Building a Robust QA System

Stanford CS224N Default Project

Ximing Wang
Department of Computer Science
Stanford University
ximing@stanford.edu

Abstract

The robustness to domain shifts is very important for NLP, as in real world, test data are rarely IID with training data. This NLP task is to explore a Question Answering system that is robust to unseen domains with few training samples. In this task, three out-of-domain datasets show very different characteristics and they are trained with different in-domain datasets which are more beneficial for their challenges. Multiple transfer learning models are mixed in different ways: mixture of logits, mixture with custom output, and mixture with more features. Three majority vote strategies were taken to ensemble the models. The results achieved 43.739 EM and 60.930 F1 on test leaderboard from two majority vote strategies respectively.

1 Introduction

In machine learning and NLP problems, training data and test data are generally required to have similar distribution to make reliable inference. However, it's usually not the case in practice. The robustness to domain shifts is very important for NLP tasks and other machine learning problems. Humans could easily understand out-of-domain knoweldge, while it's very difficult for machines. This project is to explore a Question Answering system that is robust to unseen domains with few training samples.

2 Related Work

For out-of-domain and fewshot learning, researchers have worked on transfer learning, mixture-of-experts, data augmentation, in-context learning, and more. Transfer learning[1, 2, 3] is shown to work well for domain adaptation from limited data. Mixture-of-experts[4, 5] trains several expert models along with a gating function to control the mixture, and the grating function could be learned from a neural network. Data augmentation could be done using word substitution[6, 7] and back-translation[8, 9], as well as meaning preserving perturbations[10]. In-context learning[11, 12] uses additional context information to learn better on fewshots.

3 Approach

3.1 Baselines

The baseline model provided by the course is given a context and a question as input, the output is to select a span of text from the context that answers the question. In the baseline model, each (question, context) is converted into multiple chunks of size 384 with stride of 128. So the multiple chunks have large overlaps. For long context, many of them may not contain the answer span. The tokenized and preprocessed data are then cached.

The baseline model finetunes DistilBERT[13] on all in-domain training data on the sum of negative log-likelihood loss for the start and end locations, and validated on in-domain val data to save the model with the highest F1 score. This model is then applied to out-of-domain val data.

The provided baseline model is only trained on in-domain data with 3 epochs, so four more baseline models were made, with more epochs, fine-tune on all out-of-domain data, train and validate on both in-domain and out-of-domain data, and train on in-domain and out-of-domain data while validate on out-of-domain data. These models are used in further experiments as starting model.

3.2 Transfer Learning

Inspired by massive multilingual translation[3] that the insights learned from translation of high-resource languages can be transferred to low-resource languages, training on both in-domain and out-of-domain datasets might help out-of-domain performance. Baseline models are mainly trained on in-domain data, they are used as the starting models for further experiments. For out-of-domain datasets, more similar in-domain datasets should help them more. Therefore, data is analyzed in section 4.1 for dataset characteristics. And each out-of-domain dataset has different further co-train in-domain datasets based on the analysis.

Mixture and ensemble of multiple models could make results more stable and help reduce overfitting. These multiple models could be obtained using different batch size, learning rate, random seed, etc., while data splits might take more effect. In the experiments, training data and validation data are used in both directions: train as train, val as val; and val as train, train as val. Besides these, training data and validation data are mixed together and re-split into several random splits with equal number of samples in train and val. For each out-of-domain dataset, baseline model is trained on each of the first half data splits with two steps: first step is to train on out-of-domain dataset with its co-train in-domain datasets, and validate on out-of-domain dataset; second step is to train and validate on out-of-domain dataset only.

It's found that some models achieved best performance at step 0, this means it's only trained for one step, so very slight difference from starting model. To make the starting model not take too much weight in mixture and ensemble, only models that achieved best performance after step 0 were kept for further steps.

3.3 Model Mixture

3.3.1 Mixture of Logits

After obtained models on different data splits, the qualified models were mixed into one model. The mixture takes the start logits and end logits from multiple models, and learned linear layer on them to get the final start and end logits. For the linear layer, negative weights are not wanted as they would flip the values, so instead of regular linear layer, a softmax on parameters is used to guarantee positive weights. The mixture models are trained and validated on the other half data splits.

3.3.2 Mixture with Custom Output

In DistilBertForQuestionAnswering, the start and end positions are obtained independently, while they might be correlated. The hidden states of the final Transformer block are the source to compute the logits for start and end positions. Using hidden states and start/end logits from the models in section 3.2, three different mixtures were tried here. The first is to fix start logits, and compute end logits based on (hidden states, start logits, end logits given by models). The second is to fix end logits, and compute start logits based on (hidden states, start logits given by models, end logits). The third is to compute both start and end logits from (hidden states, start logits given by models, end logits given by models).

