Optimizing T+DAPT: Effective, Domain-Specific, Zero-Shot Question Answering without Labeled Target Data

Hisham Aziz and **Katherine Voss-Robinson**
University of California, Berkeley
hisham.aziz@berkeley.edu and kvossrobinson@berkeley.edu

Abstract

Over the past several years, pre-trained language models have driven substantial performance gains across many natural language processing (NLP) tasks. However, these models have historically required large quantities of labeled data to effectively learn downstream tasks. This limits the utility of pre-trained models in domains lacking sufficient labeled data. In a 2022 paper, Pan et al. introduced one possible methodology to remedy this called Task and Domain Adaptive Pretraining (T+DAPT). This technique combines task transfer from named entity recognition (NER) to question answering (QA) with domain adaptation to fine-tune a pre-trained RoBERTa model without labeled data from the target task. We seek to further improve upon the success of T+DAPT by employing architectural modifications, early stopping, L2 regularization, and continuous learning. Our approach outperforms both a generic QA baseline and the T+DAPT results from Pan et al.

1 Introduction

Despite the prevalence today of effective pre-trained language models trained on large, diverse corpora (like RoBERTa (Liu et al., 2019)), domain-adaptive pre-training consistently improves performance on target domain tasks (Gururangan et al., 2020). Domain adaptation is a type of transfer learning that aims to ensure that pre-trained models generalize well into specific domains of interest (Farahani et al., 2020). One effective method for implementing domain adaptation involves fine-tuning a pre-trained model on labeled data from the target domain. However, many domains of interest are lacking in the labeled data required for this, which is time and resource intensive to generate.

Pan et al. (2022) proposed one possible solution for cases in which generic training data for the target task and in-domain training data for another task are available but in-domain training data for the target task is not. Their method, Task and Domain Adaptive Pretraining (T+DAPT), drove robust performance improvement in zero-shot domain-specific QA across several domains over models that used general large-scale fine-tuning. In their study, Pan et al. suggest that T+DAPT's performance could be enhanced with modifications to minimize catastrophic forgetting and optimize knowledge transfer. In this paper, we aim to do exactly this by 1) exploring architectural improvements to better facilitate learning from diverse data sources and 2) implementing regularization techniques, early stopping, and continuous learning to prevent overfitting to any specific domain or task and combat catastrophic forgetting.

We evaluate the effectiveness of our modifications on T+DAPT in zero-shot domain specific machine reading comprehension (MRC) using F1 score in alignment with the testing and evaluation methodology of Pan et al. We will measure our results against two baselines: RoBERTa trained on the Stanford Question Answering Dataset (SQuAD) (the same baseline utilized by Pan et al.) and the T+DAPT model proposed by Pan et al. We summarize our contributions as follows:

- Custom implementation of T+DAPT to allow for further architectural modifications;
- Implementation of L2 regularization in both the NER and QA steps of T+DAPT;
- Implementation of early stopping in the domain-specific NER step; and
- Implementation of continuous learning by fine-tuning on multiple source domains.

2 Background

Today, most NLP research consists of two-stage training. In the first, a large neural language model is trained on large, unlabeled corpora to learn word representations. This creates a pre-trained model

that the second stage uses in supervised training for a downstream task. The second stage may also fine-tune the representations generated in the first stage (Gururangan et al., 2020). RoBERTa, developed by Liu et al. (2019), is one such pre-trained language model. Though RoBERTa was trained on a massive, diverse corpus, studies have shown that RoBERTa's performance on specific domains can be improved by continuing to fine-tune it on task or domain specific datasets.

This fine-tuning can take multiple forms depending on data availability in the target domain. Some, like Shakeri et al. (2020) found success generating synthetic domain-specific data to use in fine-tuning. Others have attempted to make do without labeled data. Gururangan et al. (2020) proposed Domain-Adaptive Pretraining (DAPT), in which they pre-train RoBERTa on a large corpus of unlabeled domain-specific text from four domains. Zhang et al. (2020) proposed another method, creating auxiliary synthetic tasks to help the language model transfer to downstream tasks. Pan et al. (2022) created yet another, using task transfer and labeled data from the target domain but a non-target task (NER, chosen because labeled NER data is widely available across domains). These studies collectively illuminate the path forward in the ongoing evolution of pre-trained large language models, advocating for the nuanced customization of models to fit the unique requirements of specific tasks and domains. As this body of work expands, it continuously enriches the toolkit available to researchers and practitioners aiming to harness the full potential of pre-trained models in diverse NLP applications, specifically for low resource domains.

