CROSS DOMAIN QUESTION ANSWERING USING DOMAIN ADAPTATION

INLP Final Project Report

Ishan Kavathekar 2022121003
Swarang Joshi 2022114010
Hari Shankar 2022114008

Contents

1 Project Description

1.1 Problem statement

Cross-domain question answering refers to the task of developing systems or models that can effectively answer questions across different domains or topics. A "domain" typically represents a specific subject area, and each domain may have its own set of linguistic nuances, terminologies, and contextual intricacies. Cross-domain question answering aims to create robust models capable of understanding and responding to queries seamlessly, irrespective of the domain from which the questions originate.
This task poses a considerable challenge as it involves adapting language models to diverse and potentially unfamiliar domains. Our study aims to investigate different adaptation methods and assess their effectiveness in the context of question answering tasks.

1.2 Overview of upcoming sections

In the upcoming sections, we provide a detailed exploration of our experimental setup, methodology, results, and analysis in the context of cross-domain question answering. Section 2.1 elaborates on the employed datasets, highlighting their diversity and relevance. Section 2.2 introduces the utilized models, discussing their architectures and adaptability. Section 2.3 delineates the methodology for evaluating adaptation methods. Section 3 focuses on experiment results and analysis. Through this examination, we aim to offer a comprehensive understanding of cross-domain question answering and the strategies to address its challenges.

2 Experimental Setup

In the experimental setup section, we outline the foundational components essential for our investigation into cross-domain question answering. This includes delineating the datasets utilized, detailing the models employed for analysis, and the methodology employed to evaluate the effectiveness of various adaptation techniques.

2.1 Datasets

The following are the datasets used for the experiments:

1. **TriviaQA** (Base domain):[1]
   TriviaQA is a large-scale dataset for evaluating the performance of question answering systems. It consists of over 650,000 question-answer pairs that are based on web pages and Wikipedia articles. The questions cover a diverse range of topics, including history, science, arts and culture, sports, and more. TriviaQA is designed to assess a system's ability to understand and reason over long passages of text to find the correct answer.

2. **SQUAD** (Target domain):[2]
   SQuAD (Stanford Question Answering dataset) is one of the most widely used benchmarks for question answering systems. It contains over 100,000 question-answer pairs based on Wikipedia articles. The questions require reading comprehension skills and the ability to locate the relevant information in the provided text passages.

3. **NewsQA** (Target domain):[3]
   NewsQA is a question answering dataset created from CNN news articles. It contains over 100,000 question-answer pairs that test a system's ability to comprehend and reason about real-world news content. The questions cover a wide variety of topics discussed in the news, such as current events, politics, business, and science.

2.2 Models

We investigate the various model architectures on the question answering task. The models under consideration are:

1. BERT: [4]
   BERT is a powerful language model developed by Google that has become widely adopted for a variety of natural language processing tasks, including question answering. BERT uses a transformer-based architecture and is trained on a massive amount of text data in an unsupervised manner. This allows the model to learn rich contextual representations of words and passages.

2. T5 (Text-to-text model): [5]
   The T5 model, developed by researchers at Google, is a large language model that is particularly well-suited for text-based tasks like question answering. T5 uses an encoder-decoder architecture, which allows it to effectively transform input text into the desired output. This makes T5 a versatile model that can be fine-tuned for a wide variety of NLP applications.

3. Gemma-2B (Open source LLM): [6]
   Gemma is a family of lightweight, state-of-the-art open models from Google, built from the same research and technology used to create the Gemini models. Is is a text-to-text, decoder-only large language model. Gemma is well-suited for a variety of text generation tasks, including question answering, summarization, and reasoning. Their relatively small size makes it possible to deploy them in environments with limited resources democratizing access to state of the art AI models and helping foster innovation for everyone.

2.3 Methodology

In this section, we detail the methodology employed in our study to investigate cross-domain question answering using adaptation methods. We describe the experimental setup, delineate the diverse settings explored, and explain the process by which adaptation techniques were applied.

