Exploring Question-Specific Rewards for Generating Deep Questions

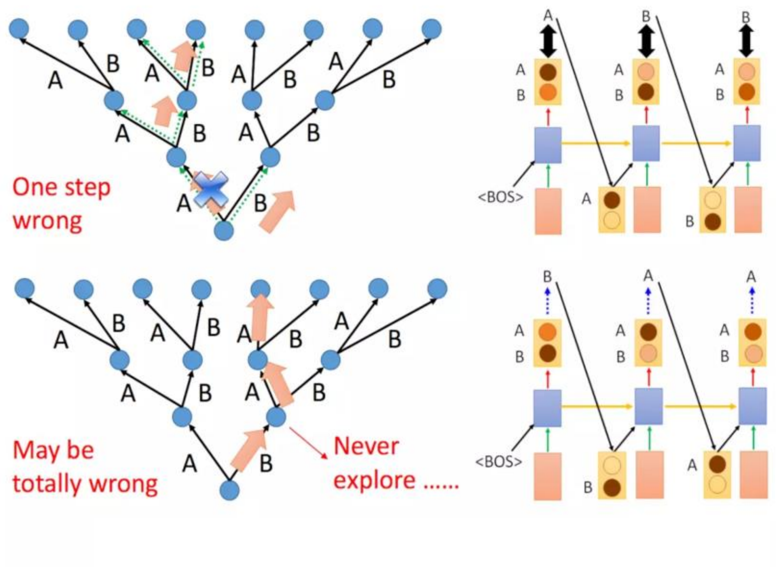
Yuxi Xie[1] Liangming Pan[2,4] Dongzhe Wang[3]
Min-Yen Kan[2] Yansong Feng[1,5]

[1]Wangxuan Institute of Computer Technology, Peking University
[2]School of Computing, National University of Singapore, Singapore
[3]Zhuiyi Technology
[4]NUS Graduate School for Integrative Sciences and Engineering
[5]The MOE Key Laboratory of Computational Linguistics, Peking University

Introduction

Challenges in Question Generation

- ***Exposure Bias***
  - *Inconsistence between training objectives and targets*
  - *Hard to measure the global quality of the generated questions*



Introduction

Challenges in Question Generation

- ***Evaluation***
  - *Current n-gram based evaluation metrics cannot properly evaluate a question*
  - *Problems in Fluency, Relevance and Answerability still remain unsolved*

Lawrence Ferlinghetti is an American poet, he wrote a short story named what?

Lawrence Ferlinghetti is an American poet, **he is a short story written by who**?

0.54

What mine was operated at an earlier date, Kemess Mine or Colomac Mine?

Between Kemess Mine and Colomac Mine, which mine was operated earlier?