Though these methods have been largely successful, many suffer from a phenomenon called *catastrophic forgetting* (in which a model forgets previously learned information upon learning new information). T+DAPT in particular relies on fine-tuning RoBERTa first on a domain-specific NER task and then on a generic QA task, making it susceptible to catastrophic forgetting of the target domain information. Xu et al. (2020) proposed several potential methods to combat this forgetting, including regularization via EWC, L2, and cosine distance. In our paper, we explore these and other methods to combat catastrophic forgetting in T+DAPT and improve its overall performance.

3 Methods

We aim to improve T+DAPT's zero-shot domain-specific QA performance through a combination of architectural improvements, regularization, early stopping, and continuous learning. We initialize all of our models (RoBERTa baseline, Pan et al. T+DAPT baseline, and T+DAPT with our improvements) with pre-trained RoBERTa weights.

3.1 Datasets

To allow for direct comparison with Pan et al.'s results, we employ some of the same datasets that they did in their paper. For generic QA training, we use the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016). Pan et al. tested the efficacy of T+DAPT across four domains, but we explore just one here: film. We conduct domain-specific NER training with labeled data from the MIT movie corpus (MIT) and employ labeled data from news and biomedical corpora ((Tjong Kim Sang and De Meulder, 2003); (Doğan et al., 2014)) for continuous learning. To test the efficacy of our model improvements on zero-shot question answering, we use labeled QA film data from Xu et al. (2020).

3.2 Experiments

3.2.1 Architectural Improvements

Our first objective is to implement T+DAPT for ourselves. In T+DAPT, Pan et al. initialized an instance of pre-trained RoBERTa, fine-tuned it on domain-specific NER, and then further fine-tuned that on generic QA (see Figure 1). In their NER and QA tasks, Pan et al. made exclusive use of AutoModels from Hugging Face's Transformers library (Wolf et al., 2020). To implement transfer learning between NER and QA, they used the instance of RobertaForTokenClassification that they had fine-tuned on domain-specific NER combined with a config file designed for QA. We have two concerns about the efficacy of their methodology. First, by using pre-trained AutoModels in their generic QA fine-tuning step, they overwrite any weights learned in the domain-specific NER fine-tuning step, erasing the model's learnings from this part of the process. Second, by feeding a token classification model with a QA config file into their QA step, they do not use QA-specific architecture.

To overcome these concerns, we create a custom implementation for T+DAPT. First, we construct our own instances of RoBERTa for token classifica-

