Task-Adaptive Tokenization: Enhancing Long-Form Text Generation Efficacy in Mental Health and Beyond

Siyang Liu[1], Naihao Deng[1], Sahand Sabour[2], Minlie Huang[2], Rada Mihalcea[1]
[1]University of Michigan [2]Tsinghua University

Motivation

Task-specific difficulty

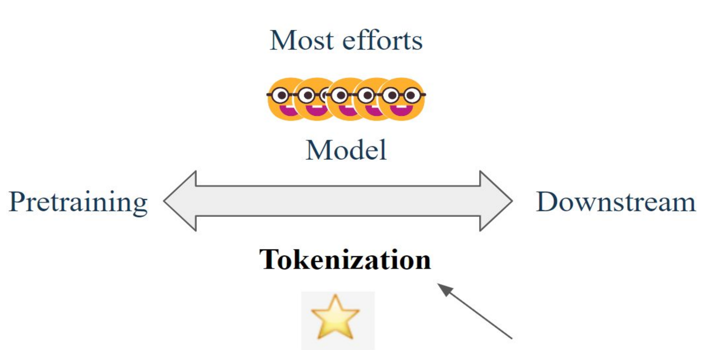
- Structure: Long-form
- Domain feature: Domain terminology, text structure and language style aligned with MHPs[1]

Why is it not easily addressed by prompting?

length restriction

unprofessional & unwanted behavior

Task-adaptive tokenization as a strategy to adapt a model to a downstream task



Don't ignore this component!

Good Compositionality
- Effectiveness
- Efficiency

Post

A 20-year-old male with bipolar disorder and symptoms of OCD, who is afraid of attractive women, is seeking treatment?

I have bipolar disorder and OCD symptoms, and I fear attractive women. I feel nervous and sweaty on the subway, and anxious when passing by attractive women on the street ... Psychoanalysis suggests fear of being seen through or assigning sexual meaning to interactions. I have experienced trauma from the opposite sex in college. I tried exposure therapy but it didn't work. Is hypnotherapy effective? Are there other treatment options like medication? I am anxious.

Keywords: diagnosis, treatment, counseling

Response

Hello! It is crucial to seek professional psychological therapy and counseling for your condition, as bipolar disorder with symptoms of obsessive-compulsive traits and opposite-sex fear is a severe mental health issue that requires proper evaluation and treatment by a qualified mental health professional. The fear of the opposite sex may stem from past traumatic experiences and involve issues related to self-disclosure, sexual desires, and identity transformation.

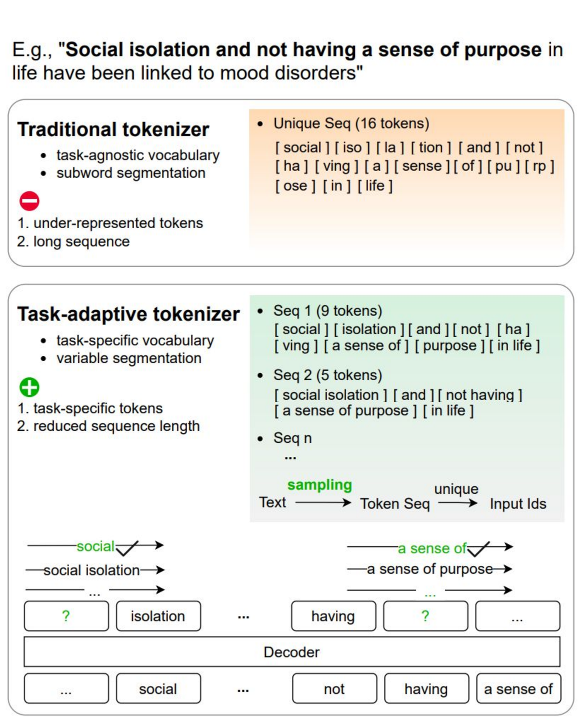
Cognitive-behavioral therapy (CBT) can be an effective approach for addressing opposite sex fear. CBT is a common therapeutic approach that focuses on identifying and changing irrational thoughts and behavior patterns to improve mental health. Strategies may include recognizing and understanding the fear, gradual exposure to the opposite sex, cognitive restructuring to change negative thought patterns, and learning relaxation techniques to manage anxiety.

