Don't Stop Pretraining: Adapt Language Models to Domains and Tasks

Suchin Gururangan[†] **Ana Marasović**[†◇] **Swabha Swayamdipta**[†]
Kyle Lo[†] **Iz Beltagy**[†] **Doug Downey**[†] **Noah A. Smith**[†◇]

[†]Allen Institute for Artificial Intelligence, Seattle, WA, USA
[◇]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA
{suching,anam,swabhas,kylel,beltagy,dougd,noah}@allenai.org

March, 11, 2021
구재원

Background: Pretraining

>> Learning for most NLP research systems consists of training in two stages.

LM
(pretrained)

→

Fine-tuning

First, you guys treat me like an object.

And now you want to change(tune) me?!

Background: Pretraining

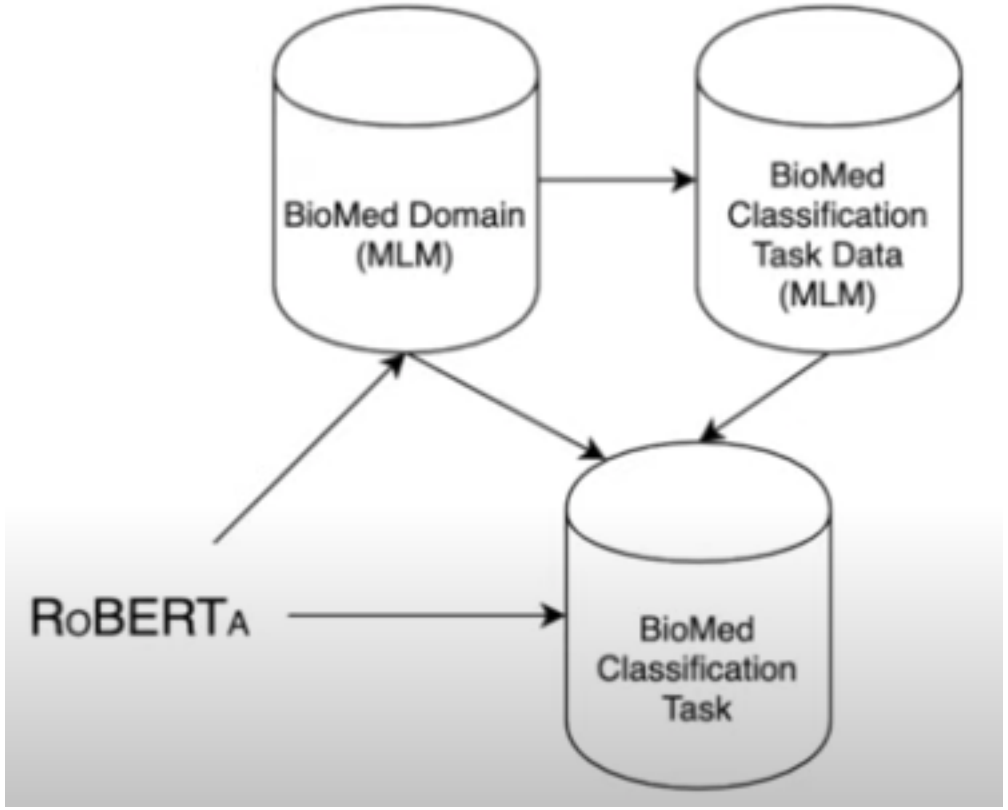
>> RoBERTa is pre-trained on over 160GB of uncompressed text (Encyclopedic, new articles, literary works, web content, …) and attains better performance that its predecessors.

Q1. Do the latest pretrained models work universally?

Q2. Is it still helpful to build separate pretrained models for specific domains?

