IITB-Monash Research Academy

An Indian-Australian research partnership

Indian Institute of Technology Bombay

MONASH University

SCAI 2019, 12/08/2019

A FRAMEWORK FOR AUTOMATIC QUESTION GENERATION FROM TEXT USING DEEP REINFORCEMENT LEARNING

VISHWAJEET KUMAR[1,2,3], GANESH RAMAKRISHNAN[2], YUAN-FANG LI[3]

[1]IITB-MONASH RESEARCH ACADEMY, [2]IIT BOMBAY, [3]MONASH UNIVERSITY

OUTLINE

- Introduction & motivation
- The generator-evaluator framework
- Evaluation
- Conclusion

WHEN/WHERE/WHY DO WE ASK QUESTIONS?

- **Organisation**: policies, product & service documentation, patents, meeting minutes, FAQ, …
- **Education**: reading comprehension assessment
- **Healthcare**: clinical notes
- **Technology**: chatbots, customer support, …

THE QUESTION GENERATION TASK

- **Goal**
  - Automatically generating questions
    - From sentences or paragraphs
- **Challenges**
  - Questions must be well-formed
  - Questions must be relevant
  - Questions must be answerable

MOTIVATION

- QG: a (relatively) recent task: a Seq2Seq problem
  - RNN-based models with attention perform well for short sentences
  - However for longer text they perform poorly
- Cross-entropy loss may make the training process brittle: the exposure bias problem

EXAMPLE GENERATED QUESTIONS

Example text: "new york city traces its roots to its 1624 founding as a trading post by colonists of the dutch republic and was named new amsterdam in 1626 ."

MODEL	QUESTION
Seq2Seq with cross-entropy loss	what year was new york named?
Copy-aware seq2seq	what year was new new amsterdam named?
GE (Seq2seq with BLEU)	what year was new york founded?

TO BE MORE SPECIFIC

- QG performance is evaluated using discrete metrics like BLEU, ROUGE etc., **not** cross-entropy loss
- Need for a mechanism to deal with relatively **rare word** and important words
- Need to handle the **word repetition** problem while decoding

