Robust Domain Adaptation by Adversarial Training and Classification

Elliot Epstein
ICME
Stanford University
epsteine@stanford.edu

Nicolas Aagnes
ICME
Stanford University
naagnes@stanford.edu

Abstract

Despite impressive improvements in the performance of NLP models in recent years, commonly used benchmarks often neglect performance on out-of-distribution data. This misalignment can become problematic when trying to use these models in production because inference data has a tendency to change over time and thus be different than the training data. To tackle this problem we extend a method proposed by Lee et al. [1] whose novel idea is to jointly train a question-answering (QA) model together with a discriminative model that forces the encoder to learn domain-agnostic embeddings. While Lee et al.'s model is designed for QA datasets that only contain answerable questions, we include an additional classifier module to determine whether a question is answerable or not. Furthermore, we enrich the training procedure by including various data augmentation strategies such as backtranslation. Our findings show how effective good data augmentation strategies can be, but our discriminative model disappointingly fails to yield any substantial increase in performance.

1 Introduction

Large, pretrained models have become the de facto starting point for most algorithms that achieve state of the art results in many natural language processing tasks. These models are often trained in an unsupervised manner by artificially creating language tasks that force the models to learn feature rich embeddings. This allows for training on large corpora that contains hundreds of millions of lines of text, which, together with massively scaled models containing possibly billions of parameters, has shown to be a driving force behind many cutting edge results and applications in NLP.

However, one critical aspect often overlooked in many common NLP benchmarks, which are typically used as testing playgrounds for developing new models, is a model's capability to generalize well to out-of-domain data distributions at inference time. This issue is particularly prelevant for deploying state of the art models in the real world, as in this case the data distribution often tends to be different from the training distribution. Hence, it is of critical importance that models designed for real world usage are trained by also taking into consideration its robustness to distributional shifts in the data.

Unfortunately, state of the art question answering (QA) models have a tendency to overfit to the training data as opposed to learning more general QA attributes that translate well to other domains without the need of fine-tuning. Methods must therfore be explicitly put in place during the training phase to make the model domain-agnostic.

In the paper *Domain-agnostic Question-Answering with Adversarial Training* [1] by Lee et al., the authors present a novel method to train a domain-agnostic QA model that can handle out-of-domain data distributions. This method does not only address the distributional shift problem, but its generality ensures that it can be applied to almost any deep QA model without making any modifications to the model itself.

2 Related Work

Our approach relies on several key concepts within the field of NLP but also on other machine learning techniques that are commonly used with other data modalities as well.

2.1 Neural Machine Translation

In recent years models such as transformers have achieved amazing state-of-the-art results on many benchmarks in NLP [2]. However, this is often a result of model sizes being increased and much more data being available for training. In a low-data environment however, it is more challenging to use these enormous models as it will quickly lead to the model overfitting. Hence, it is necessary to use data augmentation techniques to artificially create more data and increase the variance in the data distribution.

One common data augmentation technique is backtranslation. Backtranslation consists of translating a sentence back and forth between two languages, with the hope that the translation models will end up slightly changing the words and phrasing of the sentences, but not their inherent meaning. Translating from one language to another can be automatically computed at scale with pretrained neural machine translation models.

Modern neural machine translation models use a probabilistic framework to learn a target sentence that maximises the conditional probability given a source sentence. By leveraging large amounts of annotated data translated between two languages, we can use an off the shelf, pretrained translation model to artificially augment our datasets that are of small size.

2.2 Discriminative Models

Discriminative models became widely popularized through the invention of generative adversarial networks (GANs) [3]. Interestingly, they first had their breakthrough in the field of computer vision, but since then they have had profound impact in many other domains as well.

The basis of a discriminative model is to be able to clearly distinguish input samples from one another. In the case of discriminative models in computer vision for example, their purpose is to detect if an image belongs to a desired distribution (e.g. real images) or if it is outside of that distribution as in the case with fake images for example.

Discriminative networks can also be used to train domain-agnostic models by using a discriminator network to differentiate which domain an input sample belongs to. This approach requires that the dataset contains datapoints from multiple domains, however if it does then such an approach can be used to create a more robust model that can handle out of domain data at inference time since it has been trained to not depend on any domain specific features.

