Findings-#2836

Adaptive BPE Tokenization for Enhanced Vocabulary Adaptation in Finetuning Pretrained Language Models

Gunjan Balde*, Soumyadeep Roy*, Mainack Mondal, and Niloy Ganguly

Indian Institute of Technology Kharagpur

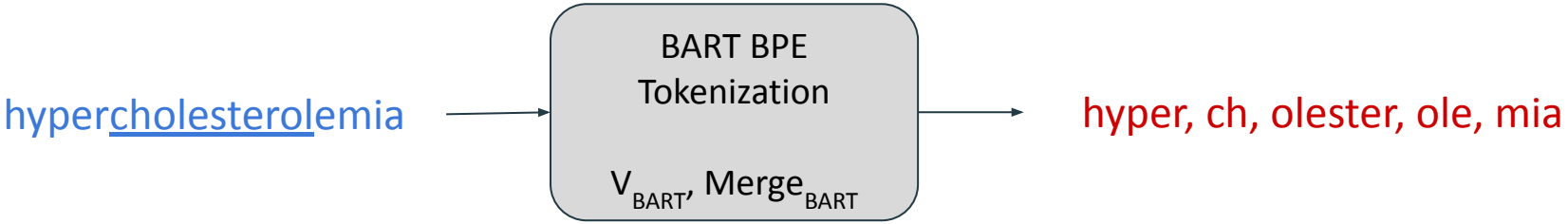
***Equal Contribution**

balde.gunjan0812@kgpian.iitkgp.ac.in

EMNLP 2024

Byte-Pair Encoding (BPE[1])

- Popular tokenization algorithm used in BART[2], LLama[3], and Mistral[4]
- Utilizes concept of ranked-based merge rules to tokenize a word



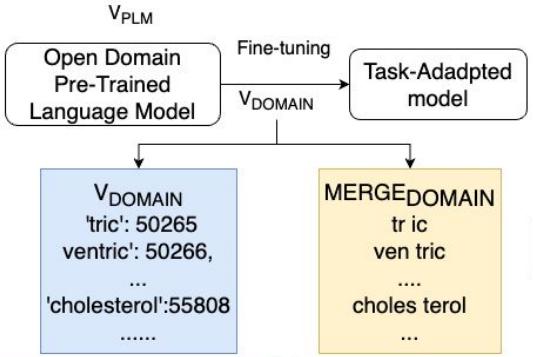
[1] Neural Machine Translation of Rare Words with Subword Units (Sennrich et al., ACL 2016)
[2] BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension (Lewis et al., ACL 2020)
[3] Llama 2: Open foundation and fine-tuned chat models. (Hugo,et al., 2023)
[4] Mistral 7B (Jiang et al., 2023)

Vocabulary Adaptation for PLM using BPE

- Strategy to adapt PLM to a domain ***during fine tuning***
- Works best when adapting to ***expert domain***
- Build a domain-specific vocabulary ($\boldsymbol{V}_{\boldsymbol{DOMAIN}}$)
- And learn corresponding merge rules ($\boldsymbol{MERGE}_{\boldsymbol{DOMAIN}}$)
- Appends it at **the end** of existing PLM Vocabulary and Merges

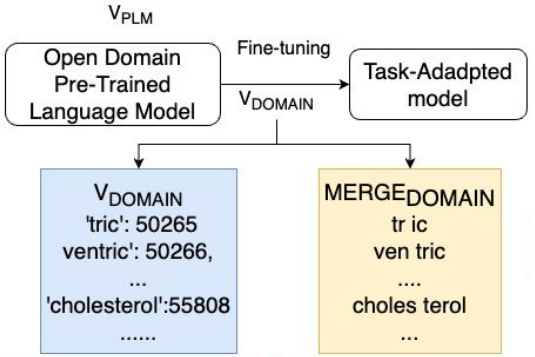
Vocabulary Adaptation for PLM using BPE

- Strategy to adapt PLM to a domain ***during fine tuning***
- Works best when adapting to ***expert domain***
- Build a domain-specific vocabulary ($\boldsymbol{V_{DOMAIN}}$)
- And learn corresponding merge rules ($\boldsymbol{MERGE_{DOMAIN}}$)
- Appends it at **the end** of existing PLM Vocabulary and Merges



Vocabulary Adaptation for PLM using BPE

- Strategy to adapt PLM to a domain ***during fine tuning***
- Works best when adapting to ***expert domain***
- Build a domain-specific vocabulary (V_{DOMAIN})
- And learn corresponding merge rules ($MERGE_{DOMAIN}$)
- Appends it at **the end** of existing PLM Vocabulary and Merges