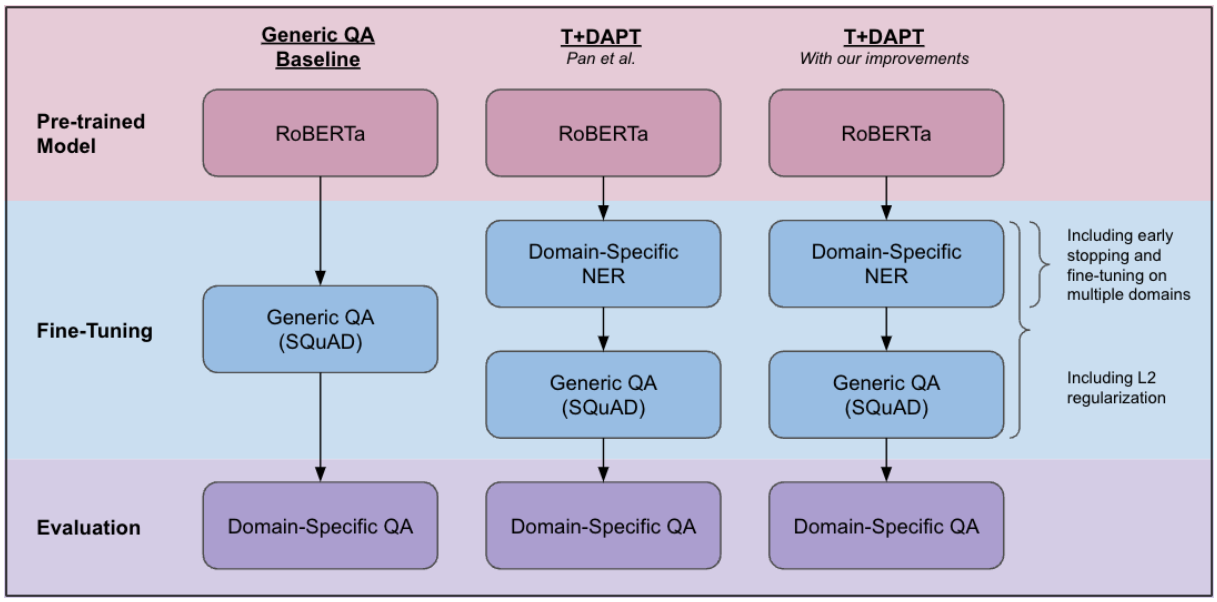

Figure 1: Sequential transfer learning procedures of our generic QA baseline, our T+DAPT baseline from Pan et al., and T+DAPT with our improvements

tion and question answering (rather than importing Hugging Face's existing AutoModels) so that we can make modifications to RoBERTa's architecture. This flexibility is essential to the rest of our experiments. Second, when implementing transfer learning between NER and QA, we initialize a custom instance of RobertaForQuestionAnswering and import only the model weights learned during our NER step. Through these steps, we adhere to the foundational principles behind T+DAPT but implement them in a novel way intended to improve performance. Like Pan et al., we build our implementation upon a foundation of Hugging Face code (Wolf et al., 2020), but unlike Pan et al., we customize it to facilitate granular architecture modifications and to ensure that we are effectively transferring the information our model learned from one step to the next.

3.2.2 Early Stopping

As an initial proof of concept, we use 2 epochs each for NER and QA in our baseline T+DAPT model. With our early stopping experiment, we seek to balance incremental learning from additional epochs of NER against overfitting and increased catastrophic forgetting. We test the efficacy of the EarlyStoppingCallback object from Hugging Face's Trainer class to determine the optimal number of training epochs for our NER task. Though we hypothesize that the greatest risk from overfitting and forgetting comes from the domain-specific task, we also evaluate the impact of changing the number of epochs on the QA fine-tuning step.

3.2.3 Regularization

In an effort to address the challenge of overfitting while still leveraging the benefits of extended training epochs for NER fine-tuning, we test the impact of L2 regularization, as proposed by Xu et al. (2020). We test L2 regularization during both the NER and QA fine-tuning steps. To integrate L2 regularization into our training regimen, we introduce additional warm-up steps and applied weight decay. We conduct a series of tests applying combinations of warm-up steps and weight decay (specifically (100, 0.01) and (500, 0.05)) to the T+DAPT model. We explore all permutations by applying L2 regularization independently to the NER segment, the QA segment, and both segments together, using each set of hyperparameters.

3.2.4 Continuous Learning

We also follow the recommendation of Pan et al. to supplement T+DAPT with continuous learning methods to improve its performance. We implement a technique proposed by Xu et al. (2020) in which we fine-tune the model on multiple domain-specific NER datasets in sequence. We used the biomedical and news domains employed by Pan et al. for this and tested permutations of domain order and number, holding epoch count constant across all NER fine-tuning steps.

RoBERTa Fine-Tuning Procedure	Pan et al. Paper	Our Results	Improvement
Generic QA Baseline	67.10	69.07	1.97
T+DAPT	68.00	**71.95**	3.95

Table 1: A comparison of F1 scores of fine-tuned RoBERTa models between Pan et al.'s results and our own custom implementations evaluated on labeled film QA data

3.2.5 Combining Experiments

Finally, we combine the optimal results from each of the above experiments to determine whether they drive even stronger performance when applied together.

4 Results and Discussion

4.1 Architectural Improvements

Our custom implementation of T+DAPT drove substantial F1 score improvement over that of Pan et al., validating our hypothesis that their fundamental architecture had room for improvement. We also found improvement in our generic QA baseline over theirs despite the fact that our implementation in this case closely matches Pan et al.'s. These results are illustrated in Table 1. We hypothesize that improvements to the RoBERTa model in the two years since Pan et al.'s paper came out drove both the improvement in the generic QA baseline and in T+DAPT. However, we note that our methodology drove much greater improvement in the T+DAPT model's performance (nearly 2x). We expect that our use of RoBERTa architecture specific to the given task and our use of custom models, rather than AutoModels, led to much of this F1 improvement.

