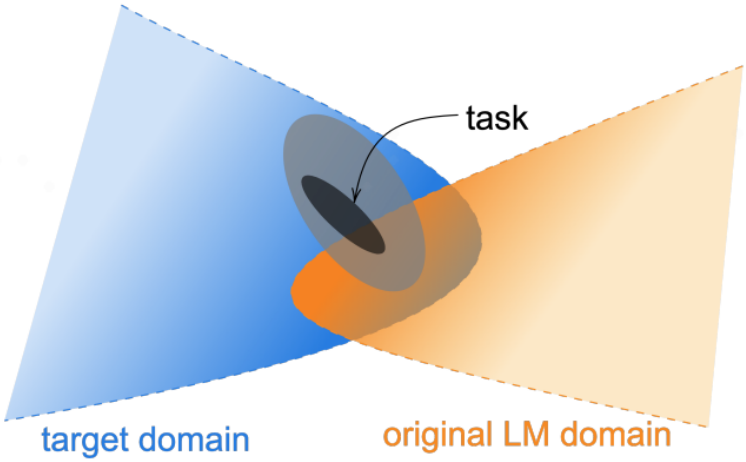
# Further Pretrain 简介

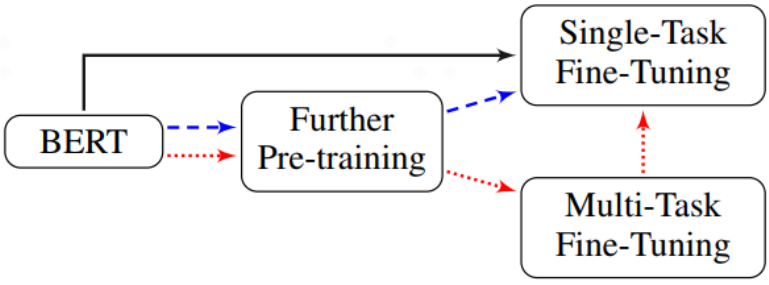
作者：纽约的自行车

日期：2022 年 10 月 10 号

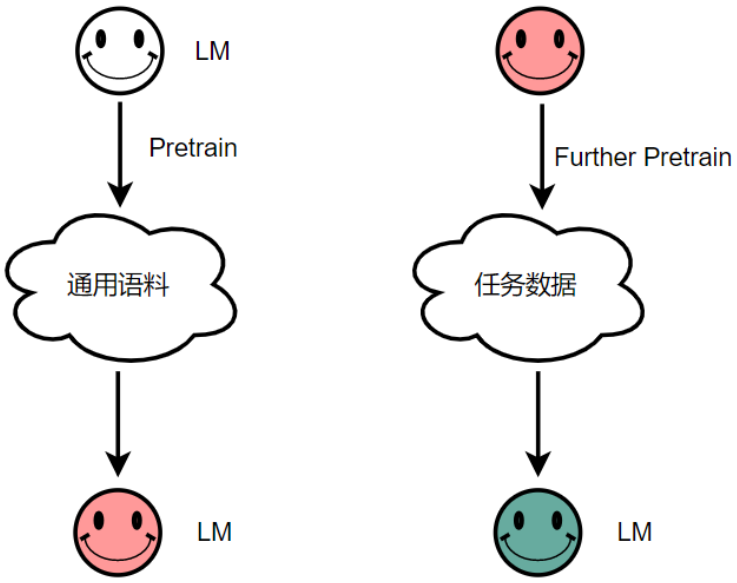
训练语言模型使用的通用语料通常与下游任务语料在领域上不同，如下图所示，黄色部分表示通用语料覆盖的领域，黑色部分表示下游任务数据覆盖的领域，灰色部分表示任务数据扩充所覆盖的领域，蓝色部分表示任务所在领域。下游任务数据和通用语料之间存在差异，那么是否有必要将语言模型迁移到下游任务领域上?



