Using Synthetic Data for Domain Adaptation of Language Models

NLU Project Partial Draft

Shubhankar Ranade **Akash Gupta** **Abhinav Gupta** **Victor Yu Cui**
New York University
{shubhankar.r, aksg, gupta.abhinav, vyc8567}@nyu.edu

Abstract

In recent studies, Pretrained Language Models (PLMs) have demonstrated remarkable human level performance across multiple downstream tasks when fine-tuned on a large amount of task-specific training data. But getting large amount of data is often expensive and time consuming, and whatever annotated target data is available may not be of the same domain as source data which was used to train the pretrained language model. In this paper, we utilize the few-shot learning capabilities of a large PLM to augment low-resource datasets to fine tune a domain adapted pretrained model (Gururangan et al.) on synthetically generated data. We finally compare our model's performance with several baseline models like RoBERTa and DAPT.

1 Introduction

Natural Language Processing (NLP) research these days has shifted from designing task-specific architectures to using large scale architectures pretrained on multiple huge, general domain corpora. Pretrained language models (PLMs) (Brown et al., 2020, Devlin et al., 2019, Liu et al., 2019) have demonstrated remarkable human level performance across multiple downstream tasks when fine-tuned on a large amount of task-specific training data. (Wang et al., 2018, Howard and Ruder, 2018). However, procuring such a large amount of data is often expensive and time consuming. Therefore, it becomes important to develop PLMs that can extract maximum information from few labeled data samples.

In recent studies, it has been revealed that PLMs exhibit intriguing few-shot learning potential (Radford et al., 2019; Brown et al., 2020; Gao et al., 2021; Scao and Rush, 2021) and are able to leverage task-specific information. But when the target task is available in low quantity and has a different domain than the PLM's source data (i.e. pretraining dataset), then performing well on downstream task becomes extremely difficult. Domain Adaptation as well as Data Augmentation techniques are widely used for tackling such issues.

Work done by Gururangan et al. and Lee et al., 2019 shows that additionally pretraining PLM on domain-specific corpus can help PLMs to adapt well on downstream tasks. However, these DAPT models require fine-tuning on moderately sized task-specific datasets, which can be scarce in many scenarios. This motivates us to look forward to several data augmentation methods and utilize these DAPT models. Unidirectional PLMs have demonstrated strong text generation power (Brown et al., 2020.) One of the ways to do controlled generation is prompt-based approach, where desired context is provided to the generator model through carefully designed prompts (Schick and Schütze, 2020). Building on this prompt-based approach, Meng et al., 2022 (SuperGen) explored the possibility of creating synthetic task-specific training data using a unidirectional PLM (e.g. CTRL Keskar et al., 2019a and Radford et al., 2019) as generator, then fine-tune a bidirectional PLM as classifier on the target task using the generated synthetic dataset.

Our work[1] relies directly on the SuperGen pipeline, but with the motivation to tackle the lack of task-specific data, we explore its potential in boosting performance of DAPT models (as in Gururangan et al.) and replacing expensive task-specific data. We decided to feed few-shot prompts to the generator because it can provide more contexts of the task domain to the generator thus steer the generated texts closer to the domain of the task (Brown et al., 2020; Meng et al., 2022).