Our architectural modifications make our subsequent experiments possible, and they also open the door to future work improving T+DAPT (as they allow for granular architectural modifications that were not possible with Pan et al.'s implementation).

4.2 Early Stopping

We found that 5 epochs of NER produced the highest F1 score without overfitting. We observed that training beyond 2 epochs on the QA task led to a decrease in overall F1 scores. This strategy, with longer fine-tuning on NER and shorter on QA, suggests that in-domain training holds greater relevance, even if it is not directly task-specific. Together, 5 epochs of NER and 2 epochs of QA achieved a modest F1 improvement over the T+DAPT baseline.

4.3 Regularization

Building on our early stopping experiment, we implemented our L2 experiments using 5 epochs of NER and 2 epochs of QA. We achieved our most favorable outcome with the hyperparameters (500, 0.05) applied to both fine-tuning steps. However, none of our model runs with L2 regularization outperformed the T+DAPT baseline or the early stopping model run without the regularization. Due to the lack of substantial evidence supporting the effectiveness of L2 regularization, we discontinued further experimentation in this area.

4.4 Continuous Learning

Though we tested many permutations, (e.g., NER fine-tuning on film first and then biomedical and/or news, NER fine-tuning on other domains before movies), we achieved our strongest results fine-tuning first on biomedical NER data and then on film NER data.

4.5 Combining Experiments

Putting the above together, we tested a model employing fine-tuning with early stopping on biomedical and then film NER data before fine-tuning on generic QA. This model trained for 4 epochs on biomedical NER, 5 on film NER, and 2 on QA. However, the F1 score of this model came in lower than the either successful experiment (early stopping and continuous learning) did on its own. This led us to conclude that while these strategies are marginally useful on their own, they are not effective when used simultaneously.

4.6 Winning Model

Ultimately, we found the greatest F1 score improvement when moving from our generic QA baseline to our implementation of T+DAPT. From there, we saw very modest improvement from both early stopping and continual learning. See Table 2 for a summary of our most effective model in each testing scenario. Our most effective model overall came from our continual learning experiment. We compared the predictions of this model and our

Experiment	NER	QA	F1 Score
Generic QA Baseline (Baseline 1)	-	2 epochs	69.07
T+DAPT (Baseline 2)	2 epochs	2 epochs	71.95
Early Stopping	5 epochs	2 epochs	72.09
L2 Regularization	5 epochs, 0.05 weight decay, 500 warm-up steps	2 epochs, 0.05 weight decay, 500 warm-up steps	71.22
Continuous Learning	2 epochs each of fine-tuning on biomedical and then film domain	2 epochs	**72.32**
Combining Experiments	4 epochs of fine-tuning on biomedical; 5 epochs of fine-tuning on film	2 epochs	66.47

Table 2: F1 scores of our two baseline models and the most successful models in each of our experiments evaluated on labeled film QA data

generic QA baseline against the ground truth of our labeled film QA data.

In cases in which our winning model predicts the correct answer and the baseline gets it wrong, it is common for the baseline to fail at properly identifying a named entity. The generic baseline also shows more tendency to data dump in answers, including additional words that are, at best, unnecessary, and at worst, incorrect. Alternatively, in cases where our winning model fails and the generic QA model succeeds, our model often selects the wrong named entity. The answers are plausible and represent success identifying a named entity, but our model picks the wrong one. There are also cases in which the ground truth answer is actually incorrect and one or both of our models predict the true correct answer. See Table 3 in the Appendix for examples of these scenarios.

Overall, these results support the conclusion that our implementation of T+DAPT drives substantial improvement over a generic baseline. However, given the modest improvement driven by our other experiments, we cannot conclude that the additional complexity introduced by regularization and continuous learning is worthwhile.

5 Conclusion

To improve on the performance of the T+DAPT model outlined by Pan et al. in their 2022 paper, we employ a combination of architectural modifications, early stopping, L2 regularization, and continuous learning on other domains. We find the largest performance improvement from our architectural modifications, with modest benefits from early stopping and continuous learning and no discernible improvement driven by L2 regularization. Our results indicate that our modified T+DAPT is a valuable approach to domain adaptation in low-resource settings for pre-trained language models.

