← Back to **Author Console** (/group?id=EMNLP/2023/Conference/Authors#your-submissions)

Quality > Quantity: Synthetic Corpora from Foundation Models for Closed-Domain Extractive Question Answering

PDF (/pdf?id=OxoP1qFotz)

Saptarshi Sengupta (/profile?id=~Saptarshi_Sengupta1), Connor Heaton (/profile?id=~Connor_Heaton1), Shreya Ghosh (/profile?id=~Shreya_Ghosh3), Preslav Nakov (/profile?id=~Preslav_Nakov2), Prasenjit Mitra (/profile?id=~Prasenjit_Mitra1)

16 Jun 2023 (modified: 28 Jul 2023) EMNLP 2023 Conference Submission Conference, Senior Area Chairs, Area Chairs, Reviewers, Authors, Publication Chairs Revisions (/revisions?id=OxoP1qFotz)

Keywords: Closed Domain Question Answering, Prompt Engineering, Foundational Models, Domain Adaptation

TL;DR: PreTraining on corpus from generative text models is helpful for in-domain pre-training for extractive question answering.

Abstract:

Domain adaptation, the process of training a model in one domain and applying it to another, has been extensively explored in machine learning. While training a domain-specific foundation model (FM) from scratch is an option, recent methods have focused on adapting pre-trained FMs for domain-specific tasks. However, our experiments reveal that either approach does not consistently achieve state-of-the-art (SOTA) results in the target domain. In this work, we study extractive question answering within closed domains and introduce the concept of targeted pre-training. This involves determining and generating relevant data to further pre-train our models, as opposed to the conventional philosophy of utilizing domain-specific FMs trained on a wide range of data. Our proposed framework uses Galactica to generate synthetic, ``targeted'' corpora that align with specific writing styles and topics, such as research papers and radiology reports. This process can be viewed as a form of knowledge distillation. We apply our method to two biomedical extractive question answering datasets, COVID-QA and RadQA, achieving a new benchmark on the former and demonstrating overall improvements on the latter. Code available upon publication.

Submission Type: Regular Long Paper
Submission Track: Question Answering
Confirmation Of Submission Requirements: Your submission has at most 8 pages (long) or 4 pages (short) of content., Your submission uses the unchanged Word or LaTeX template for EMNLP 2023., You have read our Multiple Submission Policy and your paper does not violate it, Your submission is fully anonymized and observed the anonymity period rules., You have read our Code of Ethics and your paper does not violate the policy., All authors of this paper have provided their up-to-date OpenReview profiles.
Submission Number: 4092

Filter by reply type... Filter by author... Search keywords...

Sort: Newest First

Everyone | Program Chairs | Senior Area Chairs | Area Chairs | Reviewers Submitted

4 / 4 replies shown

Authors | Reviewer KF6E | Reviewer TY4M | Reviewer KEWd | Reviewer grqj | ✖

Add: **Withdrawal** **Author License Task** **Official Comment**

Official Review of Submission4092 by Reviewer KF6E

Official Review ✎ Reviewer KF6E 14 Aug 2023, 13:41 (modified: 22 Aug 2023, 14:01)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer KF6E

Revisions (/revisions?id=IfCXH5VLXu)

Paper Topic And Main Contributions:
The paper proposes a framework to generate synthetic corpora for pre-training foundational models (FMs). These generated corpora align well with the writing styles and topics of the downstream tasks resulting in better overall performance. The authors have focused the efforts over two extractive question answering datasets of COVID-QA and RadQA.

Reasons To Accept:
The reasons for accepting this paper are outlined below:

1. The paper is well-structured and easily comprehensible.
2. The paper introduces a framework addressing a crucial NLP issue that resonates with researchers in specialized domains like clinical, finance, or legal.
3. The proposed framework is robust, supported by demonstrable results illustrating its effectiveness across various clinical QA downstream tasks. I have some questions regarding the experiment design and results which I've mentioned in the weaknesses below.