3 Approach

To build a model that has a strong performance on out of domain question answering two main approaches have been utilized, both based on a pretrained DistillBERT model. In the discriminative adversarial training approach, the aim was to create domain invariant features that generalize well to out of domain data. The goal with backtranslation of the out of domain training set with finetuning was to increase the size of the out of domain dataset by creating paraphrases of the context strings for each data sample, making finetuning more effective.

3.1 Discriminative Adversarial Training

We build a domain-agnostic QA model trained by using an auxiliary discriminative model which forces the learned embeddings to become invariant to the input domain. The algorithm comprises of two models: a question QA model and a discriminative model. The QA model serves both as the generative network for the discriminator and as the predictor for the output segments. The discriminator is trained to differentiate between embeddings originating from different domains. This forces the generator, which is the QA model, to project the question passage into an embedding space where the discriminator cannot differentiate between domain-specific embeddings.

The two models are trained in an iterative fashion as commonly done with adversarial networks. Following [1], the loss we use for the QA model is

$$\mathcal{L} = \mathcal{L}_{QA} + \lambda \mathcal{L}_{Adv}, \tag{1}$$

where

$$\mathcal{L}_{QA} = -\frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N_k} [\log P_\theta(\mathbf{y}_{i,s}^{(k)} | \mathbf{x}_i^{(k)}, \mathbf{q}_i^{(k)}) + \log P_\theta(\mathbf{y}_{i,e}^{(k)} | \mathbf{x}_i^{(k)}, \mathbf{q}_i^{(k)})], \tag{2}$$

and

$$\mathcal{L}_{Adv} = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N_k} KL(\mathcal{U}(l) \| P_\phi(l_i^{(k)} | \mathbf{h}_i^{(k)})). \tag{3}$$

In the equations above, $\mathcal{L}_{QA}$ represents the loss term associated with the QA task, and $\mathcal{L}_{Adv}$ is the loss incurred by the discriminator's ability to classify correctly the question-prompt embedding. λ is a hyperparameter that tunes the relative importance of the adversarial loss compared to the QA loss. A higher λ value places more weight on the adversarial loss, thus resulting in more domain-agnostic embeddings, however this incurs the risk of restricting the embeddings too much and hence hurt the model's performance on the QA task. On the other hand, a too low of a λ value may place too little importance on the adversarial loss, and consequently make the model suffer from domain-shift inaccuracies at inference time.

More specifically, P_ϕ is a multi-layer perceptron, P_θ is a pretrained BERT model, K is the number of in domain datasets $\{D_k\}_{k=1}^{K}$, where $D_k = \{\mathbf{c}_i^{(k)}, \mathbf{q}_i^{(k)}, \mathbf{y}_i^{(k)}\}_{i=1}^{N_k}$, $\mathbf{c}_i^{(k)}$ is the i^{th} context in dataset k, $\mathbf{q}_i^{(k)}$ is question i of dataset k, and $\mathbf{y}_i^{(k)} = (\mathbf{y}_{i,s}^{(k)}, \mathbf{y}_{i,e}^{(k)})$ is answer i of dataset k with start index $\mathbf{y}_{i,s}^{(k)}$ and end index $\mathbf{y}_{i,e}^{(k)}$. Furthermore, $\mathbf{h}$ is the [CLS] token representation from BERT, $\mathcal{U}$ is the discrete uniform distribution and $l_i^{(k)}$ is the domain category of sample i in dataset k.

$\mathcal{L}_{QA}$ is the typical QA loss function, and minimizes the negative log likelihood of predicting the correct start and end positions. The adversarial loss $\mathcal{L}_{Adv}$ is designed to maximise the entropy of $P_\phi(l_i^{(k)} | \mathbf{h}_i^{(k)})$, as this would mean that the discriminator is totally confused as to which domain the embedding vector $\mathbf{h}_i^{(k)}$ belongs to. Mathematically, this is equivalent to minimizing the Kullback-Leibler (KL) divergence between a uniform distribution over K classes, denoted as $\mathcal{U}(l)$, and the discriminator's prediction.

The discriminator loss is defined as

$$\mathcal{L}_D = -\frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N_k} \log P_\phi((l_i^{(k)}) | \mathbf{h}_i^{(k)}). \tag{4}$$

which is the standard cross-entropy loss function for a multinomial distribution.

