Wisdom of Ignorance: BERT-wBiDAF for SQuAD v2.0

Siyi Tang[1], Ruge Zhao[2], Meixian Zhu[2]

[1] Department of Electrical Engineering, Stanford University
[2] Department of Statistics, Stanford University

Stanford

Problem

- **Question answering (QA) system**: Automatically answer questions posed by humans in a natural language
- **QA tasks are challenging**: Requires both understanding of natural language and knowledge about the world
- **Goal**: Improve BERT [Devlin et al., 2018] for SQuAD v2.0 [Rajpurkar et al., 2018]
- **Task**: Given a question and a paragraph from Wikipedia, predict whether or not the question is answerable; and if yes, predict the answer text span in the paragraph

Data

- **SQuAD v2.0:** 100k answerable questions & 50k unanswerable questions written adversarially by crowd-workers
- **Adversarial dataset**: Distracting sentences added after normal context [Jia and Liang, 2017]

Baseline

- **First baseline**: BiDAF [Seo et al., 2016]
- **Second baseline**: BERT(base, un-cased) fine-tuned for question answering task
  - BERT pre-trained embedding + one linear layer

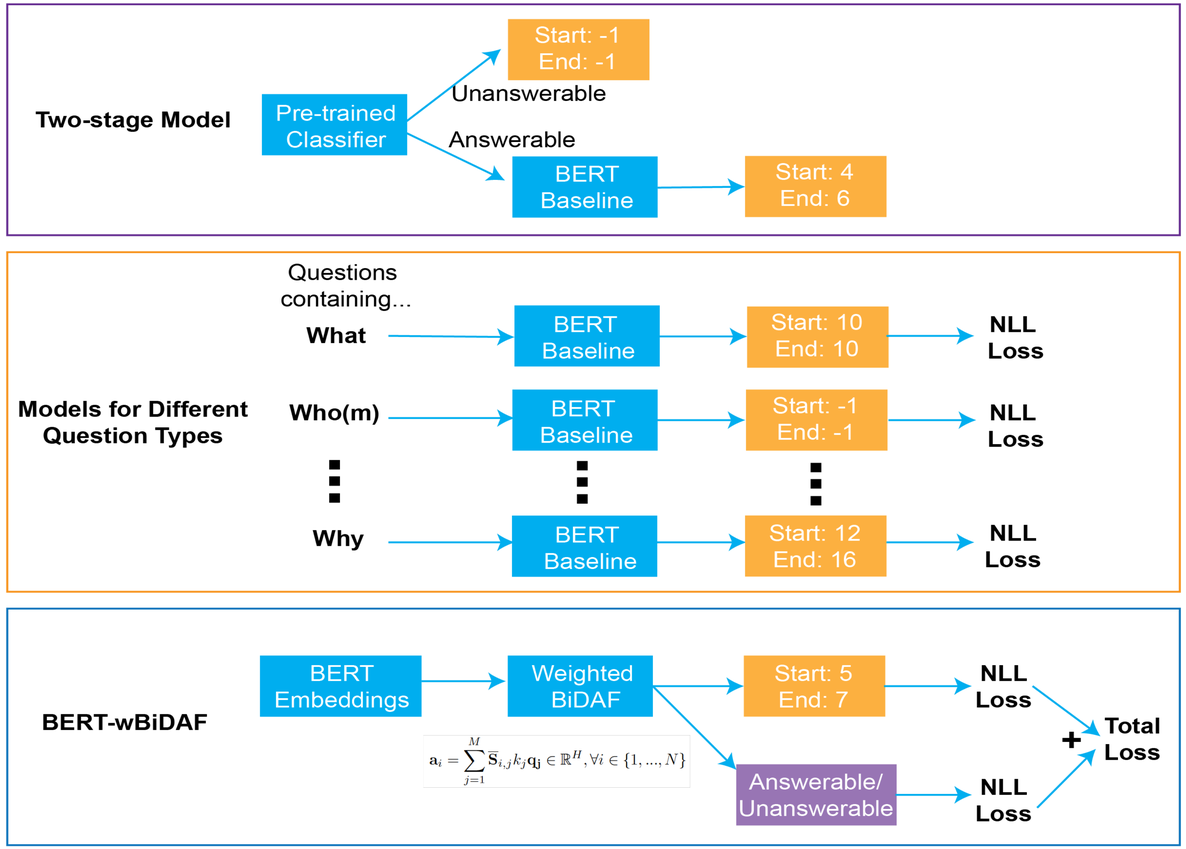

Figure 1. Overview of our approaches

Approach

- **Two-stage model**

- Train classifier (linear layer, CNN etc.) for answerable/unanswerable prediction
- At prediction time, if predicted answerable, use pre-trained BERT baseline to predict start & end positions

