香港中文大學
The Chinese University of Hong Kong

Difficulty Controllable Generation of Reading Comprehension Questions

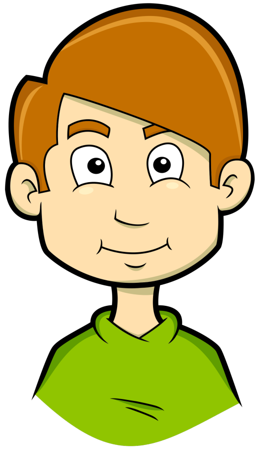
Yifan Gao[1], Lidong Bing[2], Wang Chen[1], Michael R. Lyu[1], Irwin King[1]

[1]The Chinese University of Hong Kong [2]Alibaba DAMO Academy

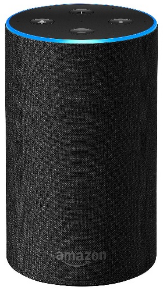
IJCAI 2019, Macao

Question Generation: Background

Tell me how to bake a cake



Sure! Which cake do you want to bake?



Alexa

What is the atomic number of the element oxygen?

Information Gathering

Knowledge Testing

Question Generation: Related Work

- Dialogue
  - Seeking Information in Task-oriented Chatbot
  - Asking Clarification Questions (Rao and Daume, 2018)
  - Interactiveness and Persistance (Wang et al, 2018)

Question Generation: Related Work

- Dialogue
    - Seeking Information in Task-oriented Chatbot
    - Asking Clarification Questions (Rao and Daume, 2018)
    - Interactiveness and Persistance (Wang et al, 2018)
- Question Answering
    - Reading Comprehension Question Generation (Du et al., 2017)
    - Harvesting Question Answer Pairs (Du et al., 2018)
    - Improving QA model (Yang et at., 2017)

Question Generation: Related Work

- Dialogue
  - Seeking Information in Task-oriented Chatbot
  - Asking Clarification Questions (Rao and Daume, 2018)
  - Interactiveness and Persistance (Wang et al, 2018)
- Question Answering
  - Reading Comprehension Question Generation (Du et al., 2017)
  - Harvesting Question Answer Pairs (Du et al., 2018)
  - Improving QA model (Yang et at., 2017)

Our Focus

Question Generation: Previous Setting

$\mathbf{S_1}$: Oxygen is a chemical element with symbol O and atomic number 8.

$\mathbf{S_2}$: 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.

Question Generation: Previous Setting

$\mathbf{S_1}$: Oxygen is a chemical element with symbol O and atomic number 8.
$\mathbf{Q_1}$: What is the atomic number of the element oxygen?

$\mathbf{S_2}$: 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.
$\mathbf{Q_2}$: What instrument is usually at the center of a hard rock sound?

Question Generation: Motivation

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

S_1 : Oxygen is a chemical element with symbol O and atomic number 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.
Q_2: (Hard) What instrument is usually at the center of a hard rock sound?

Difficulty Controllable Question Generation

- A New Task:
  - 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

Difficulty Controllable Question Generation

- A New Task:
  - 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
- Application Scenarios
  - 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

Difficulty Controllable Question Generation

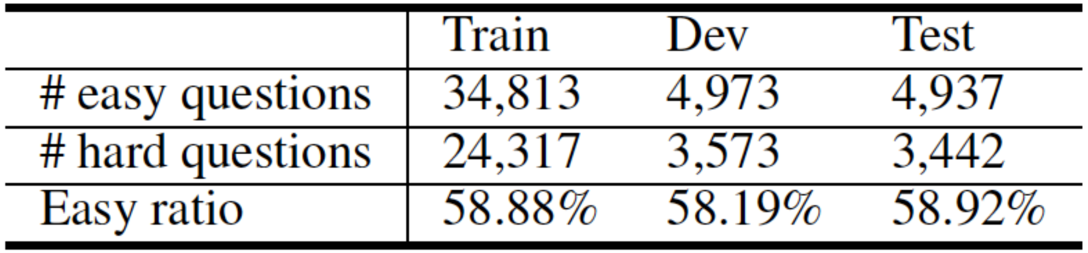
- A New Task:
  - Given a sentence, a text fragment (answer) in the sentence, and a **difficulty level**
  - To generate a question that is asked about the fragment and satisfy the difficulty level
- Challenges:
  - No existing QA dataset has difficulty labels for questions
  - For a single sentence 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
  - Syntax complexity, coreference resolution and elaboration (Sugawara et al., 2017)
- Our Protocol
  - Focus on generate SQuAD-like questions with diverse difficulty levels
    - Difficulty of RACE (Lai et al., 2017) questions mostly come from the understanding of the story but not from the way how the question is asked
  - Two difficulty levels: Easy and Hard
  - Develop an automatic labelling protocol
  - Study the correlation between automatically labelled difficulty with human difficulty

Data Preparation

- Automatic labelling protocol
  - Employ two reading comprehension systems, namely R-Net (Wang et al., 2017) and BiDAF (Seo et al., 2017)
  - 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



	Train	Dev	Test
# easy questions	34,813	4,973	4,937
# hard questions	24,317	3,573	3,442
Easy ratio	58.88%	58.19%	58.92%

Human Rating on 100 Easy & Hard Questions:

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

Exploring a few intuitions···

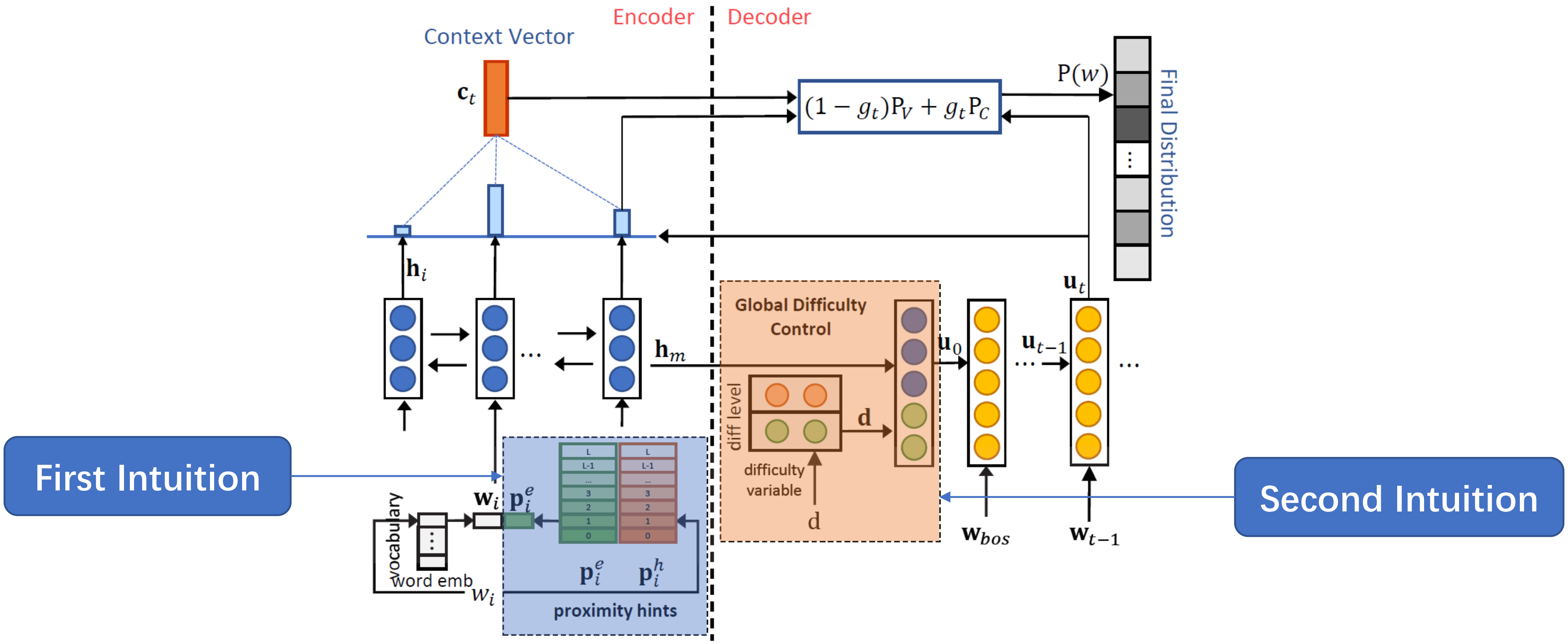
- If a question has more hints that can help locate the answer fragment, it would be easier to answer

