Introduction

Problems: Accurate question and answering systems are crucial to Web search engines to serve information needs.

Objectives: Produce a model which outperforms the baseline Bidirectional Attention Flow (BART) on SQuAD 2.0 introduced in [Lee et al., 2018] [1].

Data & Approach

- Official SQuAD 2.0 dataset + new SQuAD 2.0 example produced by the teaching team.
- Train (20,000) official SQuAD 2.0 training set.
- Dev (5,000) roughly half of the official dev set, randomly selected.
- Test (5,000) remaining examples from the official dev set, plus hand-labeled.

Two neural network structures:
- Self-attention [4], transformer [4], and the R-NET model [5].
- QANet model [4].

Feed QANet with:
- Data augmentation through backtranslation with Neural Machine Translation (NMT) models [5].
- Finally pooling models together with:
  - Ensembling techniques [5].

Example from dataset (correct, question, answer):

Question: Why was Tesla returned to Google?

Context paragraphs: On 23 March 1979, Tesla was returned to Google under police guard for not having a residence permit. On 27 April 1979, Tesla died at the age of 69 after contracting tonsillitis (although some sources report he died of a stroke). During his lifetime, Tesla taught a large class of students in the old Institut Höhere Tafel Grammar School, in Google.

Answer: not having a residence permit.

QANet Architecture

- Pointer-Generator network.
- Can train quickly, but converges early.
- Volatile loss during training.

Self-Attention and R-NET

- Construction didn’t improve upon the baseline.
- R-NET didn’t improve upon the baseline.
- Self-attention improved upon the baseline.

Results

- Six sets each on an expansion of the self-attention model, the BART model with the 3-layer embedding, and the BART model with the full augmented dataset.
- The ensemble achieves a new state-of-the-art for the development set that each individual model on its own.
- The ensemble achieves a new state-of-the-art in F1 score of 82.62 and an EM score of 65.70.

Data Augmentation

- NMT: Transformer and Neural Machine Translation models to perform data augmentation.
- Backtranslation of questions from training dataset for Arabic, Chinese, French, Finnish, Italian, Russian, and Spanish.
- Hans and Spanish show the strongest potential for improving model performance.

Ensemble Methods

- Multi-model: Feed a 0/1 mask to each model, and all combinations of models, trained in all languages.
- Smaller model: Opted for a single convolutional neural network, and trained models, trained on a single language.

Conclusions

- Best performing models are ensembles of full augmentation, self-attention, and the BART model.
- Languages that do not train in English generally have worse performance on the question-answering task.
- 10% of small augmentation models better than one model trained on all data.

Thanks:

- Thanks QANet, first paper for pass performance than original paper.
- Problematic architecture of R-NET model, on all self-attention layer.
- To different language families, such as Korean and Hindi language.