[1] https://github.com/ag2307/NLU-Project

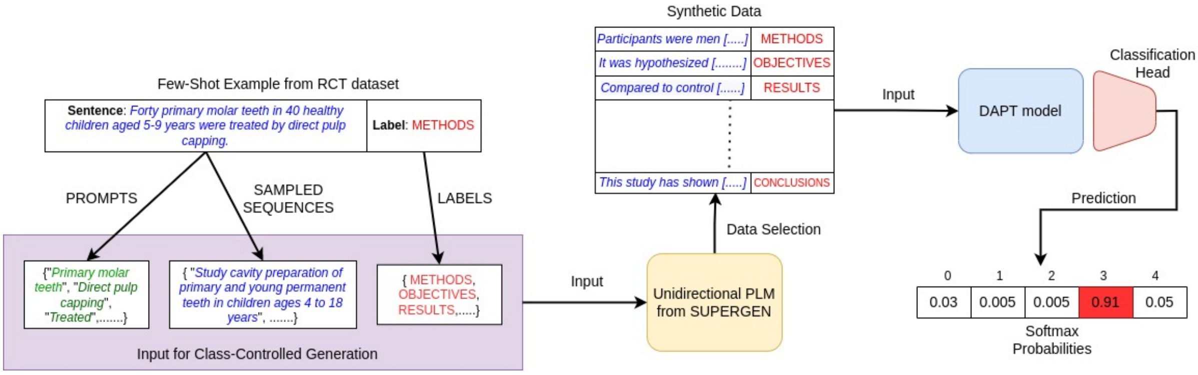

Figure 1: Pipeline for fine-tuning a DAPT model on synthetic data. Few-Shot examples from the RCT dataset are used to get prompts, sampled sequences and labels. Then a Unidirectional PLM from SuperGen generates synthetic examples which are used in the DAPT model fine-tuning. A feedforward layer is attached on to the DAPT for classification

2 Related Work

2.1 Domain Adaptation

Domain adaptation (Ganin et al., 2015, Tzeng et al., 2017, Khaddaj and Hajj, 2020) is a type of transfer learning, which aims to bridge the gap between the source and target domains by learning domain-invariant feature representations. Domain Adaptation methods allow us to fine-tune models using easily accessible data that is not in the target domain of the task and allow models to generalize better on out-of-domain data. In recent years, many such adaptive techniques (Daumé III; Ramponi and Plank, 2020) with data selection strategies (Gururangan et al.) and representation learning (Ganin et al., 2015), have demonstrated good performance in fine-tuning PLMs without sufficient data in target domain. We closely follow the work done by Gururangan et al.. and keep it as one of our baseline models, since they show that domain adaptive pretraining improves performance.

2.2 Few-shot Learning

Modern day PLMs depend on a large amount of labeled and unlabeled data for solving NLU tasks. In comparison, few-shot learning methods try to make the best use of a small amount of task-specific training data, which is a more realistic scenario. We generate data in a true few-shot setting (Perez et al., 2021), which means we don't have access to any unlabeled data or large held-out datasets for prompt and hyperparameter tuning. In this scenario, to improve data efficiency, prompt based methods are largely used (Brown et al., 2020, Gao et al., 2021, Liu et al., 2021a, Schick and Schütze, 2021a). Many approaches have used a meta-learning strategy to tackle data sparsity (Finn et al., 2017, Mishra et al., 2018). These approaches assume access to some auxiliary tasks, which helps them extract some transferable knowledge to learn the target problem. We use prompt based methods in a true few-shot learning setting.

2.3 Controlled Data Generation

Controlled text generation (Hu et al., 2017) aims to generate text that follows the distribution of the task dataset. PLMs can be made to generate desired sentiments or topics, also called high-level control (Ziegler et al., 2019), or they can be made to generate specific words or phrases (Chan et al., 2021), also called low-level control. Both high-level and low-level control can also be achieved together (Khalifa et al., 2021). It is also possible to do controlled text generation at inference time, without further training the PLM (Dathathri et al., 2020; Krause et al., 2021; Kumar et al., 2021) Natural language prompts can be used as a guidance source for the model to generate text with desired attributes (Schick and Schütze, 2021b). We use prompts to guide text generation in this work.

3 Methods

In this work, we begin with synthetic data generation for fine-tuning on a task. We assume access to a unidirectional PLM $\mathcal{G}$ (Meng et al., 2022). The technique uses carefully crafted prompts for different downstream tasks to generate sequences. The PLM generates augmented task dataset $\mathcal{T}$ which, after cleaning, is used to fine-tune a domain adapted PLM $\mathcal{D}$.

We plan to compare the baselines ROBERTA model and DAPT, with a DAPT model (Gururangan et al.) fine tuned using SuperGen (Meng et al., 2022) in a few-shot setting on 8 different tasks. The above helps to discern about how much does training on synthetic data improves performance of a domain adapted model. Another comparison of interest for this study is between ROBERTA model fine-tuned using SuperGen and the DAPT model. This tells us about how does training on only task-specific synthetic data compare to a domain adapted model.