Reasons To Reject:
Some weaknesses of the paper are as follows:

1. The authors haven't shown any qualitative analysis by showing examples to support the hypothesis in the paragraphs starting lines: 449 and 460. Both these hypothesis can be validated by examining the examples and presenting them in the discussion section.
2. The authors haven't shown the results over multiple runs with the mean and standard deviation of the performance over multiple runs.
3. The results observed in line 491 could also be because of bad train : dev : test splits. This would have been mitigated to certain extent if experimentation was performed over multiple runs.
4. The results across the two datasets are not consistent using different strategies. So, if a researcher is applying this framework to a completely new problem - what would be the suggested strategy to start with? Hence, a small recommendation sub-section would also be need in Section 4.
5. Base BERT and RoBERTa models are used for experimentation. Why not further fine-tune their clinically fine-tuned versions for experimentation? Also, any specific reason why longformers / BigBird weren't used for experimentation? These models have consistently shown better performance in clinical domain where global context changes the overall outcome of a question significantly.

Questions For The Authors:
Please refer to the weaknesses shared above.

Missing References:
N/A

Typos Grammar Style And Presentation Improvements:
Actually, I would like to commend the authors for their pristine figures and tables.

Soundness: 3: Good: This study provides sufficient support for its major claims/arguments, some minor points may need extra support or details.
Excitement: 3: Ambivalent: It has merits (e.g., it reports state-of-the-art results, the idea is nice), but there are key weaknesses (e.g., it describes incremental work), and it can significantly benefit from another round of revision. However, I won't object to accepting it if my co-reviewers champion it.
Reproducibility: 3: Could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined; the training/evaluation data are not widely available.
Ethical Concerns: No
Reviewer Confidence: 4: Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Add: **Official Comment** **Rebuttal**

Official Review of Submission4092 by Reviewer TY4M

Official Review Reviewer TY4M 12 Aug 2023, 05:13 (modified: 22 Aug 2023, 14:01)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer TY4M
Revisions (/revisions?id=G1npskvKO7)

Paper Topic And Main Contributions:
This paper proposes a novel method to generate pre-training data for closed domains and demonstrates it's effectiveness by rich experiments, setting up a new SOTA system on the COVID-QA dataset. By using LLM to generate corpora from entities in QA tasks, the system can effectively generate a small but useful corpus from LLMs to pre-train smaller models like BERT. This work provides a novel and effective knowledge distillation method.

Reasons To Accept:
The method is novel. The detailed result shows the effectiveness of their method.

Reasons To Reject:
Whether the resulting model will be influenced by LLM's hallucination problem is to be studied.

Soundness: 4: Strong: This study provides sufficient support for all of its claims/arguments.
Excitement: 4: Strong: This paper deepens the understanding of some phenomenon or lowers the barriers to an existing research direction.
Reproducibility: 4: Could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.
Ethical Concerns: No
Reviewer Confidence: 4: Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Add: **Official Comment** **Rebuttal**

Official Review of Submission4092 by Reviewer KEWd

Official Review Reviewer KEWd 06 Aug 2023, 19:38 (modified: 22 Aug 2023, 14:01)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer KEWd
Revisions (/revisions?id=aMk30o6HJU)

Paper Topic And Main Contributions:
This paper proposes to re-pretrain PLMs for downstream close-domain EQA tasks with synthetic data generated by prompting generative LLMs. The paper investigates the idea that by prompting generative LLMs with well-designed prompts and named entities from the target domain, the synthetic data would be close to the target domain in terms of content, style and structure. The paper evaluates its proposed method on two benmark datasets and analyzed their performance and identify critical factors.

Reasons To Accept:

1. The generative LLMs do not perform well on many close-domain EQA tasks, especially when the domains are critically different. This paper proposes an interesting method to utilize LLMs' generations for re-pretrain the PLMs for downstream tasks by proper prompting.
2. The paper conducts comprehensive experiments and analysis over two datasets and a series models, which is inspiring and valuable for future research.
3. This paper presents several actionable findings for applying similar pipeline for this task.