Continued Pretraining

>> We explores the benefits of continued pretraining on data from the task distribution and the domain distribution



Domains and Data Distributions

>> How does it benefit from…

1. Amount of labeled task data

2. Proximity of target domain to original pretraining corpus

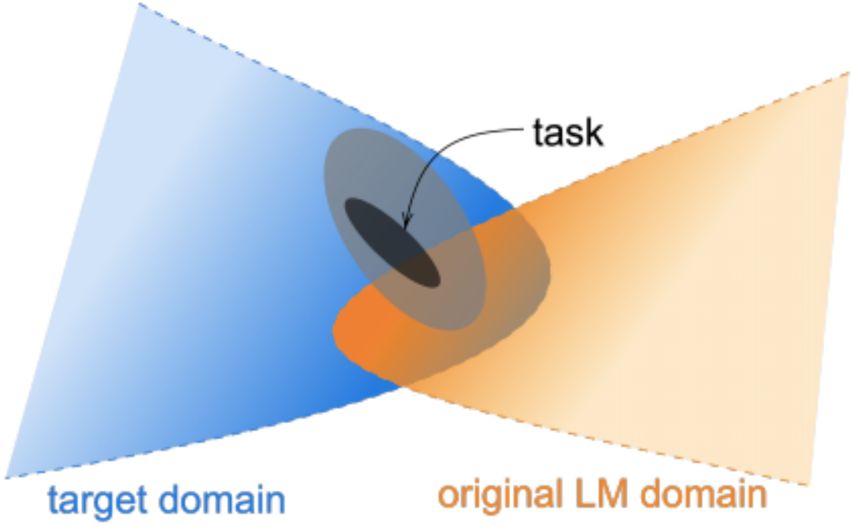


Figure 1: An illustration of data distributions. Task data is comprised of an observable task distribution, usually non-randomly sampled from a wider distribution (light grey ellipsis) within an even larger target domain, which is not necessarily one of the domains included in the original LM pretraining domain – though overlap is possible. We explore the benefits of continued pretraining on data from the task distribution and the domain distribution.

Domains and Data Distributions

>> 4 domains, 8 tasks

Domain	Pretraining Corpus	# Tokens	Size	$\mathcal{L}_{\text{RoB.}}$	$\mathcal{L}_{\text{DAPT}}$
BIOMED	2.68M full-text papers from S2ORC (Lo et al., 2020)	7.55B	47GB	1.32	0.99
CS	2.22M full-text papers from S2ORC (Lo et al., 2020)	8.10B	48GB	1.63	1.34
NEWS	11.90M articles from REALNEWS (Zellers et al., 2019)	6.66B	39GB	1.08	1.16
REVIEWS	24.75M AMAZON reviews (He and McAuley, 2016)	2.11B	11GB	2.10	1.93
ROBERTA (baseline)	see Appendix §A.1	N/A	160GB	‡1.19	-

Table 1: List of the domain-specific unlabeled datasets. In columns 5 and 6, we report ROBERTA's masked LM loss on 50K randomly sampled held-out documents from each domain before ($\mathcal{L}_{\text{RoB.}}$) and after ($\mathcal{L}_{\text{DAPT}}$) DAPT (lower implies a better fit on the sample). ‡ indicates that the masked LM loss is estimated on data sampled from sources *similar* to ROBERTA's pretraining corpus.

Domain	Task	Label Type	Train (Lab.)	Train (Unl.)	Dev.	Test	Classes
BIOMED	CHEMPROT	relation classification	4169	-	2427	3469	13
	†RCT	abstract sent. roles	18040	-	30212	30135	5
CS	ACL-ARC	citation intent	1688	-	114	139	6
	SCIERC	relation classification	3219	-	455	974	7
NEWS	HYPERPARTISAN	partisanship	515	5000	65	65	2
	†AGNEWS	topic	115000	-	5000	7600	4
REVIEWS	†HELPFULNESS	review helpfulness	115251	-	5000	25000	2
	†IMDB	review sentiment	20000	50000	5000	25000	2

Table 2: Specifications of the various target task datasets. † indicates high-resource settings. Sources: CHEMPROT (Kringelum et al., 2016), RCT (Dernoncourt and Lee, 2017), ACL-ARC (Jurgens et al., 2018), SCIERC (Luan et al., 2018), HYPERPARTISAN (Kiesel et al., 2019), AGNEWS (Zhang et al., 2015), HELPFULNESS (McAuley et al., 2015), IMDB (Maas et al., 2011).

