Domain adaptation and pre-trained models

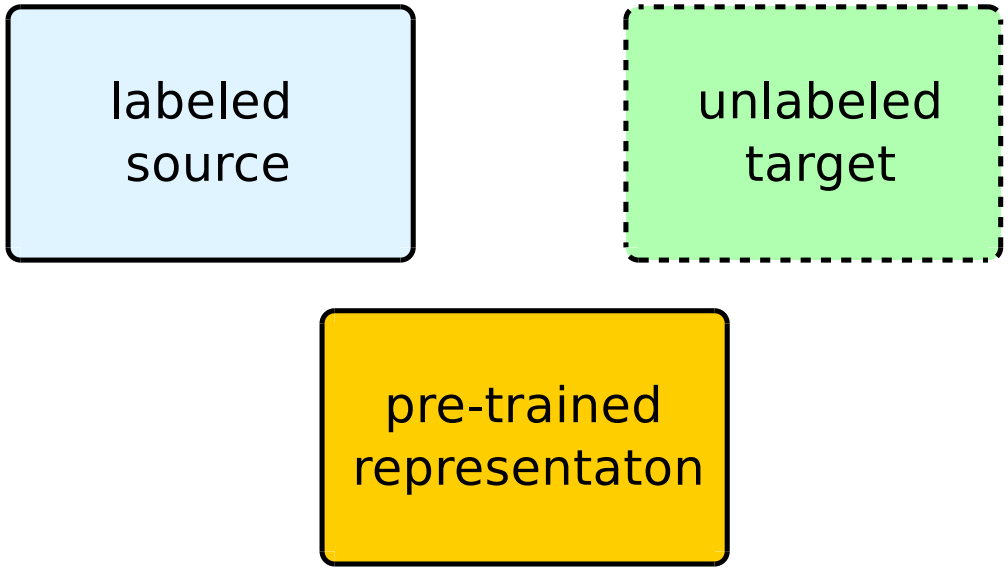
UNIVERSITY OF GOTHENBURG

CHALMERS

Richard Johansson

richajo@chalmers.se

unsupervised DA
+ representation



domain-specific pre-trained models

if you can find a representation trained to your domain, **use it**!

for instance, **SciBERT** (Beltagy et al., 2019) is a BERT-like model trained on scientific papers

SciBERT: A Pretrained Language Model for Scientific Text

Iz Beltagy **Kyle Lo** **Arman Cohan**
Allen Institute for Artificial Intelligence, Seattle, WA, USA
{beltagy,kylel,armanc}@allenai.org

there are also models for finance, clinical notes, patents, Twitter, ...

target domain fine-tuning

alternatively, we may **fine-tune an general pre-trained model to a target domain**

for instance, **BioBERT** (Lee et al., 2019) is a regular BERT model fine-tuned to the biomedical domain

Bioinformatics, 2019, 1–7
doi: 10.1093/bioinformatics/btz682
Advance Access Publication Date: 10 September 2019
Original Paper
OXFORD

Data and text mining

BioBERT: a pre-trained biomedical language representation model for biomedical text mining

Jinhyuk Lee [1,†], Wonjin Yoon [1,†], Sungdong Kim [2], Donghyeon Kim [1], Sunkyu Kim [1], Chan Ho So [3] and Jaewoo Kang [1,3,*]

[1]Department of Computer Science and Engineering, Korea University, Seoul 02841, Korea, [2]Clova AI Research, Naver Corp, Seong-Nam 13561, Korea and [3]Interdisciplinary Graduate Program in Bioinformatics, Korea University, Seoul 02841, Korea

better to train from scratch?

Jansson (2019) investigates BERT models for patent clause classification and finds that in this case,

custom pre-training > domain fine-tuning > no domain model

more domain and task fine-tuning

Gururangan et al. (2020) investigate several ways to fine-tune RoBERTa models

they showed that even small amounts of target data can be useful

if target domain data is not available for fine-tuning, it can be crawled

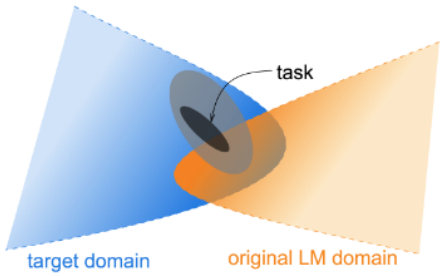
Don't Stop Pretraining: Adapt Language Models to Domains and Tasks

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

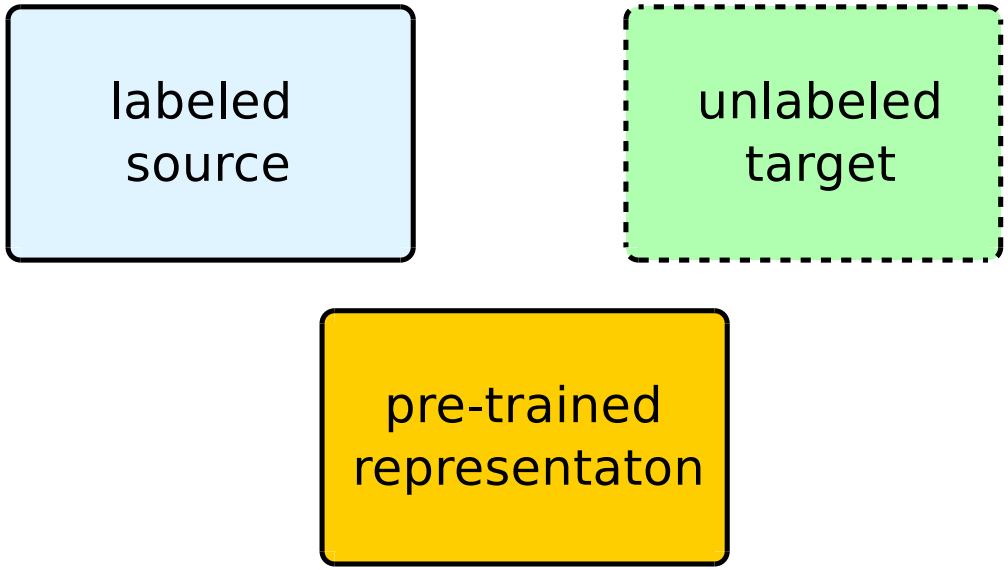
[†]Allen Institute for Artificial Intelligence, Seattle, WA, USA
[◇]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA
{suching,anam,swabhas,kylel,beltagy,dougd,noah}@allenai.org

Abstract

Language models pretrained on text from a wide variety of sources form the foundation of today's NLP. In light of the success of these broad-coverage models, we investigate whether it is still helpful to tailor a pretrained model to the domain of a target task. We present a study across four domains (biomedical and computer science publications, news, and reviews) and eight classification tasks, showing that a second phase of pretraining in-



unsupervised DA
+ representation



references

I. Beltagy, K. Lo, and A. Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *EMNLP*.

S. Gururangan, A. Marasovic; S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *ACL*.

E. Jansson. 2019. Domain adapted language models. Master's Thesis, Chalmers University of Technology.

J. Lee, W. Yoon, and S. Kim et al. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36(4):1234–1240.