ACL 2025
VIENNA
JULY 27 - AUGUST 1

Unsupervised Morphological Tree Tokenizer

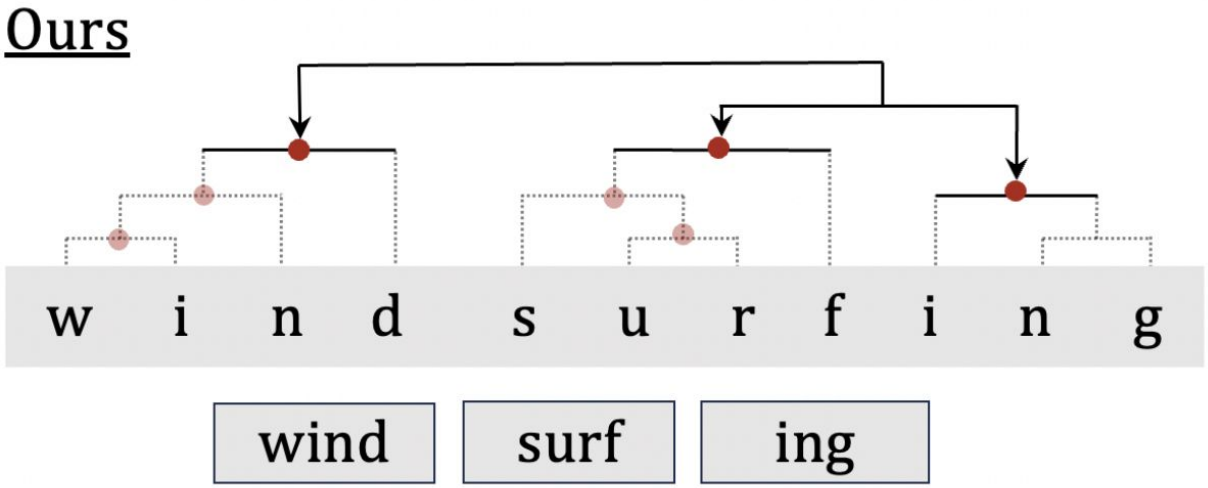
Qingyang Zhu*, Xiang Hu*, Pengyu Ji, Wei Wu, Kewei Tu