The three mixture models were experimented and it turned out that the third one had the best performance for the experiment. So the third mixture model is trained and validated on the other half data splits for each out-of-domain dataset.

3.3.3 Mixture with More Features

From data analysis in section 4.1, context length and question type might be useful features to add into the model. The number of words in context is divided into bins to feed into embedding layer. Question type is categorical, it also contributes another embedding. These embeddings together with the hidden states as well as start/end logits given by models in section 3.2 were used to compute final start and end logits. As the added features are new information, this mixture model is trained on out-of-domain dataset with its co-train in-domain datasets, and validated on out-of-domain dataset.

3.4 Model Ensemble

Previous steps produce multiple models. These models were ensembled using majority vote. So the most popular answer from multiple models was chosen as the final answer. The computation of the evaluation scores took some normalization on the answer text. This was also taken into consideration when picking the most popular answer.

4 Experiments

4.1 Data

There are 3 in-domain datasets: SQuAD, NewsQA, NaturalQuestions; 3 out-of-domain datasets: DuoRC, RACE, RelationExtraction.

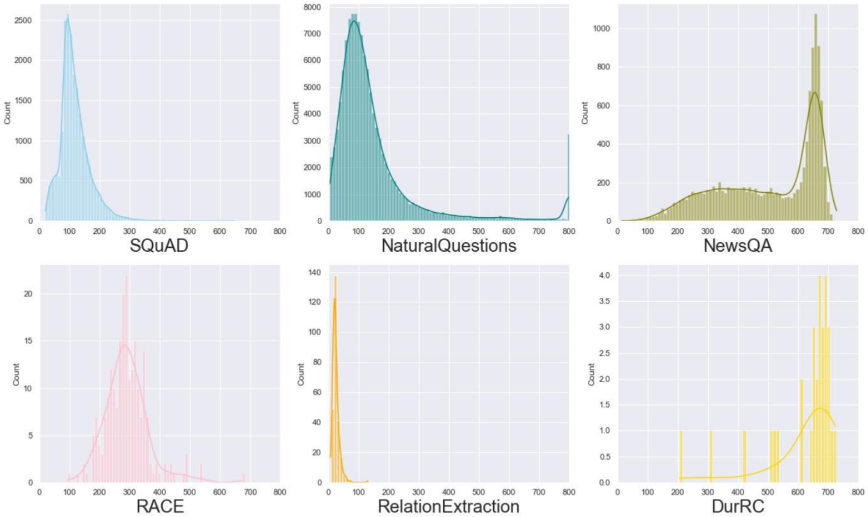

Figure 1: Distribution on number of words per context for each dataset

Data distribution is very different across datasets. Figure 1 shows the distributions on number of words per context for each dataset including both train and val. The first row shows in-domain datasets, the second row shows out-of-domain datasets. DurRC and in-domain NewsQA both have long context with peak around 670 words. RelationExtraction has very short context with peak round 20 words. The questions are generally short in all datasets with peak around 8 words. In pre-processing, the chunk size is 384. So (question, context) in DurRC and NewsQA will be converted to multiple chunks, while in RelationExtraction will be converted to one chunk. This will make RelationExtraction easier to extract answer span since all information are within one chunk and the possible answer span candidates are fewer.

The characteristics of out-of-domain datasets are summarized in Table 1. RelationExtraction has the most advantages as it has both common context source with in-domain datasets and short context.

The biggest challenge for DuoRC is its long context, for RACE is that it has no common data source with in-domain datasets.

Dataset	Question Source	Context Source	Context Length	Challenge
Relation Extraction	Synthetic -	Wikipedia SQuAD, NaturalQuestions	Short -	-
DuoRC	Crowdsourced SQuAD, NewsQA	Movie Reviews -	Long NewsQA	Long Context
RACE	Teachers -	Examinations -	Medium -	Data Source

Table 1: Out-of-domain dataset characteristics

As a result, the three out-of-domain datasets each has different challenges. It would get better result focusing on each dataset than one model for all.

RelationExtraction has similar context source as SQuAD and NaturalQuestions. It also has short context, while NewsQA has very long context. So RelationExtraction is further trained with SQuAD and NaturalQuestions.

DuoRC has very long context like NewsQA. So DuoRC is further trained with NewsQA. Besides data, sequence length might also be a factor. Long context would be converted to multiple chunks, and this would make finding answer more difficult. Current chunk size is 384, i.e. max sequence length is 384. In experiments, maximum sequence length was tried to set to 512 (the limit on BERT), and a new model was trained starting from DistilBertForQuestionAnswering on NewsQA and further on DuoRC. However, the results were not as good as when 384. So the further experiments all used chunk size as 384.

