A Morpheme-Aware Child-Inspired Language Model

Necva Bölücü (CSIRO Data61) and Burcu Can (University of Stirling)

Motivation

- Tokenization affects how models learn from text.
- Small LMs need efficient and meaningful units to learn effectively from limited data.
- BPE (Byte-Pair Encoding) splits words by frequency, ignoring morphology.

In this study:

- Morpheme-aware tokenizer is used for an efficient learning.
- Curriculum learning with double-stage training training on data without and with morphology.

Tokenizers

- **BPE:** Splits words based on frequency. Efficient but ignores word structure.
- **Simple (Rule-based):** Splits words into prefixes, stems, and suffixes. Interpretable and linguistically meaningful. Language-dependent and requires a morpheme dictionary.
- **Morfessor (Unsupervised):** Learns morphemes from data. Morphology-aware and language-agnostic.

Word	BPE	Simple	Morfessor
run	r, un	run	run
dog	d, og	dog	dog
redo	red, o	re, do	re, do
cats	c, ats	cats	cats
jumping	j, ump, ing	jump, ing	jump, ing
played	play, ed	play, ed	played
unhappy	un, happy	un, happy	un, happy
happiness	ha, pp, iness	happi, ness	happiness
friendliness	friend, l, iness	friendli, ness	friendliness
undeniable	un, deniable	un, deniable	undeniable
counterattack	counter, att, ack	counterattack	counter, attack
unbelievably	un, bel, ie, v, ably	un, believab, ly	unbeliev, ably
reconsideration	re, c, ons, ider, ation	re, considera, tion	re, consideration
misunderstanding	m, is, under, standing	misunderstand, ing	misunderstand, ing

Table 1: Comparison of tokenization outputs for selected words by BPE, Simple, and Morfessor tokenizers.

Architecture

- **GPT-2 (Radford et al., 2019)**
- **GPT-BERT (Charpentier and Samuel, 2024)**

Experiments & Results

STRICT-SMALL track (10M words)

Model	Tokenizer	BLiMP	BLiMP Supplement	EWoK	Eye tracking	Self-paced Reading	Entity Tracking	WUG
GPT-2	BPE	65.77	62.40	49.82	0.73	0.03	21.93	52.00
GPT-2	SimpleTokenizer	53.04	44.40	53.55	0.74	0.08	40.66	100.00
GPT-2	Morfessor	65.10	49.20	68.45	0.08	0.12	59.65	100.00
GPT-2 (curriculum)	Morfessor	63.19	48.80	69.64	0.09	0.26	59.82	100.00
GPT-BERT	BPE	68.70	61.50	50.40	6.20	**4.45**	25.30	44.50
GPT-BERT	SimpleTokenizer	56.45	49.18	53.18	0.91	0.05	42.18	100.00
GPT-BERT	Morfessor	69.10	50.08	70.01	0.09	0.06	62.17	100.00
GPT-BERT (curriculum)	Morfessor	**72.10**	52.12	**71.15**	0.12	0.36	**63.25**	100
babylm-baseline-10m-gpt2	BPE	66.36	57.07	49.90	8.66	4.34	13.9	52.5
babylm-baseline-10m-gpt-bert-causal	BPE	65.22	59.49	49.47	**9.52**	3.44	30.60	68.00
babylm-baseline-10m-gpt-bert-mntp	BPE	70.36	**63.71**	49.95	9.40	3.37	40.02	57.5

STRICT track (100M words)

Model	Tokenizer	BLiMP	BLiMP Supplement	EWoK	Eye tracking	Self-paced Reading	Entity Tracking	WUG
GPT-2	BPE	75.24	62.80	51.00	2.70	0.43	25.48	47.00
GPT-2	SimpleTokenizer	71.10	48.56	59.17	0.76	0.32	63.10	100.00
GPT-2	Morfessor	64.60	55.20	67.45	0.81	0.28	67.45	100.00
GPT-2 (curriculum)	Morfessor	63.12	49.60	67.82	0.69	0.32	49.47	100.00
GPT-BERT	BPE	79.60	42.60	52.00	6.20	3.05	25.30	45.00
GPT-BERT	SimpleTokenizer	69.18	58.17	69.18	1.05	0.35	67.56	100.00
GPT-BERT	Morfessor	70.12	56.18	69.56	0.98	0.32	**68.48**	100.00
GPT-BERT (curriculum))	Morfessor	73.36	58.43	**71.15**	1.09	0.46	60.21	100.00
babylm-baseline-100m-gpt2	BPE	74.88	63.32	51.67	7.89	3.18	31.51	35.5
babylm-baseline-10m-gpt-bert-causal	BPE	74.56	63.63	51.57	8.80	3.30	30.82	59.00
babylm-baseline-10m-gpt-bert-mntp	BPE	**80.75**	**75.34**	51.77	**9.34**	**3.34**	41.15	55.00

Table 2: Performance of different models across multiple evaluation benchmarks.

Findings

Morpheme-based tokenizer outperforms BPE for some tasks, such as EWoK and entity tracking by a substantial margin.

The morpheme-based tokenizer improves all the scores, including BLIMP, BLIMP Supplement, EWoK, eye-tracking, and entity tracking, when used with the GPT-BERT architecture, whereas curriculum learning does not help as desired when used with the GPT-2 architecture.

References

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Charpentier, L. G. G., & Samuel, D. (2024, November). GPT or BERT: why not both?. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning* (p. 262).