Contributions

>> Second phase of pre-training in-domain leads to gains in high and low resource settings

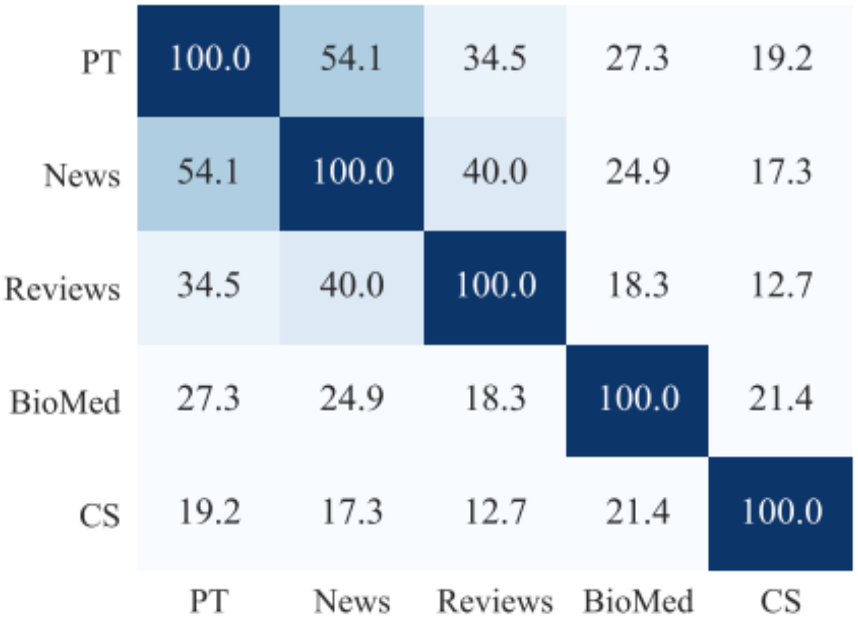
>> Adapting to the task's unlabeled data improves performance even after domain adaptive pretraining

>> unlabeled data가 없을 때 사용할 수 있는 data selection strategy 제안

Domain-Adaptive Pretraining

>> RoBERTa를 unlabeled domain-specific한 large corpus에 추가 pretrain 진행

1. Analyzing Domain Similarity



>> Domain 유사도 분석
>> RoBERTa의 사전학습 도메인과 New와 reviews는 많이 유사함

Domain-Adaptive Pretraining

>> RoBERTa를 unlabeled domain-specific한 large corpus에 추가 pretrain 진행

Dom.	Task	ROBA.	DAPT	¬DAPT
BM	CHEMPROT	$81.9_{1.0}$	**$84.2_{0.2}$**	$79.4_{1.3}$
	†RCT	$87.2_{0.1}$	**$87.6_{0.1}$**	$86.9_{0.1}$
CS	ACL-ARC	$63.0_{5.8}$	**$75.4_{2.5}$**	$66.4_{4.1}$
	SCIERC	$77.3_{1.9}$	**$80.8_{1.5}$**	$79.2_{0.9}$
NEWS	HYP.	$86.6_{0.9}$	**$88.2_{5.9}$**	$76.4_{4.9}$
	†AGNEWS	**$93.9_{0.2}$**	**$93.9_{0.2}$**	$93.5_{0.2}$
REV.	†HELPFUL.	$65.1_{3.4}$	**$66.5_{1.4}$**	$65.1_{2.8}$
	†IMDB	$95.0_{0.2}$	**$95.4_{0.2}$**	$94.1_{0.4}$

Table 3: Comparison of ROBERTA (ROBA.) and DAPT to adaptation to an *irrelevant* domain (¬ DAPT). Reported results are test macro-F_1, except for CHEMPROT and RCT, for which we report micro-F_1, following Beltagy et al. (2019). We report averages across five random seeds, with standard deviations as subscripts. † indicates high-resource settings. Best task performance is boldfaced. See §3.3 for our choice of irrelevant domains.

