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

Approach

1) Generate unsupervised cloze (fill-in-the-blank) questions

Generate (Question, Answer, Passage) tuples from unlabeled text articles [2]

- Questions are cloze-style fill-in-the-blank sentences taken from introduction
- Find exact sequence match between question and passage sentences: If match is noun phrase, verb phrase, or Named Entity then Answer = Match
- Remove Answer from Question; Passage = context passage ⟶ (Question, Answer, Passage) tuple

2) Pretrain custom Bi-Directional Attention Flow (BiDAF) model using cloze dataset

BiDAF model [2] contains added character- and part-of-speech embeddings

1) Concatenate word-, character- and POS- embeddings
2) Bi-directional LSTM for contextual embedding
3) Attention layer
4) Bi-directional LSTM Modeling layer
5) Output layer (softmax)

3) Fine-tune model on a small set of supervised QA pairs

Further train BiDAF model on small splits of SQuAD data.

Data

SQuAD 2.0:

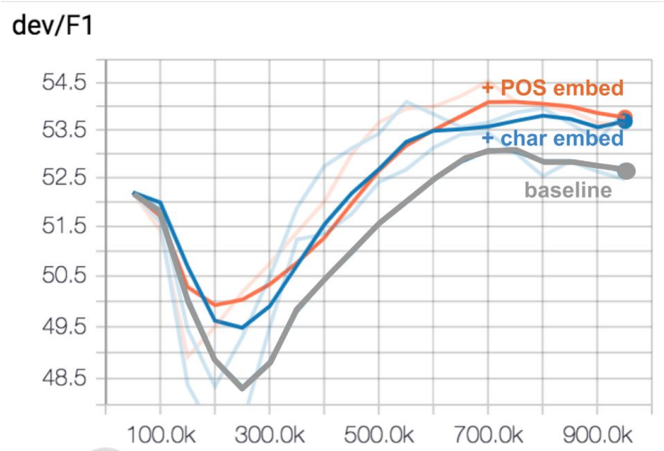
- Annotated corpus of Wikipedia articles
- Split into training, test, and validation data
- Randomly sampled 25% and 10% of training data to simulate low-resource domains

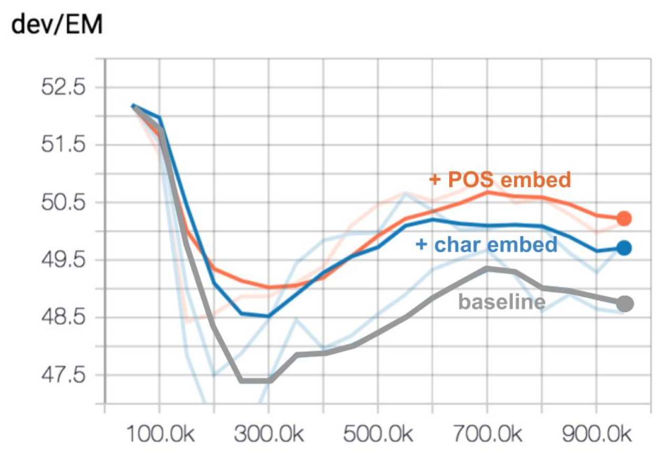
Pre-training Dataset:

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

Results

Model Improvements with Embeddings

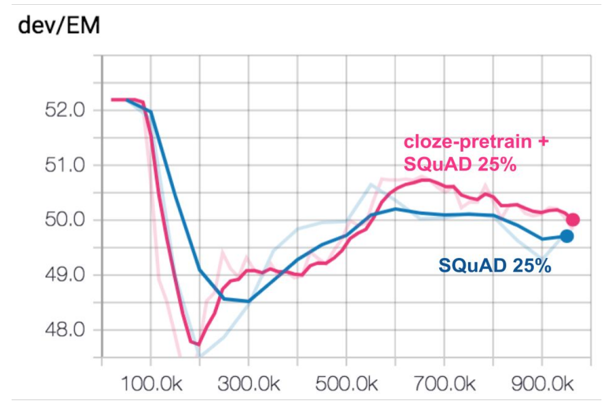


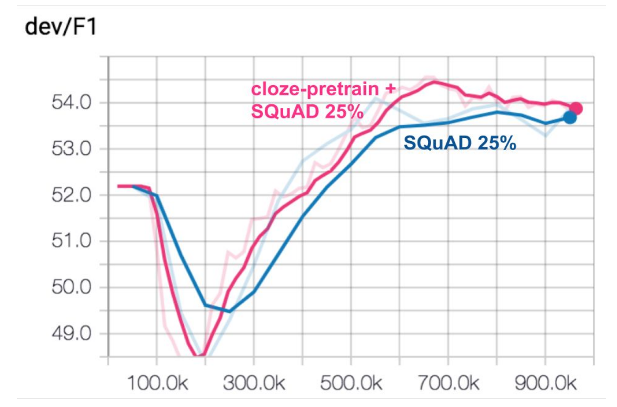


Performance gains on 25% split of SQuAD dataset:

