Robust QA using Adversarial Methods with Multiple Discriminators

Shaan Gill sgill36@gatech.edu
Morgan Byrd abyrd45@gatech.edu
Alex Li alexli1998@gatech.edu

1 Introduction

The ability of models to automatically answer questions has risen sharply in recent years, with the proliferation of information and the Internet. The earliest question-answering programs, such as BASEBALL, were limited to very specific topics.(et al., 1961). Since then, question answering has been extended to many new domains, including sentiment analysis and semantic parsing. By 2011, IBM Watson was able to win against human Jeopardy players. More recently, the advent of pretrained models, including embeddings such as BERT (Devlin et al., 2019), as well as application-specific models, has allowed newer models to incrementally build on previous models.

The ability of deep learning based systems to generalize to information out the the domain of the training data is extremely important. One of simplest methods to improve this domain generalization for QA systems would be to follow the method of (Xu et al., 2014), where they trained a separate classifier for each desired target domain and then interpolated between them for new data. Other methods to improve performance across domains include basic domain augmentation as in (Wei and Zou, 2019), meta learning based methods as in (Kim et al., 2021), and contrastive methods as in (Yue et al., 2021). The primary method we use as reference is in (Lee et al., 2019), which used adversarial methods to make the embeddings more invariant to domain shift.

The goal of our project is to improve the robustness of a QA system. While modern systems can score very well on in-domain data, with the top scores of 90.939 EM and 93.214 F1 on the SQuAD 2.0 dataset (Rajpurkar et al., 2018), accuracy can be dramatically reduced when tested on out-of-domain data (Kamath et al., 2020), due to overreliance on domain-specific features and domain shift from training to testing. Our goal is to attempt to reduce this performance gap by extending the adversarial methods of (Lee et al., 2019) by including data augmentation, and latent representation splitting.

2 Technical Approach

We define the question answer task in the following way: given a collection of tuples (y_t, p, q), where y_t is the target answer, p is the context passage, and q is the question, the goal is to train a model that returns the correct answer y_t for a given context-question pair (p, q). Specifically, our goal NLP task is to improve the performance of a QA system on out-of-domain data.

For our in-domain training data, we use the SQuAD, HotpotQA (Yang et al., 2018), and Natural Questions (Kwiatkowski et al., 2019) datasets. We initially intended to match the reference paper and also include NewsQA (Trischler et al., 2016), SearchQA (Dunn et al., 2017), and TriviaQA (Joshi et al., 2017), but were unable to do so because of hardware limitations. This limitation both reduces the amount of training data and the number of domains used for discrimination, which both negatively impact the performance.

The baseline we are comparing against is the work in (Lee et al., 2019). Figure 1 shows an overview of the classification process. The input data question and context are first encoded using the BERT encoder. For the basic QA portion, the resulting encodings corresponding to the context are given to a separate classification network where they are trained to minimize the negative log-likelihood loss of the predicted answer vs the actual answer:

$$L_{QA} = -\frac{1}{N} \sum \log P(y|p, q) \tag{1}$$

For the adversarial portion, the goal is to minimize the ability of a discriminator network to estimate the domain D based on features z of the question and passage by minimizing the KL diver-