Further, we extend the approach by [1] whose model was specifically only designed for datasets containing only answerable questions. We incorporate a classifier network whose objective is to determine whether a question is answerable or not. To this end, we add an extra module head on top of the [CLS] token to classify whether a question contains an answer inside the prompt or not. Only if the question contains an answer inside the prompt do we backpropagate the loss from the answer span classifier through the network. Our answerable classifier consists of a simple multilayer perceptron model receiving as input the latent embedding of the [CLS] token, and its loss function is the binary cross-entropy loss. Our complete model is shown in Figure 1.

3.2 Backtranslation of out of domain dataset

We have used backtranslation of enlarge the size of the out of domain training dataset ten times. The backtranslation was done with 10 reference languages: Swedish, Norwegian, Danish, Dutch, German, Italian, Russian, Spanish, Portugese, and French. For each reference language each of the out of domain datasets (DuoRC, RelationExtraction, and RACE), were backtranslated. Hence in total, we had 33 datasets available for out of domain fine tuning (3 original+30 backtranslated datasets). The out of domain validation datasets were not backtranslated.

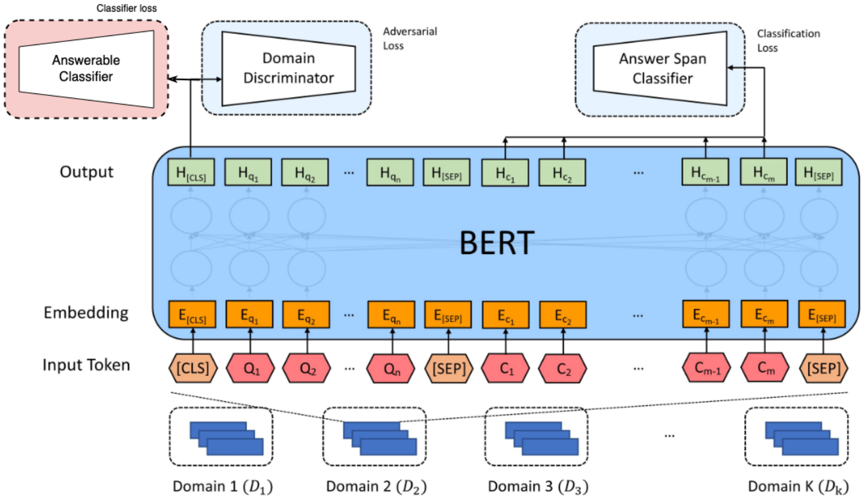
Classifier loss
Answerable Classifier
Domain Discriminator
Adversarial Loss
Answer Span Classifier
Classification Loss
Output
$H_{[CLS]}$ H_{q_1} H_{q_2} ... H_{q_n} $H_{[SEP]}$ H_{c_1} H_{c_2} ... $H_{c_{m-1}}$ H_{c_m} $H_{[SEP]}$
BERT
Embedding
$E_{[CLS]}$ E_{q_1} E_{q_2} ... E_{q_n} $E_{[SEP]}$ E_{c_1}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

Figure 1: Our proposed QA model. Our contribution is an additional classifier network for detecting unanswerable questions. This is shown in the red box in the top left corner of the figure.

Now we describe how a given out of domain dataset was backtranslated using a given reference language. First, the question paragraph pairs were split up so that each paragraph corresponded to exactly one question (originally some paragraphs had several questions and answers associated with them). Subsequently, each context paragraph was split into three parts corresponding to the part before the answer to the question, the answer itself, and the part of the context after the answer to the question. After this, the answer in the context was left untranslated, to ensure that the question answer would still be present in the context. Both the part before the answer and the part after the answer in the context was then translated to the target language and then translated back to English. The translation model used was the googletrans python API. It was decided to not back-translate the question since it has been showed to not be beneficial in [4]. We also did not backtranslate the answer in the context, as we wanted it to exactly match the answer to the question.

