

Multi-Task Learning and Domain-Specific Models to Improve Robustness of QA System

Stanford CS224N Default Project, Mentor: Daniel Do

Anmol Gupta

Department of Computer Science
Stanford University
anmolg@stanford.edu

Meredith Xu

Department of Computer Science
Stanford University
merexu@stanford.edu

Abstract

In this project, we aim to build a Question Answering (QA) language model that can perform well on data that is not in the same domain as the majority of its training data. This task is useful for building robust QA language models. Our approach is to take the pre-trained DistilBERT [1] model on high-resource in-domain dataset and then perform further pre-training using multi-task learning objective. We also explore data augmentation techniques such as synonym replacement, random word deletions and insertions, word swapping [2], and back-translation [3] to expand our data. Finally, we use Domain-Specific Models for different datasets and observe that we get the best results on different datasets using different strategies. As a result, we achieve **59.20 F1** and **42.36 EM** on the test set.

1 Introduction

Question Answering is the task of reading a paragraph and finding the answer to a question from the paragraph. It is one of the most popular downstream NLP tasks and there has been a lot of state-of-the-art models such as BERT [4] that can perform well on the Stanford Question Answering Dataset (SQuAD) [5]. However, models that perform well on SQuAD tend to not generalize well when evaluated on datasets that are not in the same domain as SQuAD [6]. Oftentimes, data is collected from multiple sources, so the setting where the domain of the training set is different from that of the testing set is more consistent with what we would encounter in the real world. Therefore, it is crucial to build a QA system that can perform well on data from different domains. In our project, we use data from SQuAD [5], Natural Questions [7], and NewsQA [8] as our in-domain data and DuoRC [9], RACE [10], and Relation Extraction [11] as our out-of-domain (OOD) data. The majority of our training data is from in-domain with a few examples from OOD and we build a system that can perform well when evaluated on OOD. For the baseline, we fine-tune a pre-trained DistilBERT model from Hugging Face [12] on in-domain data. Similar to task-adaptive pre-training (TAPT) mentioned in this paper [13], we take the baseline and perform multi-task training [14] [15] which jointly minimizes the QA loss on in-domain data and masked language modeling (MLM) loss on out-of-domain text. This approach tends to work well with large amounts of training data. Therefore, we have experimented with data augmentation techniques including EDA [2] and back-translation [3] in order to address the issue of insufficient OOD training data. Furthermore, in our experiments, we find that we get the best result on different OOD datasets using different models. As a result of that, we have employed Domain-Specific Models and have different models for the different OOD datasets. We also explore hyper-parameter tuning such as layer freezing in fine-tuning and it has improved our result on RACE. Our work results in **59.20 F1** and **42.36 EM** on the test set, **54.41 F1** and **41.62 EM** on the validation set with **+5.98 F1** and **+8.38 EM** improvement from the baseline.

2 Related Work

In NLP, a common approach for language models is to pre-train a model on a large corpus of unlabeled data from a variety of sources. These pre-trained models are then fine-tuned in a supervised setting for downstream tasks. [13] shows that downstream tasks can perform better if these models undergo an additional pre-train phase on data from the domain of this downstream task. The authors of the paper propose task-adaptive pre-training (TAPT) which is to further pre-train a model with task-specific data. It is challenging to directly adapt this method to our project because we do not have enough OOD unlabeled text so we need to make use of both in-domain data and OOD data to train a downstream QA model. Therefore, we extend TAPT and use multi-task training to ensure that the model can both learn QA related information and adapt to the domain of task-specific data.

MT-DNN[14] is a method that has pre-training followed by multi-task training on downstream GLUE tasks where each mini-batch is task specific and the model is updated according to the task specific objective. They show that this approach allows domain adaptation with substantially fewer in-domain labels and demonstrates better generalization ability via multi-task learning. Ernie[15] is a framework for continual multi-task training that uses previously learned parameters to initialize the model and then trains the new pre-training task together with the original task to ensure learned parameters do not forget previously learned knowledge. After that, they use this pre-trained model for fine-tuning on different downstream language understanding tasks. We use similar approach in our multi-task model of initializing most of the model parameters with in-domain fine-tuned model. We then jointly train Masked Language model on unlabeled out-of-domain text and Question-Answering (QA) task on labeled in-domain dataset. Finally, we fine-tune this model for QA task on out-of-domain dataset.

[16] shows that data augmentation can improve the robustness of a model. We studied data augmentation techniques including EDA [2] and back-translation [3]. EDA performs random insertion, random deletion, random swap, and synonym replacement to generate more data. It was originally proposed for text classification task. Back-translation is a method where we translate the original data to some other language and then back to the original language. To adapt these techniques in our QA system, we made modifications to make sure that the answer text does not change and the position of the answer in the context is updated correctly after modifying the context as mentioned in section 3.3

In [17], the authors propose Mixture of Experts. It shows that having separate networks in a model where each network handles a subset of training cases can be beneficial. Inspired by this idea, we have separate models for different OOD datasets to take advantage of the differences between the OOD datasets.

3 Approach

3.1 Baseline

The baseline and starting point of our system uses the pre-trained DistilBERT from Hugging Face [12] and performs question-answering fine-tuning on the in-domain training data. Baseline code has been provided to us for this project. We ran the code to reproduce the results on in-domain validation and OOD validation set.

3.2 Baseline + OOD Fine-tuning

The baseline model isn't exposed to any OOD information, so it doesn't score well on the OOD data. This led to our first experiment where we took the baseline and further fine-tuned it on all out-of-domain training datasets. This is relatively straight-forward as we added another case in the code to allow loading the baseline model parameters from the previous experiment as the starting point for fine-tuning on OOD dataset. This approach outperforms the baseline on the OOD validation set.

