Retro-BiDAF: A Retrospective Reader over BiDAF

Yu Yang

School of Statistics
University of Minnesota

December 16, 2020

Outline

Introduction

Machine reading comprehension (MRC) with unanswerable questions is a central task in NLU. It requires the the machine to predict the correct answer as well as to determine the answerability on the given passage.

Question: Who can be in the Victorian cabinet?
Context Paragraph: The Premier of Victoria is the leader of the political party or coalition with the most seats in the Legislative Assembly. The Premier is the public face of government and, with cabinet, sets the legislative and political agenda. Cabinet consists of representatives elected to either house of parliament. It is responsible for managing areas of government that are not exclusively the Commonwealth's, by the Australian Constitution, such as education, health and law enforcement. The current Premier of Victoria is Daniel Andrews.
Ground Truth Answers: (1) representatives; (2) representatives elected to either house of parliament; (3) representatives elected to either house of parliament

Figure 1: An example of SQuAD2.0 ⟨Question, Context, Answer⟩ triple[1]

- The paragraphs are from Wikipedia and half questions are not answerable.
- If a question is answerable, then the answer is a span of text in the context paragraph.

[1] Check more examples on the challenge website:
`https://rajpurkar.github.io/SQuAD-explorer/explore/v2.0/dev/`

Motivation

Retrospective Reader: proposed by Zhang et al. (2020). It's shown that the retro-reader over ELECTRA backbone improves both the EM and F1 significantly and achieves the state-of-art results. It mimics human reading in the following way:

- scan the text to get a coarse judgement $\rightarrow$ sketchy reading module
- read the text again to determine the final answer $\rightarrow$ intensive reading module + rear verification

Goal: examine the effectiveness of the retrospective reading idea along the non-PCE(Pre-trained Contextual Embeddings) track.

Retro-BiDAF: retrospective reading + BiDAF (Seo et al., 2016) backbone + GloVe word embeddings.