S_1 : Oxygen is a chemical element with symbol O and **atomic number** 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.
Q_2: (Hard) What **instrument** is usually at the center of a hard rock sound?

- Performing difficulty control can be regarded as a problem of sentence generation towards a specified attribute or style

Proposed Framework



First Intuition

Second Intuition

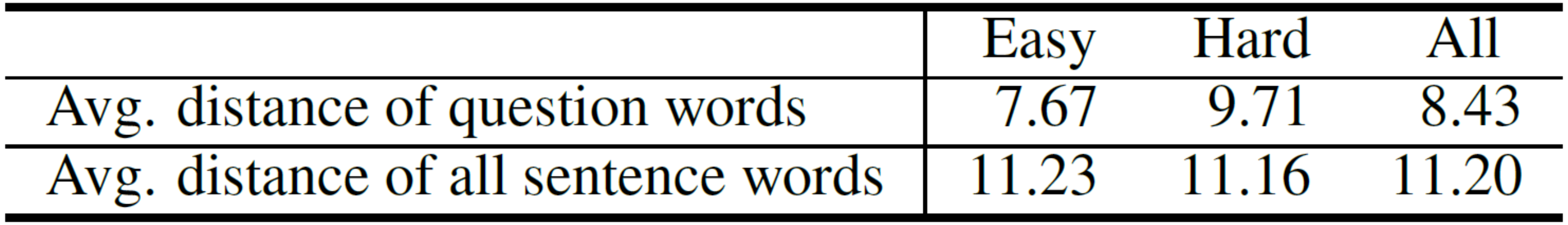
Exploring Proximity Hints

- We examine the average distance of those nonstop question words that also appear in the input sentence to the answer fragment

Question: What is the **atomic number** of the **element oxygen**?

Sentence: **Oxygen** is a chemical **element** with symbol O and **atomic number** 8.

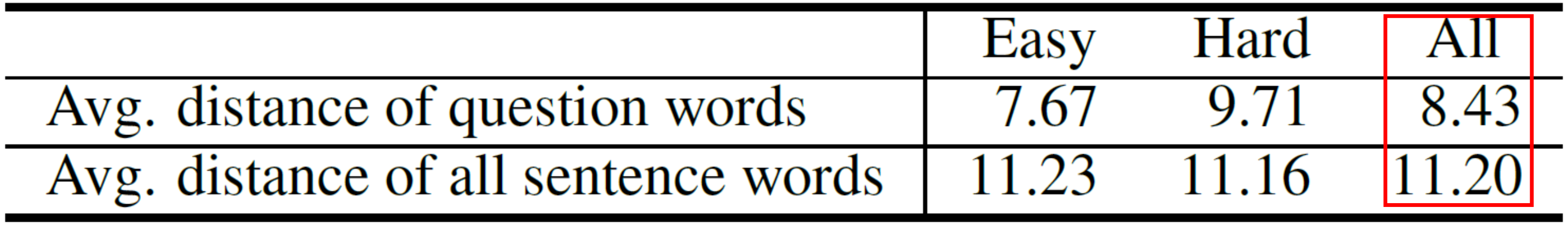
Distance: 11 7 2 1



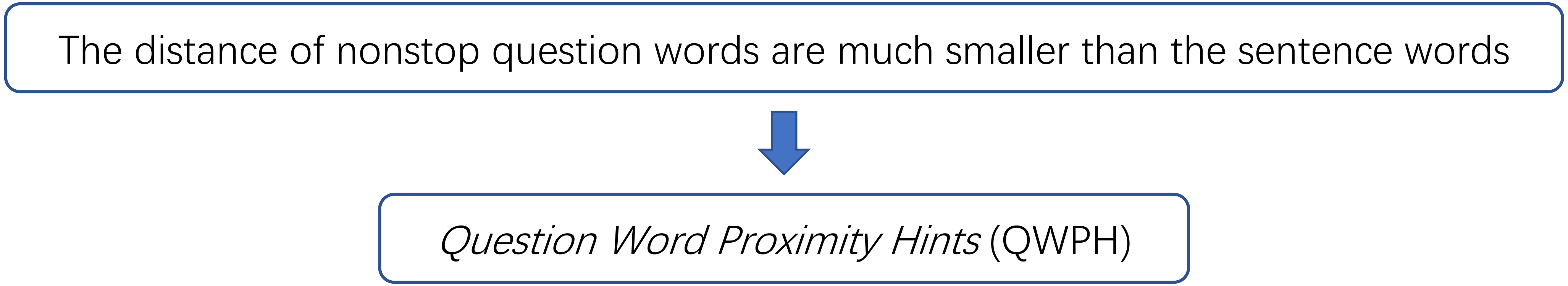
	Easy	Hard	All
Avg. distance of question words	7.67	9.71	8.43
Avg. distance of all sentence words	11.23	11.16	11.20

