Robust Question Answering on Out-of-Domain Data

Ranchao Yang

[1]Computer Science, Stanford [2]Institute of Computational & Mathematical Engineering, Stanford

Problem

Good performance of question answering models largely depends on the fact that train and test data come from the same distribution. However, real-world applications usually fail to satisfy this assumption. For example, in the field of question answering, train and test data often come from distinct user interactions.

Figure 1. Picture from [3]

Background

To solve this problem, people have experimented with various methods to improve model robustness: for example, **mix-of-expert systems**, **adversarial training**, **task-adaptive fine-tuning**, **data augmentation**, and **meta-learning**.

In this study, we apply some of these methods to a baseline question answering model. Our baseline model simply fine-tunes DistilBERT on in-domain data. Then we experiment with these methods and investigate how they contribute to our baseline model both individually and collectively.

Methods

We focus on the following methods to improve the robustness of baseline model.

- **Domain adversarial training** (DAT) [2] framework includes two parts: a QA model and a domain discriminator. The QA model trains domain-invariant features that can hide domain label from the discriminator, while the discriminator attempts to identify the correct domain label.
- **Task-adaptive fine-tuning** (TAFT) [1] refers to further pre-training on the unlabeled out-of-domain datasets .
- **Data augmentation**: we fine-tune our model on augmented out-of-domain data
  - **Back-translation** refers to first translating into a pivot language and then translating it into the original language
  - **Word substitution, random swap, and random deletion**

Experiments

Task

Given a three-tuple $(Q, P, (p_s, p_e))$ where Q represents the **question**, P represents the **context paragraph**, and p_s and p_e represent start and end positions of the actual answer. Our goal is to predict the answer span $\{p_s, p_{s+1}, ..., p_e\}$ where p_i is a token in the context paragraph that corresponds to the answer.

Data

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	-
`out-of-domain datasets`					
DuoRC	Crowdsourced	Movie reviews	127	126	1248
RACE	Teachers	Examinations	127	128	419
RelationExtraction	Synthetic	Wikipedia	127	128	2693

Table 1. Dataset Statistics.

Evaluation Methods

- Exact match (EM) evaluates whether the model output matches the ground truth answer exactly.
- F1 is the harmonic mean of the precision and recall.

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

Experimentation Results

Method	EM	F1
Baseline	30.63	47.72
+DAT -TAFT -finetune	32.20	47.19
-DAT +TAFT -finetune	30.89	48.24
+DAT +TAFT -finetune	31.68	48.33
-DAT -TAFT +finetune	34.82	50.9
+DAT -TAFT +finetune	35.60	51.75

Table 2. Results on validation datasets.

Analysis

Following is some examples of incorrect predictions.

Question	What work of fiction is Jack Harkness located in?
Context paragraph	Captain Jack Harkness is a fictional character played by John Barrowman in Doctor Who and its spin-off series, Torchwood.
Answer	Torchwood
Prediction	Doctor Who
Question	Due to which disease did Julius Garfinckel die?
Context paragraph	Julius Garfinckel died on his 64th birthday of pneumonia in Washington, D.C. His funeral was held two days later at All Souls Unitarian Church.
Answer	pneumonia
Prediction	Julius Garfinckel died on his 64th birthday of pneumonia
Question	On what instrument is Hungarian Rhapsodies played?
Context paragraph	The Hungarian Rhapsodies, S.244, R.106 (..., Hungarian: Magyar rapszódiák), is a set of 19 piano pieces based on Hungarian folk themes, ...
Answer	piano
Prediction	"Magyar rapszódiák), is a set of 19 piano"

Table 3. Examples of incorrect predictions.

In the first example, the predicted answer and ground-truth answer are both TV programs, and the model fails to identify the actual answer. In the other two examples, the model correctly finds the span that contains the answer but the prediction is much longer and contains much useless information.

Conclusions

We incorporate domain adversarial training, task-adaptive fine-tuning and data augmentation into our model to improve its robustness.

We conclude from our results that the first two methods lead to only limited success without fine-tuning on out-of-domain data. Although the model still has a number of drawbacks, the combination of DAT and fine-tuning demonstrates a significant improvement of model performance on out-of-domain data.

References

[1] 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.

[2] Seanie Lee, Donggyu Kim, and Jangwon Park. Domain-agnostic question-answering with adversarial training, 2019.

[3] Mark Newton. Dawn of the chatbots: What do consumers want and expect?, Mar 2021.