....

Exposure therapy and hypnotherapy can also be effective treatment approaches. Exposure therapy involves gradual exposure to the fear object to reduce fear, while hypnotherapy uses hypnosis to modify thoughts and behaviors. However, it is important to undergo these therapies under the guidance of a qualified professional.

In addition to psychotherapy, medication can be a viable treatment option for bipolar disorder. Common medications used include lithium salts, antidepressants, and antipsychotics. However, medication should be prescribed and monitored by a qualified healthcare professional due to potential side effects.

It is crucial to seek professional mental health care and counseling promptly. Mental health issues require timely treatment to prevent further impact on well-being. Wishing you a speedy recovery.



An example of our envisioned tokenization (the green case)

A data example of PsyQA

Design Principles for task-adaptive tokenization

Design Principle 1 - "Exploiting Task-specific Data"

❑ A cognitive linguistics perspective[1]

Human's productive vocabulary refers to words actively used when writing/speaking

Productive & Receptive Vocabulary

❑ An optimization perspective

How text is segmented into tokens influences the optimization outcomes

E.g., bert vocabulary which is optimized from GNMT

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}}\, L(\theta),$$

where

$$L(\theta) = \sum_{s=1}^{|D|} logP(r^{(s)}|c^{(s)};\theta)$$

Objective with tokenizer determined

$$L(\theta) = \sum_{s=1}^{|D|} E_{\substack{r\sim p(r|R^{(s)})\\ c\sim p(c|C^{(s)})}} logP(r|c;\theta),$$

where

$$\{< C^{(s)}, R^{(s)} >\}_{s=i}^{|D|} \overset{\text{tokenizer}}{=} \{< c^{(s)}, r^{(s)} >\}_{s=i}^{|D|},$$

Objective without tokenizer determined

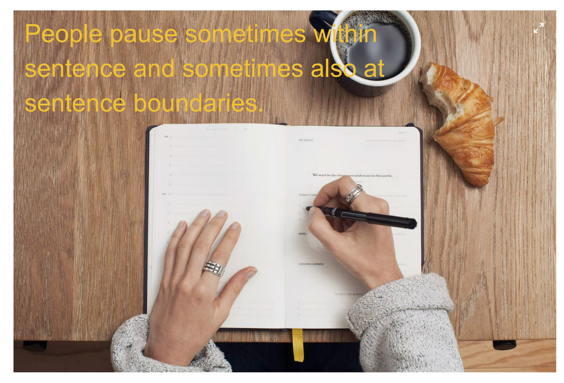
❑ Summary

A downstream vocabulary optimized from the task dataset is beneficial

Design Principle 2 - "The Importance of Variable Segment"

❑ A cognitive linguistics perspective[1]

We could write by letter, word, phase, and even sentence. If an expression is actively used, we store them as a whole in the memory.