NYU
蚂蚁集团 ANT GROUP
上海科技大学 ShanghaiTech University

Current Tokenizers

Bottom-Up Tokenization

Our Tokenizer



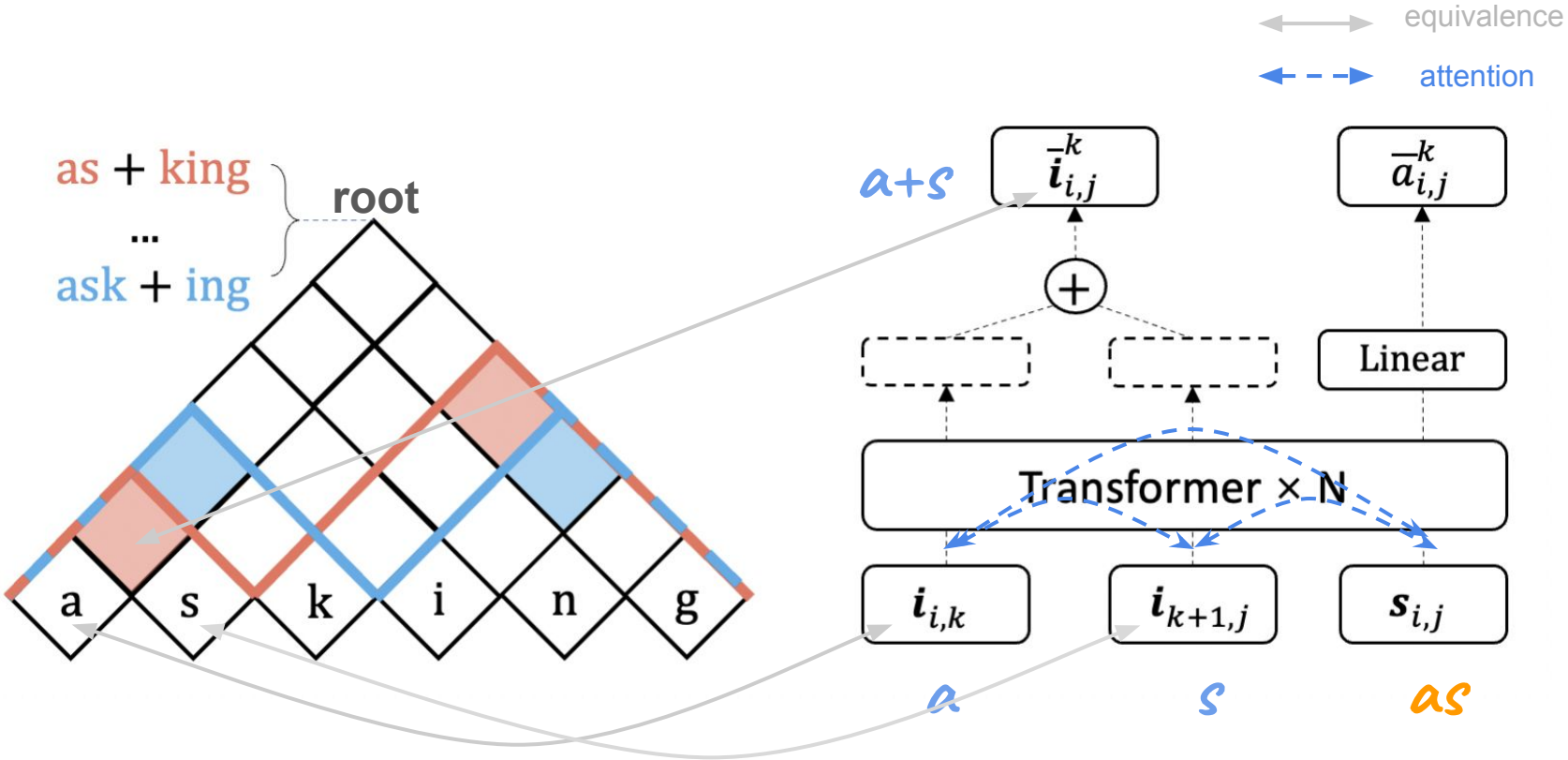
vocab

...
w, i, n, d, s, u, r, f, i, n,
g, wi, win, wind,
ur, sur, surf, ing,
...

Top-Down Tokenization

Parse Trees for each Word:

Unsupervised Character-level Structure Induction



Training objectives

1. Auto-Encoding loss (intra-word)

Predicting each character or character span from the rest of a word (neighbor context representations).

2. Auto-Regression loss (inter-word)

Predicting the next word in the sentence using the root representations of the preceding words, contextualized by GPT layers.

Two Ingredients of A Tokenizer

1. **Vocabulary Construction.**
    - Most of the tokenizers are named under this step (BPE, WPE, Unigram).
    - Our method runs a **tree-constrained BPE** to create an initial vocabulary, then apply a **tree-aware Unigram pruning** to remove meaningless fragments.
    - Being constrained by tree reduces unmeaningful vocabulary candidates.
2. **Segmentation/Tokenization** (using the vocabulary from 1)**.**
    - Use the pretrained composition model to parse each word into a tree and traverse it top-down, yielding a token whenever we hit a known morpheme.

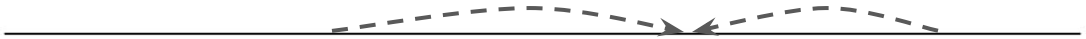
Experiment Results: Tokenization Quality

- **Datasets:** Morpho Challenge 2010 Workshop (Morpho), CompoundPiece (Compound) (Minixhofer et al. 2023).
- **Metric:** The ratio of examples that are correctly segmented (Acc.).

