Multi-hop Question Answering on HotpotQA

Stanford University

Chris Wang, Yilun Xu, Qingyang Wang

{chrwang, ylxu, iriswang} @ stanford.edu

Introduction

- ***Multi-hop question answering*** on HotpotQA dataset
- SOTA structures achieve ***close to human performance*** on the **single-hop** area, while the generalization into the multi-hop area is impeded by complex query structure and long, noisy context

Question understanding | *Context denoising* | *Multi-hop reasoning*

- We explore both ***BERT*** and ***RNN-based*** structures

Data Set

HotpotQA: 113K valid examples of question-answer pairs

- Question type: bridge entity, comparison, etc.
- Supporting fact: provided as extra supervision

Q: The director of "Big Stone Gap" is based in what New York city?	*Q: Adriana Trigiani is based in what New York city?*
C1: Big Stone Gap is a 2014 American drama romantic comedy film written and directed by Adriana Trigiani… *C2: Adriana Trigiani is … based in Greenwich Village, New York City.*	*C: Adriana Trigiani is … based in Greenwich Village, New York City.*
A: Greenwich Village, New York City	

Example: Bridge entity question vs. single hop

Two sets of dev/test datasets:

- Full wiki: n paragraphs as context for each question
- Distractor: exactly 10 paragraphs as context for each question

Our focus

We use 10% of the training dataset to test ideas and only train our final model on the full dataset

Evaluate models by 3 pairs of metrics:

- Answer EM/F1, Supporting fact EM/F1, Joint EM/F1

Our focus

Model Overview

No evidence shows that BERT performs predominantly well in multi-hop QA area. We try both BERT and RNN-based methods

BERT with classifier

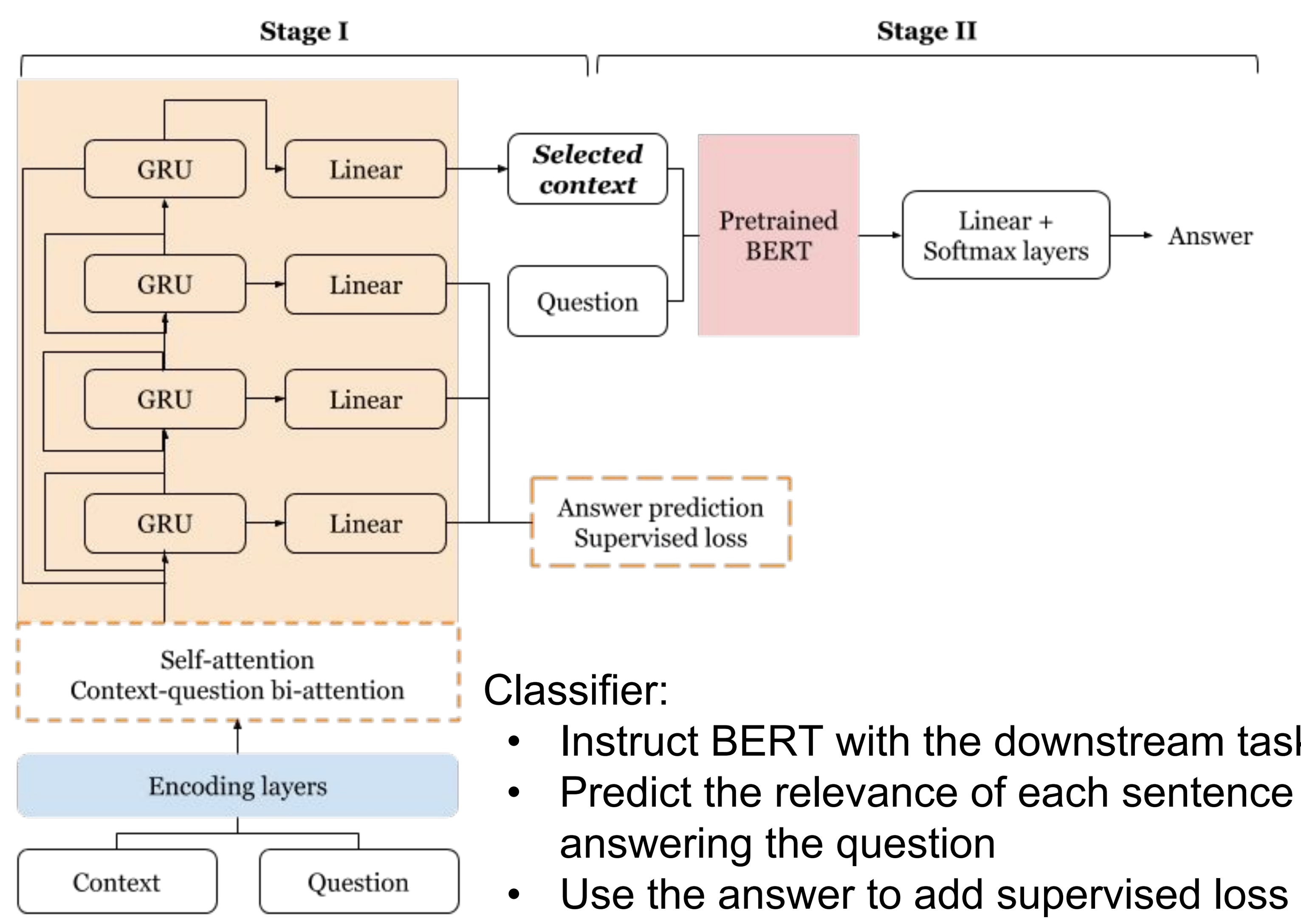
- Pipeline method
- RNN-based classifier
- Pre-trained BERT (small)