Exploring Proximity Hints

- We examine 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

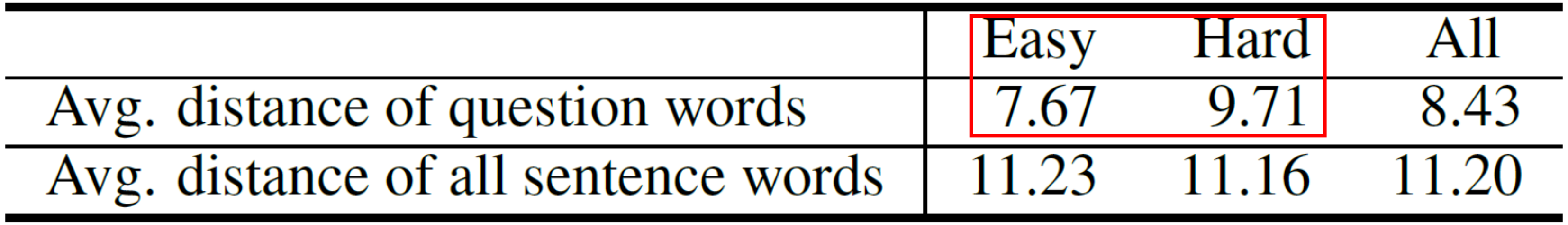


The distance of nonstop question words are much smaller than the sentence words

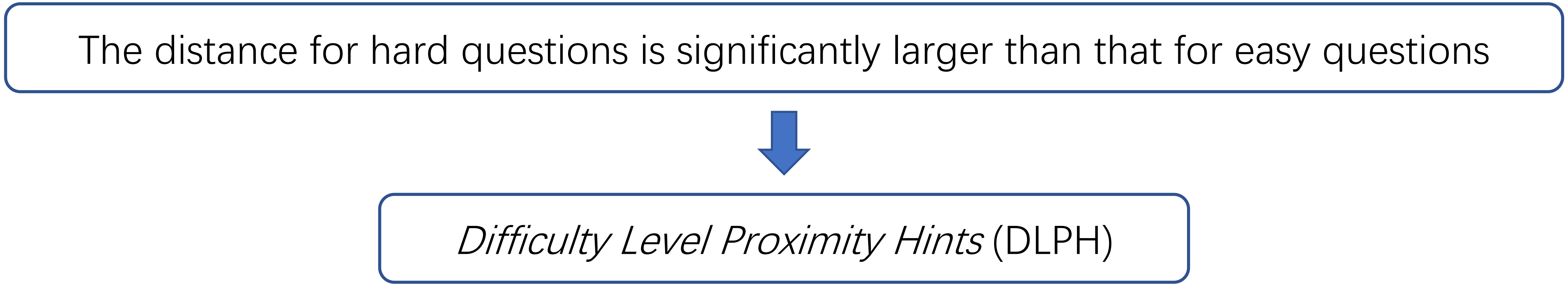
Question Word Proximity Hints (QWPH)

Exploring Proximity Hints

- We examine 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

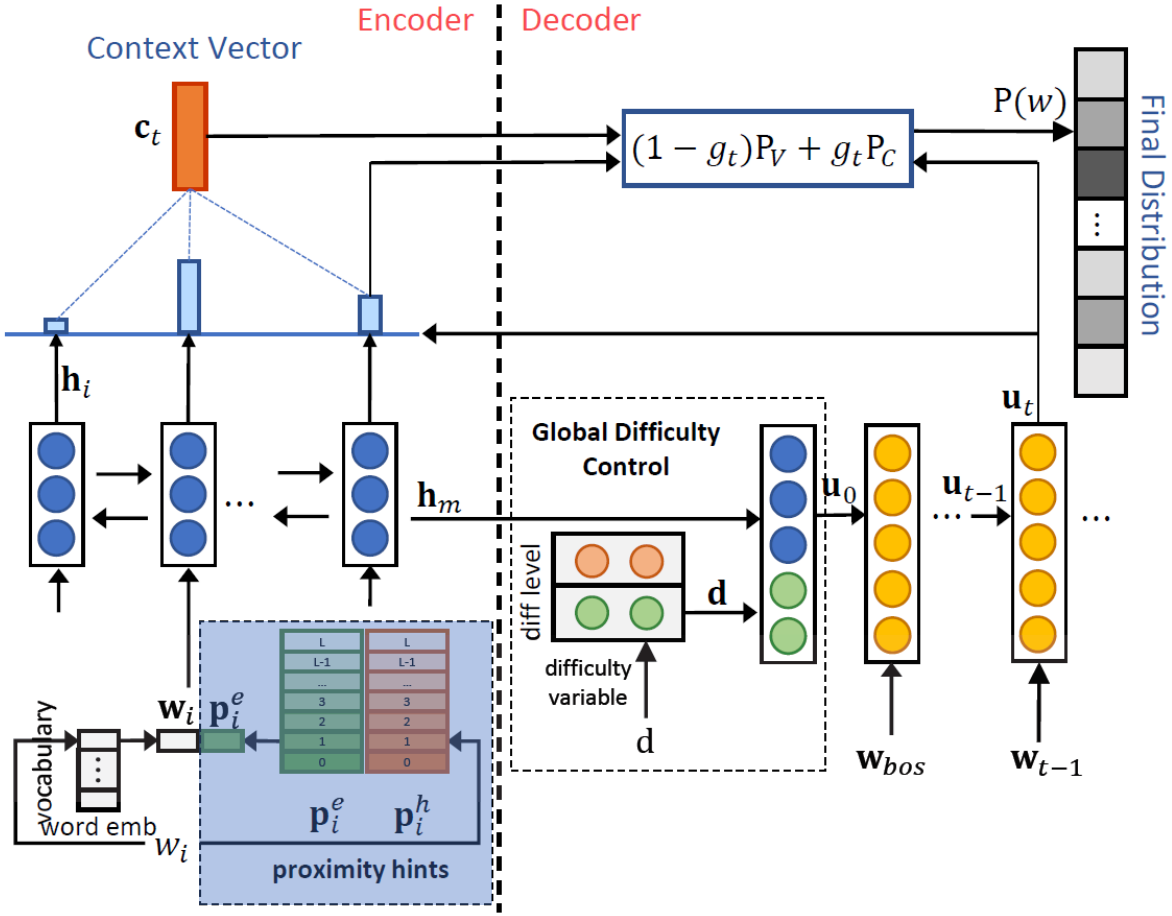


The distance for hard questions is significantly larger than that for easy questions

Difficulty Level Proximity Hints (DLPH)