	Morpho (Acc.) ↑	Compound (Acc.) ↑	$\|\mathbb{V}\|$
BPE	19.50	62.98	30,000
WordPiece	26.20	62.19	30,000
Unigram	27.10	53.10	30,000
TreeTok	**37.9**	**68.07**	30,000

Experiment Results: Tokenization Quality

- **Metric:** Rényi efficiency (Zouhar et al., 2023), sentence-level perplexity, BLEU, avg number of tokens per sentence.
- **Datasets:** Wikitext-103, WMT14 de-en (BLEU only).



correlation

	Rényi↑	PPL↓	BLEU↑	avg. #tokens
BPE	44.66	107.76	26.55	26.58
WordPiece	44.54	110.97	-	26.60
Unigram	45.07	106.91	-	31.68
TreeTok	44.82	107.26	26.68	25.99

Efficiency and Quality: TreeTok yields minimal tokens— 18% fewer than Unigram. Despite fewer tokens, TreeTok outperforms BPE on perplexity and yields better BLEU.

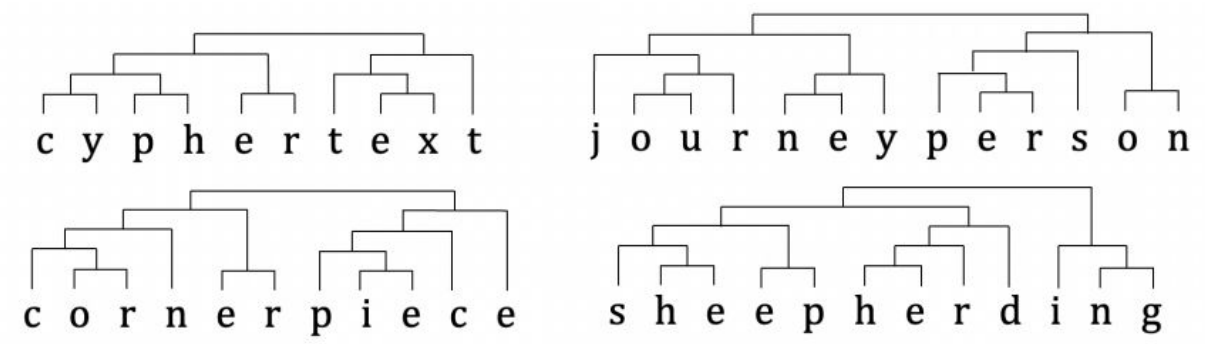
Experiment Results: Tree structure quality

- **Metric:** Morpheme Recall Rate (the percentage of morphemes in the gold segmentation that can be found in the spans of the evaluated tree).

	Morpho EN.	Compound EN.
Fast R2D2	67.69	48.96
Neural PCFG	39.87	58.33
TreeTok	**90.10**	**86.20**
w/o context	70.00	63.02
w/o *MorphOverriding*	75.99	46.35
w/o span loss	86.79	73.70

Case Study

original word	bed	commonly	windsurfing	tricycles	uniquenesses
BPE	bed	commonly	wind/sur/fing	tric/y/cles	uniqu/eness/es
Unigram	b/e/d	common/ly	wind/surf/ing	t/r/i/cycle/s	unique/ness/e/s
WordPiece	bed	commonly	winds/ur/fing	tric/y/cles	unique/ness/es
TreeTok	bed	commonly	wind/surf/ing	tri/cycles	unique/ness/es



Word-level Tree Samples

Conclusion

- TreeTok can induce **morphology-aligned** word-internal tree structures in a fully **unsupervised** way.
- We discovered that recognizing the indecomposability of morphemes is key, and to address this, we developed a composition model with a *MorphOverriding* mechanism alongside two self-supervised objectives.
- TreeTok induces tree structures that closely match human-labeled morphology and consistently outperforms baselines like BPE and WordPiece across various task.