Reasons To Reject:

1. Wiki papges are used as a baseline method to compare the generations' performance. However, the whole wiki articles are noisy and contain contents that are not relevant to the target domain, which has been well known and thus makes the comparison not fair enough. There should be baselines about retrieval methods, and some other domain adaptation and data augmentation methods like back-translation etc.
2. The research presentation, especially the table position and its analysis locate in several different papges, which makes it hard to follow.
3. The quality evaluation and filtering process is not well presented. There should be more details or source code for this essential process.

Typos Grammar Style And Presentation Improvements:

1. Line 089, grammar issue.

Soundness: 3: Good: This study provides sufficient support for its major claims/arguments, some minor points may need extra support or details.
Excitement: 2: Mediocre: This paper makes marginal contributions (vs non-contemporaneous work), so I would rather not see it in the conference.
Reproducibility: 3: Could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined; the training/evaluation data are not widely available.
Ethical Concerns: No
Reviewer Confidence: 3: Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math, experimental design, or novelty.

Add: **Official Comment** **Rebuttal**

Official Review of Submission4092 by Reviewer grqj

Official Review Reviewer grqj 05 Aug 2023, 23:53 (modified: 22 Aug 2023, 14:01)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer grqj
Revisions (/revisions?id=xnUgAY8jxQ)

Paper Topic And Main Contributions:
The authors propose targeted pre-training as a solution for a closed-domain extractive QA task. Using a generative LM they generate task focused pre-training data for two biomedical extractive question answering datasets, COVID-QA and RadQA. The pre-trained model achieves a new benchmark on the former and demonstrates overall improvements on the latter.

Reasons To Accept:

1. The submission proposes an end-to-end flow for creating an LM for closed-domain tasks. It is well written and well

structured.

2. The authors use a generative LM for dataset generation in a closed-domain task where data privacy is a concern. Although numerous recent works make use of LLMs for dataset generation, the authors clearly outline the motivation to do so for the medical extractive QA task pre-training.
3. The work also clearly outlines the motivation for redefining the domain-specific pre-training paradigm and carry out experiments evaluating the same.
4. The authors also carry out extensive ablation studies to evaluate the impact of different prompts, corpus size, context length, and wikipedia articles pre-training baseline.
5. The results show that the proposed solution performs comparably or better than the current solutions/baselines.
6. Overall the paper is well written and well structured. The authors corroborate all contributions outlined in the introduction section.

Reasons To Reject:

1. The authors did not provide any motivation for carrying out token filtering in the experiments section. Was it done due to the observations from Table 4, row/section 4 experiments?
2. In section 4.3, the authors carry out experiments to evaluate if there was any information gain and the results show a significant drop the in the performance of the proposed pipeline. Why was this approach not considered in the first place? The difference clearly shows that using generative LM's for dataset creation can lead to leakage. It would be nice if the authors can discuss additional measures to handle this problem in a closed-domain setting.
3. Why was the number of entities increased from 1 to 10 in entity based dataset generation? What was the motivation for choosing the same?

Soundness: 3: Good: This study provides sufficient support for its major claims/arguments, some minor points may need extra support or details.
Excitement: 4: Strong: This paper deepens the understanding of some phenomenon or lowers the barriers to an existing research direction.
Reproducibility: 3: Could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined; the training/evaluation data are not widely available.
Ethical Concerns: No
Reviewer Confidence: 3: Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math, experimental design, or novelty.

Add: **Official Comment** **Rebuttal**

About OpenReview (/about)
Hosting a Venue (/group?id=OpenReview.net/Support)
All Venues (/venues)
Sponsors (/sponsors)

Frequently Asked Questions (https://docs.openreview.net/getting-started/frequently-asked-questions)
Contact (/contact)
Feedback
Terms of Service (/legal/terms)
Privacy Policy (/legal/privacy)

OpenReview (/about) is a long-term project to advance science through improved peer review, with legal nonprofit status through Code for Science & Society (https://codeforscience.org/). We gratefully acknowledge the support of the OpenReview Sponsors (/sponsors).