Motivation:
- Question answering (QA) is an important end-user task for NLP and IR.
- QA systems propose complex architectures; do the aforementioned architecture justify their empirical results?

Goal:
- Create a relatively simple QA system to answers questions correctly from the SQuAD 2.0 dataset.
- Extend a BiDAF baseline by employing an additional input feature (word-in-question) and self-attention.

Data
- Question: What president eliminated the Christian position in the curriculum?
  - Context: Charles W. Eliot, president 1869–1893, eliminated the feared position of Christianity from the curriculum while opening it to student self-direction. While Eliot was the most crucial figure in the secularization of American higher education, he was motivated not by a desire to secularize education, but by Transcendentalist Unitarian convictions. Derived from William Ellery Channing and Ralph Waldo Emerson; these convictions were focused on the dignity and worth of human nature, the right and ability of each person to perceive truth, and the indwelling God in each person.
  - Answer: Charles W. Eliot
- SQuAD 2.0: composed of question, context, answer triples
  - Contains answerable and unanswerable questions
  - Train: 129,941 example, Dev: 6078 examples, 5915 examples

Approach
- Heuristic: Word-In-Question (wiq) Feature
  \[ wiq = I(\exists k: x_k = q) \]
- Self-Attention: Multi-Headed Attention
  - Scaled Dot-Product Attention

Results + Analysis

Analysis Cont’d
- The sizeable improvements in ‘why’ and ‘other’ question types performance could be because of lack of training and dev data.
  - High variance, possibly misleading estimation of model’s performance.
- Most models overall showed improvement over baseline on dev set.

Conclusion
- Character-level embeddings slightly improved on the performance of the baseline at the cost of a minimal increase in training time.
- Self-attention performed minimally worse than baseline.
  - Implementation potentially unsuited for QA systems.
- Holistically, baseline + char embeddings is the best model.
  - Wiq feature has some potential merit given dev set performance.
  - Can explore further into, additional wiq feature types?

References