4 Experiments

4.1 Data

The data consist of in-domain and of out of domain data. Each data sample is a dictionary consisting of a context and at least one question-answer dictionary. An question-answer dictionary contains a question, a question id, and an answer dictionary. The answer dictionary is either empty if the question is not answerable or contains an answer string, and the start character index of where the answer string can be found in the context. The in-domain datasets consists of the three large QA datasets: SQuAD, NewsQA and Natural Questions containing 50000 training samples each. The three out-domain datasets are RACE, RelationExtraction, and DuoRC, each with only 127 training samples. The out-domain training data is also augmented by using backtranslation as described in section 3.2. A summary of all the data used is provided in Table 4.1.

4.2 Evaluation method

The performance of all the models that have been experimented with have been evaluated with respect to both the EM and F1 scores, on the out of domain validation dataset. When hyperparameters have been optimized it has been with respect to the maximal EM score.

Dataset	Question Source	Passage Source	Train	Dev	Test
in-domain dataset					
SQuAD	Crowdsourced	Wikipedia	50000	10507	-
NewsQA	Crowdsourced	News articles	50000	4212	-
Natural Questions	Search logs	Wikipedia	50000	12836	-
out-domain datasets					
DuoRC	Crowdsourced	Movie reviews	127	126	1248
DuoRC BT × 10	Crowdsourced	Movie Reviews	127×11	126	-
RACE	Teachers	Examinations	127	126	419
RACE BT × 10	Teachers	Examinations/Synthetic	127×11	126	-
RelationExtraction	Synthetic	Wikipedia	127	128	2693
RelationExtraction BT × 10	Synthetic	Wikipedia/Synthetic	127×11	128	-

Table 1: A summary of all the data sources used during experiments. BT is an abbreviation for backtranslation.

4.3 Experimental details

Training the baseline model

The baseline model is a pretrained DistillBERT model, loaded from Huggingface that is subsequently trained on the in domain training dataset (Squad, News QA, and Natural Questions). The hyperparameters are the same as those provided in the GitHub repository https://github.com/michiyasunaga/robustqa. Using the same hyperparameters as used to train the baseline, we also tried to train the baseline for longer, but this did not yield improved results.

Finetuning the baseline model using backtranslated out of domain dataset

The baseline model was finetuned for 10 epochs on the out of domain training set, saving the model with the highest EM score on out-domain validation. The learning rate used was 10^{-5}. This was then extended using the out of domain training set obtained through backtranslating the out of domain datasets (DuoRC, RACE, RelationExtraction) in 10 different languages, the baseline model was finetuned for 10 epochs with a learning rate of 10^{-5}, saving the model with the highest EM score on the out-domain validation set. The validation set was not backtranslated.

Domain Adversarial Training

The domain adversarial network was trained by iteratively doing a gradient step on the QA model and discriminator network respectively. The initial learning rate was the same for both models, but a learning rate scheduler was used to reduce the learning rate of the QA model over time. Subsequently, we finetuned the model on the out-domain dataset for three epochs with a learning rate of 10^{-5}, but this made the performance of the model worse.

4.4 Results

A comparison of the experiments on the validation dataset is shown in 2. Based on these results, the best model was the baseline model finetuned on the out of domain training set with backtranslations in 10 languages. Based on this, we merged the train and validation out of domain sets, and did backtranslation with 10 pivot languages. This was the final model that we submitted to the test leaderboard, and we obtained a F1 score of 58.988 and an EM score of 41.330. Interestingly, we note that the discriminative adversarial trained model performs worse when further finetuned on the out-domain training set. We hypothesize that the reason for this is that the embeddings outputted by the domain adversarially trained model contain less information, as they have been trained to be hard to discriminate.

Model	EM	F1
Baseline	31.41	47.57
Baseline+FT	34.555	49.881
Baseline+Train Longer	31.41	47.57
Baseline+FT+BT× 10	**37.435**	**51.558**
Dom. Adv. Training	34.50	50.48
Dom. Adv. Training +FT	32.41	48.97

Table 2: The best F1 and EM scores on the out-domain validation datasets. FT is an abbreviation for finetuning and BT is an abbreviation for backtranslation.

(en) ... published by **Sony Computer Entertainment** of America, released on the PlayStation 2.