In this study, our focus centers on two distinct domain adaptation methodologies: fine-tuning and data augmentation coupled with fine-tuning. Our aim is to investigate the efficacy of these methods in enhancing the performance of models across varying domains. Initially, we train the models on the base domain, namely TriviaQA, and subsequently evaluate their performance on target domains, specifically SQUAD and NewsQA. Given the dissimilarities between the source and target domains, we anticipate sub-optimal performance from the models due to the lack of domain-specific training. Subsequently, we apply

domain adaptation techniques with the expectation of improving their performance on the target domains.
We now focus on the specific experimental settings of the domain adaptation methods:

1. **Fine-tuning**:
   Fine-tuning involves the process of taking a pre-trained model and further training it on domain-specific data. By fine-tuning, the model adapts its parameters to better capture the nuances and patterns present in the target domain. This method leverages the pre-existing knowledge encoded in the model's parameters while allowing for adjustments that cater to the characteristics of the new domain. Fine-tuning is particularly effective in domain adaptation as it enables the model to learn domain-specific features, thereby enhancing its performance on tasks within the target domain. Through fine-tuning, the model can generalize better to previously unseen data in the target domain, ultimately improving its overall performance.
   We select 10,000 data points from the training set of both the base domain and each target domain for fine-tuning the models. It is important to note that each model employed in our study possesses a distinct architecture, resulting in varying context sizes. For example, models like BERT and T5 operate within a relatively short window size of 512 tokens, whereas Gemma-2B has a larger context window of 8192 tokens. While the context of data points from the SQUAD and NewsQA datasets typically adhere to these constraints, the context of data points from TriviaQA show considerable variation. Some samples contain contexts exceeding 10,000 tokens. Addressing such instances by arbitrarily truncating the context would result in information loss which is undesirable. Therefore, we summarize the context to condense the it while ensuring it remains within the specified constraints. We use the python library sumy to summarize the contexts.

2. **Data augmentation**:
   Data augmentation is a technique used to artificially increase the size and diversity of the training data by applying various transformations to the existing examples. These transformations may include techniques such as synonym replacement, paraphrasing, or adding noise to the text. By augmenting the training data, the model is exposed to a wider range of variations and scenarios, which helps it become more robust and adaptable to different domains. In the context of domain adaptation, data augmentation can be particularly beneficial when the available training data in the target domain is limited. By augmenting the training data with synthetic examples generated from the existing data, the model can learn more effectively from the available information, leading to improved performance when applied to the target domain.
   We utilize paraphrasing as our chosen method for data augmentation. Specifically, we paraphrase the questions associated with each data point in our experiments. This process yields a diverse set of data points, enriching the training dataset with variations of the original questions. We refrain from paraphrasing the context of the questions themselves to avoid potential information loss, which could adversely affect the quality of the dataset. To facilitate the paraphrasing process, we leverage the PARROT library. It is important to acknowledge that PARROT may occasionally

fail to produce suitable paraphrases for certain questions, in which case we opt to skip the question altogether

3 Results and Analysis

3.1 Results

3.1.1 Vanilla model

Test dataset	Exact match	F1 score
TriviaQA	0.592	0.956
SQUAD	0.442	0.913
NewsQA	0.438	0.881

Table 1: Evaluation metrics of **BERT** trained on TriviaQA(base domain) and tested on base and target domains

Test dataset	Exact match	F1 score
TriviaQA	0.615	0.812
SQUAD	0.489	0.536
NewsQA	0.31	0.662

Table 2: Evaluation metrics of **T5** trained on TriviaQA(base domain) and tested on base and target domains

Test dataset	Exact match	F1 score
TriviaQA	0.0	0.801
SQUAD	0.0	0.714
NewsQA	0.0	0.723

Table 3: Evaluation metrics of **Gemma-2B** trained on TriviaQA(base domain) and tested on base and target domains

The observed sub-optimal performance of models when tested on domains they were not trained on, aligns with intuitive expectations. This can be seen in fig 3, 1 and 2 . It's notable that this sub-optimality persists consistently across various architectures. Consequently, our focus now shifts to evaluating the efficacy of domain adaptation techniques. By examining these techniques, we aim to mitigate the performance gap observed when models are applied to target domains.