For RACE, the biggest challenge is that it's a completely out-of-domain dataset with no common data source with in-domain datasets. The question type might provide some insights on RACE dataset. Figure 2 shows the frequency of question type for each dataset. RACE has more questions on *What* and *Which*, and SQuAD has more questions with these two types as well. Therefore, SQuAD might help RACE. So RACE is further trained with SQuAD.

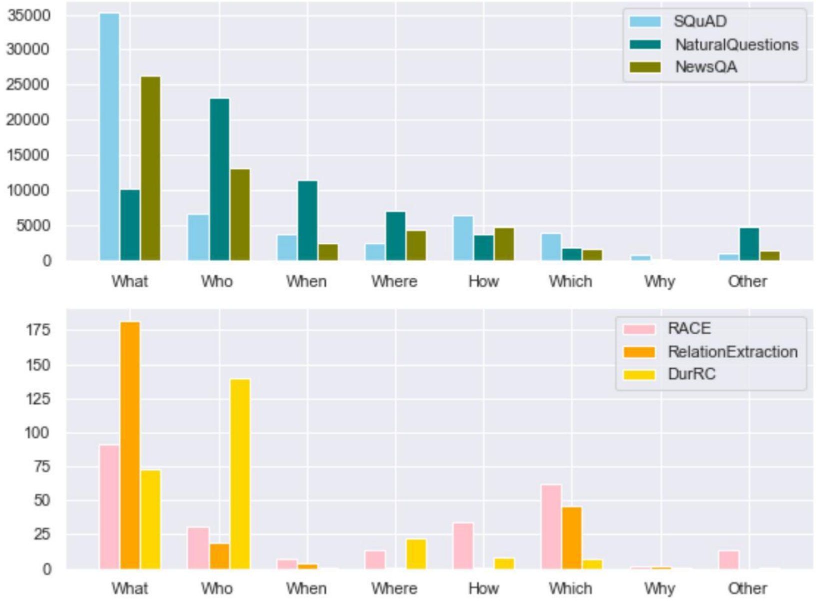

Figure 2: Frequency of question type for each dataset

4.2 Evaluation Method

There are two metrics for performance evaluation: Exact Match (EM) score and F1 score. Each question has one or more human-provided answers. The maximum F1 and EM scores across the provided answers is taken, then the average across the entire evaluation dataset is the final score.

4.3 Experimental Details

4.3.1 Transfer Learning

For the two steps in transfer learning, the first step to train on out-of-domain dataset with its chosen in-domain datasets and validate on out-of-domain dataset took evaluation every 100 steps for 1 epoch and learning rate 1e-5. The second step to train and validate on out-of-domain dataset took evaluation every 10 steps for 6 epochs and learning rate 1e-5. Only models that achieved the best performance after training step 0 were kept. And if the models from two steps both qualify, only the second model is kept. So each data split has at most one transfer learning model, to avoid some data split taking too much weight in mixture and ensemble.

4.3.2 Model Mixture

For mixture of logits, the qualified models were fed into mixture to train and validate on out-of-domain dataset with evaluation every 10 steps for 10 epochs. Different learning rates were tried, and 1e-2 was chosen. This learning rate is relatively large compared to learning rate when training DistillBERT, since the mixture is only using the outputs from DistillBERT models, and mixture training is to train parameters on top of the outputs.

For mixture with custom output, three different custom outputs were tried: fixed start logits and custom end logits, fixed end logits and custom start logits, both custom start and custom end logits. The last one was chosen. Models from transfer learning were fed into mixture with custom start/end logits to train and validate on out-of-domain dataset with evaluation every 10 steps for 20 epochs and learning rate 1e-2.

For mixture with more features, the number of words in context is divided into bins to feed into embedding layer. The bin width is narrower for small numbers, and wider for large numbers. The maximum number of words in context for all datasets is 800. Based on the distribution in Figure 1, the bin width is set to 10 when context words is less than 250, 15 when context words is between 250 and 400, and 20 when context words is between 400 and 800. Question type is categorical, it includes [*What*, *Who*, *When*, *Where*, *How*, *Which*, *Why, Other*]. Many questions in *Other* type are *Whether* questions, but not all of them. Two different mixtures were tried: more features with fixed start and end logits, more features with custom start and end logits. The second one was chosen. Since the added features are new information wanted to learn from in-domain data, this mixture was trained on out-of-domain dataset with its co-train in-domain datasets, and validated on out-of-domain dataset, with evaluation every 100 steps for 1 epoch and learning rate 1e-2.