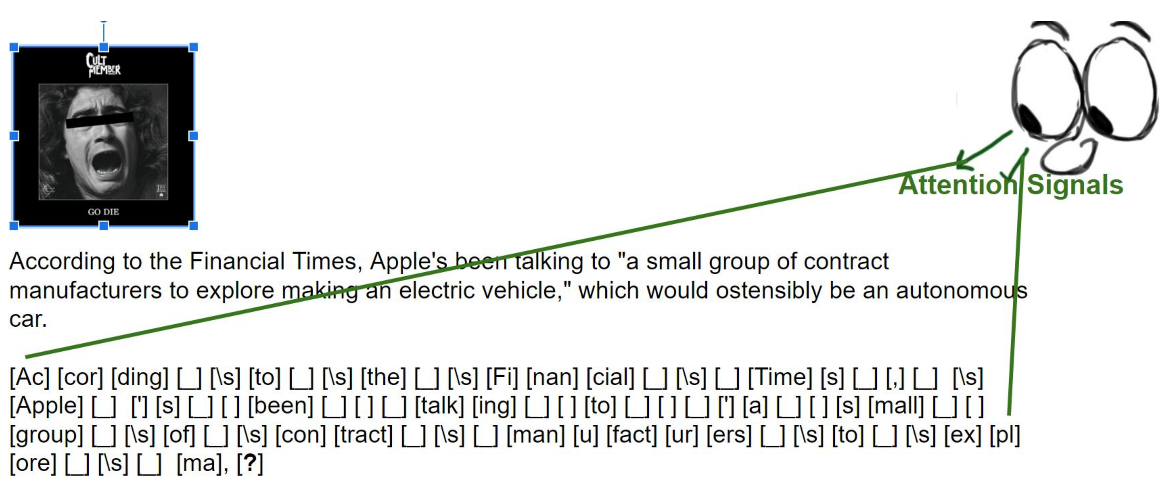


People think and produce spans with variable length

❑ An NLP engineering perspective

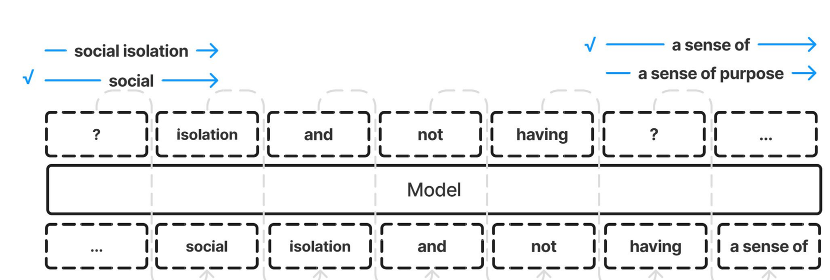
If segmentation is simply fine-grained and not optimized:

- Degradation in token representation[2]
- Generation inefficiency and wasted token cost[3][4]



❑ Summary

Allow a vocabulary to entail any granularity, and so the generation like below is available:



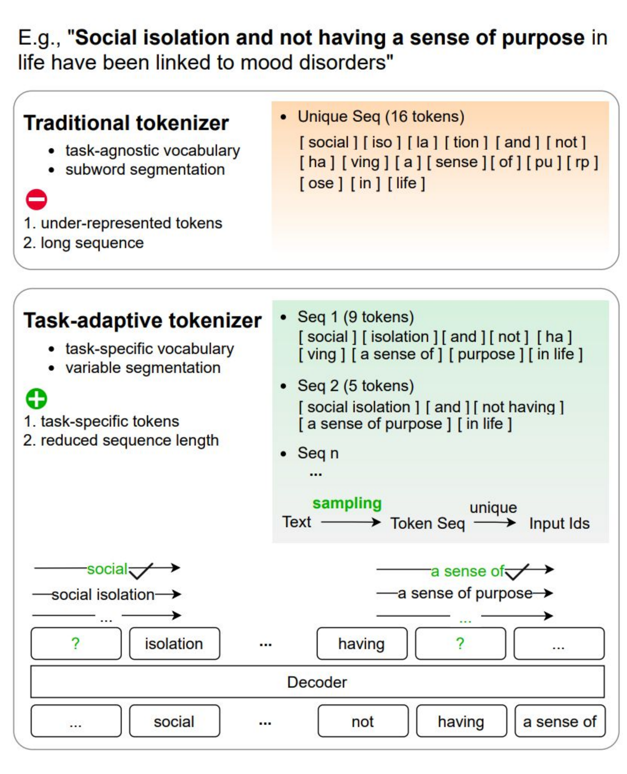
Task-adpative tokenizer construction

Step 1 - "Adapting an existed work to construct downstream vocabulary adhered to our design"

❑ Our adjustion on "Unigram Model with Subword Regularization"[3]

1. Cut sentences of the downstream corpus into ~~subword pieces~~ any granularity pieces
2. According to the frequency, pick a large set of ~~subword~~ pieces to form a big seed vocabulary (e.g., 4w)
3. Build a unigram model on the corpus, by modeling all possible combination of pieces to represent a text.
4. Apply EM algorithm to maximize the likelihood of the corpus, and get the log likelihood score of each piece.
5. According to the score of each piece, truncate the big seed vocabulary into one with the expected size (e.g., 1w)

❑ Features of proposed downstream vocabulary
- ❑ Each token corresponds to a score, indicating its log likelihood contribution to modeling the corpus.
- ❑ A sentence may have multiple segmentation results. The relationship between text and token sequence is one2many.
- ❑ In training, the token sequence is sampled based on the log likelihood score of all possible token sequences.