3.3 Data Augmentation

Since we only have about 100 samples in each OOD training dataset, we implemented data augmentation techniques such as word shuffling, synonym replacement, random insertions, random deletions described in EDA[2], and back-translation[3]. To produce more data for QA fine-tuning task, data augmentation is tricky as we need to ensure the answer text and answer start position is correct after modifying the context. For insertion, deletion and synonym replacement we first split the context in two parts, one before and one after the answer string, apply data augmentation on these two parts separately using library[18], then combine the two context parts and answer string, and finally recompute the answer start token. For Back-translation, we use the machine translation model[19] in Hugging Face and some back-translation reference code [20] to translate the context to Spanish and French from English and translate it back to English. In back-translation, to ensure the answer start and answer text is correct for the translated sentence, we split the context into multiple sentences, mask the sentence that has the answer string, and perform back-translation on remaining sentences. After that we insert the sentence with the answer string back in the translated context and recompute the answer start position. For back-translation, we exclude the samples where contexts have only one sentence, where answer strings have a period "." and where questions can have multiple answers. For samples with multiple Question-Answer pairs for a single context, we first expand the dataset to have unique context for each Question-Answer pair and then perform data-augmentation for each context as described above.

3.4 Pre-Training

Following the ideas described in [13], we took the baseline and performed task-adaptive pre-training (TAPT) which is to further pre-train an existing model using unlabeled task-specific data. We extracted the context sentences from OOD training datasets and did pre-training on them by performing Masked Language Modeling (MLM) task using DistilBertMaskedLM model from Hugging Face [21]. After pre-training on OOD datasets, we performed fine-tuning on OOD datasets. The result of this method was worse than the one without pre-training described in section 3.2. To investigate the reason behind its subpar performance, we compared the performance of baseline and the model that has baseline + OOD pre-training when evaluated on in-domain validation set and found that score of the latter model is much lower than the baseline, which indicates that OOD pre-training hurts the model’s ability to do QA task.

3.5 Multi-Task Training (Pre-Training on out-of-domain + Fine-Tuning on in-domain)

To remedy the issue of QA information loss with pre-training, we implement a pre-training method that uses multi-task learning objective. Instead of only optimizing the MLM loss during pre-training, our model jointly minimizes MLM loss and QA loss as shown in figure1. We create a model where

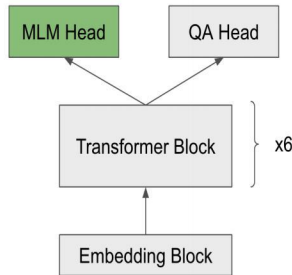


Figure 1: Multi-Task DistilBert Network Architecture

the last DistilBERT encoder transformer block has two heads, one for calculating QA loss and the other for calculating MLM loss. During training, we sample a batch for MLM which takes the labels and masked sentences and predicts the masked tokens. In the same training iteration, we sample a batch from in-domain dataset for QA prediction which takes context, question, start, and end position labels to predict the answer span start and end logits. We scale and combine these two losses and

perform back propagation on it as shown in algorithm 1. We load the model parameters from the baseline and initialize the MLM Head layers. This combined training not only prevents the model from forgetting QA related information but also allows it to adapt to the domain of OOD dataset with Masked Language Model. For the MLM task, we use data-augmentation techniques, EDA and back-translation, on the context sentences from OOD training dataset to increase the dataset size. Unlike fine-tuning data augmentation described in Section 3.3, pre-training data-augmentation is fairly simple as we don't have the issue of preserving the answer string for MLM, so we apply it directly to the entire context. After multi-task training, we take the model and perform QA fine-tuning on OOD dataset to get the best model.

Algorithm 1: Multi-Task Training

```

Load model parameters from In-Domain fine-tuned model.
Initialize MLM Head parameters.
Set the max number of epoch:  $epoch_{max}$  based on OOD MLM dataset.
for  $epoch$  in 1, 2, ...,  $epoch_{max}$  do
    for  $mlm_{batch}$  in MLM dataset do
        // $mlm_{batch}$  is a mini-batch for Masked Language Model(MLM) on OOD context.
        1. Compute MLM loss :  $L_{MLM}$ .
        2. Fetch next  $qa_{batch}$ .
        // $qa_{batch}$  is a mini-batch for QA Task from In-Domain dataset.
        3. Compute QA loss:  $L_{QA}$ 
        4. Total Loss:  $L_{Total} = \beta * L_{MLM} + \gamma * L_{QA}$ 
        // $\beta$  and  $\gamma$  are hyper parameters for scaling MLM and QA losses respectively
        5. Compute gradients.
        6. Update Model.
    end
end

```

3.6 Domain-Specific Models

Similar to the idea described in Mixture of Experts [17], we transitioned from a single fine-tuned model for all three OOD datasets to three separate models one for each of the OOD datasets: DuoRC, RACE, Relation Extraction. In addition to running the methods described above using the entire OOD dataset, we have experimented each method on DuoRC, RACE, and Relation Extraction separately and observed the best results on different datasets using different configurations. Note that unlike Mixture of Experts, we don't combine different models for three OOD datasets together using a selector gate. We keep three separate models because for the setting of this project, we know which dataset each sample comes from. To avoid confusion to the Mixture of Experts method, we will call the method described here as Domain-Specific Models.