Retro-BiDAF Model Architecture

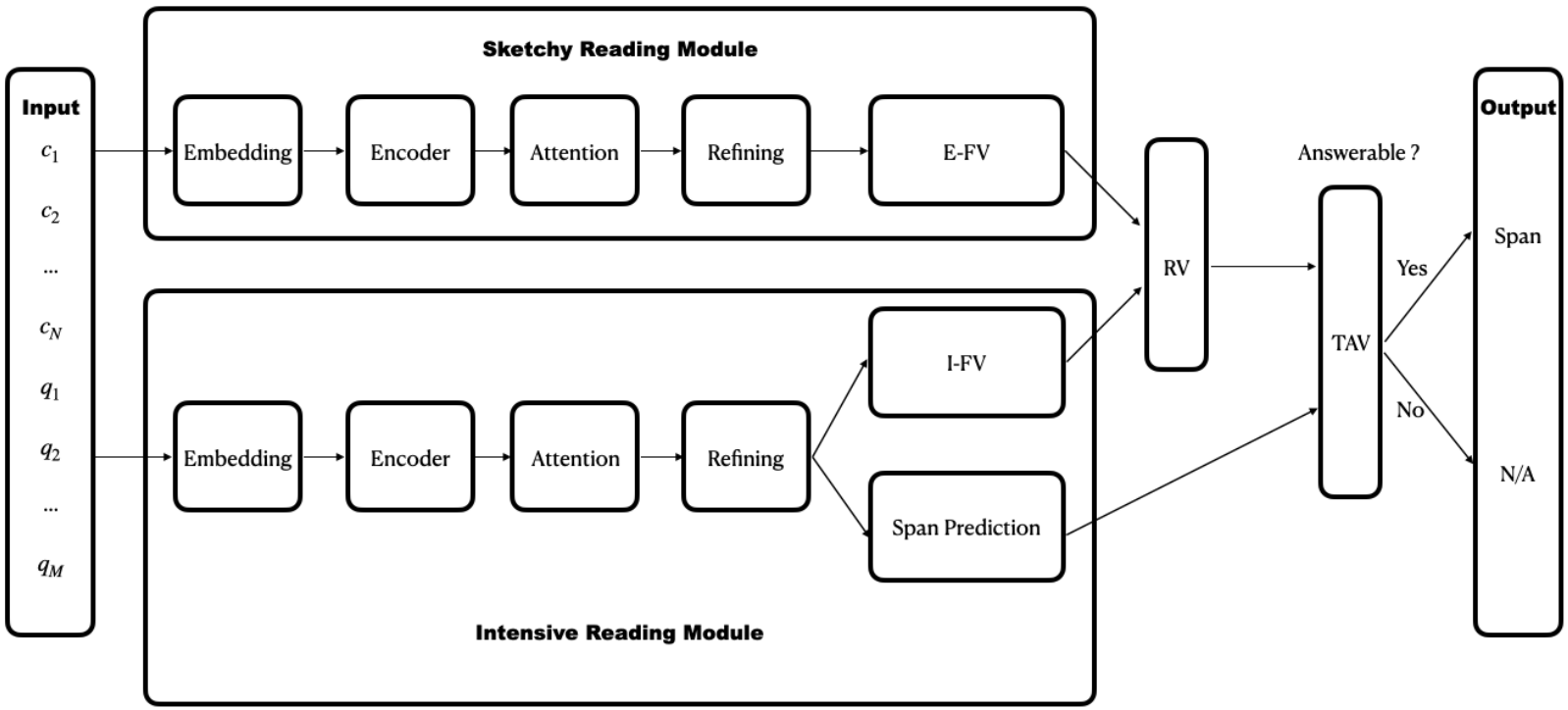

Figure 2: Retro-BiDAF Model Architecture

Sketchy Reading Module

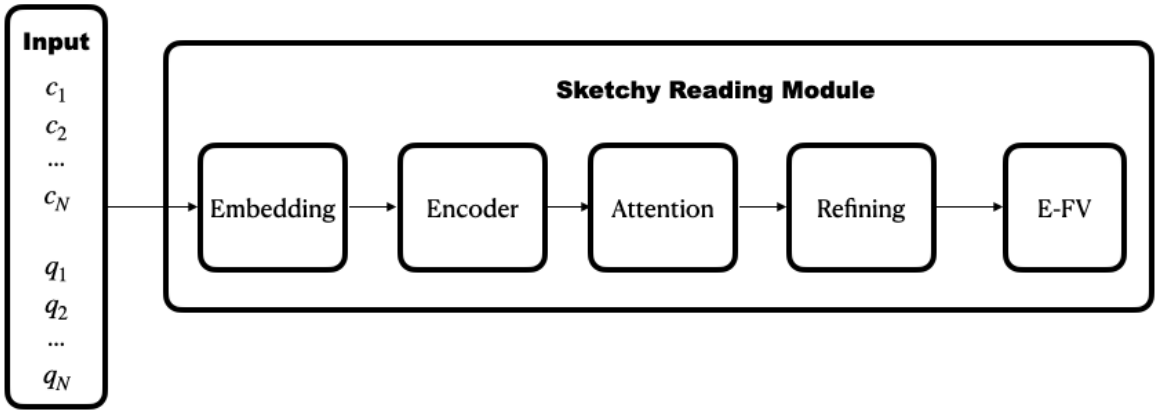

Figure 3: Sketchy Reading Module

1. **Embedding**: GloVe word embeddings + a projection layer + a Highway Network
2. **Encoder**: a one-layer bidirectional LSTM
3. **Attention**: a bidirectional attention flow layer to get better context representations.
4. **Refining**: a two-layer bidirectional LSTM to incorporate the temporal information between context conditioned on the question.
5. **External Front Verification (E-FV)**: a fully connected layer to get classification logits.

Sketchy Reading Module (Con't)

Training Objective: Binary Cross Entropy loss

$$\mathbb{L}^{ans} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)],$$

where $\hat{y}_i \propto SoftMax(Linear(m_N))$ denotes the predicted probability by E-FV, y_i is the binary target indicating the answerability, and n is the number of samples in the training dataset.