3.1.2 Finetuning

Results from fig 4, 5 and 6 show that fine-tuning aids in domain adaptation. Fine-tuning demonstrates its efficacy by enhancing not only the overall performance but also the model's ability to generalize to the target domains

	Test Dataset	Exact match	F1 score
Finetuned on	SQUAD	0.627	0.985
SQUAD	NewsQA	0.407	0.915
Finetuned on	SQUAD	0.422	0.894
NewsQA	NewsQA	0.593	0.963

Table 4: Results from fine-tuning BERT on target domains

	Test Dataset	Exact match	F1 score
Finetuned on	SQUAD	0.592	0.793
SQUAD	NewsQA	0.417	0.644
Finetuned on	SQUAD	0.433	0.683
NewsQA	NewsQA	0.624	0.813

Table 5: Results from fine-tuning T5 on target domains

	Test Dataset	Exact match	F1 score
Finetuned on	SQUAD	0.0	0.837
SQUAD	NewsQA	0.0	0.734
Finetuned on	SQUAD	0.0	0.705
NewsQA	NewsQA	0.0	0.818

Table 6: Results from fine-tuning Gemma-2B on target domains

3.1.3 Data Augmentation

	Test Dataset	Exact match	F1 score
Finetuned on	SQUAD	0.622	0.988
SQUAD	NewsQA	0.413	0.925
Finetuned on	SQUAD	0.430	0.882
NewsQA	NewsQA	0.618	0.971

Table 7: Results from fine-tuning and using data augmentation BERT on target domains

	Test Dataset	Exact match	F1 score
Finetuned on	SQUAD	0.599	0.817
SQUAD	NewsQA	0.412	0.747
Finetuned on	SQUAD	0.448	0.743
NewsQA	NewsQA	0.641	0.822

Table 8: Results from fine-tuning and using data augmentation T5 on target domains

Our experimentation with data augmentation combined with fine-tuning has yielded promising results, demonstrating a notable enhancement in the performance of our models

	Test Dataset	Exact match	F1 score
Finetuned on	SQUAD	0.0	0.845
SQUAD	NewsQA	0.0	0.734
Finetuned on	SQUAD	0.0	0.711
NewsQA	NewsQA	0.0	0.839

Table 9: Results from fine-tuning and using data augmentation Gemma-2B on target domains

across target domains. Through the augmentation of training data with diverse samples from the target domain and subsequent fine-tuning, we observed a significant improvement in model adaptability and generalization of some architectures.

3.2 Analysis

Expanding on the comparison between fine-tuning alone and the combination of data augmentation with fine-tuning, we observe a notable improvement in model performance with the latter approach. This enhancement can be attributed to the increased diversity and volume of training samples resulting from data augmentation techniques. By exposing the model to a broader range of examples through paraphrasing, the data augmentation process enables more robust learning and better generalization to unseen data in the target domains. However, despite this improvement, the models' performance gains are not as substantial as expected. This can be attributed to the summarization step implemented to adhere to context constraints. While summarization helps prevent the abrupt truncation of context, ensuring information retention, it introduces the risk of information loss and potential distortion of the text. Consequently, the summarization process may impact the quality and relevance of the generated answers, thus limiting the overall performance improvement.

Furthermore, it's essential to highlight the unique behavior of Gemma-2B in terms of exact match scores. As a generative model, Gemma has the capability to produce responses that differ significantly from the ground truth answers. This divergence can result in exact match scores of zero, as the metric evaluates whether the generated answer precisely matches the reference answer. Gemma's tendency to provide long or even irrelevant responses further contributes to the lower exact match scores observed. Hence, while Gemma-2B may excel in certain aspects of text generation, its performance in terms of exact match metrics should be interpreted in light of its generative nature and unique response characteristics.

