Ensemble BERT with Data Augmentation and Linguistic Knowledge on SQuAD 2.0

Wen Zhou, Hang Jiang, Xianzhe Zhang
{zhouwen, hjian42, xianzhez}@stanford.edu

1. Baseline: BERT
a. BERT is a multi-layer bidirectional Transformer encoder with two novel unsupervised pre-training tasks (MLM and NSP) and constructed input representation. In terms of input, BERT constructs input representation by summing the WordPiece token embeddings, segment embeddings, and positional embeddings.
b. Adaptation to SQuAD:

2. Post-processing with Linguistic Knowledge
During prediction time, the probability for a text span Text(i, j) being the answer is:

\[
P(i, j) = \text{softmax}(\text{start}_\text{logits}) + \text{end}_\text{logits}
\]
a. for ‘When’ questions: if Text(i) is one of ‘before’, ‘after’, ‘about’, ‘around’, ‘from’, ‘during’], add 0.2 to P(i, j);
b. for ‘Where’ questions: if Text(i) is one of ‘in’, ‘at’, ‘on’, ‘behind’, ‘from’, ‘through’, ‘between’, ‘throughout’], add 0.2 to P(i, j);
c. for ‘Whose’ questions: if Text(i,j) contains ’s’, add 0.2 to P(i, j);
d. for ‘Which’ questions: if Text(i) is ‘the’, add 0.2 to P(i, j);

3. Ensemble
For each question, we output the n-best predictions made by multiple models along with the probability, then for each proposed prediction, we sum up the probability from each model together, and finally output the prediction with the highest probability as the answer to that question. A weighting scheme is also used according to the performance of individual models.

Approach

Data Augmentation:

- Injecting more augmented data
d. Adding linguistic heuristics can improve the performance of BERT ensemble.

Post-processing:

- Adding linguistic heuristics can significantly boost EM, while slightly boost F1.

Ensemble

- The best ensemble model can outperform the best single model by 1.2 F1 and 2.1 EM. A proper weighting scheme is very useful (Ens3 vs Ens4).

Conclusions

In our project, we significantly improved the BERT model using a few novel NLP techniques.

1. Ensemble BERT with weighting is an effective way of improving the single BERT.
2. Our novel data augmentation algorithm based on synonym replacement is a good way of enriching the diversity of training set. Adding BERT models fine-tuned on augmented data can significantly improve the performance of BERT ensemble.
3. External linguistic knowledge such as grammar rules and common sense is helpful to correcting the predictions of the models.
4. BiDAF can contribute positively to the ensemble’s performance despite its significantly lower performance compared to BERT, meaning that it can correct some systematic errors made by BERT models.

In the future, we can probably explore a NMT-based data augmentation, which will introduce more diversity to the data. Also, more linguistic knowledge can be integrated into machine learning models. At last, ensemble different models can overcome the systematic errors of the current system.