Intensive Reading Module

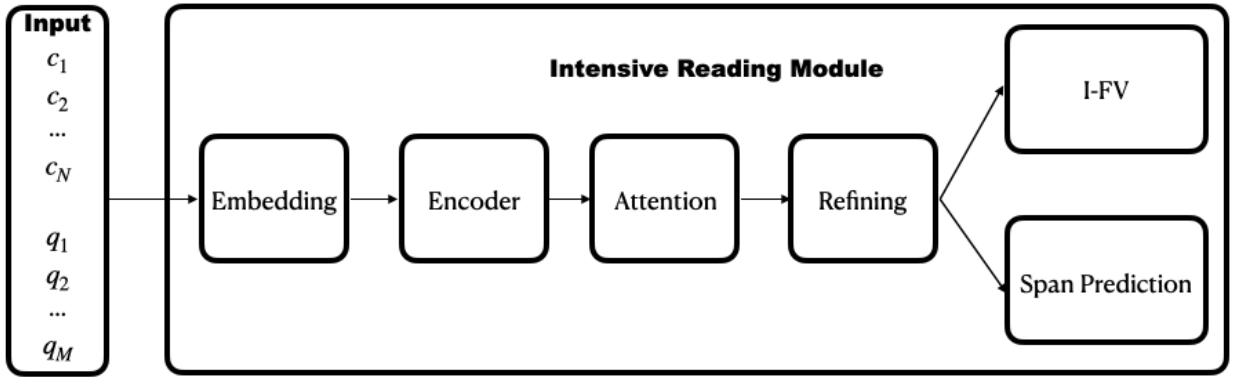

Figure 4: Intensive Reading Module

Internal Front Verification (I-FV): a fully connected layer to get classification logits.

$$\mathbb{L}^{ans} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log \overline{y}_i + (1 - y_i)\log(1 - \overline{y}_i)],$$

where $\overline{y}_i \propto SoftMax(Linear(m_N))$ denotes the predicted probability by I-FV and y_i is the binary target indicating the answerability.

Intensive Reading Module (Con't)

Span Prediction: a one-layer bidirectional LSTM + two fully connected layers + softmax to get the soft prediction probabilities $s, e \in \mathbb{R}^N$.

Training objective: the sum of the cross entropy loss for the start and end predictions,

$$\mathbb{L}^{span} = -\frac{1}{n}\sum_{i=1}^{n}[\log(s_{y_i^s}) + \log(e_{y_i^e})],$$

where s_k is the probability that the answer starts at position k in the context, and e_l is the probability that the answer ends at position l in the context, and y_i^s, y_i^e are the target start and end positions of example i.

Joint Loss: span prediction and I-FV are trained jointly.

$$\mathbb{L} = \alpha_1 \mathbb{L}^{span} + \alpha_2 \mathbb{L}^{ans},$$

where are α_1 and α_2 are weights.

Inference

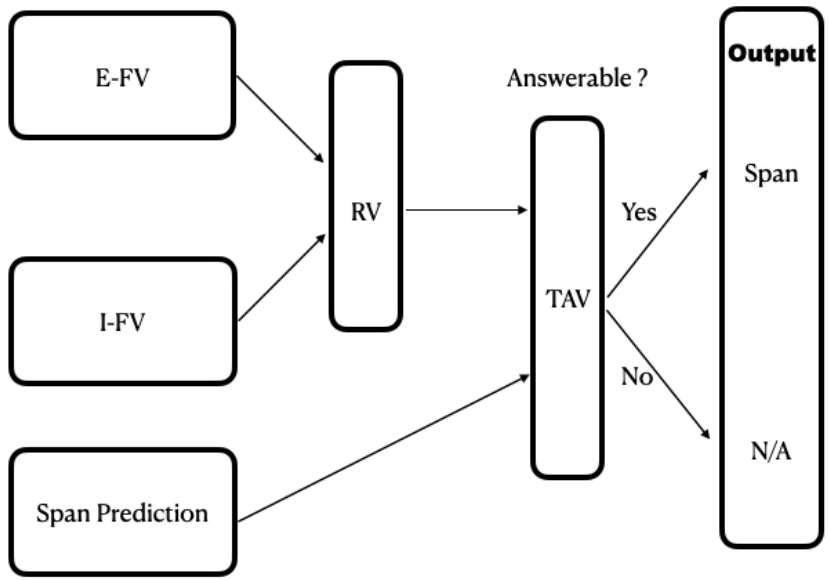

Figure 5: RV + TAV

Rear Verification (RV): aggregate the predicted probabilities given by E-FV ($\hat{y}$)and I-FV ($\overline{y}$).

$$v = \beta_1 \hat{y} + \beta_2 \overline{y},$$

where β_1 and β_2 are weights.