4 Experiments

4.1 Data

We are using the SQuAD [5], Natural Questions [7], and NewsQA [8] datasets for in-domain training and DuoRC [9], RACE [10], and Relation Extraction [11] for OOD training. These datasets are already given to us and preprocessed. Each sample in the training and validation set contains context text, question, and the correct answer to the question. We have 50000 training samples for each in-domain dataset and around 100 training samples for each OOD dataset. We are also provided with validation set and test set for each of the OOD datasets. We have around 100 samples in the validation set for each OOD dataset. In the test set, we have over 1000 samples for DuoRC, 400 for RACE, and 2000 for Relation Extraction.

4.2 Evaluation method

Project handout gives the details on the evaluation. Performance will be measured using two metrics: Exact Match (EM) score and F1 score. These EM and F1 scores will be averaged across the entire evaluation dataset to get the final scores.

4.3 Experimental details

Table 1 shows F1 and EM scores we presented in the milestone report. The highest score in this table is obtained by using the method described in section 3.2. As we can see in the table, pre-training on OOD datasets drops the final score on the OOD validation datasets. To further investigate this issue,

Model	F1	EM
Baseline	48.43	33.25
Baseline + Fine-tune on OOD	50.31	35.34
Baseline + Pre-train on OOD + Fine-tune on OOD	45.44	31.15
Pre-train on OOD + Fine-tune on OOD	31.51	20.42

Table 1: Milestone Experiment Results

we evaluate the performance of baseline and the model that performs baseline + OOD pre-training on the in-domain validation set. Baseline has 70.95 F1 and 55.07 EM whereas the latter model has only 54.78 F1, 39.66 EM, indicating that OOD pre-training hurts the QA performance.

In the next experiment, we take the baseline and fine-tune it on each of the OOD datasets separately. The result is shown in table 2. Since baseline + fine-tune on OOD gave us the best result in the milestone, we want to see what the score on each individual dataset would be if we fine-tune the baseline on each OOD dataset separately instead of using all OOD data combined. Experiment setup is described in Appendix A.1. We find that Relation Extraction F1 improves by +1.5 points, RACE F1 improves by +3 points, and DuoRC F1 score is up by +3 points. The overall validation score is also higher (F1 is +2.2, EM is +1.5) than the highlighted row in table 1, indicating that having three separate models works better than using one single model on all OOD datasets.

Dataset	F1	EM
RACE	40.04	28.12
DuoRC	41.56	29.36
Relation Extraction	75.47	56.25
Overall	52.58	36.91

Table 2: Domain-Specific Models Validation Scores

After that, we fine-tune each of the models shown in table 2 with data augmentation techniques discussed in Section 3.3. Details about the dataset size are in table 7 of Appendix A.1. We have the following result in table 3. RACE doesn’t show any improvement with augmented OOD fine-tuning. For DuoRC, only back-translation helps improve the F1 by 1 point and there’s no F1 improvement with EDA. For Relation Extraction, we couldn’t perform back-translation technique as most of the contexts were single sentences with the answer string. We can see that EDA improves the F1 and EM by around 0.5 and 1.5 points respectively for Relation Extraction.

Dataset	EDA	Back-translation	EDA+Back-translation
RACE	F1:40.04, EM:28.12	F1:40.04, EM:28.12	F1:40.04, EM:28.12
DuoRC	F1:41.54, EM:31.75	F1:42.75, EM:33.33	F1:42.11, EM:30.95
Relation Extraction	F1: 76.01, EM: 57.81	-	-

Table 3: Fine-tuning Data Augmentation

In the final experiment, we take the baseline and perform multi-task training. During the training process, we collect multiple checkpoints and analyze F1 curve, EM curve for in-domain QA task and MLM loss curve for OOD dataset on Tensorboard and pick checkpoints with relative high F1 and EM on in-domain validation set and low MLM loss on OOD validation text. We then perform OOD fine-tuning on the selected checkpoints and present the best score we get for each of the dataset. Figure 2 shows the training curves for the DuoRC dataset as an example. For DuoRC, we got the best OOD QA model performance from the checkpoint at 4500th step. For MLM during multi-task training, we use augmented data with EDA and back-translation for each of the contexts from OOD training datasets. Since the MLM Head is initialized, we scale down the MLM loss by a factor of 0.1

before combining MLM Loss and QA Loss. After multi-task training, we use data augmentation in OOD fine-tuning phase as well based on the method described in the previous section and the final results are shown in table 4. Out of the three OOD datasets, we only see F1 improvement on the DuoRC dataset by almost 3 points compared to table 3. However, the performance drops slightly for the other two datasets.

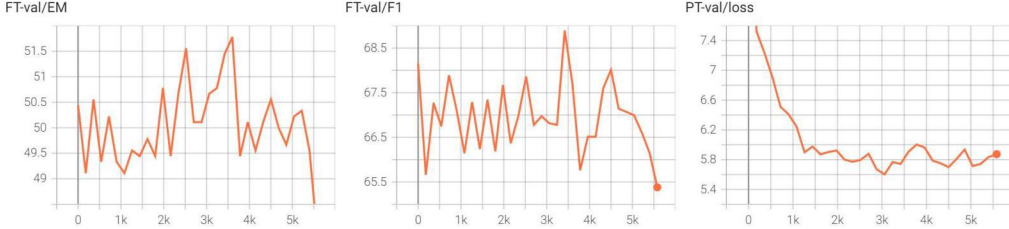


Figure 2: DuoRC Training Iteration. Here FT-Val refers to QA score on in-domain validation set and PT-Val refers to the MLM loss for OOD validation set