RNN-based model

- CoVe
- High-level attention
- Extra self-attention

BERT with Classifier

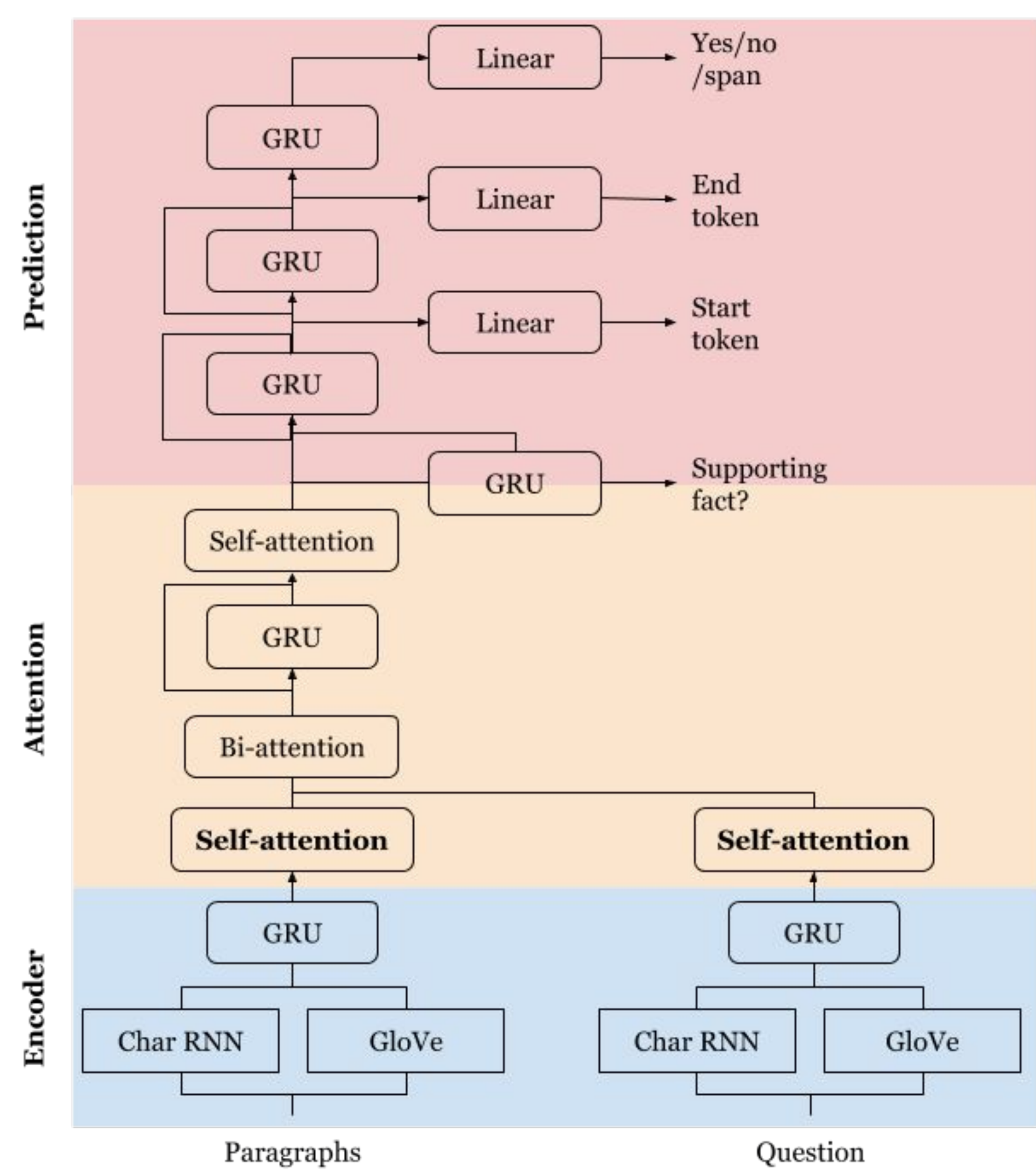
- **Directly using BERT?:** No pretrained model allows long context input; high % of noise distracts BERT; training time grows with context length
- **Pipeline**: select "supporting" sentences and feed them into BERT; fine-tune the classifier's threshold to control level of "noise" for BERT
- Apply linear layers and softmax on BERT to predict yes/ no/ answer span



Classifier:

