Machine Reading Comprehension on SQuAD 2.0 with Hierarchical Attention and Stochastic Dropout

Yancheng Li, Yiyang Li, Shichang Zhang {lycheng, yiyang7, shichang}@stanford.edu

Motivations

Machine reading comprehension (MRC) is a challenging task, where the goal is to have systems read a text passage and then answer any question about the passage. Its importance is demonstrated by its wide applications. This task is an useful **benchmark to demonstrate natural language understanding,** and is applied across industries, e.g. conversational agents and customer service support.

Recently, MRC has largely benefited from the availability of large-scale benchmark datasets and it is possible to train large end-to-end neural network models.

Dataset

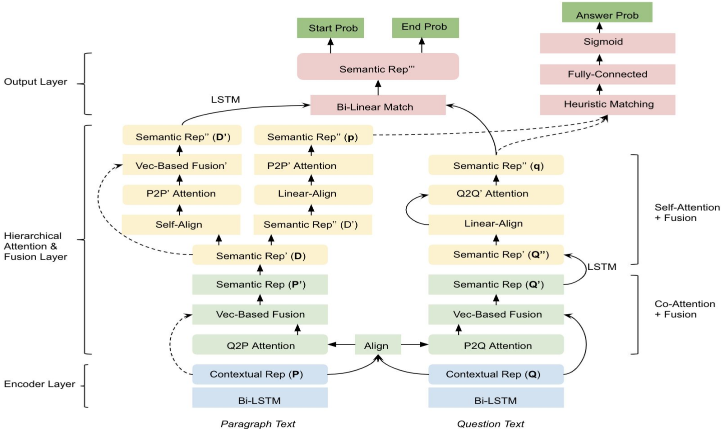
We trained, validated, and compared our model on the **SQuAD 2.0 dataset**, consisting of over **150k question-answer pairs** on 500+ Wikipedia articles where the answer to each question is a span taken from the article. The major difference as compared with SQuAD 1.1 is that the updated dataset includes over **50k unanswerable questions**. This forces the model to comprehend both questions and passages more thoroughly.

Paragraph: One of the first Norman mercenaries to serve as a Byzantine general was **Hervé** in the 1050s. By then however, there were already Norman mercenaries serving as far away as Trebizond and Georgia. They were based at Malatya and Edessa, under the Byzantine duke of Antioch, Isaac Komnenos. In the 1060s, Robert Crispin led the Normans of Edessa against the Turks. **Roussel de Bailleul** even tried to carve out an independent state in Asia Minor with support from the local population, but he was stopped by the Byzantine general **Alexius Komnenos**.

Question 1: “When did **Herve** serve as a Norman general?”
Plausible Answer: NO ANSWER

Question2: Who ruined **Roussel de Bailleul's** plans for an independent state?
Plausible Answers: Alexius Komnenos

MRC Model Architecture



Encoding Layer: employs Bi-LSTM to refine the coarse word-embeddings and obtain contextual representations of both query and passage.

Co-Attention Layer: captures relationship between query and passage by using the hierarchical fusion kernel that combines representation from multiple granularities to better the model understanding and training efficiency.

Self-Attention Layer: employs a bilinear self-attention function to address the long-distance dependency within different contexts and allow contextual information to flow between passages and queries.

Matching and Output Layer: inherits the idea of stochastic dropout and bilinear matching function into the model span detector.

AvNA Classifier: a binary classifier to predict if the a given query is answerable. We apply the idea of multi-tasking to train the MRC model and classifier simultaneously.

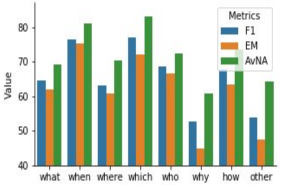
Cross Entropy Loss Function:

$$L_{span} = -\log p_{start}(i) - \log p_{end}(j)$$
$$L_{classify} = -\mathbf{1}_{\{answerable\}} \log P_{score} - (1 - \mathbf{1}_{\{answerable\}}) \log P_{score}$$
$$L_{joint} = L_{span} + L_{classify}$$

Results

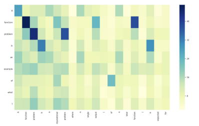
	Dev Set	Test Set
Models	EM/F1	EM/F1
BiDAF baseline	55.9/59.2	-/-
BNA	59.8/62.6	59.2/62.1
DocQA	61.9/64.8	59.3/62.3
SAN	69.3/72.2	68.7/71.4
SLQA+ (with ELMo)	-/-	71.4/74.4
Human	-/-	82.3/91.2
Our Model	**63.7/66.6**	**54.4/57.9**

Table 1: SQuAD 2.0 Models Performance Comparison



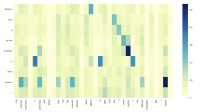
Error Analysis

Passage: A function problem is **a computational problem** where a single output (of a total function) is expected for every input, but the output is more complex than that of a decision problem, that is, it isn't just yes or no. Notable examples include the traveling salesman problem and the integer factorization problem.



Question: A function problem is an example of what?
Gold Answer: a computational problem
Predicted Answer: traveling salesman problem and the integer factorization problem

Passage: **The Internet2 community**, in partnership with Qwest, **built the first Internet2 Network**, called Abilene, **in 1998 and was a prime investor** in the National LambdaRail project.



Question: Abilene was a prime investor in what project?
Gold Answer: NO ANSWER
Predicted Answer: National LambdaRail

Conclusion

We implemented an end-to-end neural network for question answering on SQuAD 2.0 according to **Hierarchical Attention Fusion Networks** to combine representations from multi-level granularity and **stochastic span detection** module. It achieves **66.6 F1 and 63.7 EM** score on the Devset on Non-PCE leaderboard and a relatively lower score on Testset. Our hierarchical attention fusion mechanism can **capture and synthesize most related information between queries and context**. In the future, we plan to try multi-head attention on top of our model to improve model efficiency when having multi-reasoning steps, augment dev set to balance the number of sample of each type of questions of various lengths, add pre-trained embeddings like ELMo/BERT and etc.

Reference

[1] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don’t know: Unanswerable questions for squad. arXiv preprint arXiv:1806.03822, 2018.
[2] Wei Wang, Ming Yan, and Chen Wu. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. arXiv preprint arXiv:1811.11934, 2018.
[3] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543, 2014.