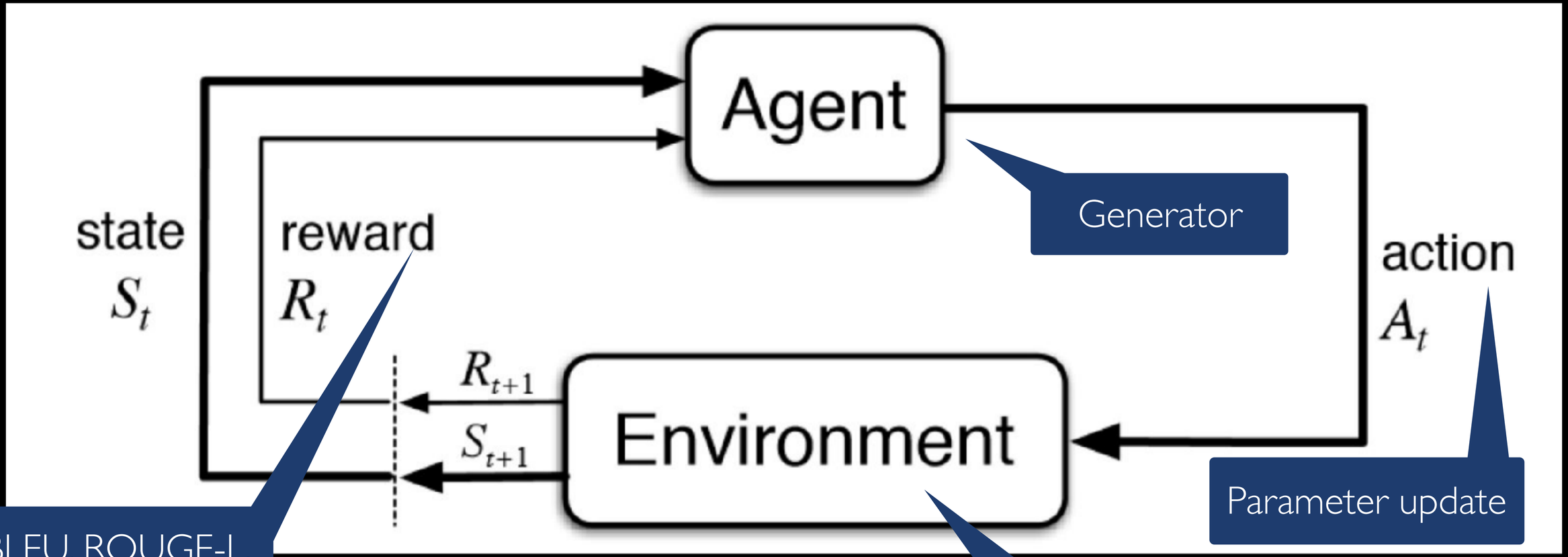
OUTLINE

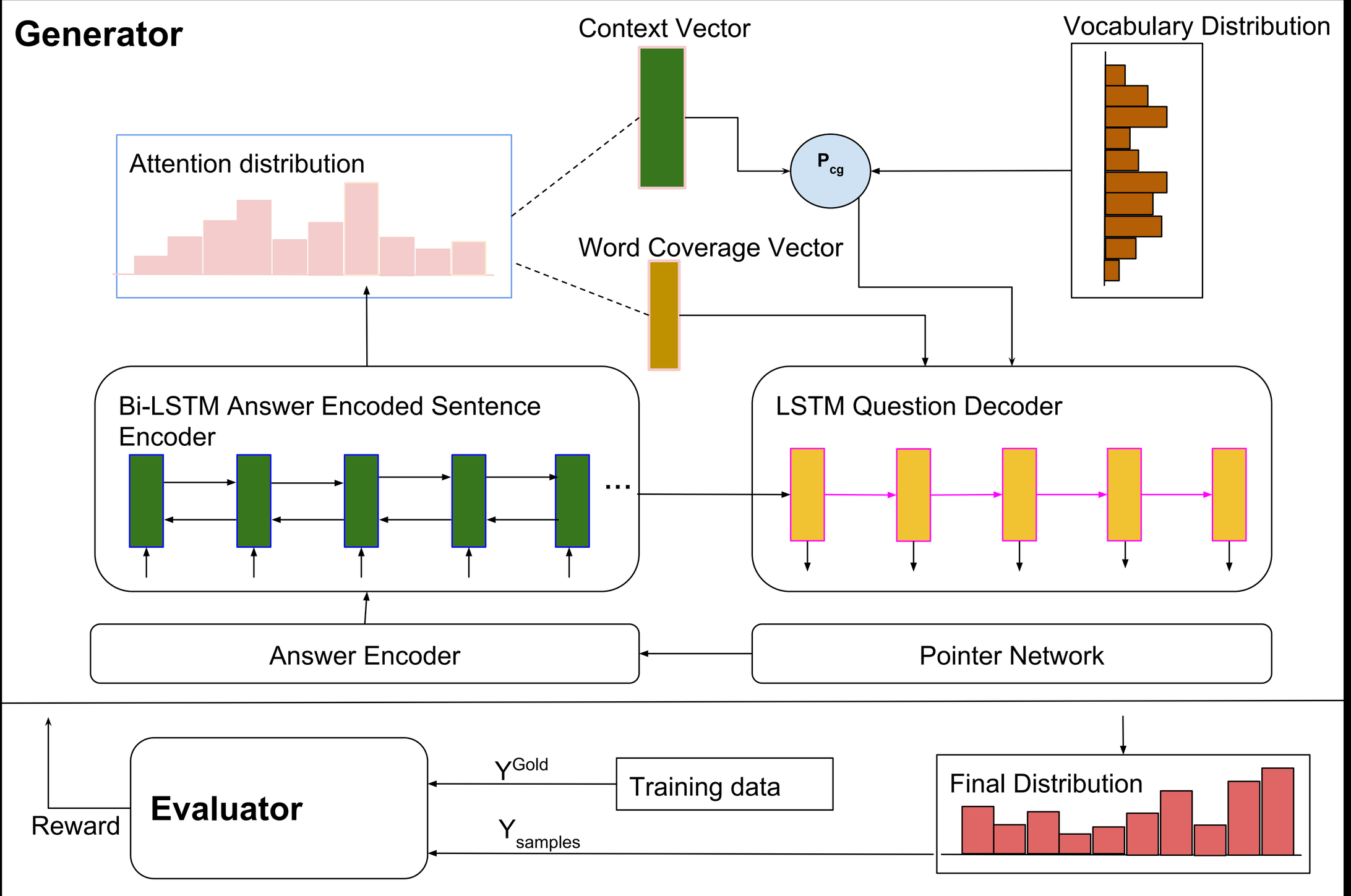
- Introduction & motivation
- The generator-evaluator framework
- Evaluation
- Conclusion

A GENERATOR-EVALUATOR FRAMEWORK FOR QG

- **Generator** (*semantics*)
  - Identifies *pivotal answers* (Pointer Networks)
  - Recognises *contextually important* keywords (Copy)
  - Avoids *redundancy* (Coverage)
- **Evaluator** (*structure*)
  - Optimises *conformity* towards ground-truth questions
  - Reinforcement learning with performance metrics as rewards

REINFORCEMENT LEARNING FOR QG





ARCHITECTURE

REWARD FUNCTIONS

- General rewards
  - BLEU, GLEU, METEOR, ROUGE-L
  - DAS: decomposable attention that considers variability
- QG-specific rewards
  - QSS: degree of overlap between generated question & source sentence
  - ANSS: degree of overlap between predicted answer & gold answer

OUTLINE

- Introduction & motivation
- The generator-evaluator framework
- Evaluation
- Conclusion