Domain-Adaptive Pretraining

>> Domain Overlap

	PT	News	Reviews	BioMed	CS
PT	100.0	54.1	34.5	27.3	19.2
News	54.1	100.0	40.0	24.9	17.3
Reviews	34.5	40.0	100.0	18.3	12.7
BioMed	27.3	24.9	18.3	100.0	21.4
CS	19.2	17.3	12.7	21.4	100.0

>> New domain의 DAPT 모델이 Review에서 괜찮았음 (HELPFULNESS: 65.5, IMDB:95.0)

>> [Future Works] domain간의 경계를 벗어난 사전 학습이 유용할 수 있음

Task-Adaptive Pretraining

>> Task-adaptive pretraining(TAPT) 는 task에 대한 unlabeled dataset으로 사전 학습하는 것을 의미

>> DAPT 보다 적은 자원으로 비슷한 효과를 낼 수 있는 효율적인 adaptation 방법

Domain	Task	RoBERTa	Additional Pretraining Phases		
			DAPT	TAPT	DAPT + TAPT
BioMed	ChemProt	$81.9_{1.0}$	$84.2_{0.2}$	$82.6_{0.4}$	$\mathbf{84.4}_{0.4}$
	†RCT	$87.2_{0.1}$	$87.6_{0.1}$	$87.7_{0.1}$	$\mathbf{87.8}_{0.1}$
CS	ACL-ARC	$63.0_{5.8}$	$75.4_{2.5}$	$67.4_{1.8}$	$\mathbf{75.6}_{3.8}$
	SciERC	$77.3_{1.9}$	$80.8_{1.5}$	$79.3_{1.5}$	$\mathbf{81.3}_{1.8}$
News	HyperPartisan	$86.6_{0.9}$	$88.2_{5.9}$	$\mathbf{90.4}_{5.2}$	$90.0_{6.6}$
	†AGNews	$93.9_{0.2}$	$93.9_{0.2}$	$94.5_{0.1}$	$\mathbf{94.6}_{0.1}$
Reviews	†Helpfulness	$65.1_{3.4}$	$66.5_{1.4}$	$68.5_{1.9}$	$\mathbf{68.7}_{1.8}$
	†IMDB	$95.0_{0.2}$	$95.4_{0.1}$	$95.5_{0.1}$	$\mathbf{95.6}_{0.1}$

Table 5: Results on different phases of adaptive pretraining compared to the baseline RoBERTa (col. 1). Our approaches are DAPT (col. 2, §3), TAPT (col. 3, §4), and a combination of both (col. 4). Reported results follow the same format as Table 3. State-of-the-art results we can compare to: ChemProt (84.6), RCT (92.9), ACL-ARC (71.0), SciERC (81.8), HyperPartisan (94.8), AGNews (95.5), IMDB (96.2); references in §A.2.

Task-Adaptive Pretraining

>> Cross task Transfer

BioMed	RCT	ChemProt
TAPT	$87.7_{0.1}$	$82.6_{0.5}$
Transfer-TAPT	$87.1_{0.4}$ ($\downarrow$0.6)	$80.4_{0.6}$ ($\downarrow$2.2)

CS	ACL-ARC	SciERC
TAPT	$67.4_{1.8}$	$79.3_{1.5}$
Transfer-TAPT	$64.1_{2.7}$ ($\downarrow$3.3)	$79.1_{2.5}$ ($\downarrow$0.2)

News	HyperPartisan	AGNews
TAPT	$89.9_{9.5}$	$94.5_{0.1}$
Transfer-TAPT	$82.2_{7.7}$ ($\downarrow$7.7)	$93.9_{0.2}$ ($\downarrow$0.6)

Reviews	Helpfulness	IMDB
TAPT	$68.5_{1.9}$	$95.7_{0.1}$
Transfer-TAPT	$65.0_{2.6}$ ($\downarrow$3.5)	$95.0_{0.1}$ ($\downarrow$0.7)

Table 6: Though TAPT is effective (Table 5), it is harmful when applied *across* tasks. These findings illustrate differences in task distributions within a domain.