Not all the added vocabulary tokens are utilized

hypercholesterolemia: *hyper, ch, olester, ole, mia*

How does BPE work

Split the input text down to character level

u-n-r-e-l-a-t-e-d

How does BPE work

Split the input text down to character level

u-n-r-e-l-a-t-e-d

Find **applicable merges** on the list and get *their rank*

u-n: 148
r-e: 6
…
e-d: 22

How does BPE work

Split the input text down to character level

u-n-r-e-l-a-t-e-d

Find **applicable merges** on the list and get *their rank*

u-n: 148
r-e: 6
…
e-d: 22

Apply the **top-ranked** merge rule

u-n-**re**-l-a-t-e-d

How does BPE work

u-n-r-e-l-a-t-e-d Merge **r-e**
u-n-re-l-a-t-e-d

Repeat merges till no merge applicable

How does BPE work

u-n-r-e-l-a-t-e-d Merge **r-e**
u-n-re-l-a-t-e-d Merge **a-t**

Repeat merges till no merge applicable

How does BPE work

u-n-r-e-l-a-t-e-d	Merge **r-e**
u-n-re-l-a-t-e-d	Merge **a-t**
u-n-re-l-at-e-d	

Repeat merges till no merge applicable

How does BPE work

u-n-r-e-l-a-t-e-d	Merge **r-e**
u-n-re-l-a-t-e-d	Merge **a-t**
u-n-re-l-at-e-d	Merge **e-d**

Repeat merges till no merge applicable

How does BPE work

u-n-r-e-l-a-t-e-d	Merge **r-e**
u-n-re-l-a-t-e-d	Merge **a-t**
u-n-re-l-at-e-d	Merge **e-d**
u-n-re-l-at-ed	

Repeat merges till no merge applicable

How does BPE work

u-n-r-e-l-a-t-e-d	Merge **r-e**
u-n-re-l-a-t-e-d	Merge **a-t**
u-n-re-l-at-e-d	Merge **e-d**
u-n-re-l-at-ed	Merge **u-n**

Repeat merges till no merge applicable

How does BPE work

u-n-r-e-l-a-t-e-d	Merge **r-e**
u-n-re-l-a-t-e-d	Merge **a-t**
u-n-re-l-at-e-d	Merge **e-d**
u-n-re-l-at-ed	Merge **u-n**
un-re-l-at-ed	

Repeat merges till no merge applicable

How does BPE work

u-n-r-e-l-a-t-e-d	Merge **r-e**
u-n-re-l-a-t-e-d	Merge **a-t**
u-n-re-l-at-e-d	Merge **e-d**
u-n-re-l-at-ed	Merge **u-n**
un-re-l-at-ed	Merge **at-ed**

Repeat merges till no merge applicable

How does BPE work

u-n-r-e-l-a-t-e-d	Merge **r-e**
u-n-re-l-a-t-e-d	Merge **a-t**
u-n-re-l-at-e-d	Merge **e-d**
u-n-re-l-at-ed	Merge **u-n**
un-re-l-at-ed	Merge **at-ed**
un-re-l-ated	

Repeat merges till no merge applicable

How does BPE work

u-n-r-e-l-a-t-e-d	Merge **r-e**
u-n-re-l-a-t-e-d	Merge **a-t**
u-n-re-l-at-e-d	Merge **e-d**
u-n-re-l-at-ed	Merge **u-n**
un-re-l-at-ed	Merge **at-ed**
un-re-l-ated	Merge **re-l**

Repeat merges till no merge applicable

How does BPE work

u-n-r-e-l-a-t-e-d	Merge **r-e**
u-n-re-l-a-t-e-d	Merge **a-t**
u-n-re-l-at-e-d	Merge **e-d**
u-n-re-l-at-ed	Merge **u-n**
un-re-l-at-ed	Merge **at-ed**
un-re-l-ated	Merge **re-l**
un-rel-ated	

Repeat merges till no merge applicable

How does BPE work

u-n-r-e-l-a-t-e-d	Merge **r-e**
u-n-re-l-a-t-e-d	Merge **a-t**
u-n-re-l-at-e-d	Merge **e-d**
u-n-re-l-at-ed	Merge **u-n**
un-re-l-at-ed	Merge **at-ed**
un-re-l-ated	Merge **re-l**
un-rel-ated	Merge **rel-ated**