0.00

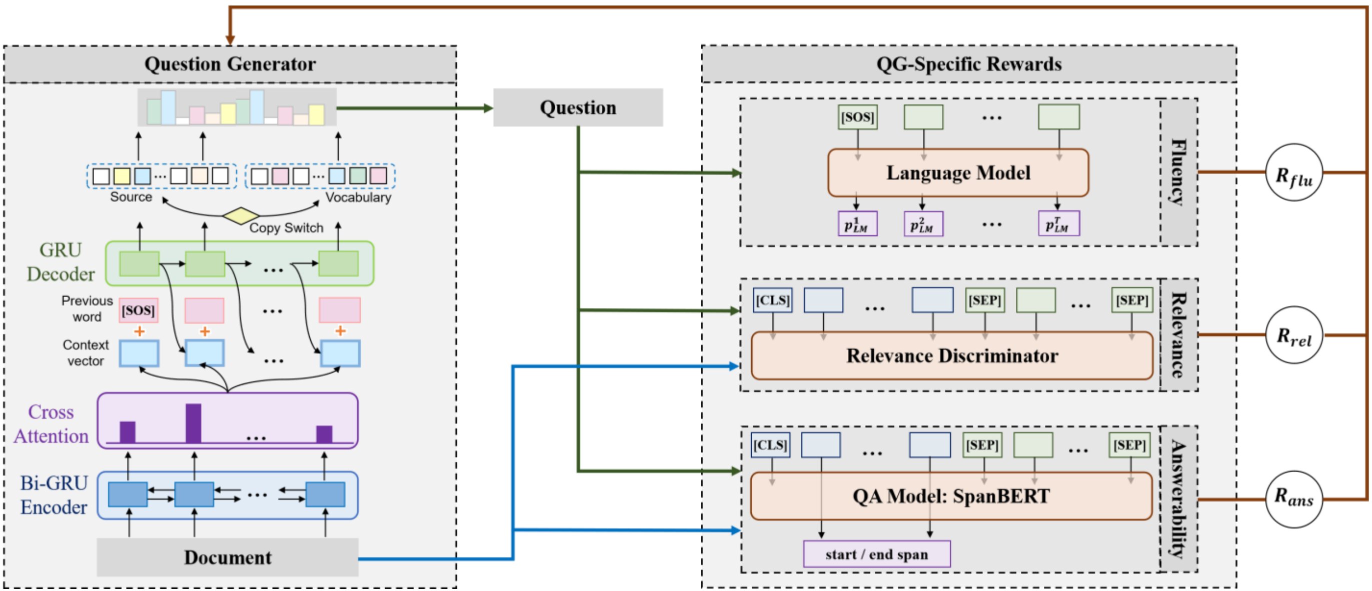
Introduction

Reinforcement Learning in Question Generation

- ***Decouple* the training procedure from the ground truth data**
  - *the space of possible questions can be better explored*
- ***Allow to target on specific properties we want the question to exhibit during training***
  - *e.g. relevant to a specific topic or answerable by the document*
- ***How to define robust and effective QG-specific rewards requires further investigation***
  - *optimizing the reward scores does not always lead to higher question quality in practice*

Methodology

Three Research Questions to Answer



- *Does optimizing RL rewards really improve the* ***question quality*** *from the* ***human standard***
- ***Which reward is more effective*** *in improving the question quality*
- *How the rewards* ***interfere*** *with each other when jointly optimized*

Methodology

Relevance Discriminator

- ***Discriminator initialization***
  - *BERT-based Sentence Classifier*

- ***Training datasets***
  - *Positive: G.T. document + question*
  - *Negative*
    - *Basic: question swap*
    - *Ghost entity: entity swap between different samples*
    - *Logic correctness: entity swap within the same sample*

Experiments

Automatic Evaluation

- *Optimizing a single reward alone (F, R, A) improves the BLEU score and its corresponding reward score.*
- *The three rewards are correlated. One improves, the other two also increase.*
- *Jointly training multiple rewards in general leads to better performance.*
- *The increase in rewards do not correlate well with improvement on automatic metrics*

Model	Rewards			Metrics						
	F	R	A	BLEU1	BLEU4	METEOR	ROUGE-L	R-FLU	R-REL	R-ANS
B1. Baseline				33.68	13.46	**21.39**	35.06	–	–	–
S1. F	√			37.59*	15.22*	19.49*	35.08	+1.48*	+0.49*	+0.03
S2. R		√		36.33*	14.83*	20.63*	**35.58***	+1.06*	**+0.61***	+0.04
S3. A			√	36.40*	13.95*	18.73*	34.07*	+1.30	+0.18*	+0.21*
E1. F + R	√	√		37.82*	15.30*	19.95*	35.48*	+1.30*	+0.60*	+0.03
E2. R + A		√	√	35.77*	14.46*	20.53*	35.26	+0.78	+0.49*	+0.36*
E3. F + A	√		√	**38.30***	14.99*	18.02*	34.50*	**+1.71***	+0.40*	**+0.51***
E4. F + R + A	√	√	√	37.97*	**15.41***	19.61*	35.12	+1.57*	**+0.61***	+0.49*

Table 1: The QG performance evaluated by automatic metrics when separately or jointly optimizing for various rewards. The last three columns show the change of reward scores compared with B1, where **R-FLU** is the fluency, **R-REL** the relevance, and **R-ANS** the answerability rewards. $*$ denotes that the increase/drop in performance compared with B1 is statistically significant for $p < 0.01$.

