Difficulty Controllable Generation of Reading Comprehension Questions

Yifan Gao[1], Lidong Bing[2], Wang Chen[1], Michael R. Lyu[1], Irwin King[1]
[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong
[2]R&D Center Singapore, Machine Intelligence Technology, Alibaba DAMO Academy
[1]*{yfgao, wchen, lyu, King}@cse.cuhk.edu.hk* [2]*l.bing@alibaba-inc.com*

IJCAI 2019 MACAO

Difficulty Controllable Question Generation: A New Task

S_1 : Oxygen is a chemical element with symbol O and atomic number 8.
A_1: 8
Q_1: (Easy) What is the atomic number of the element oxygen?

S_2 : The electric guitar is often emphasised, used with distortion and other effects, both as a rhythm instrument using repetitive riffs with a varying degree of complexity, and as a solo lead instrument.
A_2: The electric guitar
Q_2: (Hard) What instrument is usually at the center of a hard rock sound?

Motivation:

- SQuAD questions have different difficulty levels. Q_1 is easy, Q_2 is hard.
- Can we control the difficulty of generated questions?

Task Definition:

- Given a sentence, a text fragment (answer) in the sentence, and a **difficulty level**
- To generate a question that is asked about the fragment and satisfy the difficulty level

Applications:

- Balance the number of hard questions and easy questions for knowledge testing
- Test how a QA system works for questions with diverse difficulty levels
- Improve performance of QA systems

Challenges

- No existing QA dataset has difficulty labels for questions
- For a single sentence and answer pair, we want to generate questions with diverse difficulty levels, but SQuAD only has one given question for each sentence and answer pair
- No metric to evaluate the difficulty of questions

Data Preparation

Question Difficulty is a subjective notion and can be addressed in many ways:

- Some stories are inherently difficult to understand
- Questions can be difficult in different ways, such as syntax complexity, coreference resolution and elaboration

Our Method for Data Preparation:

- Focus on generate SQuAD-like questions with diverse difficulty levels
- Two difficulty levels: Easy and Hard
- Develop an automatic labelling protocol
- Study the correlation between automatically labelled difficulty with human difficulty

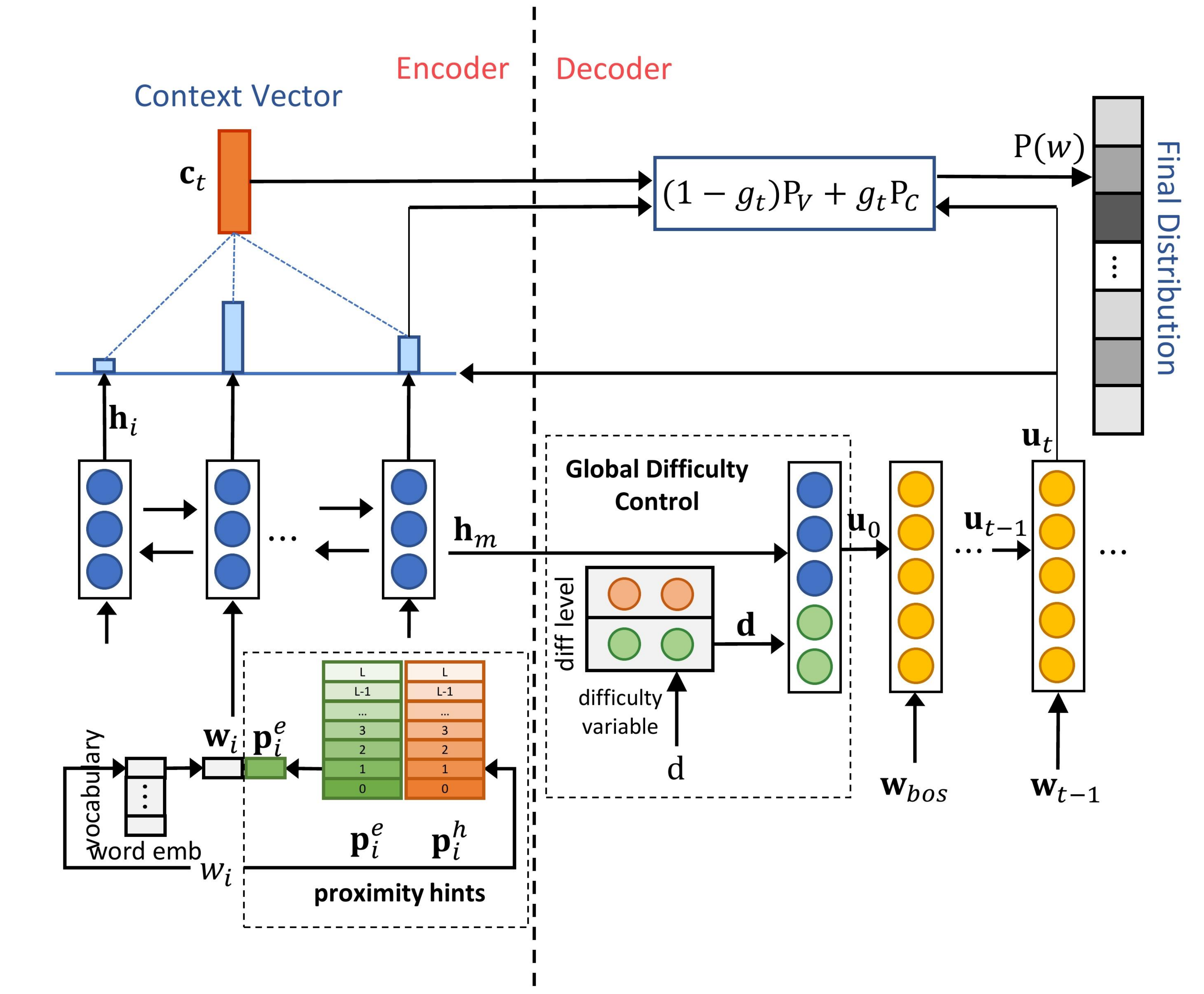
Automatic labelling protocol:

