Domain-agnostic DistilBERT for robust QA

Stanford CS224N Default Project

Medina Baitemirova
Department of Computer Science
Stanford University
medinab@stanford.edu

Abstract

This project aims to improve the robustness of DistilBERT to out-of-distribution data in a question answering task by employing multi-phase continued pretraining and data augmentation. For multi-phase pretraining, we first analyze the domain similarity between in-domain and out-of-domain datasets, and find NewsQA to be the most similar dataset to the downstream task of question answering based on examples from DuoRC, RACE, and RelationExtraction datasets. We first train the model on in-domain datasets, and call it the second-phase continued pretraining. After using NewsQA for third-phase continued pretraining, we use data augmented with synonym and antonym replacement to perform the fourth-phase pretraining. The best model achieved performance, as evaluated by EM/F1 score, of **35.60/51.23** on validation datasets and **40.39/59.42** on test datasets in comparison to the baseline of **29.06/46.14** on validation datasets.

1 Mentor

- Rachel Gardner

2 Introduction

A common practice in natural language processing has been the use of large pretrained language models, or LM. One example of such models is RoBERTa, an extension of BERT, a transformers model that is pretrained using masked language modeling.[1] These LM are first trained on heterogeneous raw data in a self-supervised way and later, in continued pretraining, fine-tuned on a specific task. Due to the strong performance of large language models on various tasks, one might conclude that these models constitute a universal LM with no further need to train on domain-specific data.

Interestingly, even large language models, such as RoBERTa, benefit from further pretraining. However, most studies have used a single domain with data that is smaller and less diverse. Examples include ULMFiT pretrained on English Wikipedia and its variation pretrained on English tweets.[2]

In addition, some work shows that LM benefit from second-phase pretraining using task-adaptive and domain-adaptive pretraining, which entails further training the LM on related domains or even on a dataset that is directly related to the downstream task.[3] However, this work doesn't evaluate those techniques on smaller models, such as DistilBERT, which is another extension of BERT that has 40% less parameters.[4]

A broader question still remains: is it possible to build a universal language model that is pretrained on an enormous body of data and generalizes beyond its training distribution? If this was accomplished, the result would be a model that somewhat resembles human generalization with no further need for domain-specialized LM models. Indeed, the general trend in NLP have been larger models pretrained on even larger amounts of data. One such example is GPT-3 trained with 175 billion parameters on 45TB of data from various sources.[5]. This model is so large that it could potentially overfit to a significantly sized corpus. Another issue stems from the computational intensity of models

such as GPT-3 that limits their usability and practicality. With larger models, pretraining requires significant computational resources. For example, in the case of GPT-3, it might take up to 355 years to train the model.[6]. Therefore, there is a critical need for smaller, domain-agnostic models, such as DistilBERT, that have the potential to be more accessible.

With smaller models, generalization is still an issue. Therefore, this project aims to explore multi-phase continued pretraining and data augmentation techniques to improve the performance of DistilBERT on out-of-domain distributions.

3 Related Work

3.1 Data Augmentation

There have been numerous studies exploring different data augmentation techniques for NLP. One common method for augmenting data in a low-resource setting is synonym replacement, where words in a particular text are replaced by their synonyms, often in a random fashion. According to work done by a group at Apple, data augmentation techniques, such as back-translation, weren't as effective at producing domain-agnostic question answering models.[7]

3.2 Multi-Phase and Domain-Adaptive Continued Pretraining

Benefits of multi-phase continued pretraining, such as domain-adaptive pretraining, or DAPT, have been established. One work evaluates the usefulness of domain-specific continued pretraining using RoBERTa as the baseline model and shows that continued pretraining on domains that are related to the downstream task improves the performance of RoBERTa on out-of-domain distributions.[3]

Another work shows the benefits of multi-stage pretraining in general and especially for low-resource datasets.[8].

4 Approach

4.1 Data Augmentation

The information about the datasets used for training, validation, and testing is summarized in Figure 1. Due to the small size of out-of-domain test datasets, we first use random synonym replacement for paragraphs, also called as contexts, to increase the amount of data by 100%. We also use antonym replacement for the context and related questions to further increase the amount of available data used for pretraining.