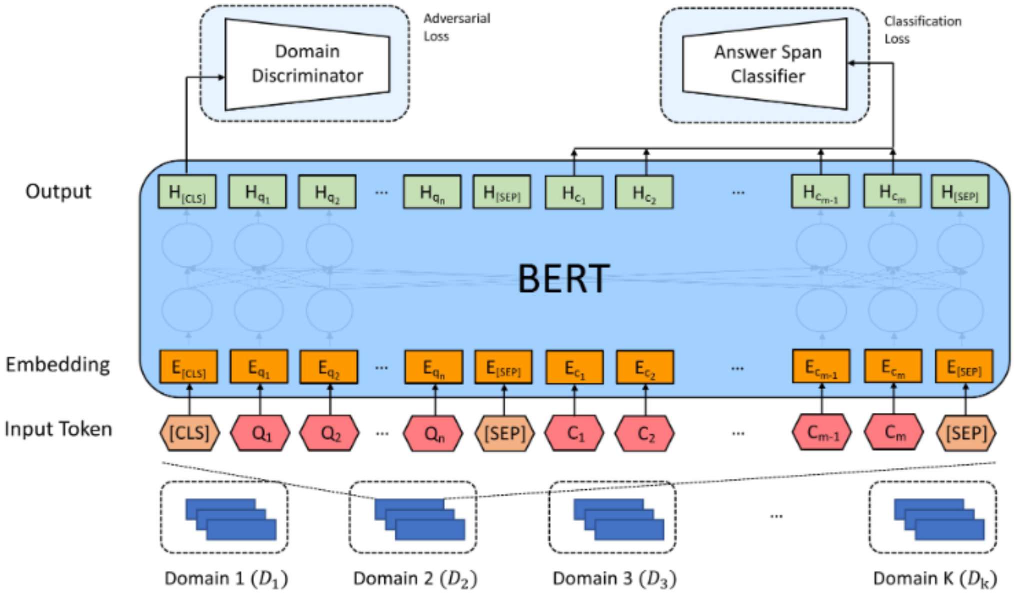

Figure 1: General model layout for baseline

gence between the discriminator prediction and a uniform distribution over domains $U(D)$ by adding an additional loss term:

$$L_{adversarial} = \frac{1}{N}\sum KL(U(D)||P(D|z)) \quad (2)$$

The [CLS] token from the BERT encoding is used as a feature vector z for the given context-question pair. The total loss function for the baseline adversarial model is:

$$L_{QA} + \lambda L_{adversarial} \quad (3)$$

Starting from their code at `https://github.com/seanie12/mrqa`, we modify it to use DistilBERT (Sanh et al., 2019) instead of the baseline BERT model, because we were unable to effectively use the original BERT-based code due to large GPU memory requirements. While the performance of this smaller model is lower, using DistilBERT allowed us to effectively train a QA model with a much larger batch size at the cost of an acceptable accuracy drop.

For our project, we have extended their model to include a second motivational discriminator network. The baseline adversarial model used its discriminator to prevent the model from discerning its domain, with the hope that this improves the out-of-domain performance of the model as it is less reliant on domain-specific features. Our extension further improves the adversarial performance via latent representation splitting (Romanov et al., 2019), where we split the latent features into two components, z and a, and attempt to make a have all the domain specific features while z has only agnostic features. This is done by adding an additional motivational loss term to maximize the ability of a second discriminator to estimate the correct domain D based only on the information in a:

$$L_{motivation} = -\frac{1}{N}\sum KL(D||P(D|a)) \quad (4)$$

Specifically, this involves splitting the [CLS] token embedding into two components, z and a, so that we can make a have all the domain-specific features for each context-question pair. Following this split, the motivational discriminator operates using only a and the original adversarial discriminator operates using only z instead of the entire [CLS] embedding. The total loss function for the model with the motivational discriminator is:

$$L_{QA} + \lambda L_{adversarial} + \eta L_{motivation} \quad (5)$$

where λ and η are hyperparameters to weight each loss term.