An example of our envisioned tokenization

Word segment occurrence probability under unigram language assumption:

$$P(x|X) = \prod_{i=1}^{M} p(x_i), \tag{2}$$

where X is a piece of text, and x is a corresponding word segment sequence $(x_1, ..., x_M)$.

Optimize $p(x)$ via the EM algorithm with maximizing L:

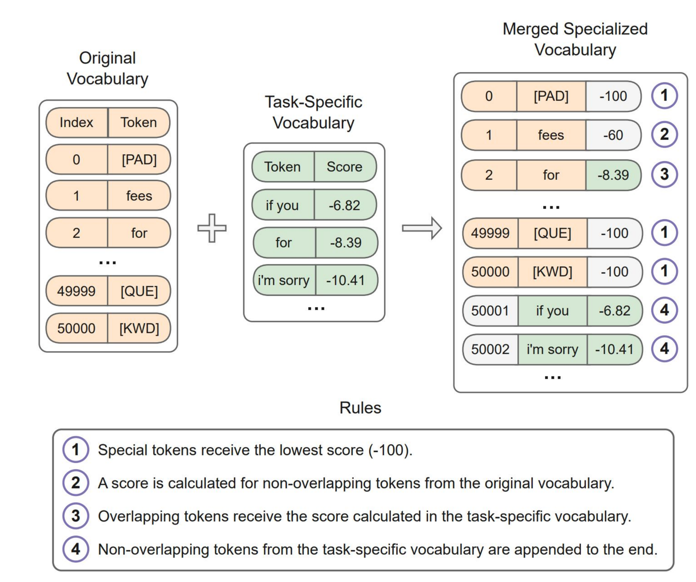
$$L = \sum_{s=1}^{|D|} log(P(X^{(s)})) = \sum_{s=1}^{|D|} log(\sum_{x\in S(X^{(s)})} P(x)) \tag{3}$$

Optimization of unigram model

Step 2 - "A protocol for merging downstream vocabulary and pre-existing vocabulary"

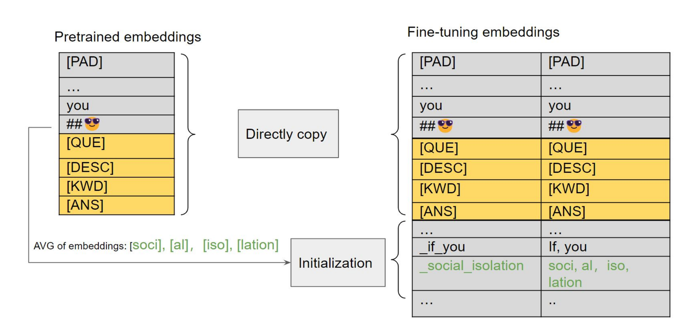
❑ A reasonable solution to order the merged vocabulary, facilitating the inheriting of the embedding matrix of the pretrained model

❑ Give moderately small scores to those tokens only existed in pre-existing vocabulary



Merging protocol

Step 3 - "A mapping mechanism for better initialization of new token embeddings"