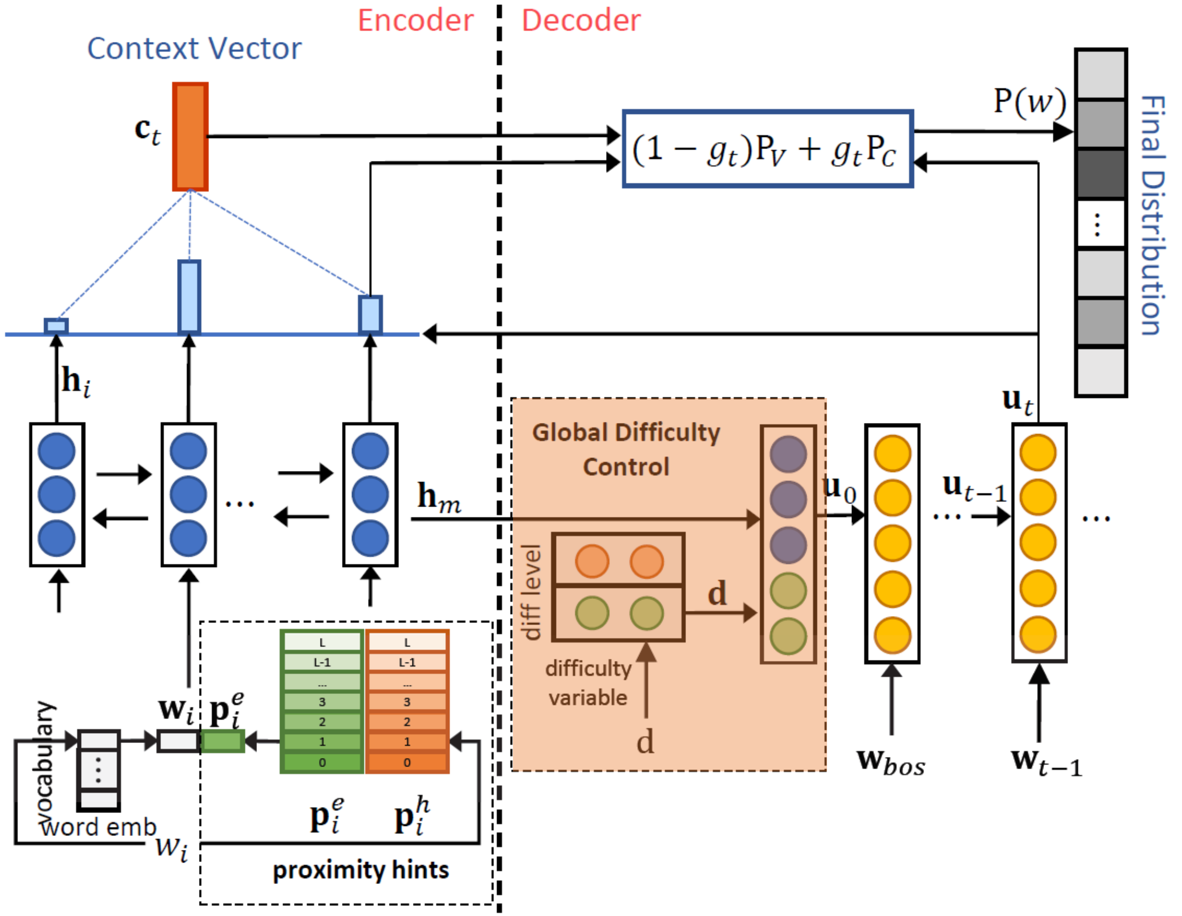
Exploring Proximity Hints

- Question Word Proximity Hints
  - 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
  - Additionally 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: $(\mathbf{p}_0^h, \mathbf{p}_1^h, \mathbf{p}_2^h, \dots \mathbf{p}_L^h)$



Proposed Framework

- Characteristic-rich Encoder
  - Concatenate word embedding and position embedding (proximity hint)
    $$\mathbf{x} = [\mathbf{w}; \mathbf{p}]$$
  - Bidirectional LSTMs encode the sequence
- Difficulty-controllable Decoder
  - **Global Difficulty Control**: use style variable to initialize decoder state
    $$\mathbf{u}_0 = [\mathbf{h}_m; \mathbf{d}]$$
  - Decoder with Attention & Copy



Evaluation Metrics

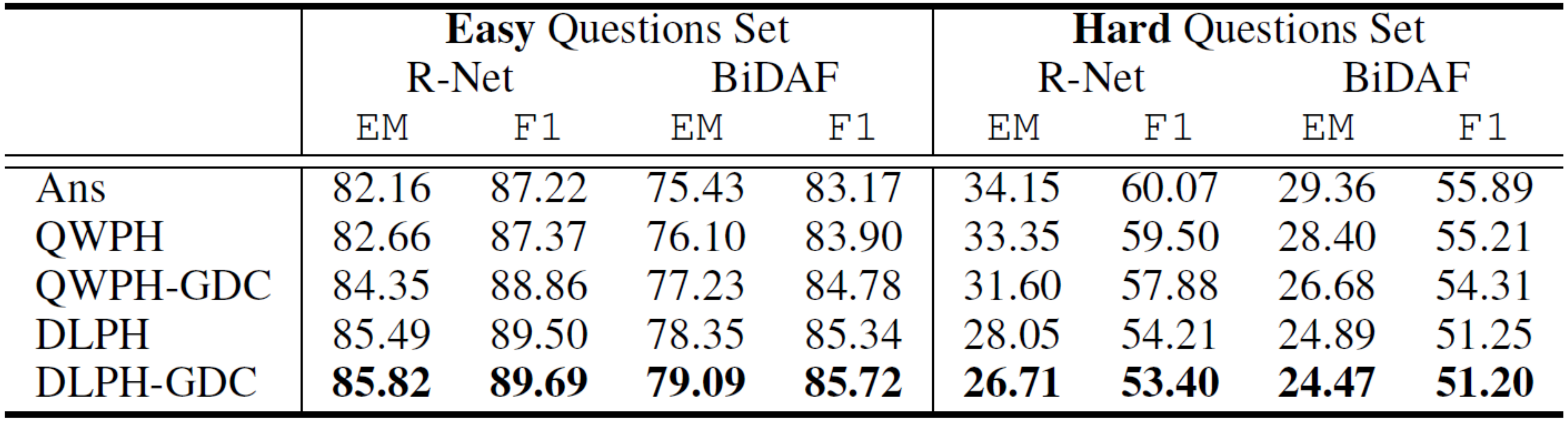
- Automatic Evaluation
  - Employ reading comprehension systems to evaluate the difficulty of generated questions
  - N-gram based similarity: BLEU, ROUGE, METEOR
- Human Evaluation
  - Fluency, Difficulty, Relevance

Baselines and Ablations

- **L2A**: Sequence-to-sequence (seq2seq) model with attention mechanism
- **Ans**: Add answer indicator embeddings to the seq2seq model
- **QWPH**: Our model with Question Word Proximity Hints
- **DLPH**: Our model with Difficulty Level Proximity Hints
- **QWPH-GDC**: Our model with QWPH and Global Difficulty Control
- **DLPH-GDC**: Our model with DLPH and Global Difficulty Control

Difficulty Control Results

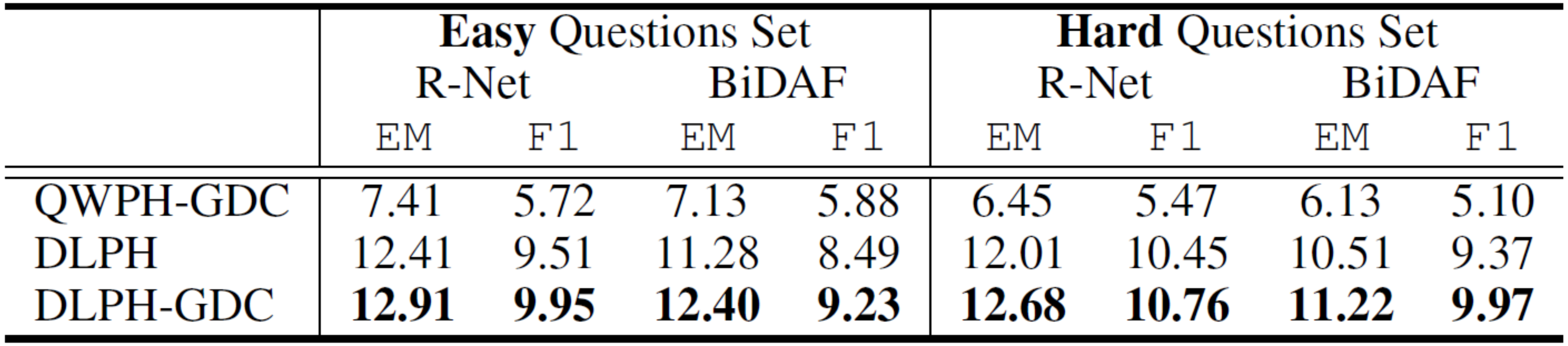
- **Difficulty of the generated questions**. For easy questions, higher score indicates better difficulty-control, while for hard questions, lower indicates better.