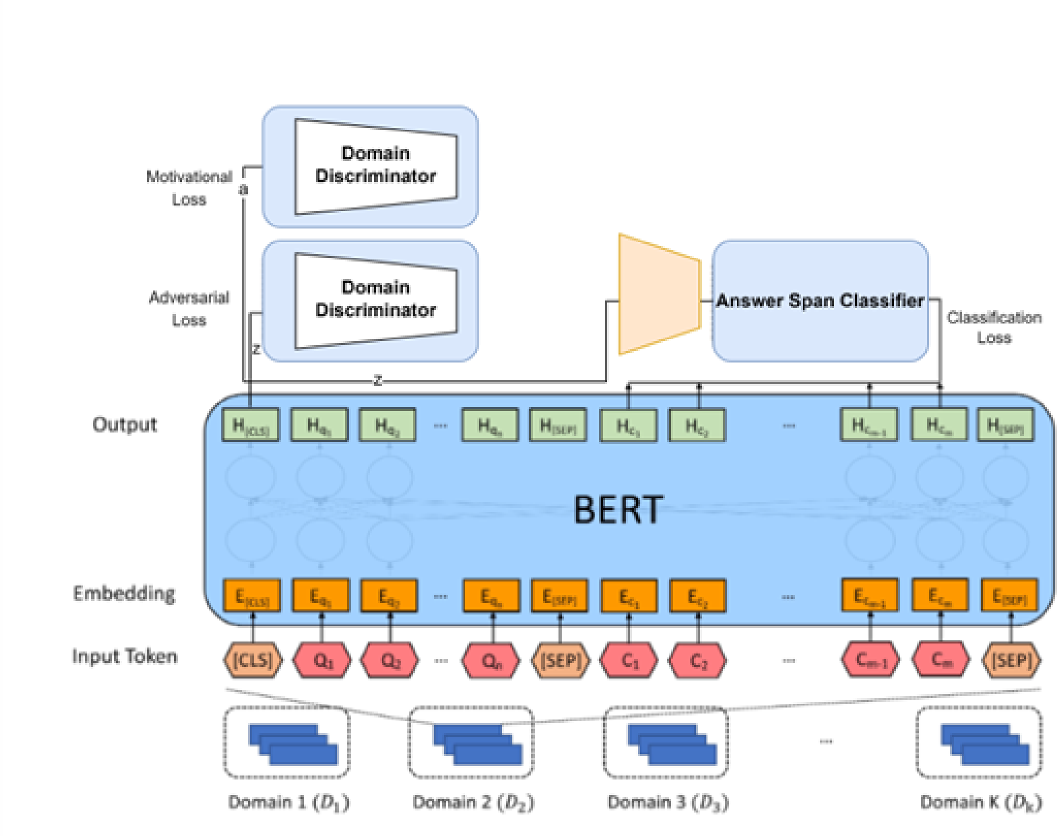
Motivational Loss
a
Domain Discriminator
Adversarial Loss
z
Domain Discriminator
Answer Span Classifier
Classification Loss
z
Output
$H_{[CLS]}$ H_{q_1} H_{q_2} ... H_{q_n} $H_{[SEP]}$ H_{c_1} H_{c_2} ... $H_{c_{m-1}}$ H_{c_m} $H_{[SEP]}$
BERT
Embedding
$E_{[CLS]}$ E_{q_1} E_{q_2} ... E_{q_n} $E_{[SEP]}$ E_{c_1} E_{c_2} ... $E_{c_{m-1}}$ E_{c_m} $E_{[SEP]}$
Input Token
[CLS] Q_1 Q_2 ... Q_n [SEP] C_1 C_2 ... C_{m-1} C_m [SEP]
Domain 1 (D_1)
Domain 2 (D_2)
Domain 3 (D_3)
...
Domain K (D_K)

Figure 2: Model layout with added motivational discriminator

Since now only domain-agnostic information should be present in z, we then incorporate this into the final QA classification prediction along with the standard embedded information by first putting it through a neural network to reshape it to the correct dimensionality and then concatenating it to the other embeddings. A diagram of the full motivational model can be seen in Figure 2.

2.1 Implementation Details

As stated above, due to hardware limitations we use DistilBERT as the encoder for our QA model. Specifically, we used "distilbert-base-uncased" from Huggingface. We chose to separate the 768 dimensional [CLS] token into a 512 dimensional vector z and 256 dimensional vector a. For each discriminator, we used a 3 layer MLP with layer sizes [768, 768, 768]. Between each fully connected layer, a ReLU activation was used and dropout (Srivastava et al., 2014) with a rate of 0.1 was applied. For fine-tuning the DistilBERT model, we trained for two epochs. We used the AdamW (Loshchilov and Hutter, 2017) optimizer with a linear warmup schedule. Both λ and η from equation 3 were set at 0.01.

3 Results

We used multiple QA datasets in order to compare results for out-of-domain performance. We used BioASQ (BA) (Tsatsaronis et al., 2015), DROP (DP) (Dua et al., 2019), DuoRC (DR) (Saha et al., 2018), RACE (RA) (Lai et al., 2017), Relation Extraction (RE) (Levy et al., 2017), and TextbookQA (TQ) (Kim et al., 2019).

