Bootstrapping Small & High Performance Language Models with Unmasking-Removal Training Policy

Yahan Yang[1], Elior Sulem[2], Insup Lee[1], Dan Roth[1]

[1]University of Pennsylvania, [2]Ben-Gurion University of the Negev

Motivation and Background

BabyBERTa [1], a smaller RoBERTa [2]-like language model trained on a 5M child-directed data corpora without using unmasked tokens during the masked language modeling training.

Examples of CHILDES:

1) there's a face with glasses . 2) there's a baby bear with his bottle .

Examples of Wikipedia:

1) it is not known which approach is more effective .
2) this feedback loop results in a reduced albedo effect .

Previous work concentrated on evaluating the zero-shot grammar ability of BabyBERTa, demonstrating that it achieves comparable performance on grammar test suites as RoBERTa but with significantly reduced training costs.

Architecture and Dataset of BabyBERTa [1] and RoBERTa [2]

	RoBERTa	BabyBERTa
layers	12	8
attention heads	12	8
hidden size	768	256
intermediate size	3072	1024
vocabulary size	50265	8192

	Dataset Size
CHILDES (CHIL)	6.5M
Wikipedia (Wiki)	15.91M
RoBERTa	30B

15 times less parameters compared to RoBERTa!

Our work

1. What is the performance for smaller models like BabyBERTa on downstream tasks that require fine-tuning?
2. How to improve the behavior of those models on downstream tasks? Can we use these models as a starting point and continually pre-train the models?

BabyBERTa for Downstream Tasks

Pre-training Recipes

Masking Policy:

- 80-10-10 Masking Policy: 80% are replaced by the <mask> token, 10% are random tokens, and 10% are kept as the same
- **Unmasking Removal Policy:** 90% are replaced by the <mask> token, 10% are random tokens

Vocabulary:

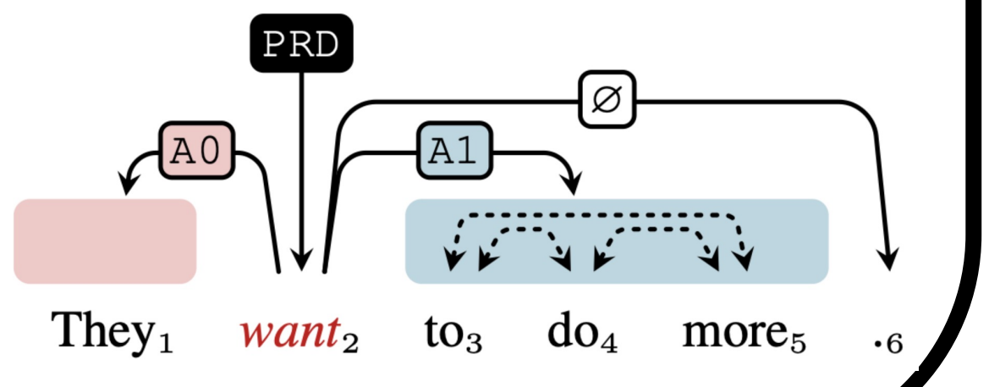
- RoBERTa-Vocabulary Size: 50265
- **BabyBERTa-Vocabulary Size:** 8192 (learned on CHILDES)

Downstream Tasks

Semantic Role Labeling (SRL): assign roles to words in a sentences to recognize the semantic predicate-argument structure

QASRL: use question-answer pairs to label verb predicate-argument structure

QAMR: use question-answer pairs to label predicate-argument structure



Baselines	SRL	QASRL	QAMR
RoBERTa-10M	79.75	90.44	80.76
RoBERTa	85	93.11	90.58

BabyBERTa-CHILDES					BabyBERTa-Wikipedia				
URPS	Vocabulary	SRL	QASRL	QAMR	URPS	Vocabulary	SRL	QASRL	QAMR
yes	RoBERTa	69.47	87.19	53.72	yes	RoBERTa	74.41	89.94	69.61
no	RoBERTa	70.03	86.54	53.57	no	RoBERTa	73.53	89.52	66.26
yes	BabyBERTa	72.38	**87.57**	**54.03**	yes	BabyBERTa	**75.96**	**90.09**	**77.43**
no	BabyBERTa	**72.44**	86.72	53.36	no	BabyBERTa	75.86	89.13	68.7

Note: URPS: apply Unmasked removal policy at starting point.

