Using Same-Domain Labelled Data to Improve Pretrained Language Model Performance in Question Answering

Aditya Singhal adis@nyu.edu
David Shimshoni ds5396@nyu.edu
Xiang Pan xp2030@nyu.edu
Alex Sheng as14565@nyu.edu

Abstract

Pretrained-language models (PLMs) have shown success in Question-Answering-Reading Comprehension (QA-RC) tasks, which are a common benchmark in the NLP research community. However, in real world applications, particularly in domains which are new or niche, data availability to train a language model, or gold-standard labelled data to effectively perform RC tasks may be limited. We study the efficacy of transfer learning techniques using auxiliary supervised learning tasks from the same domain, in this instance, Named Entity Recognition (NER). With this 'Hierarchical Fine-tuning' approach, we evaluate performance under the constraint of zero-shot learning, giving no data to the model from the target task. We find a positive effect of domain-specific auxiliary training on the downstream RC task in 3 out of the 4 domains we studied.

1 Introduction

1.1 Background

Reading Comprehension (RC) tasks, which involve language models performing Question Answering (QA) by selecting spans out of a given context, have become a familiar benchmark in NLP. But besides serving as benchmarks, they are also important for real-world applications such as tools made to answer quick questions about COVID-19 from the medical literature (Reddy et al., 2020). Pretrained Language Models (PLMs) trained on domain-data have achieved much success in these tasks (Rietzler et al., 2020).

However, in evaluating the performance on the aforementioned tasks, there is an assumption of significant data availability for the model to train and evaluate on (supervised learning). In domains that are novel or otherwise when obtaining labelled-data is too cost-intensive and time consuming, creative approaches are needed for effective Domain Adaptation. Transfer learning techniques have already shown success in improving question answering performance with limited access to training data in the target task (Hazen et al., 2019; Wiese et al., 2017).

In this paper, we aim to evaluate the effectiveness of 'hierarchical finetuning,' a transfer learning techniques used when presented with the challenge of zero shot evaluation. Here, we give the language model no examples from the target reading comprehension task to fine-tune on, and instead perform fine-tuning by gathering same domain data from a different supervised task. In our case, we choose Named Entity Recognition (NER) as the target task in each domain due to its ready availability across domains. We then train our models to generate answer spans by fine-tuning on the more general SQuAD1.1 dataset. We test the language model's performance on Reading Comprehension data taken from 4 domains: News, Movies, Biomedicine, and COVID-19. Our baseline results indicate that the standard RoBERTa-Base model hierarchically finetuned on auxiliary tasks and then trained on SQuAD1.1 slightly outperforms RoBERTa-Base models that have undergone Domain Adaptive Pretraining (DAPT) and/or no same-domain task training before the SQuAD1.1 fine-tuning regimen on domain-specific reading comprehension tasks. We see a positive effect of this training in all tested domains except News and will discuss the procedure and hypotheses about exceptions in the following sections.

1.2 Motivation

Once models have been trained in a way that reaches state-of-the-art performance in generic QA, we want to be able to use these models to answer questions in every domain. Collecting gold-standard data for each new application can be resource intensive - instead we want to 'train once, use everywhere.' This is not a trivial problem, and we try to get around it with auxiliary tasks in the

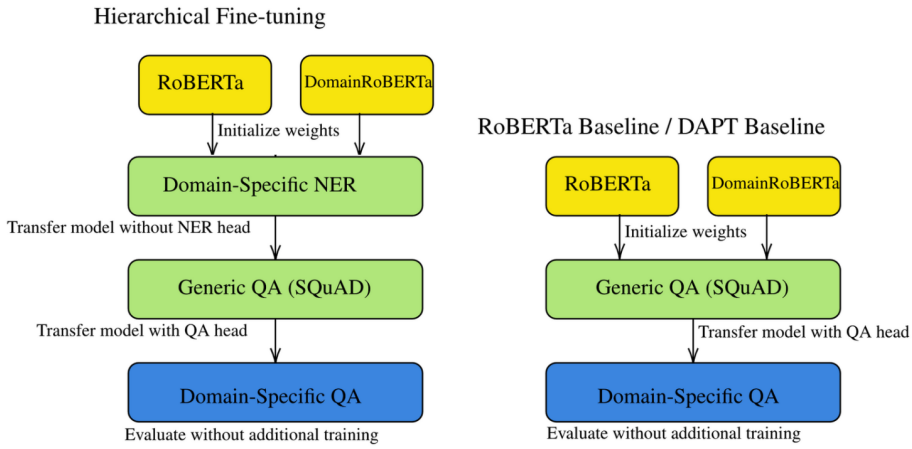

Figure 1: Hierarchical Fine-tuning for zero-shot question answering with sequential transfer learning using supervised domain-specific NER (left) vs. baseline approaches (right).

required domain and common domain adaptation techniques.