3.1 Task-specific Data

We evaluate the performance of our proposed approach in 8 different classification tasks across 4 domains, two tasks in each domain (Gururangan et al.). We use two datasets that are in the biomedical domain: ChemProt (Kringelum et al., 2016 and PubMed 200k RCT (Dernoncourt and Lee); Two in Computer Science domain: ACL articles on NLP (ACL-ARC) (Jurgens et al., 2018) and SciERC (Luan et al., 2018; Two in News articles domain: HyperPartisan News (Kiesel et al., 2019 and AGNews corpus (Zhang et al., 2015); And two in Movie Reviews domain: Helpfulness (McAuley et al., 2015) and IMDB reviews (Maas et al., 2011).

3.2 Data Generation

For each task, we randomly sample from task-specific dataset, and use these sampled data to form prompts which then fed into generator $\mathcal{G}$ to produce synthetic data. We take advantage of the few-shot examples which are used to sample these prompts, sequences and labels. Labels are passed along with prompts for class-controlled generation. The generated data then is selected based on the ranking score as stated in (Meng et al., 2022, Yuan et al., 2021) to form the synthetic training data, which later be used to fine-tune classification PLM. Due to space constraint, we do not restate SuperGen algorithms here, and treat it as black-box. Detail descriptions can be accessed in (Meng et al., 2022). However, we still fine-tune hyper-parameters including repetition penalties, temperature, and generating size. Repetition penalties prevent repetition loops in generation and controls beneficial token repetitions in sequence pair generation, temperature. Temperature varies the sharpness of generating distribution. The generating size involves trade-off between generating consumption and quality selection, given the data-selection technique remains the same.

3.3 Models for Task-specific Fine-tuning

- **Baseline** For each task, we use the DAPT model in Gururangan et al. that has been pretrained using data from the task domain, and fine-tuned on the complete task specific datasets as our baseline models. This DAPT model serves as good baseline model as it was proven to have close to state-of-art performance on the tasks with or without further pretraining, and it allows us to clearly examine the SuperGen technique.

- **SuperGen Fine-tuned** Instead of fine-tuning DAPT model with the complete task-specific dataset as in baseline, we fine-tune DAPT using synthetic data generated by SuperGen. We compare

Domain	Tasks	RoB$_A$	DAPT
BM	CHEMPROT	81.1	81.8
	RCT	86.6	87.5
CS	ACL-ARC	61.0	69.4
	SCIERC	72.3	78.1
NEWS	HYP.	85.0	86.0
	AGNEWS	92.9	93.1
REV.	HELPFUL.	65.0	65.7
	IMDB	93.4	**95.3**

Table 1: Results of the baseline models ROBERTA and DAPT on 4 different domains each containing 2 tasks. The scores reported are macro-F1. Best task performance is boldfaced.

4 Experiments

As described in section 3.1, we plan on evaluating our approach on task data used in Gururangan et al.. We use macro-F1 score for evaluating performance of our approach on classification tasks. The results for the baseline models ROBERTA and textscdapt are shown in Table 1. Following the implementation in Meng et al., 2022, we use CTRL (Keskar et al., 2019a) as our generator PLM $\mathcal{G}$. However, we will use the domain adapted model DAPT (Gururangan et al.) as the discriminator since that is our baseline.

5 Collaboration Statement

All members worked together on brainstorming the idea, held weekly discussions and drafted the report. For evaluations, Akash worked on getting the results for IMDB and RCT datasets. Shubhankar worked on CHEMPROT and SCIERC. Abhinav worked on ACL-ARC and AGNEWS and Victor worked on HELPFULNESS and HYPERPARTISAN.

References

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, et al. Don't stop pretraining: Adapt language models to domains and tasks.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, et al. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *arXiv e-prints*, art. arXiv:1804.07461, April 2018.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *ACL*, 2018.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *ArXiv*, abs/2012.15723, 2021.

Teven Le Scao and Alexander M. Rush. How many data points is a prompt worth? In *NAACL*, 2021.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz682. URL `https://doi.org/10.1093/bioinformatics/btz682`.

Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. *CoRR*, abs/2009.07118, 2020. URL `https://arxiv.org/abs/2009.07118`.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. Generating training data with language models: Towards zero-shot language understanding. 02 2022.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation, 2019a. URL `https://arxiv.org/abs/1909.05858`.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. 2015. doi: 10.48550/ARXIV.1505.07818. URL https://arxiv.org/abs/1505.07818.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971, 2017.

