Semi-Supervised Question Answering on SQuAD 2.0
Niki Agrawal, Mayuka Sarukkai
nikhar@stanford.edu, mayuka@stanford.edu

Problem
Can we create and evaluate question-answering models that perform effectively when trained on mostly unlabelled data?

Motivation
- Q&A systems help us understand and extract info from text, advance Natural Language Understanding
- In low-resource domains, large labeled training corpora do not exist

Existing Approaches
- Q&A systems using labeled training data: BiDAF, ELMo, BERT
- Dhingra, et. al: Pre-train on self-generated cloze (fill-in-the-blank) question-answer pairs

Data

SQuAD 2.0:
- Annotated corpus of Wikipedia articles
- Split into training, test, and validation data
- Randomly sampled 25% and 10% of training data to simulate low-resource domains

Pre-training Dataset:
-Parsed and stripped a random sample of ~5500 raw Wikipedia articles from WikiDumps dataset
-Generated 748 cloze question-answer pairs

Analysis

Rich Embeddings
Character and POS embeddings boost performance with negligible increase in training time for small training sets.

Cloze Pretraining
- pre-trained model's poor performance on very small 10% split suggests need for higher quality cloze generation with wider range of question types to support low-resource settings
- cloze models may better support "What" question learning and bias towards exact string matches rather than deeper semantic relations
- cloze questions lack "No answer" samples, leading to potentially higher error for questions without answers.

Sample prediction: cloze-pre-trained with 10% SQuAD finetuning:

Conclusions & Future Work

1) Complex word embeddings boost performance in low-resource settings. Character-level and part-of-speech embeddings improve performance on our BiDAF model, for both large and small training sets.

2) Pre-training on cloze may boost performance, but requires more testing and refinement of cloze question generation techniques. Our model pre-trained on cloze question-answer pairs before fine-tuning improves performance on 25% of SQuAD training data, but worsens performance on 10% of SQuAD training data. More work required to generate higher quality cloze questions.

References