Pretrained Language Models (PLM) are usually trained in general domain large corpus and when applied to target downstream tasks, they experience a domain shift. This can harm the model's performance on the source domain. Such a phenomenon will affect Zero-Shot Learning in the domain adaptation setting, which means there is no labeled task data for model fine-tuning. Motivated by Zhang et al. (2020); Madasu and Rao (2020), we would like to use more readily available labeled domain data from unrelated tasks to fine-tune our model sequentially. Specifically, we will use domain-adaptive pretraining (DAPT) (Gururangan et al., 2020) and Named Entity Recognition (NER) as the auxiliary task to help with domain adaptation. Our code is publicly available for reference.[1]

2 Related Work

2.1 Domain Adaptive Pretraining (DAPT)

Augmenting a language model which has a transformer architecture by performing a masked language modelling pretraining procedure with a substantial corpus of unlabelled domain data has been shown to significantly improve the language model's performance on downstream supervised tasks in the same domain (Lee et al., 2019; Gururangan et al., 2020). Gururangan et al. (2020) in particular has illustrated this general procedure for PLMs, known as Domain Adaptive Pretraining (DAPT). The intuition is that the specific language representations learned by the language model in DAPT for a given domain allows for the model to better adapt to supervised tasks in that domain. This procedure has been shown to improve performance on specific domain reading comprehension tasks, in particular in the biomedical domain (Gu et al., 2021). In this paper we evaluate the performance of DAPT-enhanced language models in their respective domains, both in isolation with SQuAD1.1 fine-tuning and in conjunction with transfer learning methods that incorporate the respective domain's NER task. PLMs for two of our domains, News and Bio, are taken directly from HuggingFace and were uploaded by the authors of the DAPT paper themselves. DAPT models for Movies and COVID-19 were trained by us.

2.2 Catastrophic Forgetting

The only issue with finetuning directly on tasks in the target domain is that it leads to catastrophic forgetting– a reduction in performance on the source domain. We took inspiration from previous work by Xu et al. (2020), which explored methods to reduce forgetting during fine-tuning without assuming access to data from the source domain. In our attempts, we tried to use varying amounts of training on SQuAD (with different epochs and amount of data) to see if excessive SQuAD training was leading to forgetting of domain-specific data from DAPT and auxiliary tasks. The results from these runs are in the Appendix (Table 3). The paper by Xu et al. (2020) also released 6 narrow domain data sets that can be used as reading comprehen-

[1] https://github.com/adityaarunsinghal/Domain-Adaptation

sion benchmarks. From the domains *biomedical, computing, film, finance, law and music*, we use MoviesQA and BioQA. These datasets were obtained by applying topic modelling to a dataset collected by Microsoft based on internet search queries to Bing - MSMARCO (Bajaj et al., 2018).

2.3 Zero-Shot Learning

This paper follows up on prior work at IBM on using domain adaptation for domain-specific QA in low-data settings. In Reddy et al. (2020)'s approach, synthetic examples are generated and used in a zero-shot domain adaptation setting to improve performance in neural information retrieval and machine reading comprehension. This domain adaptation approach successfully achieved state-of-the-art performance in end-to-end QA on multiple COVID-19 datasets. Our paper builds upon this usage of zero-shot domain adaptation to improve performance on domain-specific QA, and further experimentally explores domain adaptation for domain-specific QA in the COVID-19 domain in addition to other domains.

3 Experiments

We aim to achieve zero-shot transfer to an unseen domain-specific QA task by fine-tuning on separate tasks that are in the same domain as the target task. The RoBERTa model (Liu et al., 2019a) is initialized to pretrained weights, then hierarchically fine-tuned with a domain-specific task to augment domain knowledge, and finally trained on SQuAD to bestow generic QA capabilities to achieve zero-shot QA in the target domain on an unseen domain-specific QA task without explicitly training on the final task. This method is illustrated in Figure 1.