4.3.3 Model Ensemble

Three majority vote strategies were tried.

Majority Vote 1: All baseline models, transfer learning models, and mixture models are included into majority vote.

Majority Vote 2: Pick the best mixture model on each data split where the mixture model also needs to out-perform all models it mixtured upon. These mixture models and transfer learning models are included into majority vote.

Majority Vote 3: Do majority vote on models from each data split, then majority vote on the results from all data splits.

4.4 Results

The EM and F1 scores obtained on the RobustQA track test leaderboard is in Table 2. In this table, Transfer Learning was using the best transfer learning models from the default train and val split for each out-of-domain dataset. The three majority vote ensembles are described in section 4.3.3.

Ensemble	EM	F1
Transfer Learning	42.018	59.053
Majority Vote 1	**43.739**	60.813
Majority Vote 2	43.440	**60.930**
Majority Vote 3	43.670	60.881

Table 2: RobustQA track test leaderboard scores

The results show that Majority Vote 1 has the best EM score, and Majority Vote 2 has the best F1 score. The ensemble of transfer learning and mixture models on different data splits improved the Transfer Learning result by +1.721 for EM and +1.877 for F1.

5 Analysis

5.1 Dataset

Different out-of-domain datasets have very different scores. RelationExtraction performance is much better than the other two datasets. Table 3 shows the best transfer learning model score with default train and val data split for each out-of-domain dataset. For comparison, the baseline model scores are also contained in the table.

	Transfer		Baseline	
Dataset	F1	EM	F1	EM
RelationExtraction	76.39	58.59	67.48	43.75
DuoRC	46.75	34.92	40.86	31.75
RACE	44.02	27.34	37.15	21.88

Table 3: Best transfer learning model score with default train and val data split

This result aligns with the data analysis in section 4.1. RelationExtraction has common context source with two in-domain datasets, and it also has short context. These advantages make it outstanding in performance.

In experiments, it's also found that when fitting all out-of-domain datasets in one model, the improvement on one dataset might decrease the performance on other datasets, especially between RelationExtraction and DuoRC. The reason might be their context lengths are so different: RelationExtraction is super short, DuoRC is very long. This triggered the decision to focus on each out-of-domain dataset instead of one model for all.

RACE has the worst score in baseline. After trained with all in-domain and out-of-domain datasets, RACE still showed poor performance. With the data analysis in section 4.1, RACE is decided to co-train with SQuAD. Then the performance showed great improvement. So appropriate in-domain dataset is very important for out-of-domain dataset to learn from.

5.2 Mixture

The mixture model performance is compared with the transfer learning models it mixtured upon. All mixture models are obtained on the second half data splits. Each transfer learning model has its own F1 and EM score on each data split val. The mixture model score is compared with the best score from individual transfer learning model on the same data split val. The average improvements across different data splits are in Table 4. For DuoRC, as the mixture didn't show overall improvement, and the experiment of mixture with more features on one data split showed even worse results, mixture with more features was not continued on DuoRC so no numbers in the table.

	RelationExtraction		DuoRC		RACE	
Mixture Model	F1	EM	F1	EM	F1	EM
Mixture of Logits	+1.17	+3.51	-0.85	-1.12	+1.76	+2.53
Mixture with Custom Output	+2.80	+4.30	-2.26	-1.34	+3.06	+4.62
Mixture with More Features	+3..48	+5.28	-	-	+3.57	+3.69

Table 4: Mixture improvements on each out-of-domain dataset

For both RelationExtraction and RACE, mixture models showed performance improvement, while not for DuoRC. The reason might be DuoRC context is too long. So (question, context) in DuoRC is converted into multiple chunks, and it's hard to manage between different chunks with the proposed mixtures. Further work could study on how to do mixture on BERT hidden states and logits from multiple chunks.

5.3 Ensemble

For the three majority vote strategies, Majority Vote 1 and Majority Vote 2 have the best EM and F1 respectively. Majority Vote 3 has both scores in the middle. There are 134 questions showing different answers among the three majority votes: 105 out of 2693 questions in RelationExtraction, 0 out of 1248 questions in DuoRC, 29 out of 419 questions in RACE. Their question type frequencies are shown in Table 5.

