathi	Devanagari	राज्यात हळूहळू अनलॉक होण्यास सुरूवात झाली
Burmese	Burmese	အစအဦး၌ ဘုရားသခင်သည် ကောင်းကင်နှင့်
Ojibwe	Ojibwe	ᒦᑕᔥ ᐅᐅᐧᐁ ᐁᔑᐲᐃᑳᑌᓂᒃ ᐋᓃᑎ ᒋᓭᔅ x ᑳᐱᐅᒥ ᓂᑖᐧᐃᑭᒻ
Syriac	Syriac	[illegible]

Figure 3: Non-latin scripts samples.

Experimental Setup

- We obtain BPE, Unigram, and HFT tokenizers for each dataset with a vocabulary size 500->8000
- We compute and plot the metrics against the vocabulary size

freq@95%

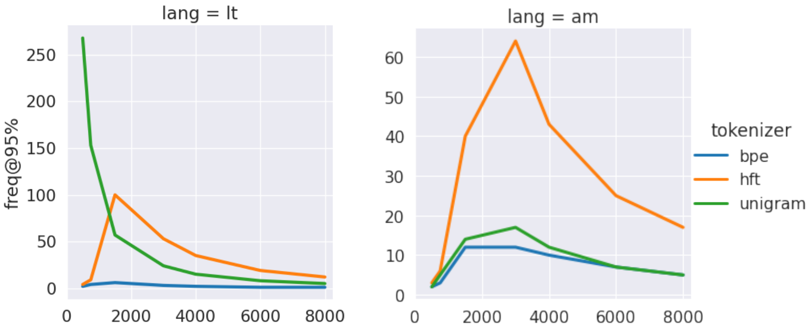


Figure 4: $F_{95\%}$ plotted against vocabulary size on the Lithuanian and Amharic.

freq@95%

- HFT outperforms the other methods by far, but in some cases after a size threshold.
- Saving at least one occurrence of each character is detrimental for very small vocabularies, and for noisy corpora

avg_len

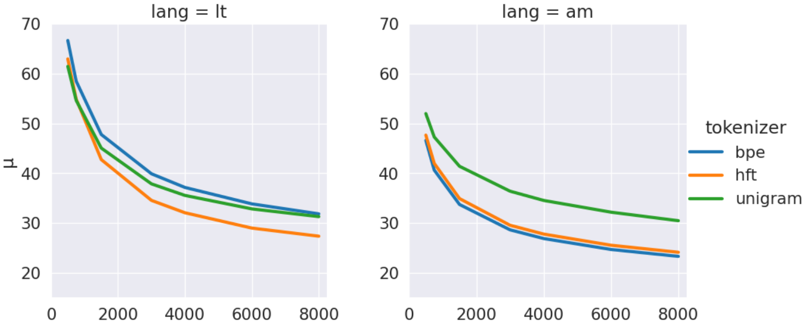


Figure 5: avg_len μ plotted against vocabulary size on the Lithuanian and Amharic datasets.

avg_len

- HFT performance is better than other methods on 15 out of 18 datasets, but
- size of the increase depends on the dataset
- and in some cases HFT fares comparably (ar, hi, mr) or worse (ja, my)

Freq-Rank Weighted Avg

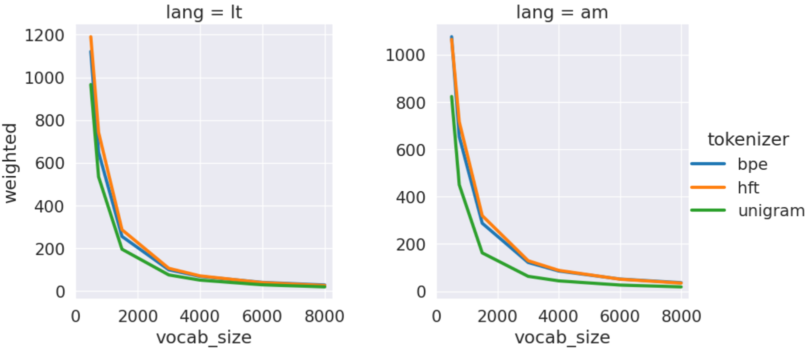


Figure 6: Weighted average ν plotted against vocabulary size on the Lithuanian and Amharic

Freq-Rank Weighted Avg

- freq@95% does not consider the whole vocabulary
- HFT fares slightly better than BPE, and much better than Unigram in some cases

Frequency Analysis

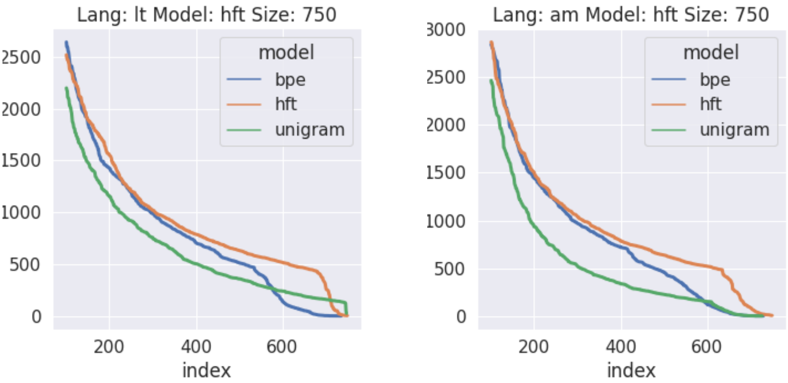


Figure 7: Frequency distribution of tokens in a sample of the 0.75k vocabularies.

Frequency Analysis

- Frequency plots show that HFT trades off frequency values between the highest occurring tokens and the least occurring ones.
- This is a favourable trade, since very frequent tokens will still be well represented.