We explore the performance of this approach in the Movies, News, Biomedical, and COVID-19 domains. Specifically, our target domain-specific QA tasks are MoviesQA (Xu et al., 2020), NewsQA (Trischler et al., 2017), BioQA (Xu et al., 2020), and CovidQA (Möller et al., 2020), respectively. We choose to use NER data for all four target domains because of the convenient availability of this kind of data for all of them. The domain-specific NER tasks are performed using supervised training data from the MIT Movie Corpus (Liu et al., 2013), CoNLL 2003 News NER (Tjong Kim Sang and De Meulder, 2003), NCBI-Disease (Doğan et al., 2014) and COVID-NER [2]. The domain-specific language modeling tasks are performed using unsupervised text from IMDB (Maas et al., 2011), the RealNews Corpus (Zellers et al., 2020), the Semantic Scholar Open Research Corpus (Lo et al., 2020) and the Covid-19 Corpus [3].

We explore the effectiveness of our approach in using either NER or language modeling for the domain adaptation task with a sequential training regime. Our experiments cover every combination of domain (Movies, News, Biomedical, or COVID) and adaptation task type (NER or language modeling).

We compare our approach to a previous approach (DAPT) as well as a baseline model. For the baseline, a pretrained RoBERTa-Base model is fine-tuned on SQuAD and evaluated on domain-specific QA without any domain adaptation. In the DAPT approach, RoBERTa-Base is first initialized with fine-tuned DAPT weights (NewsRoBERTa and BioRoBERTa) provided by Gururangan et al. (2020) or made from scratch using different Movies and Covid-19 datasets (Maas et al., 2011; Danescu-Niculescu-Mizil and Lee, 2011; Pang et al., 2019). These models are used through the HuggingFace model hub (which have been finetuned on unsupervised text corpora for domain adaptation), fine-tuned on SQuAD, and evaluated on domain-specific QA.

4 Results

We present our results in Table 1. We use F1 score to evaluate the QA performance on the target domain. From the result, we can conclude that both DAPT and NER can bring performance improvements in certain cases. However, DAPT is more limited and does not work in most domains we studied.

4.1 Analysis

Hierarchical Fine-tuning outperformed the baseline approach in three out of four domains (all but News), whereas Domain-Adapative Pretraining underperformed the baseline in three out of four domains (all but COVID-19). This method shows promising performance gains when used for zero-shot domain-specific question answering, which is particularly true in domains where domain-specific features are important, such as biomedical, movies,

[2] https://github.com/tsantosh7/COVID-19-Named-Entity-Recognition

[3] https://github.com/davidcampos/covid19-corpus

Model	MoviesQA	NewsQA	BioQA	CovidQA
RoBERTa Base + SQuAD1.1	67.0875	**56.9803**	57.9668	42.0485
RoBERTa Base + DAPT + SQuAD1.1	60.7109	54.4171	57.8325	**47.2190**
RoBERTa Base + Domain NER + SQuAD1.1	**67.9869**	55.9614	**58.8560**	42.6584
RoBERTa Base + DAPT + Domain NER + SQuAD1.1	66.3845	54.1825	55.1012	43.0710

Table 1: Performance of RoBERTa-Base models on dev sets of QA data for given domains with the stated pretraining regimens (All scores F1)

and COVID, as our experiments have shown. The lack of performance gains from Hierarchical Fine-tuning in the News domain could possibly be attributed to its broad nature, as it relies mostly on general world knowledge. This world knowledge from RoBERTa pretraining may have been partially lost due to catastrophic forgetting during the multiple fine-tuning tasks, while little applicable knowledge was accrued from domain adaptation for News. RoBERTa (Liu et al., 2019b) was also additionally trained on CC-news, a sizable news corpus, and so the baseline for this domain could have had a significant advantage in the NewsQA task that got lost when a much less relevant CoNLL-NER was added in the mix. Lastly, there has been some evidence of wrong ground-truth labels in the CoNLL dataset (Reiss et al., 2020), which could potentially explain why training on CoNLL consistently harmed QA performance in the News Domain.