- Instruct BERT with the downstream task
- Predict the relevance of each sentence in answering the question
- Use the answer to add supervised loss

RNN-based model



Attention

√ Fine-tune optimizer and regularization

x High-level attention: impede back propagation

√ Self-attention: enhance the understanding of question/context

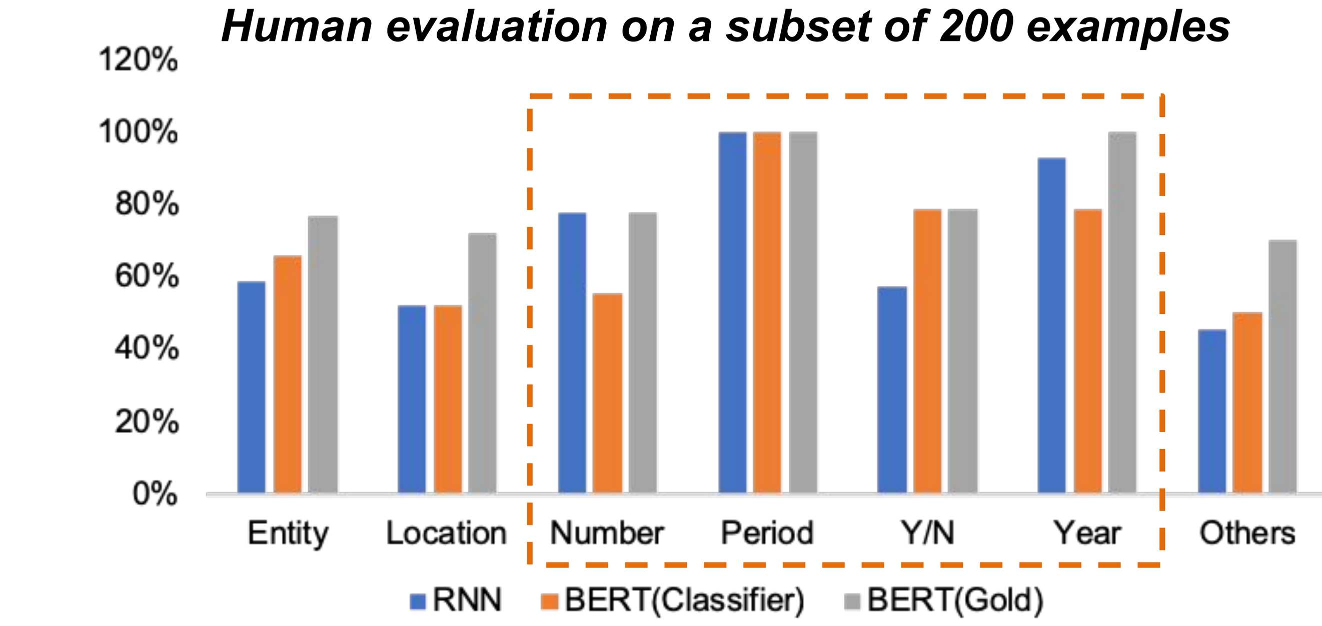
x ELMo: concerns about training time and memory usage

Results, Analysis and Conclusion

Dev set results	Setting	Ans EM	Ans F1	Sup EM	Sup F1	Joint EM	Joint F1
Baseline	Distractor	44.44	58.28	21.95	66.66	11.56	40.86
BERT w/ classifier	Distractor	42.13	53.63	22.58	62.14	12.44	36.64
BERT w/ gold para	*Distractor*	*53.67*	*66.97*				
RNN	Distractor	43.17	56.90	19.82	65.44	10.14	39.58
Baseline	Full-wiki	24.68	34.36	5.28	40.98	2.54	17.73
BERT w/ classifier	Full-wiki	24.11	32.39	5.55	39.91	3.35	17.31
BERT w/ gold para	*Full-wiki*	*29.13*	*37.43*				
RNN	Full-wiki	22.96	31.97	4.27	36.35	2.04	16.23

- Our proposed structures outperform baseline on 10% data, but fails to generalize to the whole dataset

Human evaluation on a subset of 200 examples



- BERT does surprisingly well on Y/N questions and RNN outperforms on Number and Year questions
- BERT: The connection of multiple related sentences is not exactly where BERT struggles; rather, it is having to use a context which has noise
- RNN: Good at handling noise while performance varies on different question categories. Hyper-parameters are fine-tuned on subset, which might be suboptimal for the full training set

Future Work

- **Co-train classifier and BERT**: linking the models together and co-train the two components may improve the result
- **BERT model configuration**: use large BERT w/ more epochs
- Hyperparameter fine-tuning

References and Sources

We are grateful to Peng Qi and Xiaoxue Zang for their patient guidance

Yang, Zhilin, et al. "Hotpotqa: A dataset for diverse, explainable multi-hop question answering." *arXiv preprint arXiv:1809.09600* (2018).

Vaswani, Ashish, et al. "Attention is all you need." *Advances in Neural Information Processing Systems*. 2017.