Improving Bi-Directional Attention Flow for Machine Comprehension

Bosen Ding, Yue Wang

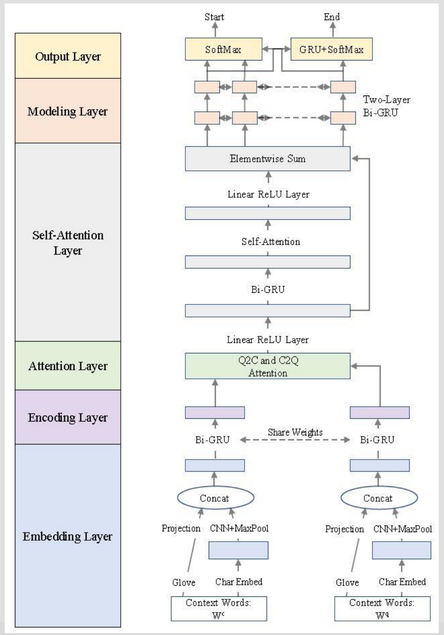
Background

- Machine comprehension (MC) requires a complex model to capture the interaction between a context paragraph and a query to answer the question correctly.
- Bi-Directional Attention Flow (BiDAF) system can obtain query-aware context representation without early summarization. However, BiDAF performs poorly on SQuAD 2.0 compared to SQuAD1.1.
- DATASET: SQuAD 2.0

BiDAF with Residual Self-Attention

Major Difference with the Baseline Model:

- character-level embedding
- the replacement of LSTM with GRU
- the addition of a complex residual self-attention layer



Experimental Details

- character-level embedding size = 32
- hidden state size = 100
- 100 1D filters for the CNN character embedding, with a kernel size = 5
- AdaDelta optimizer with a batch size = 64
- initial learning rate = 0.5
- dropout rate = 0.2

We have trained three baseline models and 20+ BiDAF variant models of different configuration.

Results

Best Performance:

F1-Score		EM	
Dev: **69.81**	Test: **68.05**	Dev: **66.78**	Test: **65.10**

Performance Comparison between Models:

Table 1: Result on SQuAD 2.0 dev set

Model	EM	F1
baseline	57.87	61.23
trained-char-embed-BiDAF-w/o-proj	56.47	60.14
rand-char-embed-BiDAF-w/-proj	56.56	60.17
trained-char-embed-BiDAF-w/-proj	59.87	63.17
rand-char-embed-BiDAF-w/-proj	60.28	63.32
residual-self-attention-char-embed-BiDAF	**64.24**	**67.27**
baseline ensemble (3 models)	60.71	63.76
residual-self-attention-char-embed-BiDAF-ensemble(3)	**66.14**	**69.00**

Performance on Answerable & Unanswerable Sets:

- Significant improvement on unanswerable questions

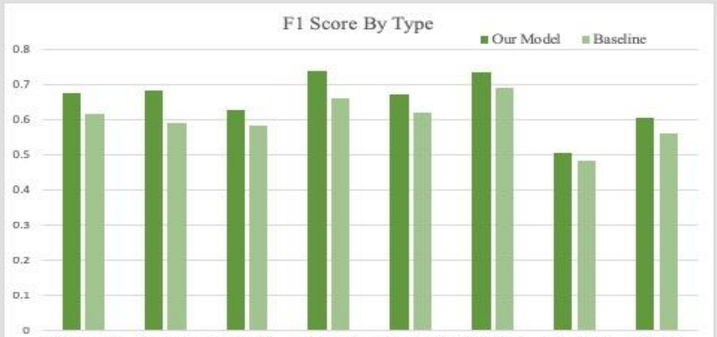
Table 2: Result on SQuAD 2.0 dev set by answerability

Model	EM	F1	A-EM	NA-EM	A-F1	NA-F1
baseline	57.87	61.23	59.66	55.82	66.90	55.82
char-embed-BiDAF	59.87	63.17	61.62	58.58	**68.68**	58.58
our final model	**64.24**	**67.27**	**61.38**	**66.87**	67.71	**66.87**

Analysis

Performance Comparison on Different Types of Questions

- - improves upon the baseline model across the question spectrum
- - the largest percentage improvement comes from who type



Example: our model succeeds where the baseline model fails

- Of the 351 unanswerable questions of who type, the baseline answers 172 questions correctly while our model answers 219 correctly.

- Context: The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. They were descended from Norse ("Norman" comes from "Norseman") raiders and pirates from Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear fealty to King Charles III of West Francia. Through generations of assimilation and mixing with the native Frankish and Roman-Gaulish populations, their descendants would gradually merge with the Carolingian-based cultures of West Francia. The distinct cultural and ethnic identity of the Normans emerged initially in the first half of the 10th century, and it continued to evolve over the succeeding centuries.
- Answer: N/A
- Prediction by baseline: West Francia
- Prediction by our model: N/A

Conclusion

- BiDAF variant neural network implementation improves more than 6 points on a single model performance of SQuAD2.0 dataset.
- Char-level embedding improves the BiDAF by 2 points
- Residual self-attention improves the BiDAF by 4 points.
- Self-attention structure improves the performance on unanswerable questions.

Reference

[1] Christopher Clark and Matt Gardner. "Simple and Effective Multi-Paragraph Reading Comprehension". In: CoRR abs/1710.10723 (2017).

[2] Min Joon Seo et al. "Bidirectional Attention Flow for Machine Comprehension". In: CoRR abs/1611.01603 (2016).