	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**

Difficulty Control Results

- **The results of controlling difficulty**. The scores are performance gap between questions generated with original difficulty label and questions generated with reverse difficulty label.



	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**

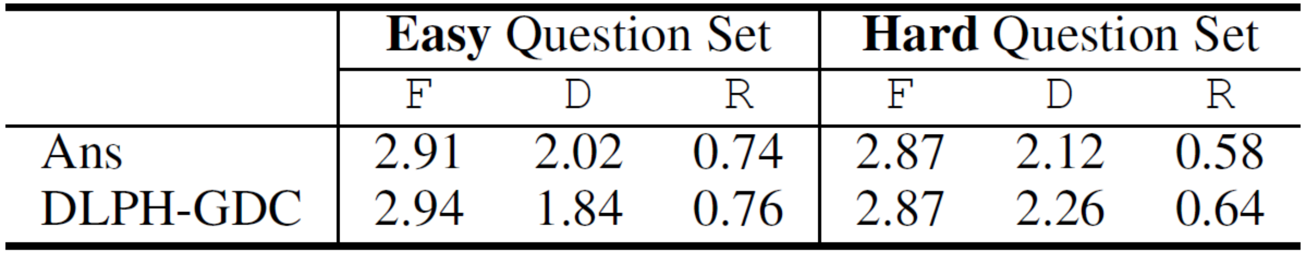
(1) DLPH-GDC has the strongest capability of generating difficulty-controllable questions.

(2) The local difficulty control (i.e. DLPH) is more effective than the global.

Human Evaluation

- 3 annotators rate the same 100 easy questions and 100 hard questions
- During annotation, difficulty labels are not shown
- Metrics
  - Fluency (F): grammatical correctness and fluency, 1-3 scale, 3 for best
  - Difficulty (D): difficulty of generated questions, 1-3 scale, 3 for best
  - Relevance (R): if the question is ask about the answer, 0-1 scale, 1 for best

Human Evaluation

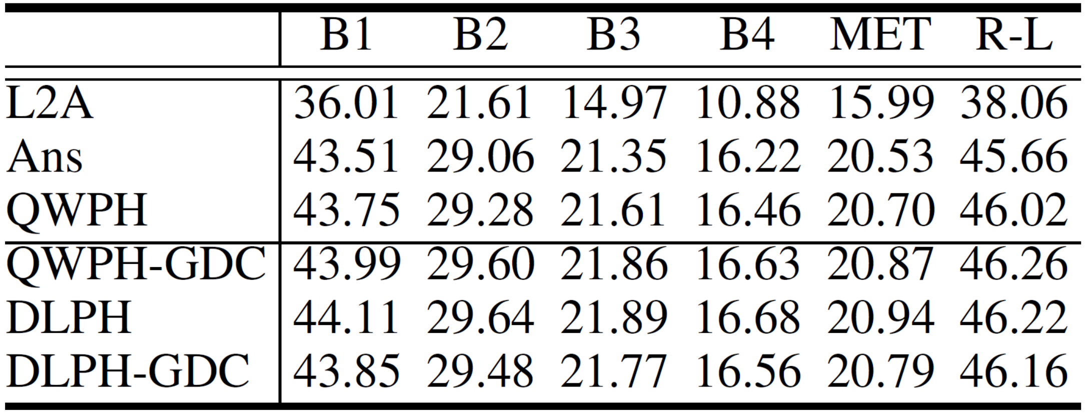


	Easy Question Set			**Hard** Question Set		
	F	D	R	F	D	R
Ans	2.91	2.02	0.74	2.87	2.12	0.58
DLPH-GDC	2.94	1.84	0.76	2.87	2.26	0.64

- **Fluency**: Both models achieve high score on fluency, owing to the strong language modelling capability of neural models
- **Difficulty**:
  - For human beings, all SQuAD-like questions are not really difficult, therefore, the difference of difficulty values is not large
  - DLPH-GDC can generate easier or harder questions than Ans
- **Relevance**: DLPH-GDC with position embedding can generate more relevant questions than answer embedding only

Automatic Evaluation of Question Quality

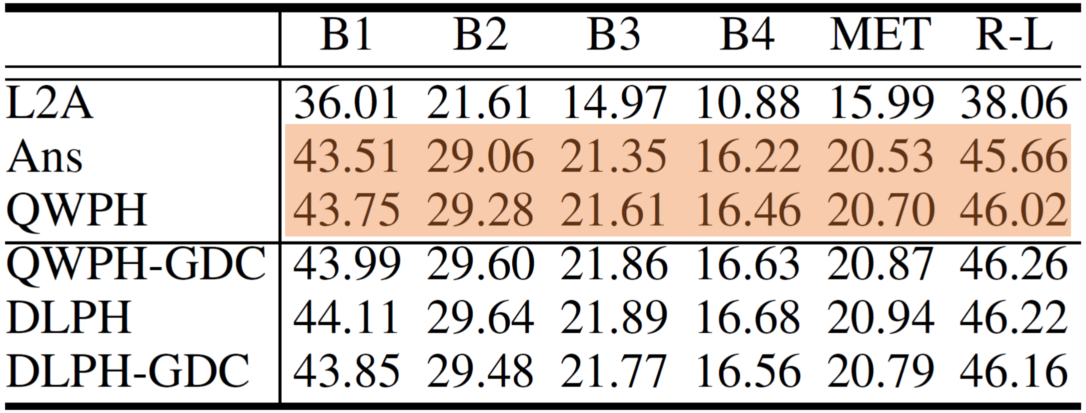
- We evaluate the similarity of generated questions with the ground truth questions by feeding the ground truth difficulty labels
- Metrics: BLEU (B), METEOR (MET), ROUGE-L (R-L)



	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

Automatic Evaluation of Question Quality

- We evaluate the similarity of generated questions with the ground truth questions by feeding the ground truth difficulty labels
- Metrics: BLEU (B), METEOR (MET), ROUGE-L (R-L)

