Learning to Ask Unanswerable Questions for Machine Reading Comprehension

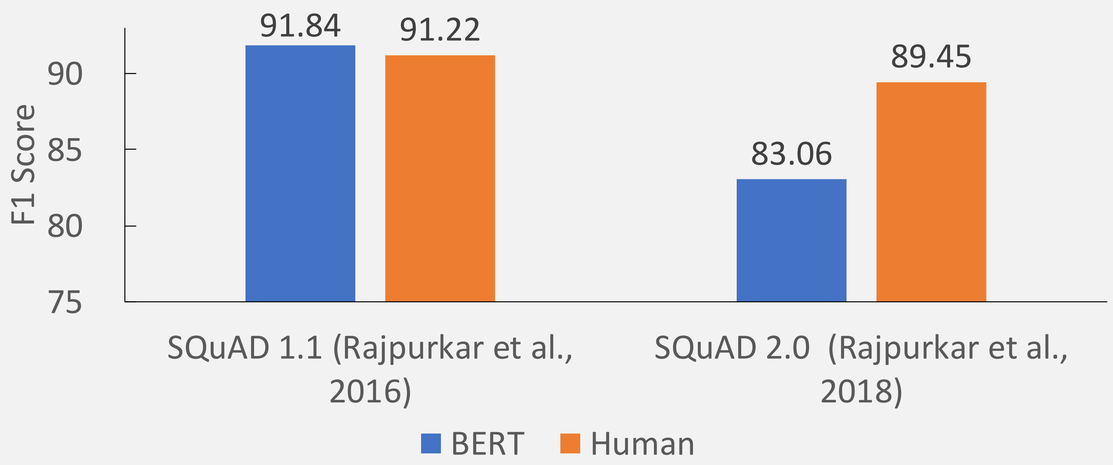
Haichao Zhu[1], Li Dong[2], Furu Wei[2], Wenhui Wang[2], Bing Qin[1], Ting Liu[1]

[1]Harbin Institute of Technology [2]Microsoft Research Asia

Machine reading comprehension with unanswerable questions is a challenging task. In this work, we propose a data augmentation technique by automatically generating relevant unanswerable questions according to an answerable question paired with its corresponding paragraph that contains the answer.

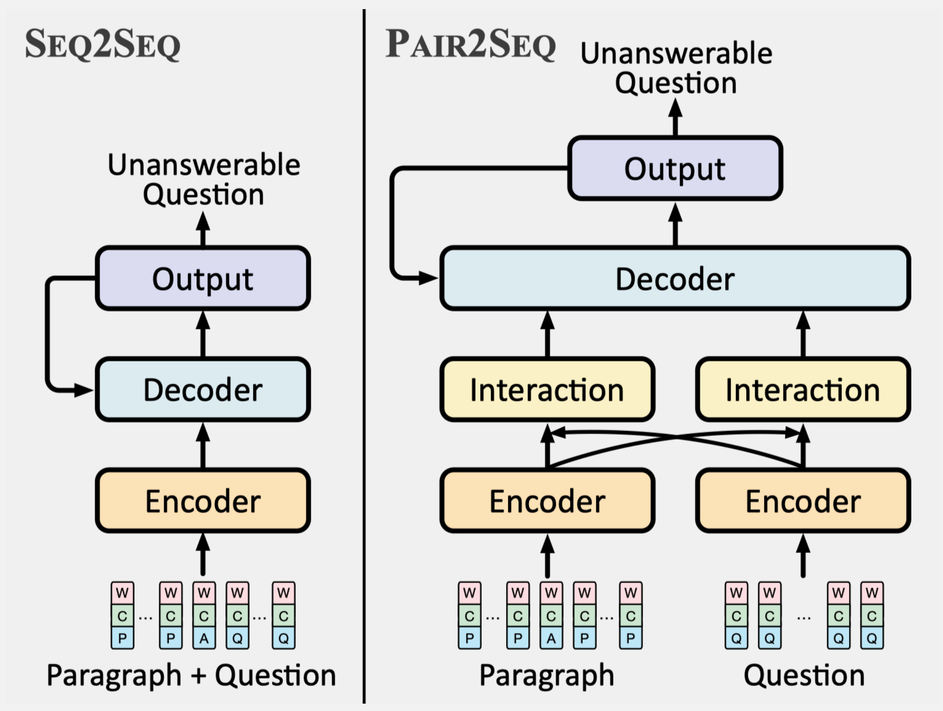
Motivation

Question answering models have **outperformed human** on the extractive benchmark SQuAD1.1. SQuAD2.0 with **unanswerable questions** poses new **challenges** to top-performance systems. **Data augmentation** of unanswerable questions can help data-driven neural models.



Question Generation

Two generation models are under **encoder-decoder framework**. In Seq2Seq model, paragraph and question pairs are packed into an ordered sequence. In **Pair2Seq** model, paragraph and question are encoded separately and an **interaction module** is used to capture interactions between them. Input embeddings are the sum of **word** embeddings, **character** embeddings and **token type** embeddings. We use copy mechanism (Gu et al., 2016; See et al., 2017) to facilitate the use of input words during the generation process.