4.2 Similarity Analysis

To take advantage of the domain-adaptive pretraining, we first analyzed the similarity between in-domain and out-of-domain datasets to identify a dataset that is the most similar to DuoRC, RACE, and RelationExtraction and can be used for the third-phase domain-adaptive pretraining. This pretraining phase also ensures that the model is not overfitting to unrelated domains.

We isolated context from 6 documents, where each document represents one of the 6 datasets: SQuAD2.0, NewsQA, Natural Questions, DuoRC, RACE, and RelationExtraction. Text containing context only was later tokenized and relevance for each word was calculated using TF-IDF, or term frequency–inverse document frequency. TF-IDF takes two metrics into account when calculating how important a word is: how frequent that word is in one document and how many documents contain that word among all documents of interest. The formula for TF-IDF is illustrated below, where we calculate the importance of word i in document j as follows:

$$w_{i,j} = tf_{i,j} * log(\frac{N}{df_i}) \tag{1}$$

Where $tf_{i,j}$ stands for number of occurrences or word i in document j, df_{ij} stands for number of documents that contain word i, and N stands for the total number of documents of interest.[9]

Later, we isolated top 1000 words from each document (excluding certain tokens, such as "EEPE" and "BPB") and calculated percent overlap between all words for all 6 documents.

4.3 Multi-Phase Continued Pretraining

4.3.1 First-Phase Training

First-phase in the context of this project refers to the training of DistilBERT on Toronto Book Corpus and English Wikipedia.

4.3.2 Second-Phase Continued Pretraining

There were three different second-phase pretraining methods:

- Further pretraining DistilBERT on the entire in-domain set
- Further pretraining DistilBERT on NewsQA only
- Further pretraining DistilBERT on the entire out-of-domain set

4.3.3 Third-Phase Continued Pretraining

We refer to these methods as third-phase continued pretraining:

- After in-domain pretraining, further pretraining DistilBERT on NewsQA only
- After NewsQA pretraining, further pretraining DistilBERT on out-of-domain distribution only, including datasets generated by data augmentation.

4.3.4 Fourth-Phase Continued Pretraining

For models that were pretrained on in-domain distributions followed by NewsQA continued pretraining, we have also performed a fourth-phase continued pretraining using out-of-domain distributions, including datasets generated by data augmentation.

5 Experiments

5.1 Data

DistilBERT was originally trained on Toronto Book Corpus and English Wikipedia, therefore, the domain distribution is heterogeneous.[4] In-domain dataset that consists of a union of SQuAD2.0, NewsQA, and Natural Questions datasets, also qualifies as a heterogeneous domain dataset. SQuAD2.0 and Natural Questions are based on Wikipedia data while NewsQA is based on articles from CNN.[10][11][12] Our target task dataset consist of RelationExtraction dataset based on Wikipedia, RACE based on passages from English examinations in China, and DuoRC is based on movie plot from Wikipedia and IMDb.[13][14][15]

Information about the data used in this project is summarized in Figure 1.

Dataset	Question Source	Passage Source	Train	dev	Test
`in-domain datasets`					
SQuAD	Crowdsourced	Wikipedia	50000	10,507	-
NewsQA	Crowdsourced	News articles	50000	4,212	-
Natural Questions	Search logs	Wikipedia	50000	12,836	-
`oo-domain datasets`					
DuoRC	Crowdsourced	Movie reviews	127	126	1248
RACE	Teachers	Examinations	127	128	419
RelationExtraction	Synthetic	Wikipedia	127	128	2693

Figure 1: Statistics for datasets that were used for this project. **Question Source** and **Passage source**data sources from which the questions and passages were obtained. Table borrowed from [16].

The task is to answer questions based on passages in DuoRC, RACE, and RelationExtraction.

An input for the question answering task is a paragraph and a question about that paragraph. Output contains the span of the answer, and the text containing that answer.

An example of an input and an output is illustrated in Figure 2.