Repeat merges till no merge applicable

How does BPE works

u-n-r-e-l-a-t-e-d	Merge **r-e**
u-n-re-l-a-t-e-d	Merge **a-t**
u-n-re-l-at-e-d	Merge **e-d**
u-n-re-l-at-ed	Merge **u-n**
un-re-l-at-ed	Merge **at-ed**
un-re-l-ated	Merge **re-l**
un-rel-ated	Merge **rel-ated**
un-related	

Repeat merges till no merge applicable

How does BPE work

u-n-r-e-l-a-t-e-d	Merge **r-e**
u-n-re-l-a-t-e-d	Merge **a-t**
u-n-re-l-at-e-d	Merge **e-d**
u-n-re-l-at-ed	Merge **u-n**
un-re-l-at-ed	Merge **at-ed**
un-re-l-ated	Merge **re-l**
un-rel-ated	Merge **rel-ated**
un-related	Merge **un-related**

Repeat merges till no merge applicable

How does BPE work

u-n-r-e-l-a-t-e-d	Merge **r-e**
u-n-re-l-a-t-e-d	Merge **a-t**
u-n-re-l-at-e-d	Merge **e-d**
u-n-re-l-at-ed	Merge **u-n**
un-re-l-at-ed	Merge **at-ed**
un-re-l-ated	Merge **re-l**
un-rel-ated	Merge **rel-ated**
un-related	Merge **un-related**
unrelated	

Repeat merges **till no merge applicable**

Issue: Added Merges receive low priority!

- Added vocabulary appended at the end of existing PLM vocabulary
- Priority of $\boldsymbol{MERGE_{DOMAIN} < MERGE_{PLM}}$
- Added merge rules are never utilized for multiple instances
- **Resulting in ill tokenization of added vocabulary tokens**

Issue: Added Merges receive low priority!

- Added vocabulary appended at the end of existing PLM vocabulary
- Priority of $MERGE_{DOMAIN} < MERGE_{PLM}$
- Added merge rules are never utilized for multiple instances
- **Resulting in ill tokenization of added vocabulary tokens**

Can we modify standard BPE to mitigate ill tokenization?

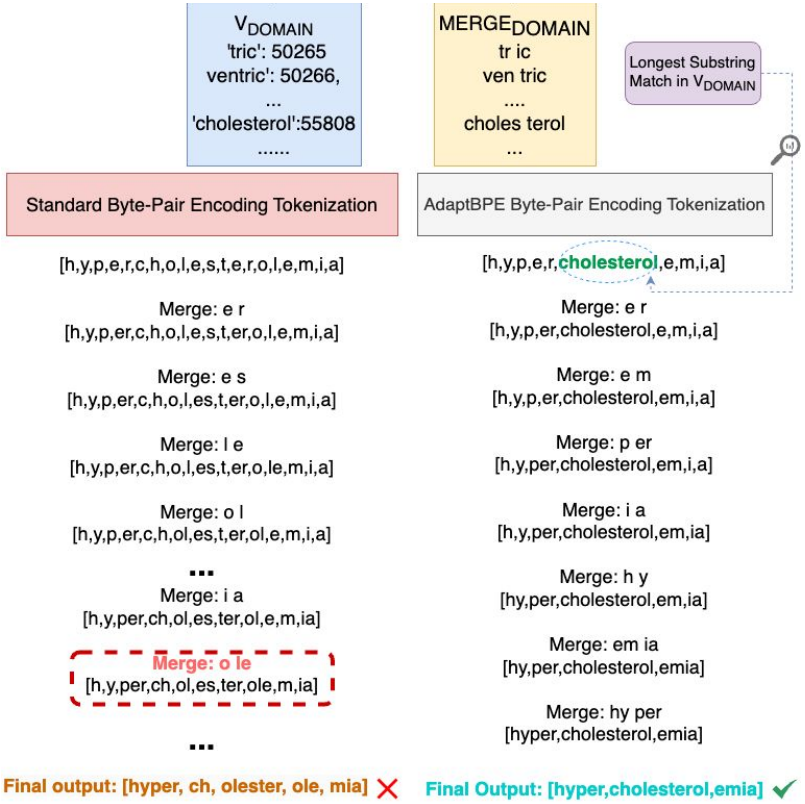
ADAPTBPE: Mitigate Ill-tokenization

We propose a **fundamental change** in BPE

- BPE starts by splitting the text to character level
- **Instead**:
  - find longest substring match in V_{DOMAIN} iteratively
  - Preserve the match as is
- Run merge operations of BPE

ADAPTBPE: Mitigate Ill-tokenization

In standard BPE because of merge rules from PLM merges, the word **cholesterol** could never be formed resulting in the ill-tokenization issue



However in ADAPTBPE, the subword **cholesterol** is correctly captured in longest substring match phase mitigating the ill-tokenization issue