Model	GLEU-3	GLEU-4	BLEU-3	BLEU-4	ROUGE-2	ROUGE-3
Seq2Seq	33.13	27.39	36.80	27.84	46.54	32.98
Pair2Seq	**35.06**	**29.43**	**37.67**	**29.17**	**47.46**	**34.18**
-Paragraph (+AS)	34.42	28.43	37.35	28.44	47.13	33.29
-Paragraph	33.58	27.54	35.89	26.99	46.14	31.45
-Question	9.40	6.21	6.7	3.1	12.64	5.63
-Copy	25.06	19.80	36.06	22.84	33.40	20.45

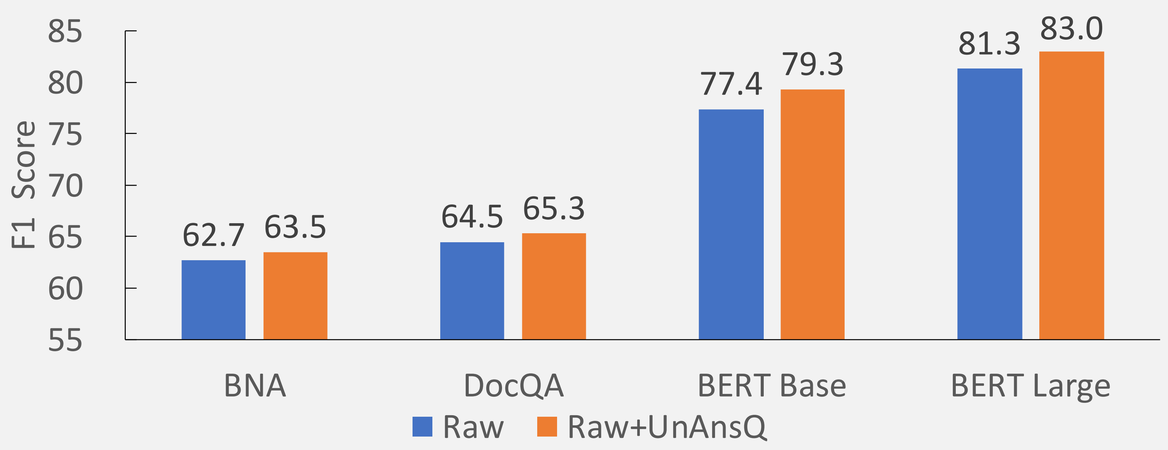
Question Generation Data Construction

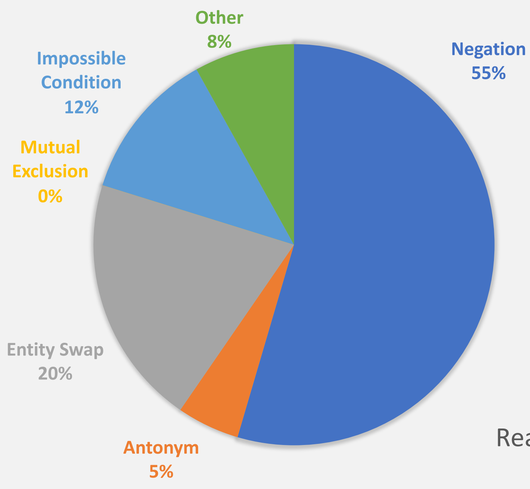
We use answer span as pivot to align questions from SQuAD2.0.

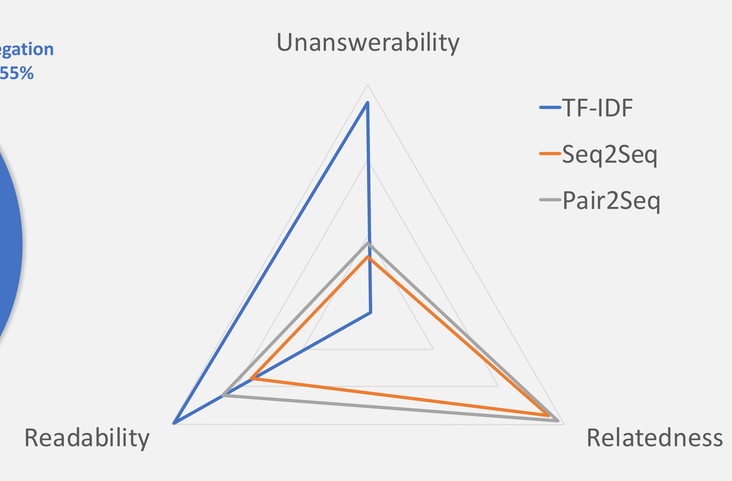
- **Paragraph:** *"Public schools, also known as state or government schools, are funded and run directly by the Victoria Department of Education. Students do not pay tuition fees, but some extra costs are levied. Private fee-paying schools include parish schools …"*
- **Ans. Question:** *"What organization runs the public schools in Victoria?"*
- **UnAns. Question:** *"What organization runs the waste management in Victoria?"*
- **(Plausible) Answer:** "Victoria Department of Education"

Question Answering

We use the **SQuAD2.0** in question answering experiments. Augmentation data is composed of unanswerable questions generated by the generation model, combined with the original data to train question answering models.







Conclusion

- Propose to **generate unanswerable questions** as a means of **data augmentation** for machine reading comprehension.
- Introduce **pair-to-sequence** generation model to capture interactions between question and paragraph.
- Present a way to **construct training data** for unanswerable question generation models.