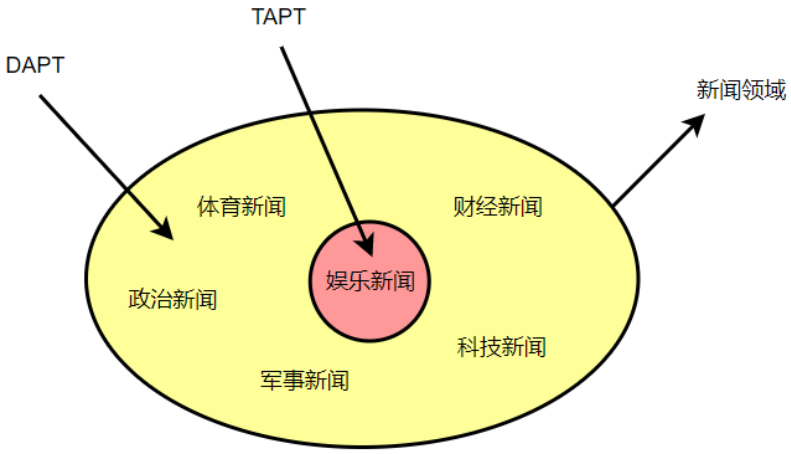

2019 年复旦大学邱锡鹏教授团队针对这个问题，尝试了 further pretrain [1]，并在文本分类任务上进行验证。相比用 Bert 直接在任务数据上 Finetune（下图黑色路径），作者增加了一个 Further Pretrain 步骤（下图蓝色路径），即在任务数据上继续做预训练，然后再 Finetune，使得流程由原来的两阶段变成三阶段。




什么是 further pretrain? further pretrain 是 pretrain 的延续，further pretrain 与 pretrain 的区别是换了个数据集，并且 further pretrain 使用的模型是在大规模语料上训练后的预训练模型，further pretrain 的训练任务可以和 pretrain 的训练任务相同，也可以不同，最常用的 further pretrain 训练任务是 mlm。




2020 年 Suchin Gururanga [2] 在 ACL 会议上发表了一篇主题相同的论文，论文中作者将 Further Pretrain 分为 Domain-Adaptive Pretraining（DAPT）和 Task-Adaptive Pretraining（TAPT），并做了更详细的实验论证。DAPT（Domain-adaptive pretraining）是在任务数据所属领域的语料上进行 further pretrain，该方法适用于任务数据量非常少的场景。TAPT（Task-adaptive pretraining）是直接在任务数据上进行 further pretrain，适用于任务数据量较大的场景。下图展示了 DAPT 和 TAPT 的区别。假设下游任务：娱乐新闻情感分类。DAPT 是指在新闻领域语料上做 further pretrain，但是需要排除娱乐新闻（下游任务数据）。TAPT 是指在下游任务数据上做 further pretrain。




作者选取低资源数据集和高资源数据集，下表中第二列左上角有十字标记的表示高资源。八个数据集分成四个领域，每个领域两个数据集，全部都是文本分类任务。

下表是实验结果，DATP+TAPT 是三阶段方式，使用 roberta 模型，现在相关语料上做 DAPT，然后在任务数据上做 TAPT，最后 finetune。虽然最终效果相比 DAPT 和 TAPT 确有提升，但是训练耗时长。

实验结论：

1. 在大数据集上 TAPT 平均提升 1.25%，在小数据集上 TAPT 平均提升 2.73%；

| Domain | Task | Label Type | Train (Lab.) |
|---|---|---|---|
| BIOMED | CHEMPROT | relation classification | 4169 |
| | †RCT | abstract sent. roles | 18040 |
| CS | ACL-ARC | citation intent | 1688 |
| | SCIERC | relation classification | 3219 |
| NEWS | HYPERPARTISAN | partisanship | 515 |
| | †AGNEWS | topic | 115000 |
| REVIEWS | †HELPFULNESS | review helpfulness | 115251 |
| | †IMDB | review sentiment | 20000 |

| | | | Additional Pretraining Phases | | |
|---|---|---|---|---|---|
| **Domain** | **Task** | **ROBERTA** | **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 | $95.0_{0.2}$ | $95.4_{0.1}$ | $95.5_{0.1}$ | $\mathbf{95.6}_{0.1}$ |

2. 在大数据集上 DAPT 平均提升 0.55%，在小数据集上 DAPT 平均提升 4.95%；
3. 在大数据集上 TAPT 优于 DAPT，在小数据集上 DAPT 优于 TAPT；
4. 在大数据集上 DAPT+TAPT 相比 TAPT 有略微提升；
5. 在 roberta 的训练语料领域 news 上，TAPT 也有明显提升；

## 参考文献


[1] How to Fine-Tune BERT for Text Classification? [J]. 2019.

[2] Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks[J]. 2020.