Experiments

Automatic Evaluation

- *If judging by automatic evaluation metrics, we find that optimizing QG-specific rewards is effective in generating deep questions, compared with other strategies.*
- *However, does optimizing rewards really improves the question quality as expected?*

Model	Features						Metrics			
	AE	**LF**	**CP**	**CV**	**SA**	**RL**	**BLEU1**	**BLEU4**	**Meteor**	**Rouge-L**
B2. Bahdanau et al. (2015)							32.97	11.81	18.19	33.48
B3. NQG++ (Zhou et al., 2017)		●	●				35.31	11.50	16.96	32.01
B4. Zhao et al. (2018)			●		●		35.36	11.85	17.63	33.02
B5. Zhao et al. (2018) + ans, cov	●		●	●	●		38.74	13.48	18.39	34.51
B6. CGC-QG (Liu et al., 2019)	●	●	●				31.18	14.36	**25.20**	**40.94**
B7. SG-DQG (Pan et al., 2020)	●	●	●	●			**40.55**	**15.53**	20.15	36.94
E4. Ours (F + R + A)			●	●		●	37.97	15.41	19.61	35.12

Table 2: Performance comparison. For all baselines, we use the reported performance from Pan *et al.* (2020). Legend: **AE**: answer encoding, **LF**: linguistic features, **CP**: copying mechanism, **CV**: coverage mechanism, **SA**: gated self-attention, **RL**: reinforcement learning.

Experiments

Human Evaluation

Model	**Flu.** (1-5)	**Rel.** (1-3)	**Ans.** (0-1)	**Cpx.** (1-3)
B1. Baseline	3.98	2.77	0.67	**1.59**
S1. F	4.07	2.78	0.61	1.50
S2. R	**4.24**	**2.83**	**0.70**	1.51
S3. A	3.82	2.63	0.46	1.55
E4. F+R+A	4.10	2.72	0.53	1.52

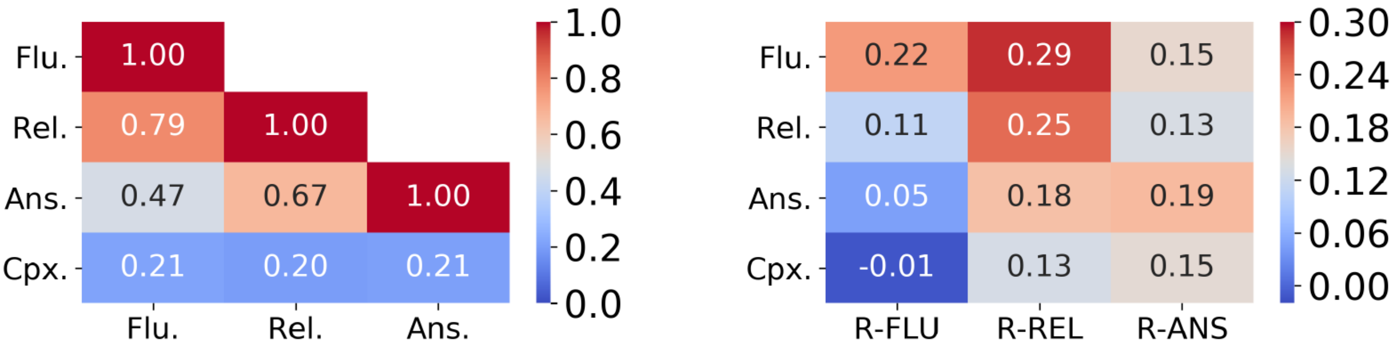
Table 3: Human evaluation results for different methods. **Flu.**, **Rel.**, **Ans.**, and **Cpx.** denote the *Fluency*, *Relevance*, *Answerability*, and *Complexity*, respectively.