Dataset	F1	EM
RACE	38.32	24.22
DuoRC	45.62	37.30
Relation Extraction	74.78	57.81

Table 4: Multi-Task Training + Fine-Tuning on OOD

Finally, we have experimented with hyper-parameter tuning using layer freezing. Since bottom layers of a network usually contain more general learned information, freezing some bottom layers can make training more stable and prevent over-fitting for low resource dataset. We take the baseline and freeze some of its bottom layers before final fine-tuning on OOD data. We evaluate different number of layers to freeze and plot the result in Figure 3, where in the horizontal axis, 0 refers to only freezing the embedding layers, 1 refers to freezing the embedding layers and 0-th transformer block’s layers, etc. 6 refers to freezing all the layers except the last QA output layer. We can see that freezing only helps with RACE but not the other two OOD datasets (only the two blue curves show increasing trend) and we get the best result of 41.47 F1 and 29.69 EM on RACE by freezing all layers except the last QA output layer. We also observe that for RACE, freezing makes fine-tuning more stable as the loss doesn’t fluctuate too much across epochs with freezing. The fact that freezing works differently on each OOD dataset provides more evidence that having three separate models for each OOD dataset is better.

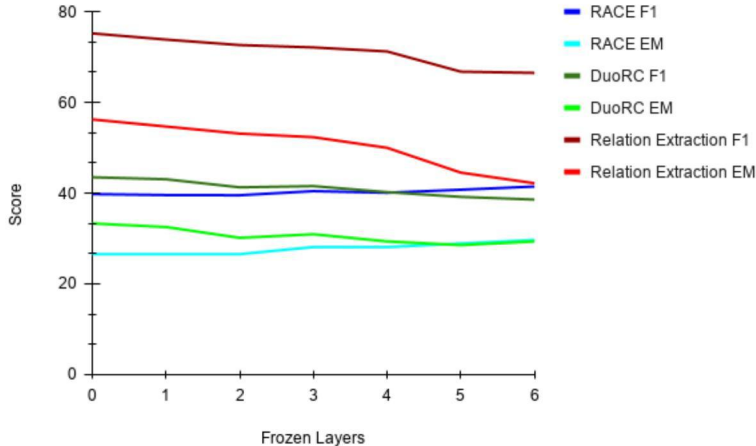


Figure 3: Layer Freezing

4.4 Results

We obtain the best result for each of the three OOD datasets using different methods and configurations. The settings for three different models are as follows:

- **RACE**: Baseline + OOD RACE fine-tuning with freezing all layers except the last QA layer of baseline before fine-tuning
- **DuoRC**: Baseline + multi-task training with EDA and back-translation + OOD DuoRC fine-tuning with EDA and back-translation
- **Relation Extraction**: Baseline + OOD Relation Extraction fine-tuning with EDA

These models give us the result in table 5 on the OOD validation set.

Dataset/Model	F1	EM
RACE	41.47	29.69
DuoRC	45.62	37.30
Relation Extraction	76.01	57.81
Our Method Overall	54.41	41.62
Baseline	48.43	33.25

Table 5: Best Result on OOD Validation Set

We got **59.203 F1** and **42.362 EM** on the test set. We’re unable to produce results on individual OOD test dataset like we did for validation set due to the constraint on test submission.

Based on our best result, the score on DuoRC has improved a lot through multi-task objective training while RACE and Relation Extraction don’t benefit from it. This is not exactly what we expected as we hoped multi-task training could help in improving performance on all OOD datasets. One possible reason could be that we don’t have enough unlabeled text during multi-task training or the texts produced by EDA and back-translation don’t have enough variation. We are also surprised that EDA and back-translation don’t help with every OOD dataset. We did notice that some sentences produced by these data augmentation techniques still look similar to the original sentence. It could be that some other pivot languages might work better than Spanish and French we chose in back-translation.

5 Analysis

Based on the result, our model performs reasonably well for Relation Extraction QA dataset which has shorter context length but shows poor performance with longer context sequences that are common in RACE and DuoRC datasets. One possible explanation for this behavior is that with longer context, it’s harder for the model to learn very long dependencies and it’s more likely for it to get confused and pick out the wrong answer from the context based on the question type. Also, we think that questions and context in the Relation Extraction dataset are more straight-forward compared to the other two datasets. Table 8 in the Appendix shows some examples from the Relation Extraction dataset.

RACE dataset has lowest performance amongst all the OOD datasets and its performance drops if we fine-tune it further on even small training set. Layer freezing helps with RACE and we get the best result when we only train the final QA head and freeze all the other transformer encoder blocks and the embedding block. We also analyzed the predicted answers for RACE on Tensorboard and observed that the dataset has several questions starting with "Which", where the question asks to find the object with a desired attribute from the list of objects. For these questions shown in Table 9 in the Appendix, the model seems to be picking any one of the object without focusing on its attributes. For the RACE dataset, we also found that some of the questions are related to counting the number of certain items mentioned in the context. The context doesn’t contain the direct answer and it requires the model to actually understand the text and perform counting in order to get the answer. One example can be found in Table 6. It’s interesting to see that the model can predict the correct sports items in this example. Another similar example is in Table 10 in the Appendix.