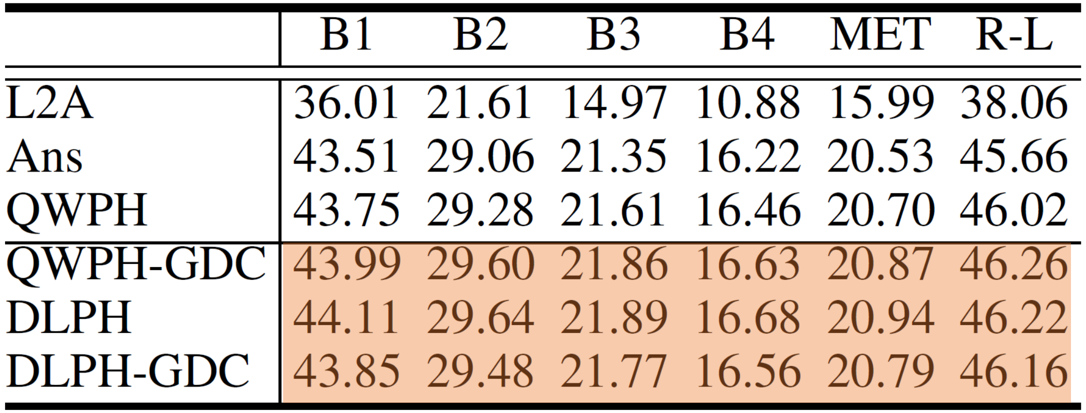


	B1	B2	B3	B4	MET	R-L
L2A	36.01	21.61	14.97	10.88	15.99	38.06
Ans	43.51	29.06	21.35	16.22	20.53	45.66
QWPH	43.75	29.28	21.61	16.46	20.70	46.02
QWPH-GDC	43.99	29.60	21.86	16.63	20.87	46.26
DLPH	44.11	29.64	21.89	16.68	20.94	46.22
DLPH-GDC	43.85	29.48	21.77	16.56	20.79	46.16

Further distinguish the different distance help generate better questions

Automatic Evaluation of Question Quality

- We evaluate the similarity of generated questions with the ground truth questions by feeding the ground truth difficulty labels
- Metrics: BLEU (B), METEOR (MET), ROUGE-L (R-L)

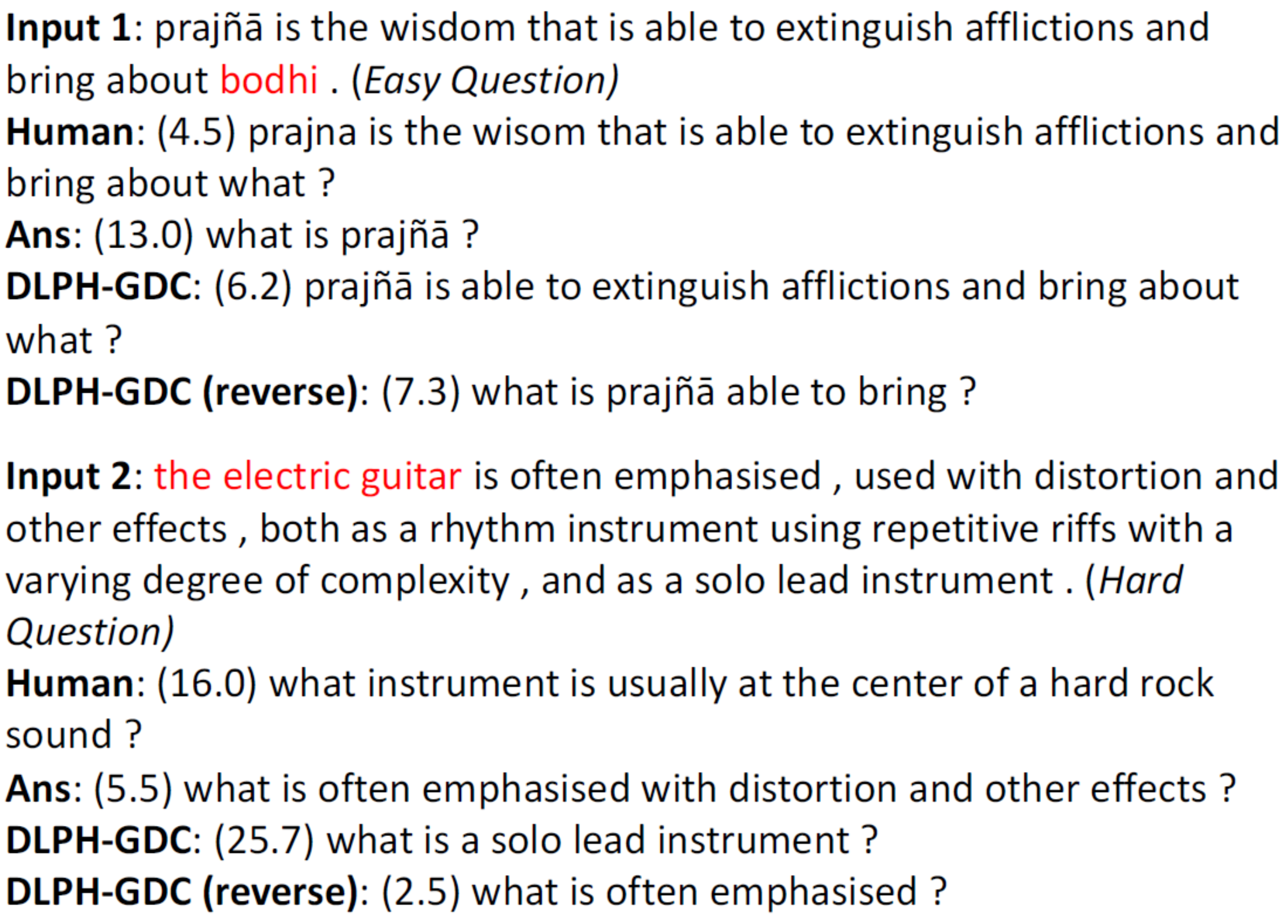


	B1	B2	B3	B4	MET	R-L
L2A	36.01	21.61	14.97	10.88	15.99	38.06
Ans	43.51	29.06	21.35	16.22	20.53	45.66
QWPH	43.75	29.28	21.61	16.46	20.70	46.02
QWPH-GDC	43.99	29.60	21.86	16.63	20.87	46.26
DLPH	44.11	29.64	21.89	16.68	20.94	46.22
DLPH-GDC	43.85	29.48	21.77	16.56	20.79	46.16

Given ground truth difficulty labels, methods with difficulty control perform better

Case Study

- Our model
  - Give more hints (shorter distance) when asking easier questions
  - Give less hints (longer distance) when asking harder questions



Input 1: prajñā is the wisdom that is able to extinguish afflictions and bring about bodhi . (*Easy Question)*
Human: (4.5) prajna is the wisom that is able to extinguish afflictions and bring about what ?
Ans: (13.0) what is prajñā ?
DLPH-GDC: (6.2) prajñā is able to extinguish afflictions and bring about what ?
DLPH-GDC (reverse): (7.3) what is prajñā able to bring ?