- **BERT for different question types**

- Fine-tune one BERT model for each question type separately

- **BERT-wBiDAF**

- BERT pretrained embeddings + BiDAF with *extra weight* given to question keyword in context-to-question attention (Figure 2)

$$\mathbf{a}_i = \sum_{j=1}^{M} \overline{\mathbf{S}}_{i,j} k_j \mathbf{q}_\mathbf{j} \in \mathbb{R}^H, \forall i \in \{1, ..., N\}$$

$$\text{loss}_{\text{QA}} = -\log \mathbf{p}_{\text{start}}(i) - \log \mathbf{p}_{\text{end}}(j)$$
$$\text{loss}_{\text{cls}} = -\log \mathbf{p}_{\text{cls}}(c)$$
$$\text{Total loss} = \text{loss}_{\text{QA}} + \gamma * \text{loss}_{\text{cls}}$$

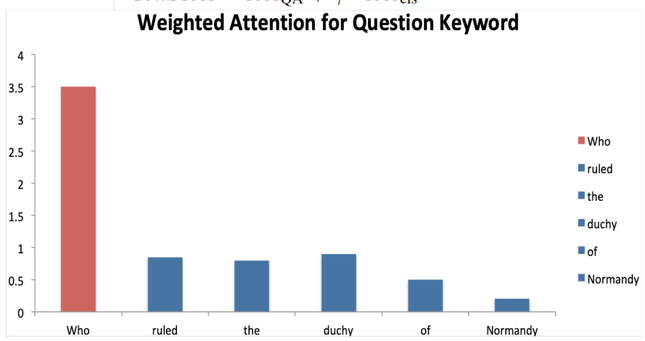

Figure 2. Demonstration of up-weighting C2Q attention for question keyword

Results

	EM	F1	AvNA
BiDAF Baseline	57.89	61.35	68.12
BERT Baseline	72.95	75.79	79.43
Two-stage Model (Classifier + BERT Baseline)	31.55	34.84	74.41
Perfect Classifier + BERT Baseline on Answerable	89.78	93.54	100
BERT + Question Type	65.94	69.13	73.76
BERT-wBiDAF without BCE loss	73.17	76.70	80.45
BERT-wBiDAF	**73.89**	**77.45**	**81.11**

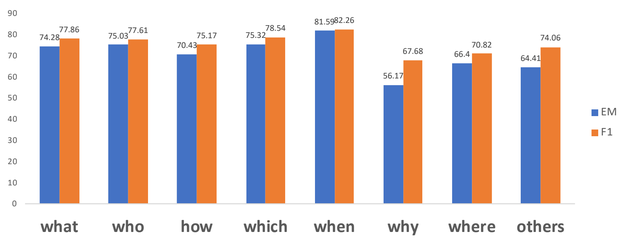

Figure 3. Dev set result breakdown by question type of BERT-wBiDAF

Analysis

- **Adversarial Example**

Context: … was a great fief of medieval France, and under ***Richard I*** of Normandy was forged into a cohesive and formidable principality in feudal tenure. The Normans are noted both for their culture, such as their unique Romanesque architecture and musical traditions, and for their significant military accomplishments and innovations. … ***Jeff Dean ruled the duchy***
Question: Who ruled the duchy of Normandy?
Correct answer: Richard I
Predicted answer: Jeff Dean

- **Train & Dev Loss**

Dev loss started increasing after around 1 epoch (overfit)

Conclusion

- Combining BERT pre-trained embeddings with weighted BiDAF and classification loss improved performance on SQuAD v2.0
- Encouraging model to focus on keywords in questions and taking answerable/unanswerable classification into account are promising directions
- **Future work**: BERT large; different attention weight for different question types

References

1. Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attentionflow for machine comprehension. 2016. URLhttps://arxiv.org/abs/1611.01603.
2. Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems.2017. URLhttps://arxiv.org/abs/1707.07328
3. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018. URL https://arxiv.org/abs/1810.04805.
4. Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. CoRR, abs/1806.03822, 2018. URL http://arxiv.org/abs/1806.03822.