We demonstrate our results for two separate baselines for comparison in Table 1. The first is for just the baseline DistilBERT model and the second is for the DistilBERT model with the additional adversarial component. The results track those in the original reference paper, with the adversarial model generally outperforming the base model, but ours are lower than the reference, likely due to the use of DistilBERT over BERT and only training on three in-domain datasets rather than six.

Our work is shown in the table as DistilBERT-motiv, and shows improvements over the base and adversarial methods in five of the six test datasets and gives an average improvement of 0.7 F1 score

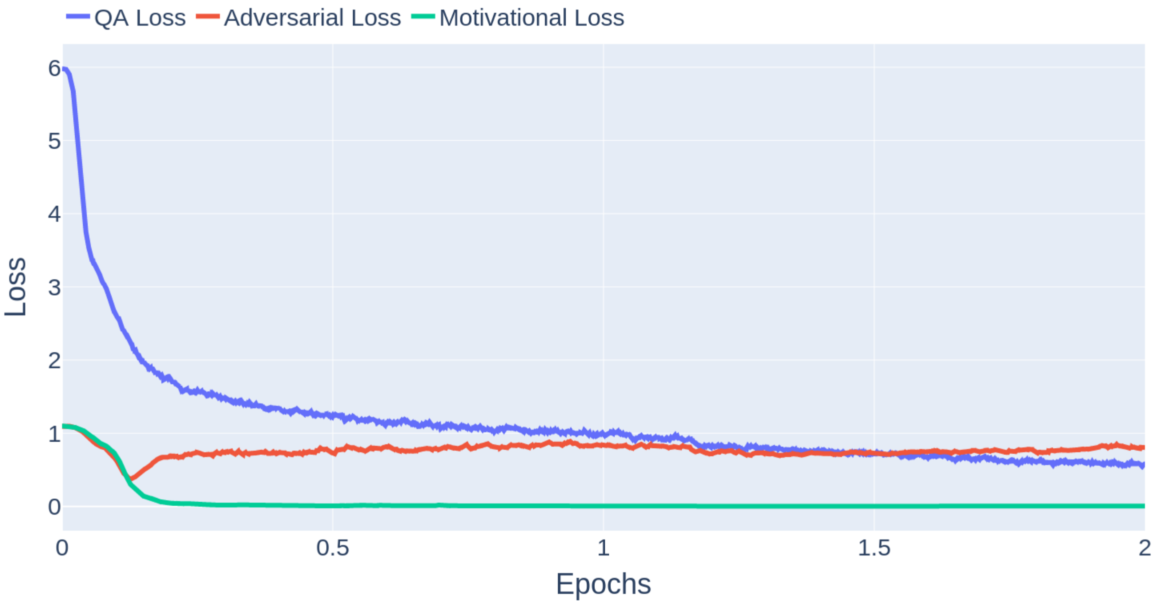

Figure 3: Learning curves

Model	BA		DP		DR		RA		RE		TQ		Avg	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
DistilBERT-base	30.4	46.9	18.5	27.5	35.4	43.6	23.3	35.8	66.2	78.8	31.9	41.9	34.3	45.8
DistilBERT-adv	**31.1**	**48.2**	19.4	29.1	39.0	48.1	23.6	35.0	68.4	80.9	32.4	41.4	35.7	47.1
DistilBERT-motiv	29.3	45.1	**20.6**	**30.3**	**40.8**	**49.7**	**24.0**	**36.4**	**70.3**	**82.1**	**34.4**	**43.0**	**36.6**	**47.8**

Table 1: Results comparison

and 0.9 Exact Match over the adversarial baseline. A plot showing each loss during training is shown in Figure 3. Given a lack of computational resources, we did not spend much effort in tuning our hyperparameters for the motivational model, namely λ and η in the loss function and the dimensionality of z and a, so this improvement could likely be further increased.

3.1 Ablation Study

3.1.1 Effect of Number of Datasets

Along with this work, to get a better idea of the impact that using only three input datasets had on our final results, we also conducted an ablation study where we trained our motivational model using only two of the three datasets. The results of this are in Table 2. Each of the three potential dataset choices is represented by a single character, where SQuAD = S, HotPotQA = H, NaturalQuestions = N. Unsurprisingly, a smaller amount of training data along with fewer different training domains led to a much lower overall performance at test time on out of domain data. Otherwise, it shows that HotpotQA is the least useful dataset if one had to be removed, which suggests that multi-hop reasoning is less useful for our test datasets.

