**Introduction**

The goal of an end-to-end QA system is to extract information from a given source (a context) such as a passage, document, image, etc. based on the user’s request (a query). The effectiveness of such a system lies in its ability to provide concise and accurate answers.

Q: Who leads the United States?

C: Barack Obama is the president of the USA.

**Experiments**

**Data**

We used the SQuAD 2.0 dataset with custom dev and test sets. (The official test set is unknown and reserved for final evaluation.)

- train: 113,941 examples
- dev: 5,058 examples
- test: 16,766 examples

The dataset contains records of (context, question, answer) triples of both answerable and unanswerable questions. The training set has one answer per question whereas the dev set has two answers for every question. In addition, 286-dimensional Glove word embeddings and 64-dimensional character embeddings are provided.

**Evaluation Method**

We used the SQuAD official Exact Match (EM) and F1 metrics for quantitative evaluation of our model. EM score measures whether the predicted answer spans exactly matches the ground truth. F1 score is the harmonic mean of precision and recall. Precision (P) is calculated as the number of correct words divided by length of predicted answer. Recall (R) is calculated as number of correct words divided by length of ground truth:

\[ P = \frac{\text{number of correct words}}{\text{length of predicted answer}} \]

\[ R = \frac{\text{number of correct words}}{\text{length of ground truth}} \]

To track the classification accuracy of no-answer predictions, we used the recommended Answer vs. No Answer (AVN) metric. It simply counts the percentage of correct predictions.

**Experimental Details**

We used the default configuration to train the baseline model:

- character embedding size: 64
- character size: 15

After applying character-level embeddings, we trained the model with the following hyper-parameters:

- hidden state size: 50
- batch size: 32
- learning rate: 0.1
- dropout rate: 0.2
- word dropout rate: 0.2
- 12 Layers: Dense 0.0003

The training time was ~15 minutes for each epoch on Tesla V100. All models were trained for a maximum of 30 epochs.

**Results**

![Figure 3: Performance scores on SQuAD 2.0 dev and test sets](image)

We saw modest improvements for both enhancements. Integrating learnable character embeddings lead to considerable improvement to the baseline. This improvement is attributable to the enhanced ability of the model to receive extra bit of signal to learn word meanings, and thus match words and infer answer spans better. The results also highlight the effectiveness of Self-Matching Attention. Scanning the entire context and aggregating signal relevant to the current context word and query, limits the information loss and produces better predictions.

**Conclusions**

The results highlight the effectiveness of Self-Matching Attention as described in R-Net. Scanning the entire context and aggregating signal relevant to the current context word and query limits the information loss and produces better predictions. Our model achieved 66.92 F1 and 63.62 EM scores on the dev set which is an 8% improvement over the baseline.

In future work we would like to explore different network structures such as GNNs to handle questions that require complex inferences.

**References**


**Acknowledgments**

Professor: Chris Manning

TA Mentor: Christopher Wolff