Input 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 . (*Hard Question)*
Human: (16.0) what instrument is usually at the center of a hard rock sound ?
Ans: (5.5) what is often emphasised with distortion and other effects ?
DLPH-GDC: (25.7) what is a solo lead instrument ?
DLPH-GDC (reverse): (2.5) what is often emphasised ?

Conclusion

- A new setting: Difficulty Controllable Question Generation
- Prepare a question generation dataset with difficulty labels
- Proximity Hints & Global Difficulty Control
- Evaluation methods for question difficulty

- Limitations and Future Work
  - Explore better definition of question difficulty
  - New evaluation methods for question difficulty

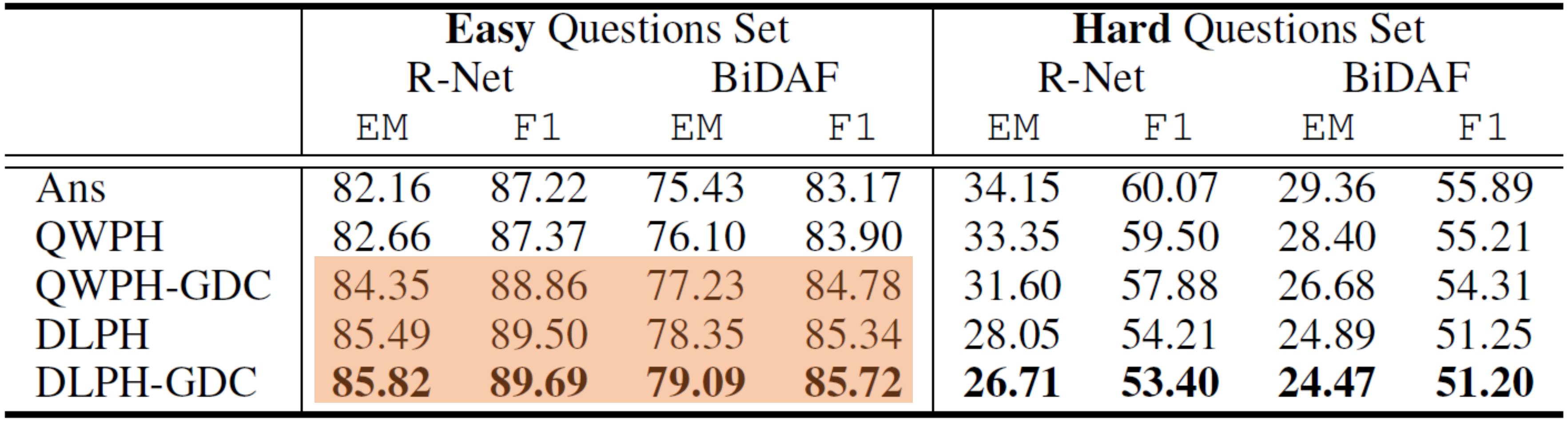
Reference

1. Sudha Rao and Hal Daumé III. Learning to Ask Good Questions: Ranking Clarification Questions using Neural Expected Value of Perfect Information. In ACL 2018.
2. Yansen Wang, Chenyi Liu, Minlie Huang, Liqiang Nie. Learning to Ask Questions in Open-domain Conversational Systems with Typed Decoders. In ACL 2018.
3. Xinya Du and Claire Cardie. Harvesting Paragraph-Level Question-Answer Pairs from Wikipedia. In ACL 2018.
4. Xinya Du, Junru Shao and Claire Cardie. Learning to Ask: Neural Question Generation for Reading Comprehension. In ACL 2017.
5. Zhilin Yang, Junjie Hu, Ruslan Salakhutdinov, William W. Cohen. Semi-Supervised QA with Generative Domain-Adaptive Nets. In ACL 2017
6. Saku Sugawara, Yusuke Kido, Hikaru Yokono, and Akiko Aizawa. Evaluation Metrics for Machine Reading Comprehension: Prerequisite Skills and Readability. In ACL 2017.
7. Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale Reading Comprehension Dataset from Examinations. In EMNLP 2017.
8. Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. Gated Self-matching Networks for Reading Comprehension and Question Answering. In ACL, 2017.
9. Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional Attention Flow for Machine Comprehension. ICLR, 2017.

Thanks

Difficulty Control Results

- **Difficulty of the generated questions**. For easy questions, higher score indicates better difficulty-control, while for hard questions, lower indicates better.

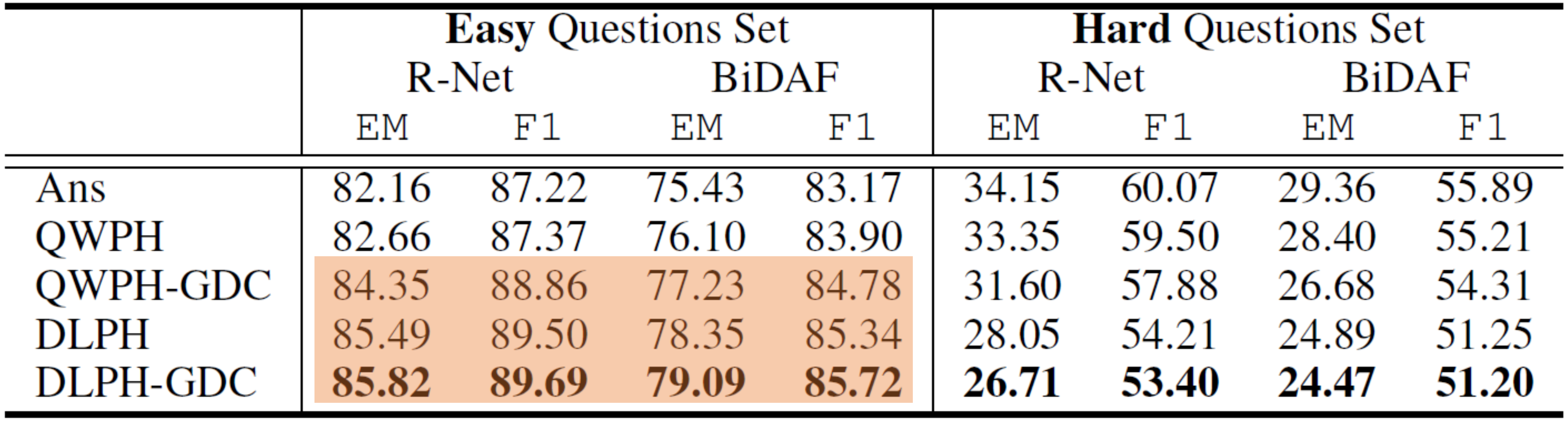


	Easy Questions Set				**Hard** Questions Set			
	R-Net		BiDAF		R-Net		BiDAF	
	EM	F1	EM	F1	EM	F1	EM	F1
Ans	82.16	87.22	75.43	83.17	34.15	60.07	29.36	55.89
QWPH	82.66	87.37	76.10	83.90	33.35	59.50	28.40	55.21
QWPH-GDC	84.35	88.86	77.23	84.78	31.60	57.88	26.68	54.31
DLPH	85.49	89.50	78.35	85.34	28.05	54.21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

Generate easier questions!