Downstream NMT Evaluation

- We evaluate HFT on downstream NMT in an experimental environment

Dataset

	DATASET	N. of SENT
en-ga	LoResMT2021	8.112
en-mr	LoResMT2021	16.748

Experimental Setup

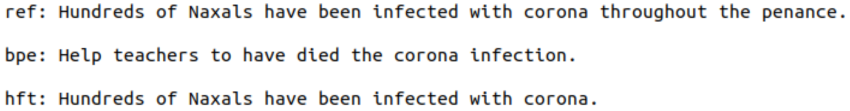
- We use Fairseq [5] to:
  - train 5 Transformer [9] models
  - for 30 epochs for en-ga, en-mr
  - each translation direction
  - for both BPE (`subword-nmt`) and HFT
- We generate translations and score the output with sacreBLEU [6] for each model
- We average and compare the results

Training Parameters

Vocabulary size	**2000, 3000, 4000**
architecture	**Transformer**
optimizer	**adam**
learning rate	**0.0005**
lr scheduler	**inverse square root**
warmup updates	**4000**
feed-forward dimension	*1024*, **2048**
attention heads	*2*, **8**
dropout	*0.0*, **0.1**, *0.3*
enc/dec layers	*5*, **6**
label smoothing	*0.0*, **0.1**, *0.5*
enc/dec word dropout	**0.0/0.0**, *0.0/0.1*, *0.0/0.2*
activation dropout	**0.0**, *0.3*
max tokens	**4096**

Results

DATASET	MODEL	BLEU						INCREMENT
		1	2	3	4	5	avg	
en-ga	t-bpe	4.46	4.54	4.06	4.69	4.73	4.50	
	t-hft	**5.34**	**5.49**	**5.95**	**5.69**	**5.59**	**5.61**	**+1.11**
ga-en	t-bpe	5.57	5.48	5.12	5.80	5.51	5.50	
	t-hft	**6.09**	**6.49**	**6.57**	**6.10**	**6.33**	**6.32**	**+0.82**
en-mr	t-bpe	7.49	7.21	6.88	6.57	6.12	6.85	
	t-hft	**7.33**	**7.99**	**8.80**	**8.31**	**8.31**	**8.14**	**+1.29**
mr-en	t-bpe	9.58	8.56	10.15	8.58	9.56	9.29	
	t-hft	**11.05**	**12.09**	**12.19**	**11.06**	**10.82**	**11.44**	**+2.15**



```
ref: Hundreds of Naxals have been infected with corona throughout the penance.

bpe: Help teachers to have died the corona infection.

hft: Hundreds of Naxals have been infected with corona.
```

Figure 8: Example from the *mr-en* translation systems. The first line gives the reference translation, the second gives the translation from a `bpe`-based system, while the last gives the translation from an `hft`-based system. The named entity *Naxals* is preserved by `hft`.

Discussion

- on average HFT leads to better NMT quality
- the increase still depends on the dataset
- the variance in the increases can be explained by the size of the training data,

Limitations and Future Work

- reduce the sensitiveness to noise
- investigate the case in which HFT underperformed
- investigate more datasets, tasks, and models (i.e. multilingual/unsupervised NMT)

Summary

- We presented a new tokenization algorithm, HFT, aimed at obtaining more meaningful subword vocabularies
- it leverages pretokenization and an iterative algorithm to search more frequent subword units
- We tested and demonstrated its efficacy on both established and new metrics
- and showed some preliminary results on downstream NMT
- Some future work still has to be done regarding some datasets and evaluation, such as other tasks and models

MUNI
FI

NLP Centre

e.signoroni@mail.muni.cz

GitHub: edoardosignoroni

Twitter: @Audhwer

References I

[1] Miguel Domingo et al. *How Much Does Tokenization Affect Neural Machine Translation?* 2018. DOI: 10.48550/ARXIV.1812.08621. URL: https://arxiv.org/abs/1812.08621.

[2] Thamme Gowda and Jonathan May. “Finding the Optimal Vocabulary Size for Neural Machine Translation”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 3955–3964. DOI: 10.18653/v1/2020.findings-emnlp.352. URL: https://aclanthology.org/2020.findings-emnlp.352.

References II

[3] Philipp Koehn and Rebecca Knowles. “Six Challenges for Neural Machine Translation”. In: *Proceedings of the First Workshop on Neural Machine Translation*. Vancouver: Association for Computational Linguistics, Aug. 2017, pp. 28–39. DOI: 10.18653/v1/W17-3204. URL: https://aclanthology.org/W17-3204.

[4] Taku Kudo. “Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 66–75. DOI: 10.18653/v1/P18-1007. URL: https://aclanthology.org/P18-1007.

References III

[5] Myle Ott et al. “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)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 48–53. DOI: `10.18653/v1/N19-4009`. URL: `https://aclanthology.org/N19-4009`.

[6] Matt Post. “A Call for Clarity in Reporting BLEU Scores”. In: *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 186–191. DOI: `10.18653/v1/W18-6319`. URL: `https://aclanthology.org/W18-6319`.

References IV

[7] Rico Sennrich, Barry Haddow, and Alexandra Birch. “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)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. DOI: `10.18653/v1/P16-1162`. URL: `https://aclanthology.org/P16-1162`.

[8] Rico Sennrich and Biao Zhang. “Revisiting Low-Resource Neural Machine Translation: A Case Study”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 211–221. DOI: `10.18653/v1/P19-1021`. URL: `https://aclanthology.org/P19-1021`.

References V

[9] Ashish Vaswani et al. “Attention is All You Need”. In: 2017. URL: https://arxiv.org/pdf/1706.03762.pdf.

MUNI
FACULTY
OF INFORMATICS