DAPT was not effective when used to fine-tune language models for zero-shot domain-specific question answering. In fact, the original DAPT paper Gururangan et al. (2020) also reports its superior performance on tasks like relation classification, sentiment analysis and topic-modelling but leaves out the standard task of QA. We suspect this may be because of the dissimilarity of the unsupervised language modeling task used in DAPT from the question answering task. Unsupervised language modeling may not provide readily transferable features for QA, as opposed to NER which classifies tokens specifically into entities. These entities are also often answer tokens in QA. Another possible factor is that RoBERTa was pretrained on the English Wikipedia corpus, the same source that SQuAD questions were drawn from. Because of this, it is possible that pretrained RoBERTa had an intrinsic advantage for question answering which was lost to catastrophic forgetting during DAPT language modeling.

In the COVID domain, we use the articles from Wang et al. (2020). These articles also make the basis for the CovidNER, and the CovidQA (Möller et al., 2020) datasets, which may have caused us to see the bump in performance when doing DAPT in this domain. We can conclude that DAPT is sensitive to its similarity to the target tasks.[1]

4.2 Conclusion

We evaluate the performance of a hierarchical learning approach with domain-specific NER in a zero-shot setting on four different domain QA datasets. We find that this approach outperforms the baseline RoBERTa-Base model which was only fine-tuned on SQuAD1.1 in 3 of the domains, and also outperforms the same approach but with DAPT in 3 out of 4 domains. We find that the combination of DAPT and a hierarchical learning approach fails to improve performance in any of the domains. Our findings indicate that in a limited data setting, where directly training on a domain-specific QA could be difficult, a hierarchical learning approach could be highly successful in bestowing the model with domain knowledge, although we are still to experiment with other types of auxiliary tasks and different domain-adaptation techniques to prevent catastrophic forgetting.

4.3 Future Work

In future work, we intend to explore various methods to improve the performance of our hierarchical domain adaptation approach by remedying catastrophic forgetting and maximizing knowledge transfer. For this we hope to emulate the regularizations used by Xu et al. (2020) and try multi-task learning and continual learning methods like AdapterNet (Hazan et al., 2018). In order to improve the transferability of learned features, we will explore different auxiliary tasks (like NLI and sentiment analysis) and few-shot learning approaches.

[1] We do some additional experiments in COVID with different auxiliary tasks and this is presented in the Appendix A.1

5 Ethical Considerations

Question answering systems are useful tools in complement to human experts, but the increasing "word-of-machine effect" (Longoni and Cian, 2020) hints at dangerous over-trust in the results of such systems. While the methods proposed in this paper would allow more thorough usage of existing resources, they also bestow confidence and capabilities to models which may not have much domain expertise. In a way, conveniently domain adapted models could be approximations of extensively domain-trained models which were themselves approximations of real experts or source documents. Use of DA methods for rapid applications could propagate misinformation. For example, while making an information retrieval system for the Constitution could become easier, we do not yet believe people should look to such systems for any form of legal advice.

6 Collaboration Statement

Alex, David, Adi and Xiang handled training and evaluation for the Bio, News, Movies and COVID-19 domains respectively. Additionally, David took charge in collaboration and formatting, Adi was instrumental with planning, formal writing and technical management, Alex helped with literature review and idea formation and Xiang was helpful with literature review and experimental design. This team is mentored by IBM's Avi Sil and Sara Rosenthal.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *arXiv:1611.09268 [cs]*.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks.

Alon Hazan, Yoel Shoshan, Daniel Khapun, Roy Aladjem, and Vadim Ratner. 2018. Adapternet - learning input transformation for domain adaptation.

Timothy J. Hazen, Shehzaad Dhuliawala, and Daniel Boies. 2019. Towards domain adaptation from limited data for question answering using deep neural networks. *arXiv:1911.02655 [cs]*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Jingjing Liu, Panupong Pasupat, Yining Wang, Scott Cyphers, and Jim Glass. 2013. Query understanding enhanced by hierarchical parsing structures.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S. Weld. 2020. S2orc: The semantic scholar open research corpus.

Chiara Longoni and Luca Cian. 2020. Artificial intelligence in utilitarian vs. hedonic contexts: The "word-of-machine" effect. *Journal of Marketing*.

Andrew Maas, Raymond Daly, Peter Pham, Dan Huang, Andrew Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis.

Avinash Madasu and Vijjini Anvesh Rao. 2020. Sequential domain adaptation through elastic weight consolidation for sentiment analysis. *arXiv:2007.01189 [cs]*.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. COVID-QA: A question answering dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. Covid-qa: A question answering dataset for covid-19.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2019. Thumbs up? sentiment classification using machine learning techniques.