DuoRC dataset is based on movie plots from Wikipedia and IMDB. The dataset has several questions starting with "Who". Upon inspecting the predictions, it looks like the context contains a lot of movie

Question	How many sports items are mentioned in the passage?
Context	Indoor and outdoor courts. Coaching from beginners to advanced, every day not evenings.Children only-Sat. mornings. SKING Dry slopes-3 levels instructors at weekends and Fridays. Daytime parcitice. 8 years upwards. SWIMMING 2 pools I heated Olympic length. Tuition available. Women: Tuesday and Thursday. Men: Monday, Wednesday and Friday. Children: Saturday Family day: Sunday GOLF 9 hole practice course. Professional Coaching. Lessons must be booked in advance in daytime. Evening practice. Minimum age - 9 years. GYMNASTICS Maximum age-18 years. Children aged 5-10. Monday and Wednesdays. 4:00-6:00 p.m. 10-18 year-old. Friday evenings. Bar work on Sunday mornings. AND MUCH MORE Table Tennis, Snooker, Darts , Chess(everyday and evening), Cafe (all day),Bar(lunch time and evenings), Nursery(weekdays and weekends, not evenings). Centre open 10:00 a.m.-10:00 p.m. Daily. Interested? More details inside. Quote Card Number: 99.
Answer	9
Prediction	Table Tennis, Snooker, Darts , Chess

Table 6: RACE Validation Set Prediction for Question type "How many"

character names and actor names. The model is unable to make correct predictions for questions where the answer is not direct and needs deeper context understanding. It could guess the "Who" questions correctly where the movie character's and actor's names are close to each other in the context. Appendix shows examples of both hard (Table 12) and easy (Table 11) "Who" type samples. We think that a bigger model might perform better for cases where we have such long dependencies.

6 Conclusion

In this project, we implemented multi-task training, added data augmentation including EDA and back-translation to our system, and employed Domain-Specific Models to have separate models for each OOD dataset. Through our experiments, we show that multi-task training is more beneficial than simply pre-training the model with OOD context texts and data augmentation can help in improving the model's performance on OOD datasets in some cases. We also demonstrate that building separate models for each OOD dataset works better than a single model. One limitation of our work is that we don't have enough unlabeled OOD context data for multi-task training. While we tried to collect some data from similar sources for RACE and DuoRC by expanding the unlabeled text for RACE using MCTest [22] from similar domain and DuoRC using datasets from Kaggle [23] [24], we didn't see any performance gain on the downstream fine-tuning QA task on these datasets. The training time was over a day, and we could only perform limited experiments. We wanted to try similar experiments for Relation Extraction but were not able to download the relevant data source. For future work, collecting more context data can help multi-task learning and potentially improve the result of multi-task learning on RACE and DuoRC. We can also look into increasing the variation of sentences produced by data augmentation either by using other pivot languages in back translation or using unsupervised Question Answer generation techniques [25] to generate more data.

7 Acknowledgments

We would like to thank the CS224N course instructors and teaching assistants for implementing the baseline model and providing guidance throughout our project development.

References

- [1] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019.
- [2] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks, 2019.
- [3] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. *CoRR*, abs/1808.09381, 2018.
- [4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.
- [5] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250, 2016.

- [6] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. *CoRR*, abs/1707.07328, 2017.
- [7] 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. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 2019.
- [8] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset. *CoRR*, abs/1611.09830, 2016.
- [9] Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. DuoRC: Towards complex language understanding with paraphrased reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1693, Melbourne, Australia, July 2018. Association for Computational Linguistics.
- [10] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.
- [11] Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. *CoRR*, abs/1706.04115, 2017.
- [12] DistilBERT Question Answering Model by Huggingface. https://huggingface.co/transformers/model_doc/distilbert.html#transformers.DistilBertForQuestionAnswering.
- [13] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don’t stop pretraining: Adapt language models to domains and tasks, 2020.
- [14] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy, July 2019. Association for Computational Linguistics.
- [15] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. Ernie 2.0: A continual pre-training framework for language understanding, 2019.
- [16] Shayne Longpre, Yi Lu, Zhucheng Tu, and Chris DuBois. An exploration of data augmentation and sampling techniques for domain-agnostic question answering. *CoRR*, abs/1912.02145, 2019.
- [17] Robert Jacobs, Michael Jordan, Steven Nowlan, and Geoffrey Hinton. Adaptive mixture of local expert. *Neural Computation*, 3:78–88, 02 1991.
- [18] EDA Implementation. https://github.com/jasonwei20/eda_nlp.
- [19] MarianMT Model by Huggingface. https://huggingface.co/transformers/master/model_doc/marian.html.
- [20] Text Data Augmentation with MarianMT. <https://amitness.com/back-translation/>.
- [21] DistilBERT Masked LM Model by Huggingface. https://huggingface.co/transformers/model_doc/distilbert.html#transformers.DistilBertForMaskedLM.
- [22] Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

- [23] Wikipedia Movie Plots. <https://www.kaggle.com/jrobischon/wikipedia-movie-plots>.
- [24] IMDB Movie Plots. <https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset>.
- [25] Patrick S. H. Lewis, Ludovic Denoyer, and Sebastian Riedel. Unsupervised question answering by cloze translation. *CoRR*, abs/1906.04980, 2019.

A Appendix

A.1 Experiment Settings

For Domain-Specific models and data-augmentation experiments, we used batch size of 16, max sequence length of 384 and stride of 128, learning rate of $3e-5$ and Adam Optimizer. We evaluated on the model for which we get the best validation score. For multi-task training, we chose batch size of 12 due to capacity limits of the GPU. Table 7 shows the number samples for each of the OOD dataset. For EDA, we generated 3 new samples for each original input sample. Based on the recommendations in [2] for a given input dataset size, we set the parameters: *synonym replacement: 0.2, random insertion: 0.05, random swap: 0.05, random deletion: 0.05* in the library’s helper function[18]. For back-translation, for each valid input we generated two back-translated samples, one from Spanish and other from French.