Examples of input and output		
	Input	Output
Example 1	**Context:** *LATAM Airlines Group S.A. is a Latin American airline holding company incorporated under Chilean law with its headquarters Santiago, Chile.* **Question:** *What city is the headquarters of LATAM Airlines Group?*	**Answer:** answer_start: 123 text: "Santiago"
Example 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.* **Question:** *Who published Galaxian 2?*	**Answer:** answer_start: 110 text: "Entex Industries"

Figure 2: Two examples containing an input (a paragraph and a question) and an output (an answer span and answer text).

5.2 Evaluation method

One metric used in this project to evaluate the performance of the model is the Exact Match, or EM, score. It measures the match between the input and the output. Another metric used for this project is the F1 score, which balances the individual impacts of precision and recall by taking a harmonic mean. F1 and EM scores can be useful when comparing the relative model performance on training and test sets (both in- or out-domain). The complete code for computing evaluation metrics has been provided by the CS224N team.

5.3 Experimental details

5.3.1 Model configurations

We used DistilBERT and the code provided by the CS224N team. The learning rate of 3e-5, random seed of 42, and training for 30 epochs with manual early stopping was kept the same throughout all experiments.

5.4 Results

5.4.1 Data Augmentation

Examples of both synonym and antonym replacements can be found in Figure 3.

According to these examples, and according to the general trend among most examples that were manually reviewed, synonym replacement resulted in a poorer quality text than antonym replacement. In this project, we selected synonyms randomly from a predefined list of synonyms, therefore, the selected word was not always preserving the original meaning of the sentence and is potentially destroying context cohesiveness. One example is when "who" was replaced with "World Health Organization" that is abbreviated as "WHO". Using other methods, such as near-synonym replacement, could have been a better option.[17] Antonym replacement was an original idea by the authors of this project and was used to test a middle ground between synonym replacement and negative sampling. Negative sampling in some contexts is creating unanswerable questions or questions with fake and unlikely answers, often manually. According to a study, negative sampling can be an effective data augmentation technique.[7]

5.4.2 Similarity analysis

Similarity analysis showed that NewsQA had the biggest percent similarity with DuoRC, RACE, and RelationExtraction datasets. Findings are summarized in Figure 4.

Examples of data augmentation		
Original context	Synonym Replacement	Antonym Replacement
In olden times, England is in turmoil. With the death of the King, noone can decide who is the rightful heir to the throne. With war threatening to tear the country asunder, a stone and anvil appear from the heavens in London town, with a sword planted firmly in the anvil	inch olden time , England be indium whirl . With the end of the King , noone can decide world_health_organization incarnate the true heir to the throne . With war endanger to rake the state_of_matter apart , a rock_candy and incus appear from the celestial_sphere indium London township , with a sword plant securely inch the incus	In olden times , England differ in turmoil . With the birth of the queen , noone hire decide who differ the rightful heir to the dethrone . With peace threatening to tear the urban_area asunder , a stone and anvil disappear from the Hell in London town , with a sword unplanted firmly in the anvil
A lonely old woman who longs for a child is given a seed by a good witch. When planted, the seed grows into a flower, and inside the blossom is a tiny girl the size of the old woman's thumb. The old woman names the girl Thumbelina and raises her as her own	A lonely previous womanhood world_health_organization hanker for a child beryllium sacrifice a seed digression a good hag . When deep-rooted , the source develop into a flower , and inwardly the flower constitute a bantam daughter the size of the erstwhile womanhood 's hitchhike . The old womanhood diagnose the female_child Thumbelina and pilfer her adenine her own	A lonely young man who longs for a parent differ take a seed by a evil witch . When unplanted , the seed grows into a flower , and outside the blossom differ a tiny male_child the size of the young man 's thumb . The young man names the male_child Thumbelina and descent her as her own

Figure 3: Examples of synonym replacement and antonym replacement.