EVALUATION: DATASET & BASELINES

- **Dataset**: SQuAD
  - Train: 70,484
  - Valid: 10,570
  - Test: 11,877

- **Baselines**
  - Learning to ask (L2A): vanilla Seq2Seq model (ACL'17)
  - NQG_{LC}: Seq2Seq + ground-truth answer encoding (NAACL'18)
  - AutoQG: Seq2Seq + answer prediction (PAKDD'18)
  - SUM: RL-based summarisation (ICLR'18)

AUTOMATIC EVALUATION

MODEL	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L
L2A	43.21	24.77	15.93	10.60	16.39	38.98
AutoQG	44.68	26.96	18.18	12.68	17.86	40.59
NQG_{LC}	-	-	-	(13.98)	(18.77)	(42.72)
SUM_{BLEU}	11.20	3.50	1.21	0.45	6.68	15.25
SUM_{ROUGE}	11.94	3.95	1.65	0.082	6.61	16.17
GE_{BLEU}	46.84	29.38	20.33	14.47	19.08	41.07
$GE_{BLEU+QSS+ANSS}$	46.59	29.68	20.79	15.04	19.32	41.73
GE_{DAS}	44.64	28.25	19.63	14.07	18.12	42.07
$GE_{DAS+QSS+ANSS}$	46.07	29.78	21.43	16.22	19.44	42.84
GE_{GLUE}	45.20	29.22	20.79	15.26	18.98	43.47
$GE_{GLUE+QSS+ANSS}$	47.04	30.03	21.15	15.92	19.05	43.55
GE_{ROUGE}	47.01	30.67	21.95	16.17	19.85	43.90
$GE_{ROUGE+QSS+ANSS}$	**48.13**	**31.15**	**22.01**	**16.48**	**20.21**	**44.11**

HUMAN EVALUATION

MODEL	SYNTAX		SEMANTICS		RELEVANCE	
	SCORE	KAPPA	SCORE	KAPPA	SCORE	KAPPA
L2A	39.2	0.49	39	0.49	29	0.40
AutoQG	51.5	0.49	48	0.78	48	0.50
GE_{BLEU}	47.5	0.52	49	0.45	41.5	0.44
$GE_{BLEU+QSS+ANSS}$	82	0.63	75.3	0.68	**78.33**	0.46
GE_{DAS}	68	0.40	63	0.33	41	0.40
$GE_{DAS+QSS+ANSS}$	**84**	0.57	**81.3**	0.60	74	0.47
GE_{GLUE}	60.5	0.50	62	0.52	44	0.41
$GE_{GLUE+QSS+ANSS}$	78.3	0.68	74.6	0.71	72	0.40
GE_{ROUGE}	69.5	0.56	68	0.58	53	0.43
$GE_{ROUGE+QSS+ANSS}$	79.3	0.52	72	0.41	67	0.41

OUTLINE

- Introduction & motivation
- The generator-evaluator framework
- Evaluation
- Conclusion

CONCLUSION

- A generator-evaluator framework for question generation from text
  - Takes into account both semantics & structure
  - Proposes novel reward functions
- Evaluation shows state-of-the-art performance

THANK YOU!

ANY QUESTIONS?

REFERENCES

- Xinya Du, Junru Shao, and Claire Cardie. Learning to ask: Neural question generation for reading comprehension. In ACL, volume 1, pages 1342–1352, 2017.
- Vishwajeet Kumar, Kireeti Boorla, Yogesh Meena, Ganesh Ramakrishnan, and Yuan-Fang Li. Automating reading comprehension by generating question and answer pairs. In PAKDD, 2018.
- Pranav Rajpurkar, Jian Zhang, Kon- stantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In EMNLP 2016, pages 2383–2392. ACL, November 2016.
- Linfeng Song, Zhiguo Wang, Wael Hamza, Yue Zhang, and Daniel Gildea. Leveraging context information for natural question generation. In NAACL, pages 569–574, 2018.
- Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. In ICLR, 2018.

SOME MORE EXAMPLES

Text: “critics such as economist paul krugman and u.s. treasury secretary timothy geithner have argued that the regulatory framework did not keep pace with financial innovation, such as the increasing importance of the shadow banking system, derivatives and off-balance sheet financing.”

MODEL	QUESTION
AutoQG	who argued that the regulatory framework was not keep to take pace with financial innovation?
GE_{BLEU}	what was the name of the increasing importance of the shadow banking system?
GE_{DAS}	what was the main focus of the problem with the shadow banking system?
GE_{GLEU}	what was not keep pace with financial innovation?
GE_{ROUGE}	what did paul krugman and u.s. treasury secretary disagree with?

“Legislative power in Warsaw is vested in a unicameral Warsaw City Council (Rada Miasta),which comprises 60 members. Council members are elected directly every four years . Like most legislative bodies, the City Council divides itself into committees which have the oversight of various functions of the city government.”

– HTTPS://EN.WIKIPEDIA.ORG/WIKI/WARSAW

1	How many members are in the Warsaw City Council?
2	How often are the Rada Miasta elected?
3	The City Council divides itself into what?