Dataset	No-data augmentation	EDA	Back-Translation
RACE	196	470	363
DuoRC	297	820	598
Relation Extraction	127	381	-

Table 7: Fine-tuning Training Samples for Domain-specific Model and Data-Augmentation techniques

A.2 Relation-Extraction Validation-Set Predictions

Question	Who published Galaxian 2?
Context	Galaxian 2 (also written as Galaxian II) is a handheld electronic game that was released in 1981 in the US by Entex Industries.
Answer	Entex Industries
Prediction	Entex Industries
Question	What city is the headquarters of LATAM Airlines Group?
Context	LATAM Airlines Group S.A. is a Latin American airline holding company incorporated under Chilean law with its headquarters Santiago, Chile.
Answer	Santiago
Prediction	Santiago, Chile
Question	Which sports team is Ali Sadiki playing for?
Context	Ali Sadiki (born 10 December 1987) is a Zimbabwean professional footballer, who plays as a defender for TP Mazembe in DR Congo.
Answer	TP Mazembe
Prediction	TP Mazembe

Table 8: Relation Extraction Validation Set Predictions

A.3 RACE Validation-Set Predictions

Question	Which hotel is close to airports?
Context	Convenient location 5 minutes from Washington ,D.C. and one block to Crystal City METRO ,Crystal City Underground. Alexandria. Free Parking, free Deluxe Continental Breakfast and complimentary Hotel Shuttle Service to Ronald Reagan Washington National Airport. 14000 Jefferson Davis Hwy(endurance on S. Eads St.) Phone (703) 979-3722 DAYS INN ARLINGTON Days Inn Arlington is located two miles from the Pentagon ,Lincoln Memorial .There is a restaurant which serves American style food ,plus two meeting rooms and an outdoor pool. Complimentary shuttle service runs to the Rosslyn Metro Station. Complimentary parking is available on the property .An experienced staff welcomes our guests. Phone(703)525-0300 THE HIGHLANDER MOTOR INN Comfortable rooms at moderate rates . Convenient location only minutes away from Washington area attraction. Near VA Square METRO Station and Ballston Commons Shopping Mall. Free continental breakfast ,free parking. Phone(703)524-4300 HYATT REGENCY CRYSTAL CITY Conveniently located next to Washington National Airport and two miles from Washington D.C. Complimentary Shuttle to and from National Airport ,metro ,local restaurants and shopping .Experience a view to remember at out rooftop restaurants "Chesapeake Grill." Call for Reservations. Phone(073)418-12341
Answer Prediction	HYATT REGENCY CRYSTAL CITY HIGHLANDER MOTOR INN
Question	Which website should you visit if you want to take on some voluntary work?
Context	Lake Forest High School Clubs Environmental Club Club members are provided opportunities to realize their goals of environmental service. The majority of projects are student-initiated. Last year's activities included: the LFHS Courtyard Garden upkeep, Bike to School Day with free cocoa and snacks, an "eat local" 100-mile dinner, recycling solutions for LFHS, beach cleanup through the Great Lakes Alliance, and Earth Week celebrations. Meetings: First Friday of the month in Room 5 at 7:00 a.m. Advisor: Ms Mary Beth Nawor, _ Peer Tutoring Students volunteer to help fellow students in improving their academic and organizational skills. Students tutor LFHS and middle school students on an individually scheduled basis. LFHS tutors must have proficiency in the academic area in which they wish to tutor, but middle school tutors only need general academic proficiency. Meetings: All tutoring is done on an individually scheduled basis. Advisor: Ms Kathy O'Hara, _ Young Idea Young Idea is LFHS's art and literary magazine. Students of all ages who love writing and art are encouraged to become part of the staff of this award-winning magazine. Young Idea encourages all students to submit art and literature works to the magazine, whether they are a part of the staff or not. From September to February, Young Idea meets on Thursdays after school in the Public Room to discuss the pieces that have been submitted to the magazine and provide feedback for the authors. Meetings: See above explanation. Advisor: Ms Debbie Zare, _ Scout Buddies Scout Buddies is a friendship club which helps build friendship among individuals with and without disabilities. Club members participate in a variety of social activities both within the school and in the community. Activities include organizing holiday parties in the school, bowling, going to the movies, and eating out in local restaurants. Meetings: Monthly meeting dates vary, but in Room 134. Advisor: Ms Donna Lovitsch, _
Answer Prediction	Peer Tutoring Lake Forest High School Clubs Environmental Club

Table 9: RACE Validation Set Predictions for Question type "Which"

Question	How many environment-friendly inventions are mentioned in the passage?
Context	Inventors are always looking for ways to make our lives easier, greener and a whole lot more fun. Take a look at the following breakthroughs in science and technology. What in the world will inventors dream up next? Superhero Suit Have you ever dreamed of becoming a superhero? Dream no more. There's a real Iron Man suit, XOS 2, which instantly transforms the person wearing it. The suit provides the power to lift 200 pounds with ease and break slabs of wood with a single karate chop. It was designed to help the military with heavy lifting. One person in the suit could do the work of three soldiers. Up, Up and Away! It took him 30 years to develop it, and now Glen Martin's invention is ready to take off. The Martin Jetpack allows its operator to fly 8,000 feet into the air. Unfortunately, you can't soar through the skies too long. The jetpack holds only 30 minutes' worth of fuel. The aircraft will sell for \$100,000. Hardworking Robot The EMIEW2 robot is the perfect office helper. Need a document delivered? No problem! It can also guide visitors to their destinations. The three-foot-tall robot can identify different human voices and respond to commands. One day, it may serve as a receptionist or a security guard. Green Machine Say goodbye to gasoline! The AirPod car runs on air power. That means this car won't pollute the environment. A high-pressure air tank can fill the car in minutes. The three-wheeler can travel about 130 miles between fill-ups. Beef Power All aboard the beef train! Amtrak's Heartland Flyer runs partly on fuel that is made from cow fat. The goal is to reduce carbon emissions by 10%. The train travels between Oklahoma City, Oklahoma, and Fort Worth, Texas. Robo-Guard EMILY is a robotic, four-foot-long buoy . She can swim through riptides at up to 24 miles per hour. Her inventor, Tony Mulligan, says that's 15 times as fast as human lifeguards! EMILY is powered by a tiny electric pump and operated by remote control.
Answer Prediction	three 15 times