4 Challenges faced

1. Summarization because of context size: Different models used in the study have varying context window sizes. While models like BERT and T5 operate within a short window size of 512 tokens, Gemma-2B has a larger context window of 8192 tokens. However, some samples from the TriviaQA dataset contain contexts exceeding 10,000 tokens, which would result in information loss if arbitrarily truncated. To address this, the study uses the Python library sumy to summarize the contexts, condensing

them while ensuring they remain within the specified constraints.

2. Compute power for finetuning: Finetuning involves the process of further training a pre-trained model on domain-specific data. This process adapts the model's parameters to better capture nuances and patterns in the target domain, improving its performance on tasks within that domain. The problem we faced with Finetuning was that the experiments were computationally intensive, especially for models with large parameter sizes, requiring significant compute power to complete the process effectively.

3. Consistency among datasets and its effect on evaluation: One of the challenges faced is the format inconsistency among datasets. The dissimilarities between these datasets can affect the evaluation. Due to varied structure of datasets, it was not trivial to evaluate the results.

5 Conclusion

In conclusion, our study delved into the realm of cross-domain question answering, exploring the efficacy of adaptation methods in enhancing model performance. Through rigorous experimentation, we observed that fine-tuning plays a pivotal role in domain adaptation, significantly improving model performance across different domains. The combination of data augmentation with fine-tuning further amplifies these gains by augmenting the diversity of training data. Despite these challenges, our findings contribute valuable insights into the strategies and considerations involved in adapting question answering models across diverse domains. Moving forward, addressing the challenges discussed in the preceding section, exploring novel adaptation techniques will be paramount in advancing the capabilities of question answering systems in real-world applications.

References

[1] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vancouver, Canada: Association for Computational Linguistics, July 2017.

[2] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," 2016.

[3] A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman, "Newsqa: A machine comprehension dataset," 2017.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[5] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2023.

[6] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, P. Tafti, L. Hussenot, A. Chowdhery, A. Roberts, A. Barua, A. Botev, A. Castro-Ros, A. Slone, A. Héliou, A. Tacchetti, A. Bulanova, A. Paterson, B. Tsai, B. Shahriari, C. L. Lan, C. A. Choquette-Choo, C. Crepy, D. Cer, D. Ippolito, D. Reid, E. Buchatskaya, E. Ni, E. Noland, G. Yan, G. Tucker, G.-C. Muraru, G. Rozhdestvenskiy, H. Michalewski, I. Tenney, I. Grishchenko, J. Austin, J. Keeling, J. Labanowski, J.-B. Lespiau, J. Stanway, J. Brennan, J. Chen, J. Ferret, J. Chiu, J. Mao-Jones, K. Lee, K. Yu, K. Millican, L. L. Sjoesund, L. Lee, L. Dixon, M. Reid, M. Mikuła, M. Wirth, M. Sharman, N. Chinaev, N. Thain, O. Bachem, O. Chang, O. Wahltinez, P. Bailey, P. Michel, P. Yotov, P. G. Sessa, R. Chaabouni, R. Comanescu, R. Jana, R. Anil, R. McIlroy, R. Liu, R. Mullins, S. L. Smith, S. Borgeaud, S. Girgin, S. Douglas, S. Pandya, S. Shakeri, S. De, T. Klimenko, T. Hennigan, V. Feinberg, W. Stokowiec, Y. hui Chen, Z. Ahmed, Z. Gong, T. Warkentin, L. Peran, M. Giang, C. Farabet, O. Vinyals, J. Dean, K. Kavukcuoglu, D. Hassabis, Z. Ghahramani, D. Eck, J. Barral, F. Pereira, E. Collins, A. Joulin, N. Fiedel, E. Senter, A. Andreev, and K. Kenealy, "Gemma: Open models based on gemini research and technology," 2024.