ADAPTBPE improves Classification (AVocaDo[1])

- SoTA in vocabulary adaptation for classification tasks
- 4 classification datasets from different domain
- Overall improvement of **3.57%**

Dataset	Tokenizer	Accuracy	Macro-F1
CHEMPROT (BIOMED)	BPE	$81.43_{0.55}$	$54.88_{1.66}$
	ADAPTBPE	$81.40_{0.40}$	$55.02_{0.47}$
ACL-ARC (SCIENTIFIC)	BPE	$69.03_{5.05}$	$55.04_{8.24}$
	ADAPTBPE	$73.02_{4.21}$	$62.00_{4.95}$
HYP (NEWS)	BPE	$77.84_{5.20}$	$74.23_{7.01}$
	ADAPTBPE	$82.16_{2.50}$	$80.64_{3.03}$
AMAZON (REVIEWS)	BPE	$83.13_{3.64}$	$68.34_{0.47}$
	ADAPTBPE	$86.26_{0.53}$	$69.90_{0.29}$

[1] AVocaDo: Strategy for Adapting Vocabulary to Downstream Domain (Hong et al., EMNLP 2021)

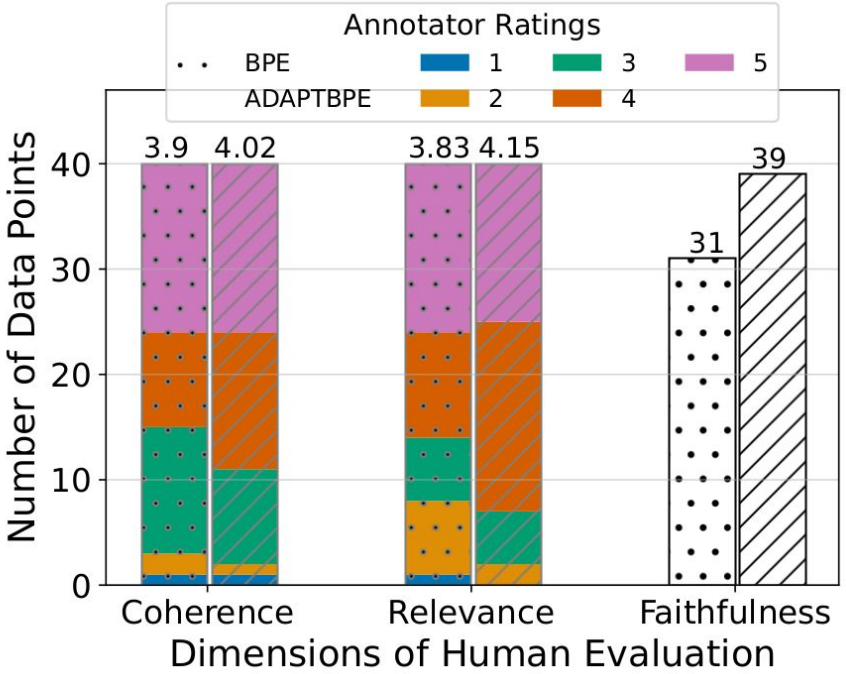
ADAPTBPE improves Summarization (MEDVOC[1])

- SoTA in vocabulary adaptation for summarization
- 4 medical summarization datasets:
  - 2 query-focussed
  - 2 consumer health query
- Overall improvement of **1.87%**
- Better in high OOV concentration

Dataset	Tokenizer	R-L (All)	R-L (H-O)
EBM	BPE	20.65	19.23
	ADAPTBPE	20.73	21.43
BioASQ	BPE	48.02	39.23
	ADAPTBPE	47.72	42.95
MeQSum	BPE	55.88	75.56
	ADAPTBPE	58.00	82.64
CHQ	BPE	40.59	33.77
	ADAPTBPE	41.92	37.60

[1] MEDVOC: Vocabulary Adaptation for Fine-tuning Pre-trained Language Models on Medical Text Summarization (Balde et al., IJCAI 2024)

Medical experts find summaries to be more relevant and faithful



Conclusion and Takeaways

- First to show **incorrect BPE tokenization issue** for vocabulary adaptation
- ADAPTBPE is applicable to any vocabulary adaptation strategy
- ADAPTBPE improves over BPE in SoTA classification and summarization

Thank You!

PMRF
Prime Minister's Research Fellows

Soumyadeep Roy Mainack Mondal Niloy Ganguly

CNeRG

Codebase

Preprint

Any Questions?

Correspondence email:
balde.gunjan0812@kgpian.iitkgp.ac.in