Table 10: RACE Validation Set Predictions for Question type "How Many"

A.4 DuoRC Validation-Set Predictions

Question	Who plays cruella de vil?
Context	American video game designer Roger Dearly (Jeff Daniels) lives with his pet dalmatian, Pongo, in London. One day, Pongo sets his eyes on a beautiful female dalmatian named Perdita. After a frantic chase through the streets and into St. James's Park, Roger and Pongo discover that Perdita likes Pongo. Her owner, Anita Campbell-Green (Joely Richardson), falls in love with Roger when they meet. After they both end up falling into the lake as a result of their dogs, they return to Roger's home, and Anita accepts his proposal. They get married along with Perdita and Pongo. Anita works as a fashion designer at the House of de Vil. Her boss, the pampered and very glamorous Cruella de Vil (Glenn Close) has a deep passion for fur, going so far as to have a taxidermist, Mr. Skinner, skin a white tiger at the London Zoo to make into a rug for her. Anita, inspired by her dalmatian, designs a coat made with spotted fur. Cruella is intrigued by the idea of making garments out of actual dalmatians and finds it amusing that it would seem as if she was wearing Anita's dog. Anita soon discovers that Perdita is pregnant, and is then informed that she is too, much to her shock. Some time later, Cruella visits their home, and expresses contempt upon meeting Roger. Her initial disgust at them having a baby turns to excitement when she finds out Perdita is expecting too. Several weeks later, she returns when a litter of 15 puppies are born and offers Roger and Anita £7,500 for them, but they refuse. She dismisses Anita and vows revenge against her and Roger. She has her henchmen, Jasper and Horace (Hugh Laurie and Mark Williams) break into their home and steal the puppies while Roger and Anita are gone for a walk at the park taking Pongo and Perdita with them. Along with 84 others that were previously stolen, they deliver them to her ancient country estate, De Vil Mansion. She also hires Mr. Skinner to kill and skin them to create her coat. With the family devastated at the loss of their puppies, Pongo uses the Twilight bark to carry the message via the dogs and animals of London, while Roger and Anita notify the Metropolitan Police Service. A dog who had witnessed the stolen puppies follows Jasper and Horace to the mansion, and finds all of them inside before helping them escape under the duo's noses. They make their way to a nearby farm, where they are later joined by Pongo and Perdita. Cruella arrives at the mansion and soon discovers what has happened. Furious, she decides to carry out the job herself, whilst Jasper and Horace attempt to search for them also. After several mishaps, Jasper and Horace discover nearby police on the hunt for Cruella and her henchmen and hand themselves in, joining Mr. Skinner who was beaten earlier while trying to kill Lucky, who had been left behind. Meanwhile, Cruella tracks the puppies to the farm where they are hiding and tries to retrieve them. However, they outwit her and cause her to fall into a vat of molasses and get thrown through a window into a pigpen. Shortly after, the fleeing dalmatians (including Lucky) are found and sent home via the Suffolk Constabulary, while those looking for Cruella arrive at the farm to arrest her. In the police van, she belittles Jasper, Horace, and Skinner for their incompetence before they are sprayed by a skunk which she had mistaken for her bag. Pongo, Perdita, and their puppies are reunited with Roger and Anita. After being informed that the remaining 84 puppies have no home to go to since they have not been claimed, they decide to adopt them, bringing the total to 101. Roger designs a successful video game featuring dalmatian puppies as the protagonists and Cruella as the villain, and they move to the English countryside with their millions.
Answer	Glenn Close
Prediction	Glenn Close

Table 11: DuoRC Validation Set Predictions for EASY Question type "Who"