- Employ two reading comprehension systems, R-Net and BiDAF
- A question would be:
  - labelled with 'Easy' if both R-Net and BiDAF answer it correctly
  - labelled with 'Hard' if both systems fail to answer it
- The remaining questions are eliminated for suppressing the ambiguity
- 44723 easy questions, 31332 hard questions

Human Rating on 100 Easy & 100 Hard Questions:

- 1-3 scale, 3 for the most difficult
- Easy: 1.90 vs. Hard: 2.52

Model



Exploring Proximity Hints:

- If a question has more hints that can help locate the answer fragment, it would be easier to answer
- The average distance of those nonstop question words that also appear in the input sentence to the answer fragment

	Easy	Hard	All
Avg. distance of question words	7.67	9.71	8.43
Avg. distance of all sentence words	11.23	11.16	11.20

- Question Word Proximity Hints
  - The distance of nonstop question words are much smaller than the sentence words
  - Learn a lookup table to map the distance into a position embedding: $(\mathbf{p}_0, \mathbf{p}_1, \mathbf{p}_2, \dots \mathbf{p}_L)$
- Difficulty Level Proximity Hints
  - The distance for hard questions is significantly larger than that for easy questions
  - Explore the information of question difficulty levels
  - Easy: $(\mathbf{p}_0^e, \mathbf{p}_1^e, \mathbf{p}_2^e, \dots \mathbf{p}_L^e)$, Hard: $(\mathbf{p}_0^h, \mathbf{p}_1^h, \mathbf{p}_2^h, \dots \mathbf{p}_L^h)$

Characteristic-rich Encoder:

- Concatenate word emb and position emb: $\mathbf{x} = [\mathbf{w}; \mathbf{p}]$
- Bidirectional LSTMs encode the sequence

Global Difficulty Control:

- Use style variable to initialize the decoder state: $\mathbf{u}_0 = [\mathbf{h}_m; \mathbf{d}]$

Decoder with Attention & Copy

Experiment Results

Automatic Evaluation:

- Employ reading comprehension systems to evaluate the difficulty of generated questions
- N-gram based similarity: BLEU(B), ROUGE-L(R-L), METEOR(MET)

Difficulty of the Generated Questions:

	Easy Questions Set				**Hard** Questions Set			
	R-Net		BiDAF		R-Net		BiDAF	
	EM	F1	EM	F1	EM	F1	EM	F1
Ans	82.16	87.22	75.43	83.17	34.15	60.07	29.36	55.89
QWPH	82.66	87.37	76.10	83.90	33.35	59.50	28.40	55.21
QWPH-GDC	84.35	88.86	77.23	84.78	31.60	57.88	26.68	54.31
DLPH	85.49	89.50	78.35	85.34	28.05	54.21	24.89	51.25
DLPH-GDC	**85.82**	**89.69**	**79.09**	**85.72**	**26.71**	**53.40**	**24.47**	**51.20**

Question Quality:

	B1	B2	B3	B4	MET	R-L
L2A	36.01	21.61	14.97	10.88	15.99	38.06
Ans	43.51	29.06	21.35	16.22	20.53	45.66
QWPH	43.75	29.28	21.61	16.46	20.70	46.02
QWPH-GDC	43.99	29.60	21.86	16.63	20.87	46.26
DLPH	44.11	29.64	21.89	16.68	20.94	46.22
DLPH-GDC	43.85	29.48	21.77	16.56	20.79	46.16

Controlling Difficulty:

	Easy Questions Set				**Hard** Questions Set			
	R-Net		BiDAF		R-Net		BiDAF	
	EM	F1	EM	F1	EM	F1	EM	F1
QWPH-GDC	7.41	5.72	7.13	5.88	6.45	5.47	6.13	5.10
DLPH	12.41	9.51	11.28	8.49	12.01	10.45	10.51	9.37
DLPH-GDC	**12.91**	**9.95**	**12.40**	**9.23**	**12.68**	**10.76**	**11.22**	**9.97**

Human Evaluation:

- Fluency (F) {1,2,3}: grammatical correctness and fluency
- Difficulty (D) {1,2,3}: difficulty of generated questions
- Relevance (R) {0,1}: if the question is ask about the answer

	Easy Question Set			**Hard** Question Set		
	F	D	R	F	D	R
Ans	2.91	2.02	0.74	2.87	2.12	0.58
DLPH-GDC	2.94	1.84	0.76	2.87	2.26	0.64