Revanth Gangi Reddy, Bhavani Iyer, Md Arafat Sultan, Rong Zhang, Avi Sil, Vittorio Castelli, Radu Florian, and Salim Roukos. 2020. End-to-end qa on covid-19: Domain adaptation with synthetic training.

Frederick Reiss, Hong Xu, Bryan Cutler, Karthik Muthuraman, and Zachary Eichenberger. 2020. Identifying incorrect labels in the CoNLL-2003 corpus. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 215–226, Online. Association for Computational Linguistics.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4933–4941, Marseille, France. European Language Resources Association.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. Newsqa: A machine comprehension dataset.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, et al. 2020. Cord-19: The covid-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Jerry Wei, Chengyu Huang, Soroush Vosoughi, and Jason Wei. 2020. What are people asking about covid-19? a question classification dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Georg Wiese, Dirk Weissenborn, and Mariana Neves. 2017. Neural domain adaptation for biomedical question answering.

Y. Xu, X. Zhong, A. J. J. Yepes, and J. H. Lau. 2020. Forget me not: Reducing catastrophic forgetting for domain adaptation in reading comprehension. *arXiv:1911.00202 [cs]*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, Yejin Choi, and Paul Allen. 2020. Defending against neural fake news.

Rong Zhang, Revanth Gangi Reddy, Md Arafat Sultan, Vittorio Castelli, Anthony Ferritto, Radu Florian, Efsun Sarioglu Kayi, Salim Roukos, Avirup Sil, and Todd Ward. 2020. Multi-stage pre-training for low-resource domain adaptation. *arXiv:2010.05904 [cs]*.

Model	CovidQA
RoBERTa Base + CovidQA(upper bound)	52.1416
RoBERTa Base + SQuAD1.1	42.0485
RoBERTa Base + DAPT + SQuAD1.1	47.2190
RoBERTa Base + Covid-NER + SQuAD1.1	42.6584
RoBERTa Base + Covid-QCLS + SQuAD1.1	42.6300
RoBERTa Base + DAPT + Covid-NER + SQuAD1.1	43.0710
RoBERTa Base + DAPT + Covid-QCLS + SQuAD1.1	**45.8314**
RoBERTa Base + DAPT + Covid-NER + Covid-QCLS + SQuAD1.1	43.0854

Table 2: Performance of RoBERTa-Base models on dev sets of QA data for COVID domain with the stated pretraining regimens (All scores F1)

Model	NewsQA
RoBERTa + SQuAD1.1 (1 Epoch, 1000 Train Examples)	19.9953
RoBERTa + SQuAD1.1 (2 Epoch, 1000 Train Examples)	35.2666
RoBERTa + SQuAD1.1 (2 Epoch, 5000 Train Examples)	47.0090
RoBERTa + SQuAD1.1 (Full Data, 2 Epochs, All layers frozen except classifier)	05.5891
NewsDAPT + SQuAD1.1 (1 Epoch, 1000 Train Examples)	17.9025
NewsDAPT + SQuAD1.1 (2 Epoch, 1000 Train Examples)	28.4453
NewsDAPT + SQuAD1.1 (2 Epoch, 5000 Train Examples)	44.1206
NewsQA Baseline (RoBERTa + Full SQuADv1 data, 2 Epochs)	**56.9803**

Table 3: Performance of RoBERTa-Base models + different amounts of SQuAD on QA data for News domain (All scores F1)

A Appendix

A.1 Experiments Details and Additional Experiments

Freezing Layer - We tried to freeze the bottom layer after NER training and only train the QA layer on SQuAD, the performance is worse than fine-tuning the whole roBERTa and QA layer. NER and QA may not rely on the exact same features for the final task which may be the reason why freezing causes a performance decrease.

Different Training Epoch and Training Examples - When selecting the best performance model, we use a validation set in target domain to evaluate the performance. From Table 3, we show our trials with different amounts of SQuAD training in the News Domain and how it affected performance in NewsQA.

Different Training order - We tried to use different training order, for example, we train on SQuAD1.1 task first and then on NER, the F1 score is 42.15 in CovidQA, which has some improvement, but QA as the last task performs better.

Another Auxiliary Task - In the Covid domain, we also do experiments on a more QA-relevant task, question classification (QCLS) (Wei et al., 2020). We show the result in Table 2appendix. The experiments show that QCLS task have more improvements than NER task. In addition, we test the model trained on CovidQA as the performance upper bound.