# Building a Robust QA System

*Amit Kumar Singh, Mohd Zahaib Mateen*

*Stanford Center for Professional Development*

Stanford
CS 224N – Winter 2022

## Abstract

Pre-trained neural models for QA Systems have shown impressive results while working with in-domain data. However, their robustness to generalize on out-of domain data has been an active area of research.

With the baseline of a pre-trained DistilBERT model, we have work on several techniques to improve the robustness of the QA system. These include:

- Data Augmentation (via Backtranslation)
- Domain Adversarial training
- Hyper-parameter fine-tuning

The motivation is two-fold:

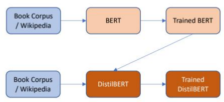
- Make the most of the limited ood data available
- To penalize the model if it tends to overfit on a specific domain

We use Exact Match and F1 scores as our evaluation metrics.

Stanford University

## Introduction

- Typically, QA systems are trained on large, homogenous, use-case specific datasets
- It is then challenging to reuse the model in cases where the data changes drastically or in cases of domain shifts
- Hence the ask is to build a "Robust" QA system that performs well on unseen data.
- The problem is formulated using *triples* of contexts, questions and answers (c,q,a). Given c,q; the goal is to find the character index for a
- The baseline model finetunes DistilBERT [1] which is a smaller, more distilled version of BERT. Given below is the architecture of DistilBERT




## Data

- QA model is trained on the in-domain reading comprehension datasets
- The evaluation is done on the out of domain datasets to test its robustness and generalizability on unseen data
- Each datapoint is a set of context c, question q and answer a, where a is represented as the character index in the context where it starts

| in domain (train size: 50000) | Out of domain (train size: 127) |
|---|---|
| SQuAD | Relation Extraction |
| NewsQA | DuoRC |
| Natural Questions | RACE |

CONTEXT

On a brief trip back to London, earnest, bookish bacteriologist Walter Fane (Edward Norton) is dazzled by Kitty Garstin...
...
Kitty meets Charles Townsend (Liev Schreiber), a married British vice consul, and the two engage in a clandestine affair. When Walter discovers his wife's infidelity, he ...
....
.Kitty rejects his overtures and walks away. When her son asks who Townsend is, she replies "No one important, darling"

QUESTION

What is the name of the man Kitty has an affair with?

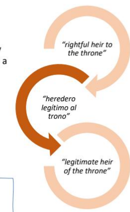
ANSWER

"answer_start": 555, "text": "Charles Townsend"

## Data Augmentation

**Motivation:** To have a larger out-of-domain dataset for training; to have the model learn semantic similarities and discourage it from learning lexicographic similarities

- Context is broken down on a sentence-by-sentence basis
- Sentences that do not contain any answers are back-translated using a "pivot" language
- That specific data-point with its new context is added to the training data.
- Other sentences are skipped




Pivot languages used
(based on best performance)

Español . français

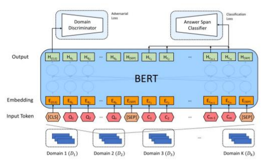
## Domain Adversarial Training

**Motivation:** To penalize the model for "memorizing" domain-specific encodings by training it in an adversarial manner

- Discriminator classifies the joint embeddings for (q,c) in d domains [2]

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

- The idea is to project (q,c) in an embedding space where the discriminator cannot categorize them based on their domains.



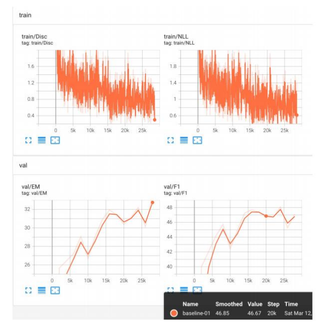

## Other approaches

We also explored the possibility of better performance by changing the learning rate, however we were unable to obtain a significant improvement in F1/EM scores.

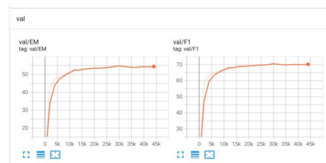
## Experiments

Baseline performance against batch#: in-domain vs out-of-domain


Domain adversarial performance at 3 epochs




Final model performance at 3 epochs




## Analysis / Challenges

**Data Augmentation by back-translation:**

- We started off with the simplest approach of back-translating questions only, but that didn't work as there needs to be lexicographical similarity between q/a.
- We then attempted to translate the entire context but that had situations of multiple answers getting missed. The quality of translation was also poor.
- We finally came up with a way to back translate on the parts of context that are unaffected by the q/a and that worked well.

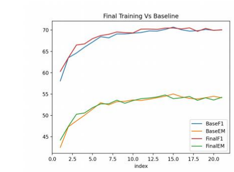
**Domain Adversarial Training:**

- We extended the DistilBERT model to include a domain discriminator
- Based on the ID, we added labels to the data points that were representative of their domains

## Final Results

Results for our model on the test set:

- F1 Score: 57.887
- EM Score: 40.092




## References

[1] Julien Chaumond Victor Sanh, Lysandre Debut and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In arXiv preprint arXiv:1910.01108, 2019

[2] Jangwon Park Seanie Lee, Donggyu Kim. Domain-agnostic question-answering with adversarial training. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 71–79, Vancouver, Canada, August 2017. Association for Computational Linguistics