Let new token attend their subword embeedings

Evaluation

Result 1 - generation effectiveness

Setting	Bleu	+pct	B-1	B-3	R-L	+pct
CN PsyQA						
gpt_{ft}+s*	20.1			-		
$GPT2_{base}$	18.2	-	55.5	2.5	15.5	-
$GPT2_{TaT}$	24.8†	+35.9%	65.7†	6.4†	**27.1**†	+74.8%
+mapping	**25.0**†	+37.1%	**66.3**†	**6.6**†	22.1†	+42.1%
$Bart_{base}$	21.6	-	62.3	4.0	21.8	-
$Bart_{TaT}$	**26.2**†	+21.3%	**69.2**†	6.7†	27.2†	+24.8%
+mapping	26.1†	+20.8%	68.8†	**6.7**†	**27.2**†	+24.8%

Setting	Bleu	+pct	B-1	B-2	R-L	+pct
MHP Reddit						
$GPT2_{base}$	3.7	-	14.0	0.6	5.7	-
$GPT2_{TaT}$	3.6	-2.7%	13.0	1.3†	8.1†	+42.1%
+mapping	**4.5**	+16.4%	**16.3**†	**1.6**†	**9.0**†	+57.9%
$Bart_{base}$	6.7	-	23.5	2.6	10.8	-
$Bart_{TaT}$	**7.6**†	+13.4%	**27.9**†	2.5	10.1	-6.5%
+mapping	6.7	+0.0%	22.9	**3.0**	**10.9**	+0.9%

Setting	Bleu	+pct	B-1	B-3	R-L	+pct
CN PsyQA						
$LLaMA_{base}$	27.9	-	64.5	**12.1**	30.1	-
$LLaMA_{TaT}$	**29.8**	+6.8%	69.6†	**12.5**	34.0†	+13.0%
+mapping	29.5	+5.7%	**69.6**†	12.3	**34.6**†	+15.0%

Results of automatic evaluation on generation effectiveness on GPT-2 small and Bart-small (left), and LLaMA-7B (right)

Result 2 - generation efficiency

Setting	#cSec	#Tok	Len	Len/#Tok ↑	Len/#cSec ↑
CN PsyQA					
$GPT2_{base}$	5.7	440.2	365.9	0.8	64.2
$GPT2_{TaT}$ + mapping	3.6	190.3	382.9	**2.0**	**106.4**
MHP Reddit					
$GPT2_{base}$	1.6	117.1	86.9	0.7	**54.3**
$GPT2_{TaT}$ + mapping	2.4	118.8	123.5	**1.0**	51.5

Table 3: Efficiency of generation. #cSec and #Tok denote the average number of centiseconds and tokens per generation on the test set respectively. Length denotes the average length of generated responses.

Result 3 - human evaluation

Metric	M vs. B Win	M vs. B Lose	NM vs. B Win	NM vs. B Lose	M vs. NM Win	M vs. NM Lose
CN PsyQA						
F	31†	15	18	24	36†	11
C	37†	9	19	19	36†	10
PE	23	20	18	22	32†	13
MHP Reddit						
F	26	20	4	43	44†	4
C	28	20	4	38	48†	1
PE	30	18	6	39	45†	3

Table 4: Human Evaluation. An explanation for abbreviations: M for $GPT2_{TaT}$ +mapping, B for $GPT2_{base}$, and NM for $GPT2_{TaT}$ w/o mapping; F for fluency, C for coherence, and PE for professional expression. Ties are not shown. † denotes a significant win (one sample sign test, p-value < 0.05).

Conclusion

1. Two key design principles for constructing downstream vocabularies
2. A protocol for merging downstream and pretraining vocabularies,
3. A mapping mechanism for new token representation learning
4. Significant improvements in both efficiency and effectivenes

References

[1] Galbraith, D., Torrance, M., & Waes, L. (Eds.). (2007). Writing and cognition: Research and applications. Elsevier.

[2] Dou, Y., Forbes, M., Koncel-Kedziorski, R., Smith, N. A., & Choi, Y. (2022, May). Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 7250-7274).

[3] Kudo, Taku. "Subword regularization: Improving neural network translation models with multiple subword candidates." arXiv preprint arXiv:1804.10959 (2018).

[4] Ahia, Orevaoghene, et al. "Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models." arXiv preprint arXiv:2305.13707 (2023).

[5]Zouhar, Vilém, et al. "Tokenization and the Noiseless Channel." Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023.