Augmenting Training Data for TAPT

>>TAPT의 성능을 바탕으로 task를 위한 dataset과 비슷한 분포의 unlabeled data를 확보할 수 있다는 환경에서 추가적으로 실험을 진행

1. Human Curated-TAPT

Pretraining	BIOMED RCT-500	NEWS HYP.	REVIEWS IMDB †
TAPT	$79.8_{1.4}$	$90.4_{5.2}$	$95.5_{0.1}$
DAPT + TAPT	$83.0_{0.3}$	$90.0_{6.6}$	$95.6_{0.1}$
Curated-TAPT	$83.4_{0.3}$	$89.9_{9.5}$	$95.7_{0.1}$
DAPT + Curated-TAPT	$\mathbf{83.8}_{0.5}$	$\mathbf{92.1}_{3.6}$	$\mathbf{95.8}_{0.1}$

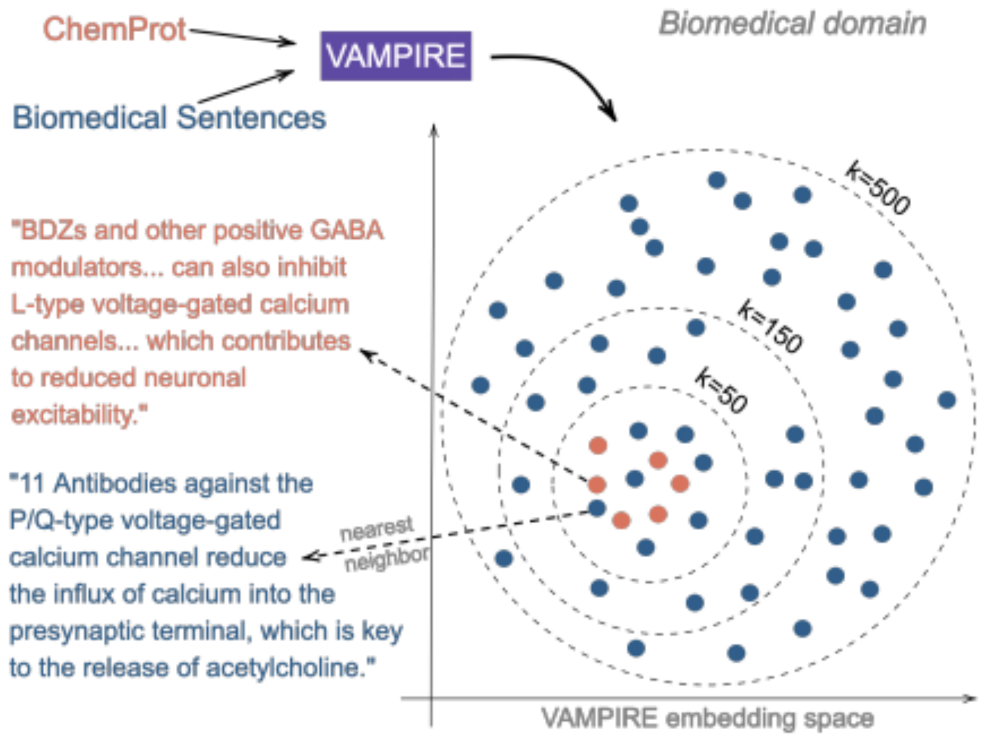
Table 7: Mean test set macro-F_1 (for HYP. and IMDB) and micro-F_1 (for RCT-500), with Curated-TAPT across five random seeds, with standard deviations as subscripts. † indicates high-resource settings.

Augmenting Training Data for TAPT

2. Automated Data Selection for TAPT

: Domain corpus에서 task dataset과 비슷한 unlabeled text를 retrieve하는 방법

: TAPT에 사용한 unlabeled data가 부족하거나 DAPT를 위한 computing resource가 부족한 경우에 효과적인 방법



Pretraining	BioMed		CS
	ChemProt	RCT-500	ACL-ARC
RoBERTa	$81.9_{1.0}$	$79.3_{0.6}$	$63.0_{5.8}$
TAPT	$82.6_{0.4}$	$79.8_{1.4}$	$67.4_{1.8}$
Rand-TAPT	$81.9_{0.6}$	$80.6_{0.4}$	$69.7_{3.4}$
50NN-TAPT	$83.3_{0.7}$	$80.8_{0.6}$	$70.7_{2.8}$
150NN-TAPT	$83.2_{0.6}$	$81.2_{0.8}$	$73.3_{2.7}$
500NN-TAPT	$83.3_{0.7}$	$81.7_{0.4}$	**$75.5_{1.9}$**
DAPT	**$84.2_{0.2}$**	**$82.5_{0.5}$**	$75.4_{2.5}$