- *Human ratings do not correlate well with automatic evaluation metrics*
- *Optimizing the relevance reward (S2) alone leads to an improvement of the human ratings for fluency, relevance, and answerability.*
- *Optimizing for answerability (S3) has a negative effect.*

- ***Conclusion**: If we want to know whether a certain reward has an effect or not, judging from automatic metrics maybe deceiving.*
- *BUT, why relevance works, but answerability fails?*

Experiments

Consistency between Rewards & Human Judgement



(a) Correlation between human ratings

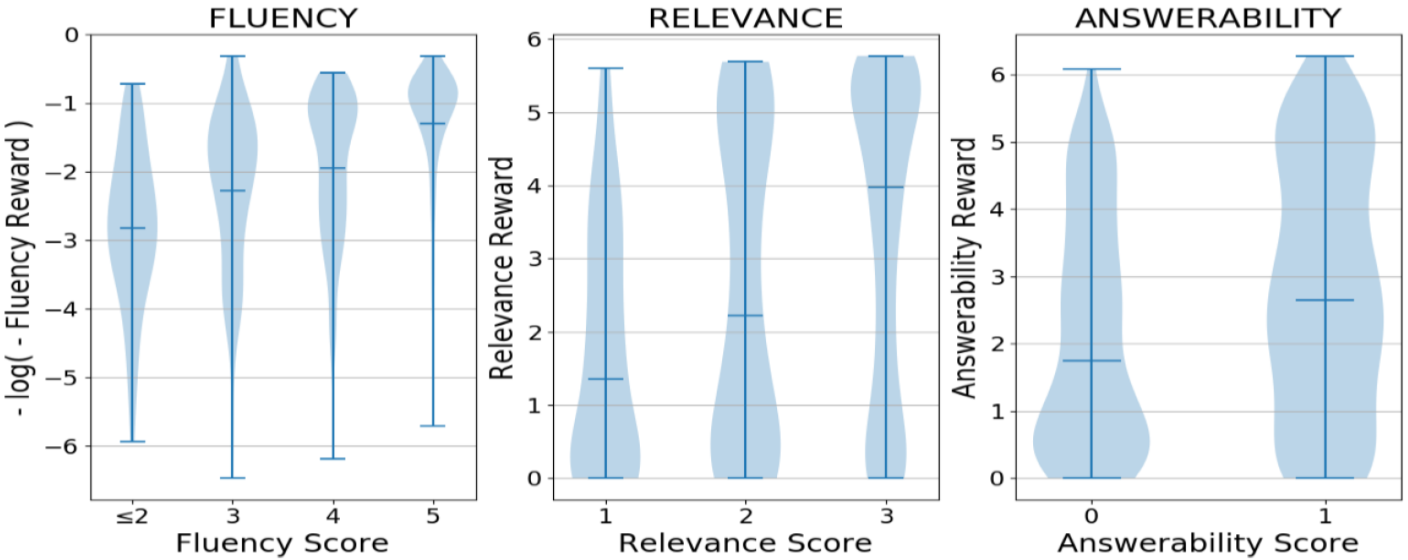
(b) Correlation between reward scores and human ratings

- *the* ***relevance rating has strong correlations*** *with both the fluency rating and the answerability rating, compared with a relatively weak correlation exists between the* ***fluency and answerability***

- *the* ***relevance reward*** *has strong correlations with all three ratings*
- *the* ***answerability reward*** *has poor correlation with fluency and relevance*

Experiments

Consistency between Rewards & Human Judgement



- *Relevance Reward: good correlation*
- *Fluency Reward: normal correlation*
- *Answerability Reward: bad correlation*

- ***Conclusion**: how well the reward score correlates with the human judgement is a good way to know whether a certain reward works or not.*