Inference (Con't)

Threshold-based Answerable Verification (TAV): with a pre-specified threshold δ, predicts the answer span if $score_{diff} > \delta$ and N/A otherwise.

$$
\begin{aligned}
score_{has} &= \max(s_k \cdot e_l), 1 \leq k \leq l \leq N, \\
score_{na} &= \lambda_1(s_0 \cdot e_0) + \lambda_2 v^2, \\
score_{diff} &= score_{has} - score_{na}.
\end{aligned}
$$

where λ_1 and λ_2 are weights.

Discretized Predictions: Choose the pair (k, l) of indices that maximizes $s_k \cdot e_l$ subject to $k \leq l$ and $l - k + 1 \leq L_{\max}$, where $L_{\max}$ is a hyperparameter.

Experiment Setup

- BiDAF baseline model: Adadelta optimizer with learning rate 0.5, no L2 penalty. The batch size per GPU is 64 and the number of epochs is 30 (plateau at epoch 22).
- Sketchy reader: Adam optimizer with a warm-up learning rate 0.02. The number of epochs is 10(plateau at epoch 5).
- Intensive reader: similar to the BiDAF baseline model except the number of epochs is 50 (plateau at epoch 47).
- Weights: $\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = \lambda_1 = \lambda_2 = 0.5$.
- The threshold $\delta = -0.006$ is tuned w.r.t. F1 using the dev set.

Experiment Results

Evaluation Metrics

- Exact Match (EM): a binary measure of whether the output matches the ground truth answer exactly.
- F1: the harmonic mean of precision and recall.
- Answer vs. No Answer (AvNA): classification accuracy when we only consider answerability prediction.

Evaluation Results

Model	*Dev*		
	EM	*F1*	*AvNA*
BiDAF baseline	58.28	55.13	**64.70**
Retro-BiDAF	**61.15**	**59.45**	63.94

Table 1: The results (%) from single models for SQuAD2.0 challenge.