→
... published by **Sony Computer Entertainment** of America, released on PlayStation 2. (sv)
... published by **Sony Computer Entertainment** of America, released on PlayStation 2. (no)
... published by **Sony Computer Entertainment** of America, released on PlayStation 2. (fr)
... published by **Sony Computer Entertainment** of America, released on the PlayStation 2. (nl)
... published by **Sony Computer Entertainment** of America, released on the PlayStation 2. (pt)
... published by **Sony Computer Entertainment** of America, released on the PlayStation 2. (ru)

(en) Two genes located near each other on **chromosome 15** (CKMT1A and...

→
Two genes that are close to each other **chromosome 15** (CKMT1A and... (sv)
Two genes located close to each other on **chromosome 15** (CKMT1A and... (no)
Two genes that are close to each other **chromosome 15** (CKMT1A and... (fr)
Two genes located close to each other on **chromosome 15** (CKMT1A and... (nl)
Two genes located next to each other in **chromosome 15** (CKMT1A and... (pt)
Two genes located next to each other on **chromosome 15** (CKMT1A and... (ru)

Figure 2: Example backtranslated sentences. To the left is a part of an context sentence containing the answer to a question in boldface. On the right hand side are several backtranslated sentences, indicating the pivot language that was used to backtranslate the sentence.

5 Analysis

5.1 Backtranslation

Here, we consider several examples of the backtranslation and evaluate how well backtranslation between English and several other languages work. We expect the backtranslation to work best when the sentence is changed by the backtranslation, but still retaining it's core meaning. Several translations from sentences in the RACE out of domain dataset are shown in Figure 2. From Figure 2 we can note several interesting things. First, many of the sentences are very similar to the original sentences. Thus, it may be beneficial to find pivot languages that gives rephrased sentences whilst preserving the meaning well. We leave it to future work to explore this at greater depth. Second, we note that it's potentially problematic to separately backtranslate the part before and the part after the answer, as the answer may provide important context for the translation.

5.2 Model errors

To better understand the behaviour of the model finetuned on backtranslated sentences, we analyse some of its errors.

Example 1

> Context: The Linth is a Swiss river that rises near the village of Linthal in the mountains of the canton of Glarus, and eventually flows into the Obersee section of Lake Zurich.

Question: What river does Linth turn into?

Correct answer: Lake Zurich

Model answer: Swiss river

This case highlights how the model has learned at least some basic understanding of language and its meaning, as Swiss river could be a plausible answer to such a question. However, it lacks the sharpness to be able to reason at a deeper level to find the correct answer. In some sense the model has outputted what a human might have guessed after taking just a very quick glance at the context.

Example 2

Context: The Yellow River Piano Concerto is a piano concerto arranged by a collaboration between musicians including Yin Chengzong and Chu Wanghua, and based on the Yellow River Cantata by composer Xian Xinghai.

Question: What instrument is Yellow River Piano Concerto scored for?

Correct answer: piano

Model answer: piano concerto arranged by a collaboration between musicians including Yin Chengzong and Chu Wanghua

In this case we see an example of how the model predicts the correct start token but fails on predicting the correct end token. Considering that most answers in the datasets are very short, adding a regularization term on the length of the answer would have probably improved upon the results.

6 Conclusion

Working on improving a pretrained DistillBERT model has been useful for learning more about transformer models. Experimenting with backtranslation and domain adversarial training has been useful for learning about both modern question answering systems at depth as well as giving the authors an increased technical expertise how to combine adversarial training approaches with natural language processing tasks.

The work has showed how backtranslation of a small out of domain dataset combined with finetuning on this augmented dataset may improve a question answering model to answer out of domain questions. This approach gave an F1 score of 58.988 and an EM score of 41.330 on the held out out of domain test set, a sizeable improvement over our considered baseline model, a pretained DistillBERT further trained on an in-domain dataset. We have also highlighted some potential difficulties with discriminative adversarial training for learning domain invariant features on this task. However, it should be noted that this work is limited in scope in that only three out of domain datasets were selected and this may not be representative of the datasets that may be encountered in a practical NLP setting where out of domain model performance is important.

We leave for future work to do a more careful ablation study of which languages are most useful to use as pivot languages for backtranslation as data augmentation. Another interesting avenue for future work is to analyze the dependence on the number of in-domain datasets available for training on the performance of the discriminative adversarial training approach. We hypothesize that a larger number of in domain training datasets will imporve the performance of the discriminative adversarial training.

References

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

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[3] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[4] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *CoRR*, abs/1804.09541, 2018.