BERT++: Reading Comprehension on SQuAD

Lirong Yuan

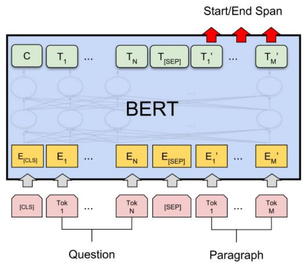
Problem

SQuAD 2.0

- Read Comprehension (RC):
  - Goal: find an answer in a paragraph or a document.
  - Required skills: logical reasoning, commonsense reasoning, understand analogy, causal relations, clause relations, and so on.
- Stanford Question Answering Dataset (SQuAD 2.0):
  - 500+ Wikipedia articles.
  - 100,000+ answerable question-answer pairs.
  - 50,000 unanswerable questions.

BERT

- Language representation model that could be fine-tuned with an additional layer to create models.
- End-to-end model for SQuAD:
  - Embedding layer
  - Linear layer
  - Loss function
  - Predictions
- Limitations:
  - Lack of generalization and real understanding.
  - System should abstain from answering when the question is unanswerable.



Approach

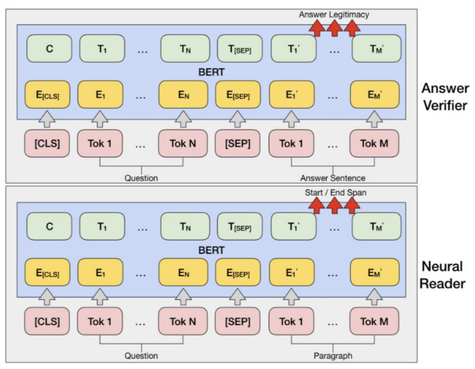
HybridBERT

- Classification Layer: predict answerability of questions
- Loss function:

$$L_{joint} = -\log\left(\frac{(1-\delta)e^{z} + \delta e^{\alpha_a \beta_b}}{e^{z} + \sum_{i=1}^{l_p}\sum_{j=1}^{l_p} e^{\alpha_i \beta_j}}\right)$$

BERT++

- Model architecture:
  - Input data
  - Neural reader
  - Answer verifier
  - Loss functions
  - Predictions
- Neural reader:
  Predict plausible answer for even unanswerable questions
- Answer verifier:
  Check whether the candidate answer is legitimate.



Results

Standard Evaluation metrics

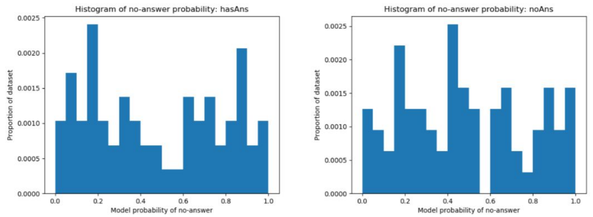
- Exact match (EM):
  - Whether the model output matches the ground truth answers exactly.
- F1:
  - Harmonic mean of precision and recall (less strict than the EM score).

Leaderboard

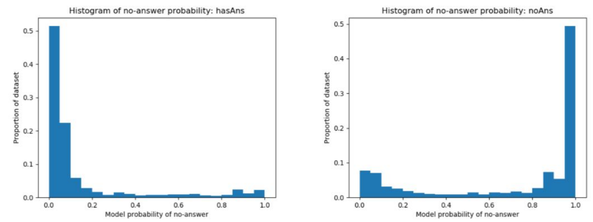
System	Dev PCE Leaderboard		Test PCE Leaderboard	
	EM	F1	EM	F1
BERT (baseline)	71.915	74.487	-	-
HybridBERT (single model)	72.474	75.158	-	-
BERT++ (single large-cased model)	76.077	78.521	71.936	74.560
BERT++ (single large-uncased model)	77.657	80.180	-	-
BERT++ (ensemble model)	78.233	80.530	75.300	77.696

Distribution of no-answer probabilities

- Baseline model:



- BERT++:



Impact of passage length, question length, answer length

Table 2: Impact of passage length on performance (Dev v2.0 data set)

Passage length	Number of questions	Baseline Model EM	Baseline Model F1	BERT++ (single model) EM	BERT++ (single model) F1
[0, 640)	2099	69.271	71.705	76.560	79.410
[640, 1280)	3424	72.663	75.010	78.417	80.636
[1280, 5120)	555	76.936	81.418	77.117	80.282

Table 3: Impact of question length on performance (Dev v2.0 data set)

Question length	Number of questions	Baseline Model EM	Baseline Model F1	BERT++ (single model) EM	BERT++ (single model) F1
[0, 40)	983	70.80	74.56	77.009	80.267
[40, 60)	2525	73.58	75.84	79.485	81.547
[60, 80)	1615	71.26	73.35	76.718	78.965
[80, 320)	955	69.52	72.53	75.078	78.532

Table 4: Impact of answer length on performance (Dev v2.0 data set)

Answer length	Number of questions	Baseline Model EM	Baseline Model F1	BERT++ (single model) EM	BERT++ (single model) F1
[0, 5)	3467	77.09	77.20	84.741	84.828
[5, 20)	1568	69.64	73.10	71.938	75.606
[20, 160)	1043	57.90	67.35	62.703	71.605

Analysis

Sample of errors for unanswerable questions

- Missing Information:
  - Question can be answered, but the information is not in the context.
- False Premise:
  - Question asserts a fact that contradicts information in the context.
- Topic Error:
  - Question references a related but different entity in the context.
- Content negation:
  - Question asks for the opposite information mentioned in the context.

Error Type	Passage	Question	Predicted Answer
Missing Information	Similarly, it is not known if L (the set of all problems that can be solved in logarithmic space) is strictly contained in P or equal to P. Again, there are many complexity classes between the two, such as NL and NC, and it is not known if they are distinct or equal classes.	What variable is not associated with all problems solved within logarithmic space?	L
False premise	... James Wolfe defeated Montcalm at Quebec (in a battle that claimed the lives of both commanders), and victory at Fort Niagara successfully cut off the French frontier forts further to the west and south...	Who was defeated by Montcalm at Quebec?	James Wolfe
Topic error	... Mercury is the working fluid in the mercury vapor turbine. Low boiling hydrocarbons can be used in a binary cycle.	What is the typical working fluid in a vapor turbine?	Mercury
Content negation	... In 1018, Roger de Tosny travelled to the Iberian Peninsula to carve out a state for himself from Moorish lands, but failed...	Who carved out a state for himself from Moorish lands?	Roger de Tosny

Conclusions

- BERT++ is better at modeling answerability of questions.
- Achieved 75.300 EM, 77.696 F1 on test set and competitive rank.

Future work

- Enhance language representation: integrate world knowledge to learn more complete semantic representation of larger language units.
- Better sentence representation: encode sentence structures with neural dependency parser.
- Build targeted models on data sets that have: short passages, very short and very long questions, or long answers and ensemble.

Acknowledgments

I would like to thank Chris Manning and all the TAs for the wonderful classes and patient help during office hours. I would also like to thank Microsoft Azure for providing resources to train the models.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805, 2018.

Minghao Hu, Furu Wei, Yuxing Peng, Zhen Huang, Nan Yang, and Ming Zhou. Read + verify: Machine reading comprehension with unanswerable questions. CoRR, abs/1808.05759, 2018.