Generating Robustness: 6 Ways to Adapt Question Answering to New Domains

Helen Gu, Quentin Hsu, Nicholas Lui

CS 224N Winter 2022, Statistics Department, Stanford University

Stanford Computer Science

Abstract

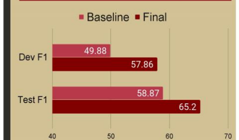
Problem

State-of-the-art QA models tend to overfit to training data and **do not generalize well to new domains**, requiring additional training on domain-specific datasets to adapt. In this project, we aim to **design a QA system that is robust to domain shifts** and can perform well on out-of-domain data.

Approach

We implement **domain adversarial training** to allow the model to learn domain-agnostic features that are robust to domain shifts. We supplement this with **finetuning on augmented data**, **improved domain alignment**, and **adding synthetic QA examples to training**. We also experiment with the **discriminator architecture** and **ensembling methods**.

Final Results



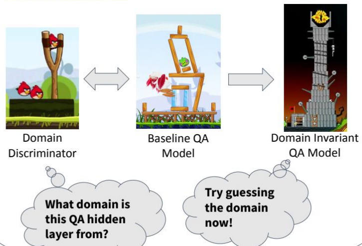
- ❖ **+16%** improvement in **Dev F1**
- ❖ **+10.8%** improvement in **Test F1**

Key Insights

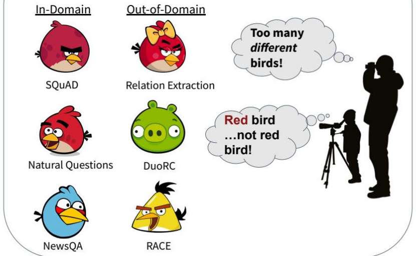
- ❖ **Finetuning** on **augmented out-of-domain data** enhances adversarial model performance
- ❖ **Well-aligned domains** improve results
- ❖ **Training** with T5 **generated synthetic QA** examples yields better generalized OOD performance
- ❖ **Ensembling** varied architectures boosts performance

Lee et al 2019 - Domain-agnostic Question-Answering with Adversarial Training

Compete with Discriminator to learn Domain Invariant Features



Wikipedia vs Non-Wikipedia Domains

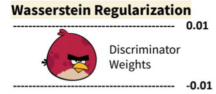


Arjovsky et al 2017 - Wasserstein Generative Adversarial Networks

Lambda Annealing (prep-school for discriminator)

Progressively train the discriminator on **harder and harder examples** by gradually annealing the adversarial penalty

Wasserstein Regularization



Clip discriminator weights to enforce the Lipschitz constraint. This **regularizes adversarial training** and **improves its stability**.

F1: 47.51 — **Domain Adversarial Training**

F1: 53.5 — **Finetuning**

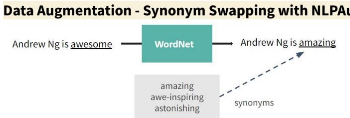
F1: 55.12 — **Domain Alignment**

F1: 55.44 — **Augment Training Data**

F1: 54.66 — **Discriminator Architecture**

F1: 57.86+ — **Ensembling**

Data Augmentation - Synonym Swapping with NLPAug



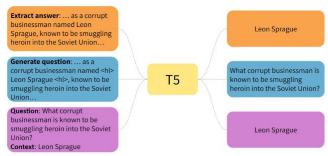
Added support for **synonym swapping** in **context** and **answer spans**

Finetuning on Expanded Out-of-Domain Examples

- ❖ **2x** the number of finetuning examples with data augmentation
- ❖ Too many augmented examples decreases performance

NLPAug - https://github.com/makcedward/nlpaug

Synthetic QA Generation with Roundtrip Consistency



1. Adapt SQuAD Trained T5 to our specific task (chunking, nearest index search)
2. Generate answer spans per sentence
3. Generate questions
4. Validate question answer pairs
5. Include in Training **~1600 synthetic QA pairs**

Alberti et al 2019 - Synthetic QA Corpora Generation with Roundtrip Consistency

Best of Each Domain - Wisdom of the Crowd

Average model logits for start and end indices prior to final prediction.

Models:

1. Best in Relation Extraction
2. Best in DuoRC
3. Best in RACE

Kitchen Sink Approach - Diversify Architectures

1. Wiki Aligned, In-Domain Trained, Aug Finetuned
2. Wiki Aligned, In-Domain Trained, Synth + Aug Finetuned
3. Wiki Aligned, Synth Aug Trained, Aug Finetuned
4. Multi Aligned, Aug Trained, Aug Finetuned
5. Updated Discriminator, Multi Aligned, Synth Aug Trained, Aug Finetuned