In future work, we propose making further architectural modifications to further improve T+DAPT's performance. These modifications could include combining early stopping and L2 regularization with continuous learning, adding additional attention heads to the NER phase of training and using alternatives for fine tuning such as the options afforded by the Transformers Adapters library.

References

R. I. Doğan, R. Leaman, and Z. Lu. 2014. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

A. Farahani, S. Voghoei, K. Rasheed, and H. Arabnia. 2020. A brief review of domain adaptation.

S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N. Smith, and Allen. 2020. Don't stop pretraining: Adapt language models to domains and tasks. pages 8342–8360.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, and P. Allen. 2019. Roberta: A robustly optimized bert pretraining approach.

MIT. The mit movie corpus.

X. Pan, A. Sheng, D. Shimshoni, A. Singhal, S. Rosenthal, and A. Sil. 2022. Task transfer and domain

adaptation for zero-shot question answering. pages 110–116.

P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. pages 2383–2392.

S. Shakeri, C. Nogueira, Santos, H. Zhu, P. Ng, F. Nan, Z. Wang, R. Nallapati, and B. Xiang. 2020. End-to-end synthetic data generation for domain adaptation of question answering systems. *Association for Computational Linguistics*, pages 5445–5460.

E. Tjong Kim Sang and F. De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. Von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, and M. Drame. 2020. Transformers: State-of-the-art natural language processing. pages 38–45.

Y. Xu, X. Zhong, Antonio Jimeno Yepes, and Jey Han Lau. 2020. Forget me not: Reducing catastrophic forgetting for domain adaptation in reading comprehension.

R. Zhang, R. Reddy, A. Sultan, V. Castelli, A. Ferritto, R. Florian, S. Kayi, S. Roukos, A. Sil, and T. Ward. 2020. Multi-stage pre-training for low-resource domain adaptation.

6 Appendix

Scenario 1: Baseline fails to properly identify a named entity
Context: Percival Graves (played by Colin Farrell) – A powerful auror and the right-hand man of the American wizarding world's president.. Set with the task of tracking down Newt. Modesty (played by Faith Wood-Blagrove) – A haunted young girl with an ability to see deep into people. **Q:** who plays percival graves **A (ground truth):** Colin Farrell **A (baseline):** a powerful auror **A (our winning model):** Colin Farrell
Scenario 2: Baseline adds extra information that makes the answer incorrect
Context: Johnny Depp is 'Edward Scissorhands'. This is a quiz on the incredible movie 'Edward Scissorhands' starring Johnny Depp and Winona Ryder.. Average score for this quiz is 5 / 10. Difficulty: Tough. **Q:** who played edward scissorhands **A (ground truth):** Johnny Depp **A (baseline):** Johnny Depp and Winona Ryder **A (our winning model):** Johnny Depp
Scenario 3: Winning model selects the incorrect named entity
Context: ABC. The season 1 finale of How To Get Away With Murder revealed that Sam Keating had ordered Frank to kill Lila Stangard, but fans were left puzzled when Rebecca Sutter was killed.Now, the mystery behind her death and the possible murderer is what seems to be the plot of the season 2.n the next season, it would not be surprising to see the writers playing on Frank's involvement with Sam that led him to kill Lila. Meanwhile, Charlie Weber, who plays Frank in the TV series, spoke about the possible killer of Rebecca Sutter. **Q:** who did wes kill on how to get away with murder **A (ground truth):** Rebecca Sutter **A (baseline):** Rebecca Sutter **A (our winning model):** Lila Stangard
Scenario 4: Ground truth answer is incorrect; our models select the correct answer
Context: Before Ledger was confirmed to play the Joker in July 2006, Paul Bettany, Lachy Hulme, Adrien Brody, Steve Carell, and Robin Williams publicly expressed interest in the role. On not being invited to reprise the Joker, Jack Nicholson remarked that he was furious. **Q:** who played the joker in the movie? **A (ground truth):** Jack Nicholson **A (baseline):** Ledger **A (our winning model):** Ledger

Table 3: Comparison of predictions of our generic QA baseline and winning model against the ground truth in different scenarios