3.1.2 Effect of Dataset Augmentation

Another side study we performed was to try to quantify the impact of dataset augmentation on improving out of domain QA performance. To determine this, we used the NLP Aug (Ma, 2019) software package to do synonym replacement of ten percent of the text of the context tokens in each of our training datasets. The results of this are in Table 3. Augmentation improved test performance for the baseline and adversarial models, while reducing it for the motivational model. The reason for the performance drop for the motivational model is unclear, but further tuning of hyperparameters for both the motivational model and the augmentation would likely lead to similar small improvements as seen in the other models.

4 Conclusion

We have successfully implemented our DistilBERT model which uses two discriminators, one adversarial discriminator to give domain agnostic features

Model	BA		DP		DR		RA		RE		TQ		Avg	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
DistilBERT-motiv-SH	30.5	45.3	17.3	27.3	38.2	47.2	21.7	33.1	66.9	79.8	27.9	36.5	33.8	44.9
DistilBERT-motiv-SN	**30.6**	**45.9**	17	26.2	**38.8**	**47.8**	**23.7**	**35.4**	**67.2**	**80.5**	**32.1**	**41.6**	**34.9**	**46.2**
DistilBERT-motiv-HN	25.7	44.2	**20.9**	**31.7**	36.9	45.4	18.5	29.7	61.3	76.9	26.5	34.8	31.6	43.8

Table 2: Dataset Ablation

Model	BA		DP		DR		RA		RE		TQ		Avg	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
DistilBERT-base-aug	31.1	47.8	20.6	29.5	37.8	47.8	21.8	33.6	67.2	80.7	31.7	41.9	35.0	46.9
DistilBERT-adv-aug	31.5	47.9	20.6	29.0	41.2	50.0	23.0	35.5	68.1	80.5	32.0	41.2	36.1	47.4
DistilBERT-motiv-aug	31.6	47.3	20.0	29.9	40.3	50.2	21.4	33.9	68.7	81.1	32.3	41.8	35.7	47.4

Table 3: Effect of Augmentation on Out of Domain Performance

and one motivational discriminator to give domain specific features. Using this, we see notable improvements in performance over the baseline fine-tuned model and the fine-tuned adversarial model. This validates the hypothesis that utilizing a second discriminator to further separate the domain agnostic and domain specific features is beneficial. Further improvements to the model could be made by included all 6 training datasets, doing more hyperparamter tuning, and using BERT over DistilBERT. Due to hardware restrictions this was not possible for us to do.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *CoRR*, abs/1903.00161.

Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new qa dataset augmented with context from a search engine.

Bert F. Green et al. 1961. Baseball: an automatic question-answerer. *AFIPS*.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.

Daesik Kim, Seonhoon Kim, and Nojun Kwak. 2019. Textbook question answering with multi-modal context graph understanding and self-supervised open-set comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3568–3584, Florence, Italy. Association for Computational Linguistics.

Jin Kim, Jiyoung Lee, Jungin Park, Dongbo Min, and Kwanghoon Sohn. 2021. Self-balanced learning for domain generalization.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Seanie Lee, Donggyu Kim, and Jangwon Park. 2019. Domain-agnostic question-answering with adversarial training. In *EMNLP*.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822.

Alexey Romanov, Anna Rumshisky, Anna Rogers, and David Donahue. 2019. Adversarial decomposition of text representation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 815–825, Minneapolis, Minnesota. Association for Computational Linguistics.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. DuoRC: Towards Complex Language Understanding with Paraphrased Reading Comprehension. In *Meeting of the Association for Computational Linguistics (ACL)*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Éric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Zheng Xu, Wen Li, Li Niu, and Dong Xu. 2014. Exploiting low-rank structure from latent domains for domain generalization. In *Computer Vision – ECCV 2014*, pages 628–643, Cham. Springer International Publishing.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering.

Zhenrui Yue, Bernhard Kratzwald, and Stefan Feuerriegel. 2021. Contrastive domain adaptation for question answering using limited text corpora.