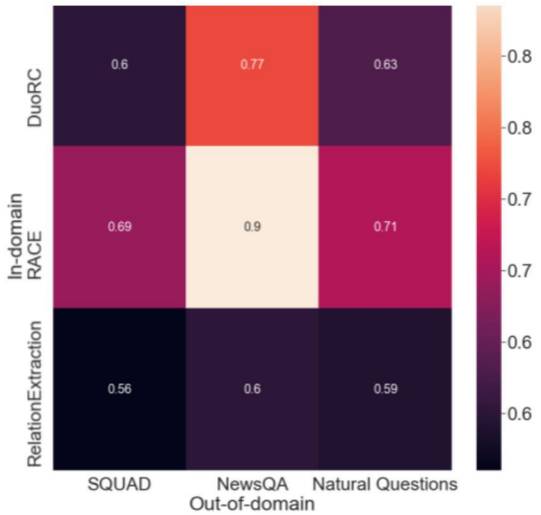

Figure 4: Similarity index, reported as percent similarity between two datasets.

It was surprising that NewsQA had the most similar domain to the downstream task. Since two in-domain and one out-of-domain datasets were based on Wikipedia, we expected SQuAD or Natural Questions datasets to be the most similar.

For this project, we used word importance and isolating top 1000 words in every document, and that method might be too crude to accurately analyze for overlaps between datasets. Another option could be using bag-of-words language models, however, the concept of isolating top most important words would remain the same.

5.4.3 Multi-Phase Continued Pretraining

The results of all multi-phase pretraining experiments are summarized in Figure 5. First three columns contain information about various datasets used for each training phase, while the last two columns show EM and F1 scores as metrics of evaluation of model's performance. The best result was achieved by the model pretrained on all in-domain datasets, followed by pretraining on NewsQA only, followed by pretraining on data that contained all out-of-domain datasets, including datasets generated by synonym and antonym replacement.

Types of Datasets			Scores	
Second Phase	Third Phase	Fourth Phase	EM	F1
in-domain all	none	none	32.46	47.67
in-domain all	NewsQA	none	34.82	50.52
in-domain all	NewsQA	out-of-domain	34	50.66
in-domain all	NewsQA	out-of-domain all with synonyms	34.82	50.05
in-domain all	NewsQA	out-of-domain all with antonyms	34.82	50.64
in-domain all	NewsQA	out-of-domain all with both	**35.6**	**51.32**
NewsQA	none	none	23.56	38.48
NewsQA	out-of-domain	none	19.9	28.49
NewsQA	out-of-domain all with synonyms	none	25.39	41.67
NewsQA	out-of-domain all with antonyms	none	23.56	39.37

Figure 5: Summary of results for multi-phase pretraining.

The results were as expected: multi-phase pretraining with the most number of training phases combined with last-phase pretraining on the largest amount of task-relevant data resulted in the best EM and F1 scores.

We used the highest performing model to evaluate the performance of DistilBERT on question answering on DuoRC, RACE, and RelationExtraction individually. The results are illustrated in Figure 6.

Dataset	EM	F1
DuoRC	34.13	44.11
RACE	22.66	36.22
RelationExtraction	47.66	71.11

Figure 6: EM and F1 scores for each individual out-of-domain datasets evaluated using the best performing model

6 Analysis

As a comparison to EM and F1 scores, we manually reviewed each answer and gave it a score of 1 if the model produced an answer satisfying the question. We found that the best performing model answered correctly in 54% of the time with DuoRC examples, in 47% of the time with RACE examples, and 89% of the time using RelationExtraction examples.

The performance of the best model on examples from RelationExtraction was as expected. This dataset contains paragraphs, questions, and answers that are less ambiguous. Typical answers were

basic knowledge that could have been retrieved from context directly, such as a year, a name, a location, among others.

Even though the DuoRC and RACE datasets were the most similar to the NewsQA dataset used for third-phase pretraining, model's performance on these datasets was poor. Both had paragraphs that were more difficult to read and answer. Issues leading to a poorer performance on DuoRC and RACE examples included the following:

6.0.1 Coreference resolution

When reading the context, the model often failed to group mentions that referred to the same entity in the context. Examples are provided in Figure 7.

Example 1	
Context	Dan is called to the square podium for he is has been chosen by a lottery to be the winner of a new Schwinn Voyager bicycle, much to his father and mother's delight
Question	Who does Jack admit he's proud of?
Correct Answer	Dan
Model Answer	The winner of a new Schwinn Voyager bicycle
Example 2	
Context	Lucy, her name taken from a Beatles song that played in a camp the night of her discovery, is part of the skeleton of what was once a 3-foot-tall ape-man
Question	What was the skeleton named after?
Correct Answer	Song
Model Answer	One of the world's most famous fossils - the 3.2 million-year-old Lucy

Figure 7: Examples from DuoRC and RACE datasets for coreference resolution

6.0.2 Question and context complexity and ambiguity

The model also failed when the context didn't provide sufficient information to answer the question, or was too complex to understand and learn.