Question	Who kills Leaven ?
Context	The film opens with a man, Alderson (Julian Richings), waking up in a cube-shaped room with glowing, computer circuit-like walls and six doors, one at the center of each wall, ceiling and floor. After recovering from his confusion, he opens two of the doors and looks into them to find rooms that differ to the one he is in only by color. He then opens and goes through a third door. He looks around and then takes a step, but is suddenly cut into large cubes. He falls apart and the rack of crosshatched wires which diced him moves into view. It folds up and retracts. Later, in another room, several people find each other: Quentin (Maurice Dean Wint), Worth (David Hewlett), Holloway (Nicky Guadagni), Rennes (Wayne Robson), and Leaven (Nicole de Boer). None of them know where they are, how they got there, or why they are there. Quentin, however, knows that there are traps, as he had looked into a room and nearly got his head cut off. The five decide to stay together and look for the way out. Rennes takes the lead. He exemplifies how to test for traps by tossing a boot into the rooms while holding onto the laces, to trigger potential traps, figuring that the trapped room contains motion detectors. Holloway speculates on several possible creators of the Cube, being aliens or the government his main ptions. Rennes remarks that staying still will not solve anything, and says that they should move in a straight line until they get to the end. The others agree, and they begin moving through the rooms. While moving, they discover a series of different numbers on the hatchways between each of the rooms. At one point, Rennes throws the boot in and does not find anything, but detects that the room has dry air, and deduces that it most likely employs an electrochemical sensor, which detects hydrogen sulphide emitted from the skin. Quentin realizes that Rennes is an escape artist who has escaped more than seven major prisons. Soon after, Rennes jumps into a room tested with a boot, and is sprayed in the face with acid. The others pull him back, but he dies as the acid corrodes his face and the inside of his head. The group decides that the room must have contained an electrochemical sensor which Rennes missed, and realize that they must find a better way of testing the rooms. Quentin asks everyone about their occupations. He says he is a police officer, Holloway says she is a doctor, and Worth says he works "in an office building, doing office building stuff." Leaven claims simply to "hang out" with her friends. Quentin believes that nothing is a coincidence, that each of them has a purpose in the cube. After Holloway talks about her rings and broaches, Quentin asks why Leaven has her glasses, while Holloway has had her jewellery taken away. Leaven reveals herself to excel at mathematics, and after looking at the numbers on a crawlspace, theorizes that when one of those numbers is prime, the room is booby-trapped. Leaven's purpose becomes attempting to "crack the Cube's code", and they progress through the cubes. When they find themselves in a room with trapped rooms all around and below, Quentin checks the door in the ceiling, through which falls a seventh person: Kazan (Andrew Miller). He appears to be mentally handicapped. At least two of the others see him as a burden, but Holloway decides to bring him along. The group starts speculating about their surroundings, which causes a conflict between Quentin and Holloway. Quentin dismisses Holloway's ideas as conspiracy theories, and Holloway thinks that Quentin is naive. Soon after this, Quentin enters a room without prime numbers and narrowly avoids death from a trap consisting of rotating razor wires. Leaven's theory that non-prime-numbered rooms are safe is shown to be incorrect. Quentin begins suspecting that Worth is a spy, and is increasingly irritated by Kazan's mental state. The group rests, while Leaven attempts to decipher the numbers
Answer Prediction	Quentin Rennes
Question	Who do Muntze and Ellis suspect to be the traitor?
Context	At the Kibbutz Stein in Israel in 1956, school teacher Rachel (Clarice van Houten) is reunited with Ronnie (Halina Reijn), a friend from The Hague during World War II. After Ronnie leaves, Rachel reflects on her adventures during the final days of the war. In 1944, Rachel hides from the Nazis in a farm in the Dutch countryside. In exchange for hiding her, they expect her to learn Bible verses. One day, a crippled Allied bomber drops its payload on the farmhouse, killing everyone except Rachel. Rob (Michiel Heisman), a young man from a neighboring farm, hides her in the family's greenhouse. That night, Van Gein (Peter Blok), a police officer, arrives to tell them that the Nazis know that Rachel is in the area and will hunt her down. He agrees to help Rachel and Rob escape to the Allied controlled southern part of Holland. Rachel visits her father's lawyer, Smaal (Dolf de Vries). He gives her enough money and jewels to live on for a year but warns her not to trust people so easily. Van Gein leads Rachel and Rob to a dock where other Jews wait to leave. Rachel is reunited with her parents and brother, who is recovering from an emergency appendectomy. Van Gein does not accompany the Jews on the boat trip. That evening, the boat is ambushed by a Nazi patrol boat. The Nazis immediately open fire; only Rachel survives the massacre. Before drifting down the river, she sees the Nazis loot the corpses. Rachel is found by Resistance fighters who smuggle her into The Hague by disguising her as a typhoid victim and placing her in a coffin that if properly "sealed" has sizable air holes. She is taken to a soup kitchen run by another member of the Resistance, Gerben Kuipers (Derek de Lint) and given the new name of Ellis de Vries. Eventually she is made a part of the Resistance's plans to smuggle in British guns and rations. The smugglers are led by Hans Akkermans (Thom Hoffman), an expert marksman. He and Ellis are to pose as husband and wife so the Nazis will not search their luggage on the train; the luggage is actually full of weaponry. But when the Nazi soldiers on the train clearly do intend to search all baggage, a new plan is needed. Ellis takes the bags and enters a private compartment occupied by SD Colonel Muntze (Sebastian Koch). The Nazis don't search Muntze's compartment. Ellis and Muntze are clearly attracted to each other and she accepts an offer to visit him at his office. Hans is clearly jealous. A truck carrying the British guns crashes in front of the soup kitchen. Kuiper's son, Tim (Ronald Armbrust) was driving and is arrested by the Gestapo. While the others go into hiding, Ellis arranges a meeting with Muntze hoping that she can persuade him to release Tim. Knowing he is an avid stamp collector, she takes some rare Dutch stamps to him. He invites her to a Nazi party. There she sees SS commandant Günter Franken (Waldemar Kobus) and recognizes him as the Nazi who led the ambush against the refugee boat. Although she is sickened at his sight, she manages to sing at the party later. She and Muntze return to his suite to make love. He intuits that she is Jewish from her dyed blond hair but has fallen in love with her. She accepts a job in his office and begins work the following day. There she meets Ronnie, who is Franken's secretary/sex partner. Franken gives them a report indicating that Tim has confessed everything and is to be executed, however, Muntze refuses to sign the execution order. Ellis sees Smaal at Nazi headquarters and learns that he and Muntze have negotiated a cease fire – if the Resistance ceases its attacks against the Nazis, the Nazis will cease its violent reprisals against Dutch civilians. However, when a hidden microphone placed by Ellis in Franken's office reveals that Franken and Van Gein have been working together to kill and rob Jews trying to escape into Allied territory, a controversy amongst the Resistance fighters ensues. Van Gein is heard -
Answer Prediction	Smaal Franken and Van Gein

Table 12: DuoRC Validation Set Predictions for HARD Question type "Who"