Building a Robust QA System
Amit Kumar Singh, Mohd Zahaib Mateen
Stanford Center for Professional Development

Abstract
Pre-trained neural models for QA Systems have shown impressive results while working with out-of-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 worked on several techniques to improve the robustness of the QA system. These include:

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

The motivation is two-fold:

• To make the most of the limited data available
• To personalize the model if it tends to overfit on a specific domain

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

Introduction

Data Augmentation

• To make use of hierarchically structured datasets
• To train the model to identify the character index for a sentence
• To train the model to identify the character index for a sentence

Domain Adversarial Training

• To provide the model with "pre-trained" domain-specific knowledge by training it in all domains
• The discriminator classifies the joint embedding with a particular language
• The discriminator learns to project the embedding space where the discriminator can separate the domain

Other approaches

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

Analysis / Challenges

Data Augmentation by back-translation

• We started off with the simplest approach of back-translation question only, but that didn’t work as there needs to be lexical-semantic similarity between the question and answer.
• We then attempted to translate the entire context but that led to situations of multiple answers getting ranked. The quality of translation was also poor.
• We finally came up with a way to back-translate on the parts of context that are not affected by the question and that worked well.

Domain Adversarial Training

• We extended the DistilBERT model to include a domain discriminator
• Based on the L1, we added labels to the data points that were representative of their domains.

Final Results

Results for our model on the test set:
• P@1 Score: 57.887
• EM Score: 48.062

References