Examples are provided in Figure 8.

Example 1	
Context	Televisions were among the most talked about items at the 2013...That glimpse into the future included a look at digital health and fitness devices, which were also big at CES 2013.
Question	At the 2013 CES, which item drew the most attention?
Correct Answer	Televisions
Model Answer	Digital health and fitness devices
Example 2	
Context	A mere hundred species are the basis of our food supply, of which but twenty carry the load.
Question	How many species are most important to our present food supply?
Correct Answer	20
Model Answer	A mere hundred

Figure 8: Examples from DuoRC and RACE datasets for context ambiguity/complexity

6.0.3 Ground truth absence

The model, as expected, failed when there was no ground truth provided by the context. Examples of such errors are provided in Figure 9.

Example 1	
Context	Ryu arrives at Dong-jin's residence in an attempt to kill him. He waits for some time, but Dong-Jin does not arrive: he is, in fact, waiting at Ryu's apartment. After Dong-Jin does not arrive, Ryu returns to his apartment.
Question	Who returns to his home first ?
Correct Answer	Ryu
Model Answer	Dong-Jin
Example 2	
Context	None provided
Question	Who does Jack admit he's proud of?
Correct Answer	Dan
Model Answer	The winner of a new Schwinn Voyager bicycle

Figure 9: Examples from DuoRC and RACE datasets for the absence of ground truth

7 Conclusion

7.1 Main Findings

We found that multi-phase pretraining does indeed improve the performance of DistilBERT on out-of-domain distributions in a question answering task. Therefore, it can be effective to perform continued pretraining of smaller models, if done in at least 4 phases and using task-relevant data in the last pretraining phase. Lastly, we found that using domains relevant to the downstream tasks in intermediate pretraining steps can boost performance as well.

7.2 Limitations

As noted in Results section, we could have tested different techniques for data augmentation to isolate techniques that result in data with the highest quality.However, we have to keep in mind that methods, such as text generation, might not be the most effective as without a clear domain for all datasets, text generation could lead to very poor quality synthetic data. Still, methods, such as negative sampling, can be effective.

Moreover, when identifying an in-domain dataset that is the most similar to the downstream task, we could have used multiple methods for analyzing domain similarity between datasets to make sure the evaluation we get is accurate and more comprehensive.

This project also doesn't perform any hyperparameter fine-tuning and keeps hyperparameters standardized across all experiments. The performance of DistilBERT could have been improved with this step.

7.3 Future Work

Future work could include repeating multi-phase pretraining experiments with data augmented using negative sampling. Moreover, domain-specific pretraining, which was done on NewsQA only, could have been repeated with each of the in-domain datasets.

References

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

[2] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification, 2018.

[3] 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.

[4] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[6] Chuan Li. Openai's gpt-3 language model: A technical overview, Sep 2020.

[7] Shayne Longpre, Yi Lu, Zhucheng Tu, and Chris DuBois. An exploration of data augmentation and sampling techniques for domain-agnostic question answering, 2019.

[8] Rong Zhang, Revanth Gangi Reddy, Md Arafat Sultan, Vittorio Castelli, Anthony Ferritto, Radu Florian, Efsun Sarioglu Kayi, Salim Roukos, Avi Sil, and Todd Ward. Multi-stage pre-training for low-resource domain adaptation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5461–5468, Online, November 2020. Association for Computational Linguistics.

[9] Tf–idf, Mar 2021.

[10] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text, 2016.

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

[12] 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.

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

[14] Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. Duorc: Towards complex language understanding with paraphrased reading comprehension, 2018.

[15] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 35–45, 2017.

[16] Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. Mrqa 2019 shared task: Evaluating generalization in reading comprehension, 2019.

[17] Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. Conditional bert contextual augmentation, 2018.