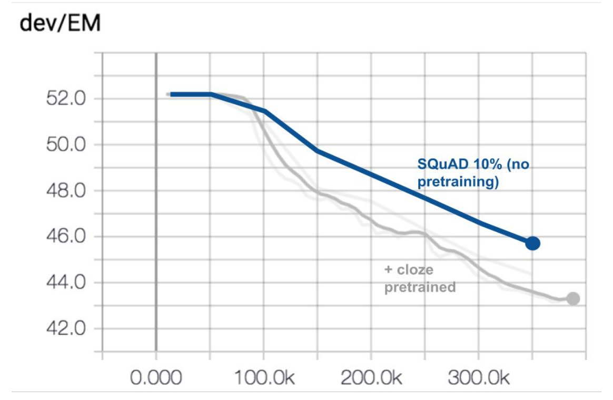
	F1	EM
Baseline (25%)	53.07	49.67
Baseline + Char Embeddings (25%)	53.80	50.20
Baseline + Char + POS Embeddings (25%)	54.09	50.97

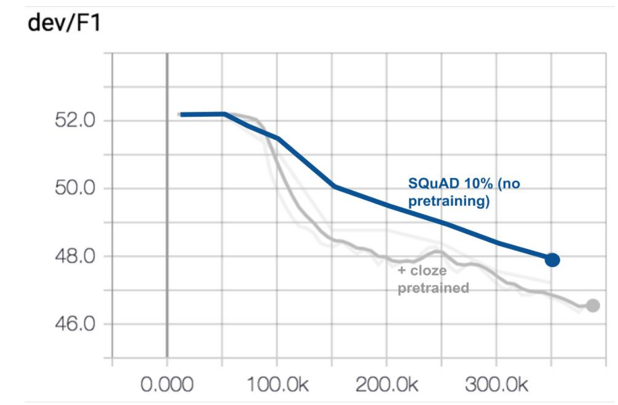
Model with Cloze Pretraining





	F1	EM
SQuAD 25%	54.10	50.65
SQuAD 25% + Cloze pretraining	54.55	50.80





	F1 (final value)	EM (final value)
SQuAD 10%	47.92	45.69
SQuAD 10% + Cloze pretraining	46.87	43.63

Analysis

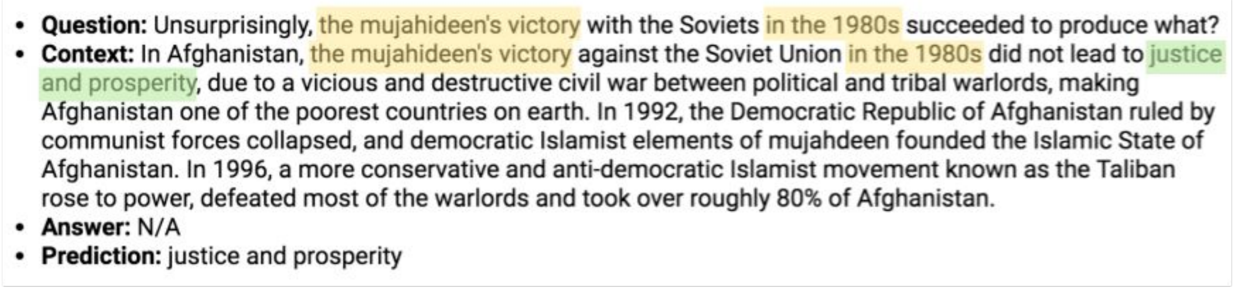
Rich Embeddings

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

Cloze Pretraining

- pretrained 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:



- **Question:** Unsurprisingly, the mujahideen's victory with the Soviets in the 1980s succeeded to produce what?
- **Context:** In Afghanistan, the mujahideen's victory against the Soviet Union in the 1980s did not lead to justice and prosperity, due to a vicious and destructive civil war between political and tribal warlords, making Afghanistan one of the poorest countries on earth. In 1992, the Democratic Republic of Afghanistan ruled by communist forces collapsed, and democratic Islamist elements of mujahdeen founded the Islamic State of Afghanistan. In 1996, a more conservative and anti-democratic Islamist movement known as the Taliban rose to power, defeated most of the warlords and took over roughly 80% of Afghanistan.
- **Answer:** N/A
- **Prediction:** justice and prosperity

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

[1] Bhuwan Dhingra, Danish Pruthi, and Dheeraj Rajagopal. Simple and effective semi-supervised question answering. CoRR, abs/1804.00720, 2018.

[2] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirizi. Bidirectional attention flow for machine comprehension. CoRR, abs/1611.01603, 2016.