Experiments

Meso Analysis - Fluency

		Samples with High / Low Rewards
F.	H.	**[FH-1] Que.** Eleven : A Music Company was born in what year ?
		[FH-2] Que. Dan Smith was born in what year ? Dan Smith
	L.	**[FL-1] Que.** Park Seo - joon starred in a South Korean television series that premiered on May 22 , 2017 every Monday and Tuesday where ?
		[FL-2] Que. Kenji Mizoguchi was born in what year ? Kenji Mizoguchi

- *sometimes the fluency reward is consistent with the human judgement on fluency*
- *the LM tends to assign low rewards to the questions with rare or unseen entities*
- *the lack of commonsense knowledge is another problem of the LM*

Experiments

Meso Analysis - Relevance

R.	**H.**	**[RH-1] Doc.** " Sk8er Boi " is a song by the singer Avril Lavigne , released as the second single from her debut album , " Let Go " (2002) . **Que.** " Sk8er Boi " is a song written by a singer born in what year ?
		[RH-2] Doc. Roy Holder then appeared in " The Taming of the Shrew " (1967) , " Here We Go Round the Mulberry Bush " (1967) , " Romeo and Juliet " (1968) . . . The Taming of the Shrew is a 1967 film based on the play of the same name by William Shakespeare about a courtship betw-een two strong - willed people . **Que.** Roy Holder appeared in a 1967 film based on the play of the same name by William Shake-speare about what?
	L.	**[RL-1] Doc.** Beitun , Xinjiang is a county - level city under the direct administration of the regional gov-ernment . Wafangdian is one of the two northern county - level cities , the other being Zhuanghe , under the administration of Dalian , located in the south of Liaoning province , China . **Que.** Are both Granly and Wafangdian located in the same country ?
		[RL-2] Doc. In physics and engineering , the Fourier number or Fourier modulus , named after Joseph Fourier , is a dimensionless number that characterizes transient heat conduction . **Que.** Joseph Fourier was named after a man born in what year ?

- *two aspects for the relevance discriminator*
    - *ghost entity*
    - *logical inconsistency*
- *it is difficult for the model to assign an appropriate relevance score when the question is asking about an unmentioned aspect of something in the document*
    - *potential solution: a good answerability discriminator*

Experiments

Meso Analysis - Answerability

A.	H.	**[AH-1] Doc.** The Worst Journey in the World was written and published in 1922 by a member of the exp- edition , Apsley Cherry – Garrard . **Que.** The Worst Journey in The Worst Journey was born in what year ? **[AH-2] Doc.** Weber ' s Store , at 510 Main St . in Thompson Falls in Sanders County (founded in 1905) , Montana was listed on the National Register of Historic Places in 1986 . **Que.** In what year was the county founded in which Terry ' s Store was listed on the National Re- gister of Historic Places ?
	L.	**[AL-1] Doc.** John Stoltenberg is the former managing editor of " AARP The Magazine " , a bimonthly publication of the United States - based advocacy group AARP , a position John Stoltenberg held from 2004 until 2012 . AARP The Magazine is an American bi - monthly magazine , published by the American Association of Retired People , AARP , which focuses on aging issues . **Que.** John Stoltenberg is the former managing editor of a magazine published by which organization? **[AL-2] Doc.** 8 Spruce Street is a 76 - story skyscraper designed by architect Frank Gehry in the New York City . The original World Trade Center was a large complex of seven buildings in Lower Manhattan , New York City , United States . **Que.** 8 Spruce Street and the original World Trade Center , are located in which country ?

- *most of the questions with high rewards are asking what year (the text highlighted in pink)*
- *when the question requires the QA model to conduct reasoning such as comparison and to utilize world knowledge, the QA model tends to give a low answerability reward*
- *to improve the answerability via a QA-based reward, it is crucial to address the QA model's bias in prediction and improve its reasoning ability*

Q & A

T H A N K   Y O U   F O R   W A T C H I N G