1. Pre-training with unmasking removal policy and smaller vocabulary set achieves the best performance
2. There is still performance gap between BabyBERTa and RoBERTa

Continually Train BabyBERTa on More Data

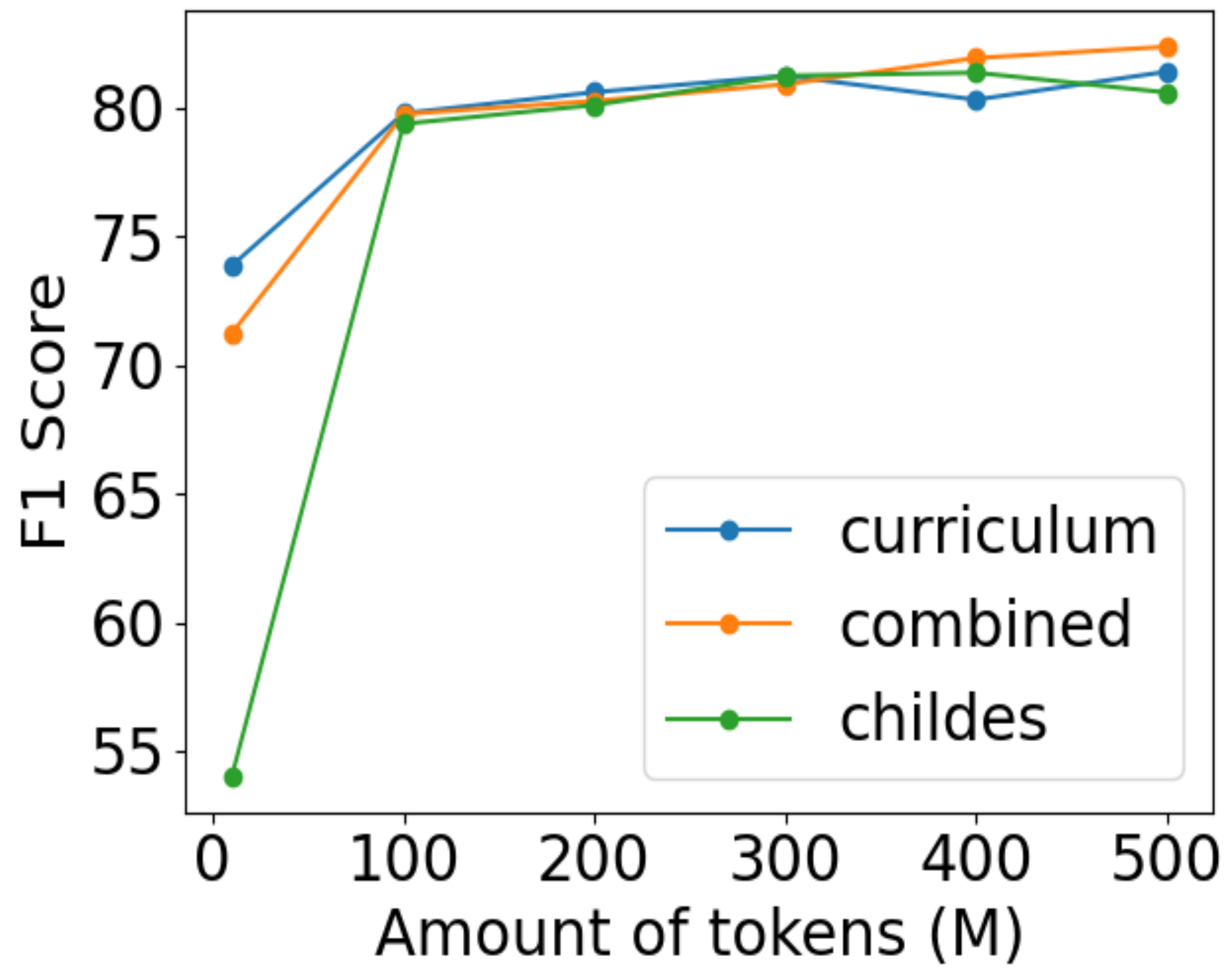
1. Continually pre-train the BabyBERTa models on 100M tokens.

Model	URPS	URPC	SRL	QASRL	QAMR
CHIL	no	no	78.04	90.48	77.6
	yes	no	78.08	90.43	77.88
	yes	yes	**78.19**	**90.56**	**78.6**
Wiki	no	no	77.95	90.4	74.83
	yes	no	78.07	90.78	79.88
	yes	yes	**78.08**	**90.93**	**80.43**

Note: URPS: apply Unmasking removal policy at starting point.
URPC: apply Unmasking removal policy at continual training stage.
We also experience with different continual pre-training datasets in our paper.

The unmasking removal policy at the starting point improves the performance after continual pre-training on downstream tasks such as QAMR.

2. How about continuing training the model on 500M, 1B data?



Model	SRL	QASRL	QAMR
BabyBERTa-Comb	79.4	91.29	82.37
RoBERTa	85	93.11	90.58

1. The performance continually improves as we keep pre-training the model.
2. However, the performance is still lower than that of RoBERTa-base.

Curriculum: CHILDES + Newsela + Wikipedia
Combined: Concatenation of 2 different Wikipedia datasets

Conclusion

- Continually pre-training when using smaller models like BabyBERTa leads to improvement on downstream performance tasks.
- Employing the unmasking removal policy and utilizing a smaller vocabulary prove advantageous for downstream tasks.

References

[1] Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. *BabyBERTa: Learning more grammar with small-scale child-directed language.* (CoNLL' 21)

[2] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. *RoBERTa: A robustly optimized BERT pretraining approach.* arXiv preprint arXiv:1907.11692.

[3] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. *Don't stop pretraining: Adapt language models to domains and tasks.* (ACL' 20)