Challenge

How did you answer the question, despite the errors? Was it:

• Proximity to the words "French Wars of Religion"?
• The preposition that opened the clause for the answer?
• The way that the preface "W" fits to the word "century"?

In our project, we present our model with the same challenges during training, but we separately encode the following intuitions to help it overcome them:

• N+2 between the start and end word in text
• Character-level indicators that load to relevant phrases

Background

We used the SQuAD2.0 dataset to train and test our model. The following related work has provided us with a foundation for our model.

• BiDAF. One of the first models used for SQuAD2.0.
  • Attention flow from question to context, and vice versa
  • It outperformed all systems in the 2016 and 2017 MRC contest.
  • UDA: Uses external data augmentation to prevent overfitting.
  • ETN: Encoder attentive non-linear transform.

• Match-LSTM with Bi-directional Attention (Pointer Net). Combined methods from previous work.
  • Match-LSTM was proposed by Luong et al. (Following above).
  • Temporal aggregation match of attention weights from each word of question.

• Pointer Net: Also attaches an attention vector to select a position from context for an answer.
  • For this paper, we choose the beam conditions and token distribution on the start token.

Methods

We employed 3 different techniques over the course of our experimentation.

1. Character Embedding: dropout, 2D convolution, max pool, concatenation with word vector.

2. Conditional loss: Adding more expressiveness to the model by removing the independence assumption between start and end tokens present in baseline BiDAF model, consequently minimizing the following negative log likelihood:

   \[ \log P(\text{start}|\text{end}) - \log P(\text{end}|\text{start}) \]

3. EDA "Shuffle": For 5% of samples each epoch, we shuffle 50% of tokens.

   Given for both word and question embeddings for both question and context.

Experiments

We ran 4 relevant experiments to completion on four different models.

1. The baseline (BiDAF) model
2. The baseline (BiDAF) model with a layer added for character embeddings
3. Adjust baseline BiDAF model where log likelihood loss changed to a conditional log likelihood loss
4. Combined character embeddings and conditional loss

(All of the experiments were trained for a total of 30 epochs, each at a learning rate of 0.5 and batch size of 64.)

Conclusions

• Subword relationships are a substantial factor in determining word meaning and function.
• The relationship between the start and end words of a context-based answer is not the strongest factor in determining the answer.
• Data augmentation requires careful tuning of proportion of changes with respect to size of dataset.

References