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

Abstract

Problem

Domain-adapted 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-invariant 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

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Final</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dev F1</td>
<td>47.51</td>
<td>53.5</td>
</tr>
<tr>
<td>Test F1</td>
<td>51.6</td>
<td>57.8</td>
</tr>
</tbody>
</table>

- +15% improvement in Dev F1
- +15.4% 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 various architectures boosts performance

Data Augmentation - Synonym Swapping with NLAug

- Added support for synonym swapping in context and answer spans

Finetuning on Expanded Out-of-Domain Examples

- In the number of finetuning examples with data augmentation
- Too many augmented examples decreases performance

Wikipedia vs Non-Wikipedia Domains

- In Domain: SQ4D, Relation-Extraction
- Out-of-Domain: Too many different birds!

Synthetic QA Generation with Roundtrip Consistency

- Adapt SQ4D T5 to our specific task (chunking, nearest index warmup)
- Generate answer spans per sentence
- Generate questions
- Validate question-answer pairs
- Include in Training ~1500 synthetic QA pairs

Best of Each Domain - Wisdom of the Crowd

- Average model logits for most and least indices prior to final prediction

Models:
1. Best in Relation Extraction
2. Best in SQ4D
3. Best in T5

Kitchen Sink Approach - Diversify Architectures

1. Multi-Align, In Domain-Train, Aug Pretrained
2. Multi-Align, In Domain-Train, Synthetic Aug Pretrained
3. Multi-Align, Synthetic Aug Train, Aug Pretrained
4. Multi-Align, Aug Train, Aug Finetuned
5. Updated Discriminator, Multi-Align, Synthetic Aug Train, Aug Finetuned

Alberti et al 2018 - Synthetic QA Corpus Generation with Roundtrip Consistency