Prediction Examples I

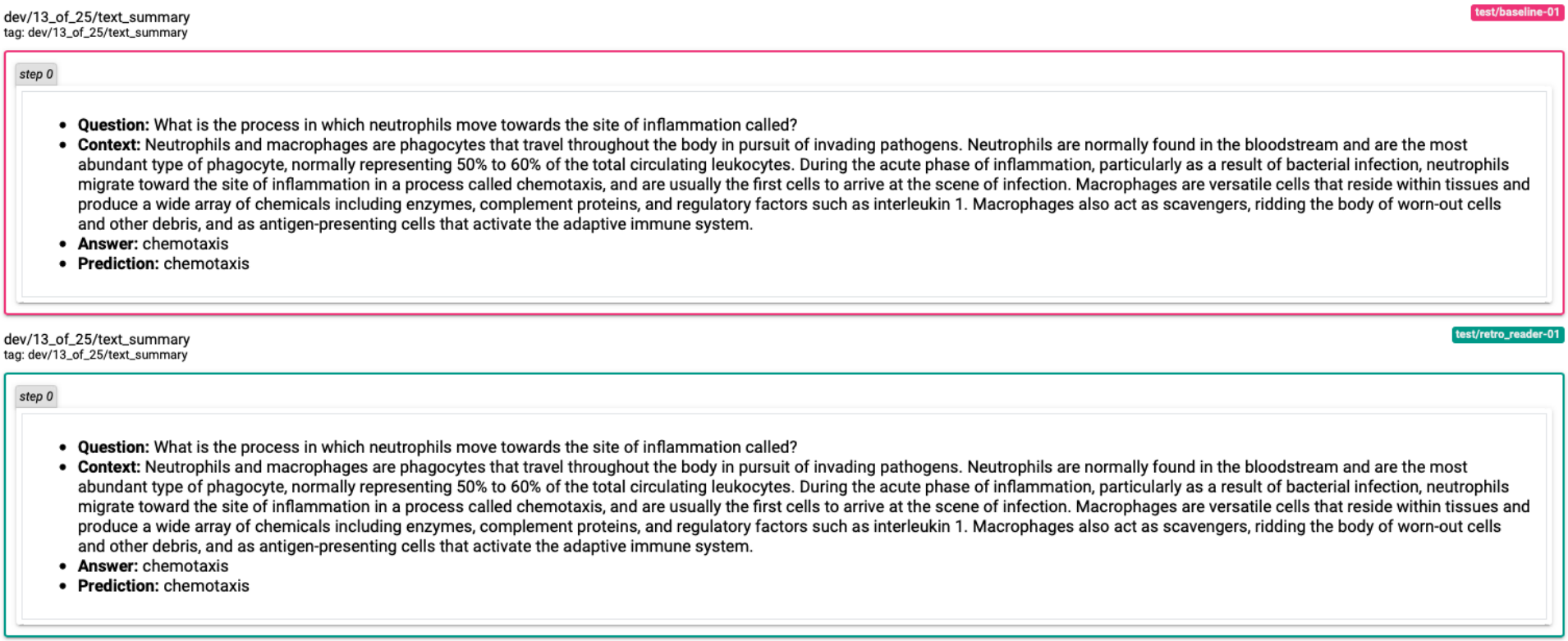

Figure 6: Both models succeed on answerable questions

Prediction Examples II

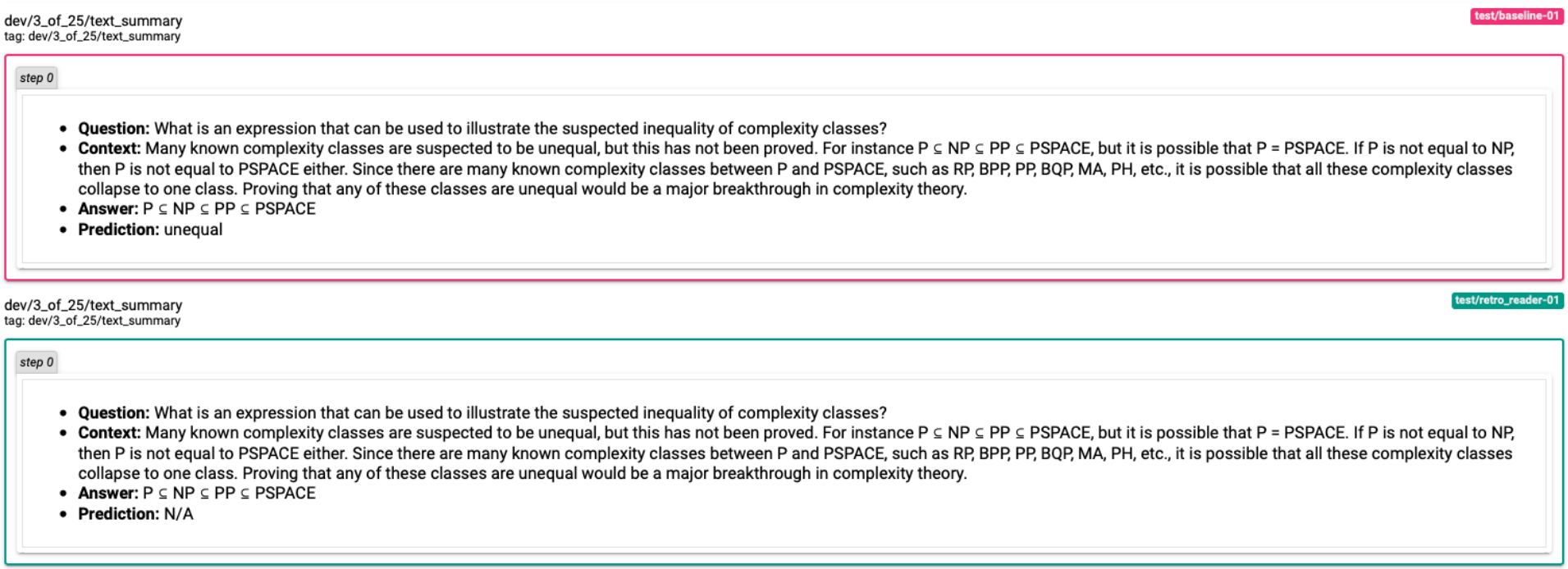

Figure 7: Both Models fail on answerable questions

Prediction Examples III

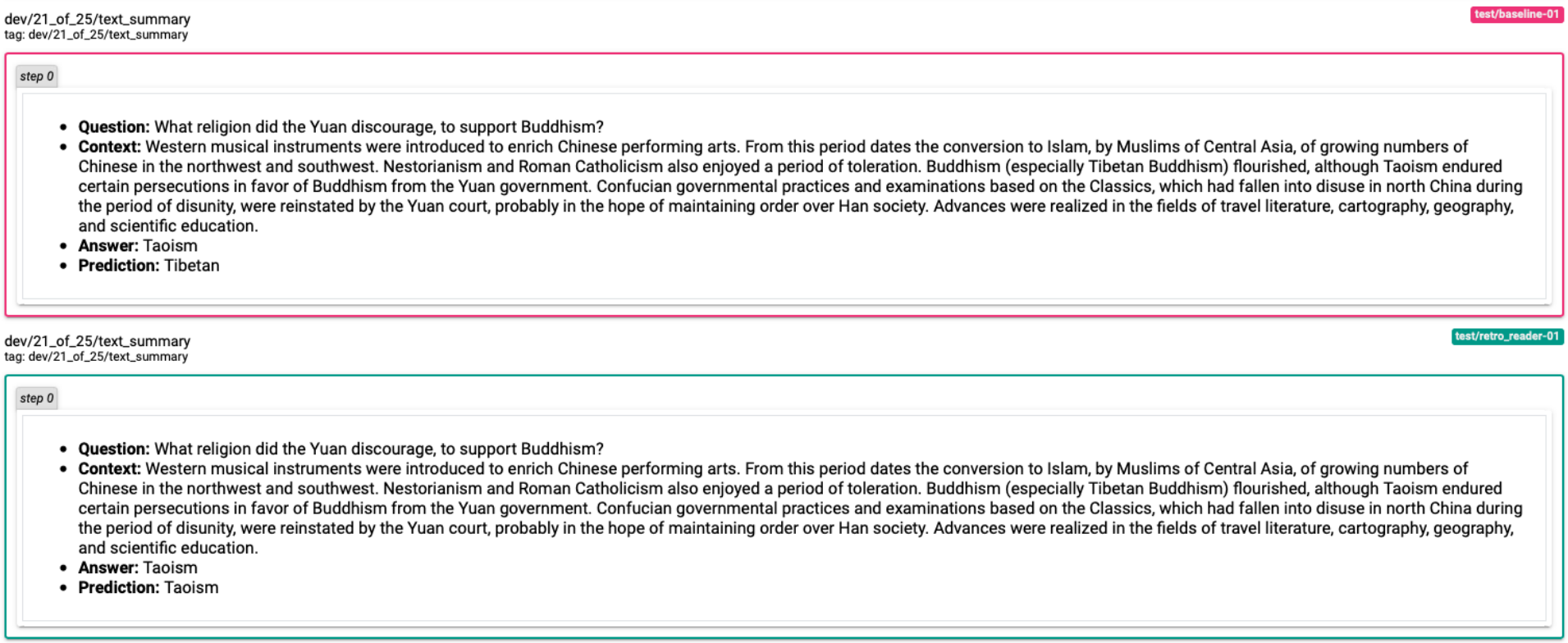

Figure 8: Retro-Reader outperforms baseline

Prediction Examples IV

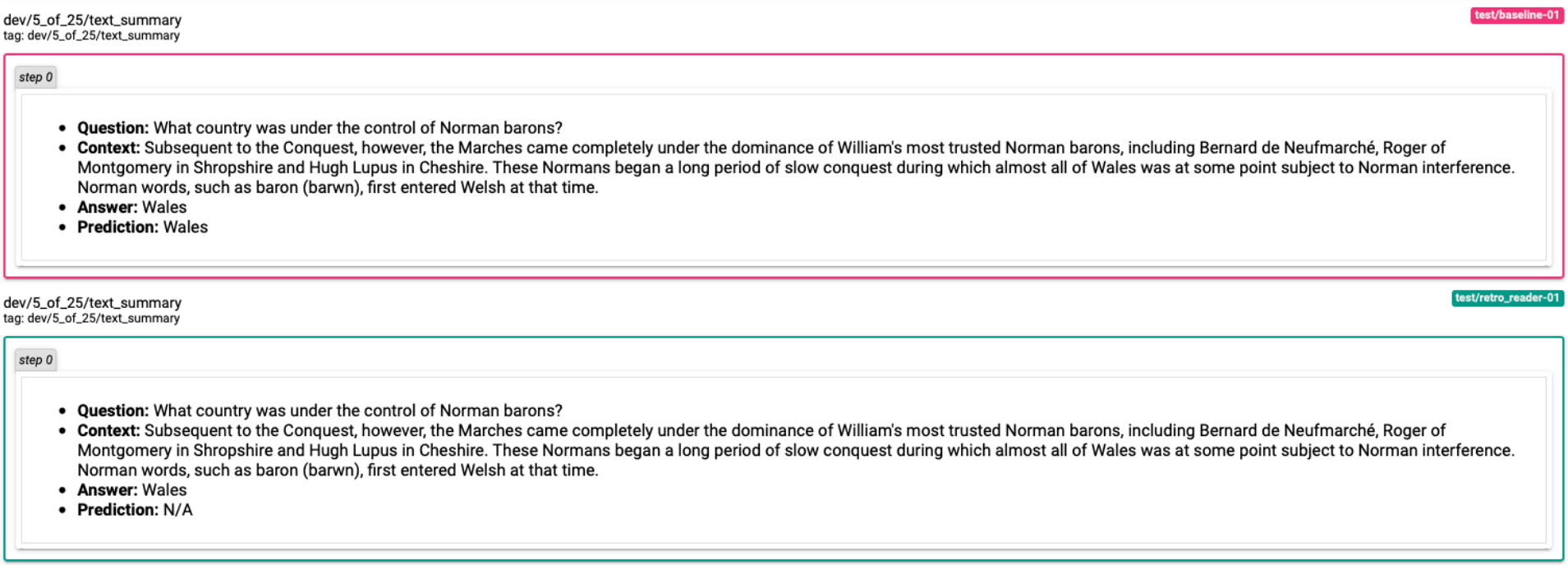

Figure 9: Baseline outperforms retro-reader

Conclusion

- Retro-BiDAF is proposed to check the effectiveness of retrospective reading along the non-PCE track.
- The idea of retrospective reading indeed helps improve the model performance with respect to EM and F1.
- More effort is needed to investigate the downgrade of AvNA.
- Code and report are accessible at: `https://github.umn.edu/YANG6367/squad`.

References I

- Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators.arXiv preprint arXiv:2003.10555, 2020.
- Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. Attention-over-attention neural networks for reading comprehension. InProceedings of the 55th Annual Meeting of the Association for ComputationalLinguistics (Volume 1: Long Papers), pages 593–602, 2017.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding.arXiv preprint arXiv:1810.04805, 2018.
- Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. Gated-attention readers fortext comprehension. InProceedings of the 55th Annual Meeting of the Association for Computational Linguistics(Volume 1: Long Papers), pages 1832–1846, 2017.

References II

- Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and PhilBlunsom. Teaching machines to read and comprehend.Advances in neural information processing systems, 28:1693–1701, 2015.
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory.Neural computation, 9(8):1735–1780, 1997.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A litebert for self-supervised learning of language representations.arXiv preprint arXiv:1909.11942, 2019.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation.InProceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages1532–1543, 2014.

References III

- Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad.arXivpreprint arXiv:1806.03822, 2018.
- Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension.arXiv preprint arXiv:1611.01603, 2016.
- Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks.arXiv preprint arXiv:1505.00387,2015.CS224n Course Staff.Cs 224n default final project:Question answering on squad 2.0.Last up-dated on February, 5, 2020.URLhttps://web.stanford.edu/class/cs224n/project/default-final-project-handout.pdf.
- Zhuosheng Zhang, Junjie Yang, and Hai Zhao. Retrospective reader for machine reading comprehension.arXiv preprintarXiv:2001.09694, 2020.