Dataset	What	Who	When	Where	How	Which	Other
RelationExtraction	55	8	5	1	0	34	2
RACE	10	5	1	0	5	7	1

Table 5: Number of questions showing different answers among three majority votes

Look at these questions, some have shorter answer correct:

- **Context**: Miklós Ambrus (born 31 May 1933 in Eger) is a Hungarian former water polo player who competed in the 1964 Summer Olympics.
- **Question**: What is Miklós Ambrus's sport?
- **Answer 1**: water polo
- **Answer 2**: water polo player
- **Answer 3**: water polo

Some have longer answer correct:

- **Context**: Willard T. Stevens was a member of the Wisconsin State Senate.
- **Question**: Which position was held by Willard T. Stevens?
- **Answer 1**: Wisconsin State Senate
- **Answer 2**: member of the Wisconsin State Senate
- **Answer 3**: Wisconsin State Senate

Some have all answers incorrect:

- **Context**: Charles McNeill Gray (March 7, 1807 in Sherburne, New York - October 17, 1885; buried in Graceland Cemetery) served as Mayor of Chicago, Illinois (1853–1854) for the Democratic Party.

- **Question**: Where was Charles McNeill Gray burried?
- **Answer 1**: Sherburne
- **Answer 2**: Sherburne, New York
- **Answer 3**: Sherburne

RACE context is longer. This is an example where there might be better answer from context:

- **Context**: British author JK Rowling was at the release of her latest Harry Potter book called "Harry Potter and the Deathly Hallows" at the Natural History Museum in London, Friday July 20, 2007. J.K. Rowling has been spotted at cafes in Scotland working on a detective novel, a British newspaper reported Saturday. The Sunday Times newspaper quoted Ian Rankin, a fellow author and neighbor of Rowling's, as saying the creator of the "Harry Potter" books is turning to crime fiction. "My wife spotted her writing her Edinburgh criminal detective novel," the newspaper quoted Rankin as telling a reporter at an Edinburgh literary festival. "It is great that she has not abandoned writing or Edinburgh cafes," said Rankin, who is known for his own police novels set in the historic Scottish city. Rowling famously wrote initial drafts of the Potter story in the Scottish city's cafes. Back then, she was a struggling single mother who wrote in cafes to save on the heating bill at home. Now she's Britain's richest woman - worth $1 billion, according to Forbes magazine - and her seven Potter books have sold more than 335 million copies worldwide. In an interview with The Associated Press last month, Rowling said she believed she was unlikely to repeat the success of the Potter series, but confirmed she had plans to work on new books. "I'll do exactly what I did with Harry - I'll write what I really want to write," Rowling said.
- **Question**: What is JK Rowling famous for?
- **Answer 1**: British author
- **Answer 2**: Harry Potter book called "Harry Potter and the Deathly Hallows"
- **Answer 3**: British author

Majority Vote 2 tends to have longer answers, it has 80 questions with longer answer than the other two majority votes, while Majority Vote 1 has only 2 and Majority Vote 3 has only 3. Majority Vote 2 ensembled the least number of models, especially less mixture models. So mixture models tend to have shorter answers.

6 Conclusion

This project explored a robust Question Answering system for out-of-domain datasets. With analyzing both in-domain and out-of-domain dataset characteristics, each out-of-domain dataset was chosen different co-train in-domain datasets for further training. Transfer learning models from different train and val data splits were mixed together in different ways: mixture of logits, mixture with custom output, and mixture with more features. The mixture models showed performance improvements on RelationExtraction and RACE. Three majority vote strategies were taken to ensemble all the models, the best EM and F1 score on test leaderboard showed in two majority vote strategies respectively. Among the three out-of-domain datasets, DuoRC didn't benefit as much from mixtures as the other two datasets. The reason might be its long context, and how to deal with long context is something worth working on. New mixtures and ensemble strategies are also promising direction for future work. With exploring more approaches and factors that might affect or improve performance, further work could provide more insights for building a robust system.

References

[1] Yu-An Chung, Hung-Yi Lee, and James Glass. Supervised and unsupervised transfer learning for question answering, 2018.

[2] Timothy J. Hazen, Shehzaad Dhuliawala, and Daniel Boies. Towards domain adaptation from limited data for question answering using deep neural networks, 2019.

[3] Ankur Bapna and Orhan Firat. Exploring massively multilingual, massive neural machine translation (https://ai.googleblog.com/2019/10/exploring-massively-multilingual.html), 2019.

[4] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

[5] Takumi Takahashi, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma. CLER: Cross-task learning with expert representation to generalize reading and understanding. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 183–190, Hong Kong, China, November 2019. Association for Computational Linguistics.

[6] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks, 2019.

[7] Siddhant Garg and Goutham Ramakrishnan. Bae: Bert-based adversarial examples for text classification, 2020.

[8] Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, Valencia, Spain, April 2017. Association for Computational Linguistics.

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

[10] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[11] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners, 2020.

[12] 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.

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