Alaa Khaddaj and Hazem M. Hajj. Representation learning for improved generalization of adversarial domain adaptation with text classification. *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, pages 525–531, 2020.

Hal Daumé III. Frustratingly easy domain adaptation.

Alan Ramponi and Barbara Plank. Neural unsupervised domain adaptation in nlp—a survey, 2020.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. True few-shot learning with language models. *ArXiv*, abs/2105.11447, 2021.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *ArXiv*, abs/2103.10385, 2021a.

Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *EACL*, 2021a.

Chelsea Finn, P. Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and P. Abbeel. A simple neural attentive meta-learner. In *ICLR*, 2018.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. Toward controlled generation of text. In *ICML*, 2017.

Daniel M. Ziegler, Nisan Stiennon, Jeff Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *ArXiv*, abs/1909.08593, 2019.

Alvin Chan, Y. Ong, Bill Tuck Weng Pung, Aston Zhang, and Jie Fu. Cocon: A self-supervised approach for controlled text generation. *ArXiv*, abs/2006.03535, 2021.

Muhammad Khalifa, Hady ElSahar, and Marc Dymetman. A distributional approach to controlled text generation. *ArXiv*, abs/2012.11635, 2021.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. *ArXiv*, abs/1912.02164, 2020.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq R. Joty, Richard Socher, and Nazneen Rajani. Gedi: Generative discriminator guided sequence generation. *ArXiv*, abs/2009.06367, 2021.

Sachin Kumar, Eric Malmi, Aliaksei Severyn, and Yulia Tsvetkov. Controlled text generation as continuous optimization with multiple constraints. *ArXiv*, abs/2108.01850, 2021.

Timo Schick and Hinrich Schütze. Few-shot text generation with natural language instructions. In *EMNLP*, 2021b.

Jens Kringelum, Sonny Kim Kjaerulff, Søren Brunak, Ole Lund, Tudor I. Oprea, and Olivier Taboureau. Chemprot-3.0: a global chemical biology diseases mapping. *Database : the journal of biological databases and curation*, 2016:bav123, Feb 2016. ISSN 1758-0463. doi: 10.1093/database/bav123. URL https://pubmed.ncbi.nlm.nih.gov/26876982. 26876982[pmid].

Franck Dernoncourt and Ji Young Lee. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406, 2018. doi: 10.1162/tacl_a_00028. URL https://aclanthology.org/Q18-1028.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1360. URL https://aclanthology.org/D18-1360.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2145. URL https://aclanthology.org/S19-2145.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification, 2015. URL https://arxiv.org/abs/1509.01626.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. Image-based recommendations on styles and substitutes, 2015. URL https://arxiv.org/abs/1506.04757.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1015.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Paper.pdf.

Galen Andrew and Jianfeng Gao. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40, 2007.

Mohammad Sadegh Rasooli and Joel R. Tetreault. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733, 2015. URL http://arxiv.org/abs/1503.06733. version 2.

Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, December 2005. ISSN 1532-4435.

Alexey Romanov and Chaitanya P. Shivade. Lessons from natural language inference in the clinical domain. In *EMNLP*, 2018.

Nikolaos Malandrakis, Minmin Shen, Anuj Goyal, Shuyang Gao, Abhishek Sethi, and Angeliki Metallinou. Controlled text generation for data augmentation in intelligent artificial agents, 2019. URL https://arxiv.org/abs/1910.03487.

Ningyu Zhang, Luoqiu Li, Xiang Chen, Shumin Deng, Zhen Bi, Chuanqi Tan, Fei Huang, and Huajun Chen. Differentiable prompt makes pre-trained language models better few-shot learners. *ArXiv*, abs/2108.13161, 2021.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. *ArXiv*, abs/1909.05858, 2019b.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. Dexperts: Decoding-time controlled text generation with experts and anti-experts. In *ACL*, 2021b.

Damian Pascual, Béni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. A plug-and-play method for controlled text generation. *ArXiv*, abs/2109.09707, 2021.

Kevin Yang and Dan Klein. Fudge: Controlled text generation with future discriminators. *ArXiv*, abs/2104.05218, 2021.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, abs/2101.00190, 2021.

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. 2019. doi: 10.48550/ARXIV.1903.10676. URL https://arxiv.org/abs/1903.10676.