Difficulty Control Results

- **Difficulty of the generated questions**. For easy questions, higher score indicates better difficulty-control, while for hard questions, lower indicates better.

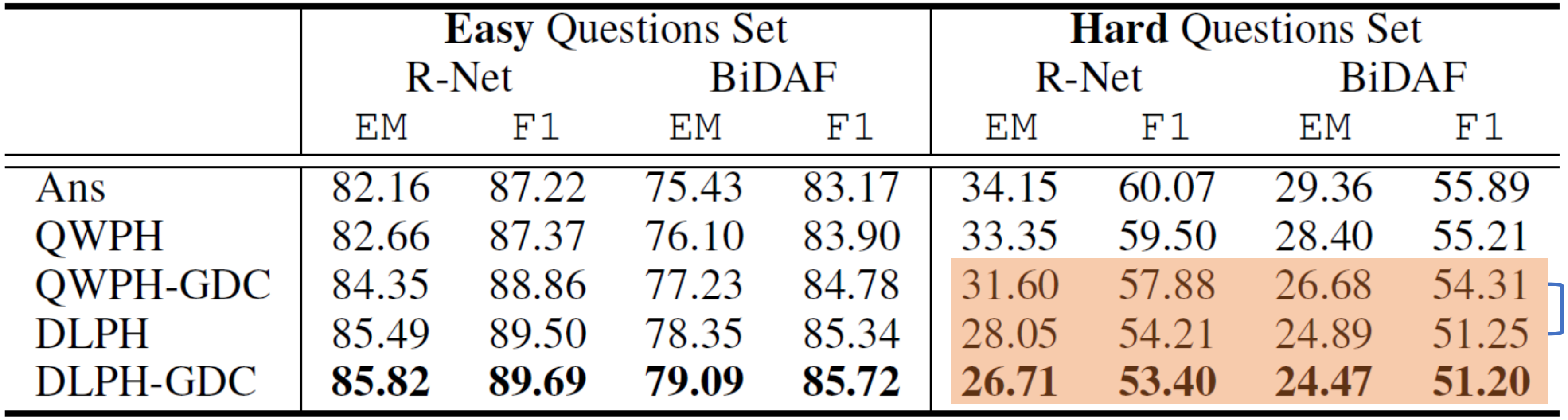


	Easy Questions Set				**Hard** Questions Set			
	R-Net		BiDAF		R-Net		BiDAF	
	EM	F1	EM	F1	EM	F1	EM	F1
Ans	82.16	87.22	75.43	83.17	34.15	60.07	29.36	55.89
QWPH	82.66	87.37	76.10	83.90	33.35	59.50	28.40	55.21
QWPH-GDC	84.35	88.86	77.23	84.78	31.60	57.88	26.68	54.31
DLPH	85.49	89.50	78.35	85.34	28.05	54.21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

- **Do our models simply produce trivial questions by having them contain the answer words?**
- **No! Only 0.09% answer words appear in generated questions.**

Difficulty Control Results

- **Difficulty of the generated questions**. For easy questions, higher score indicates better difficulty-control, while for hard questions, lower indicates better.



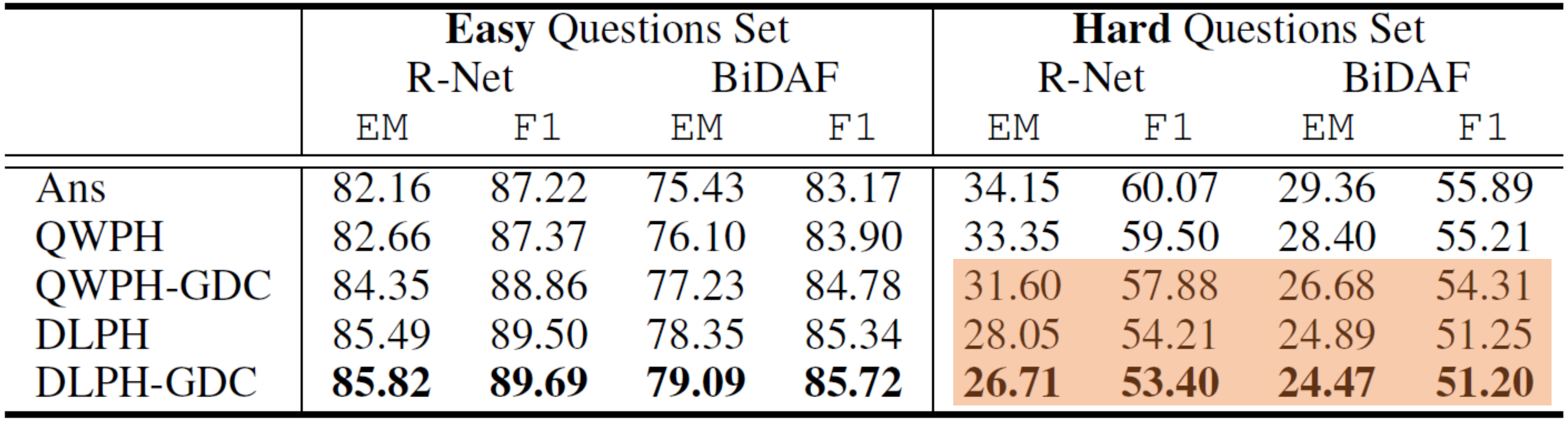
	Easy Questions Set				**Hard** Questions Set			
	R-Net		BiDAF		R-Net		BiDAF	
	EM	F1	EM	F1	EM	F1	EM	F1
Ans	82.16	87.22	75.43	83.17	34.15	60.07	29.36	55.89
QWPH	82.66	87.37	76.10	83.90	33.35	59.50	28.40	55.21
QWPH-GDC	84.35	88.86	77.23	84.78	31.60	57.88	26.68	54.31
DLPH	85.49	89.50	78.35	85.34	28.05	54.21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

Local control is more effective

For hard questions, questions irrelevant to the answer can also yield lower scores. We will discuss it in human evaluation

Recall that in Difficulty Control Results …

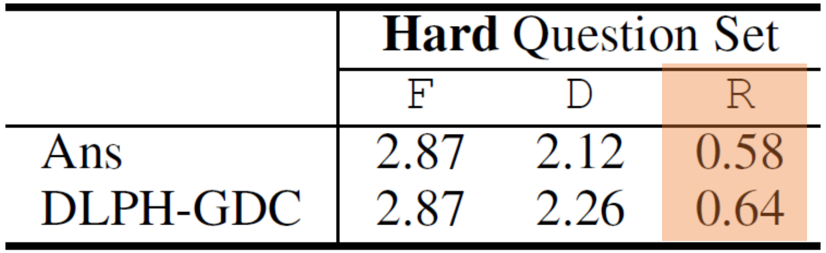
- **Difficulty of the generated questions**. For easy questions, higher score indicates better difficulty-control, while for hard questions, lower indicates better.



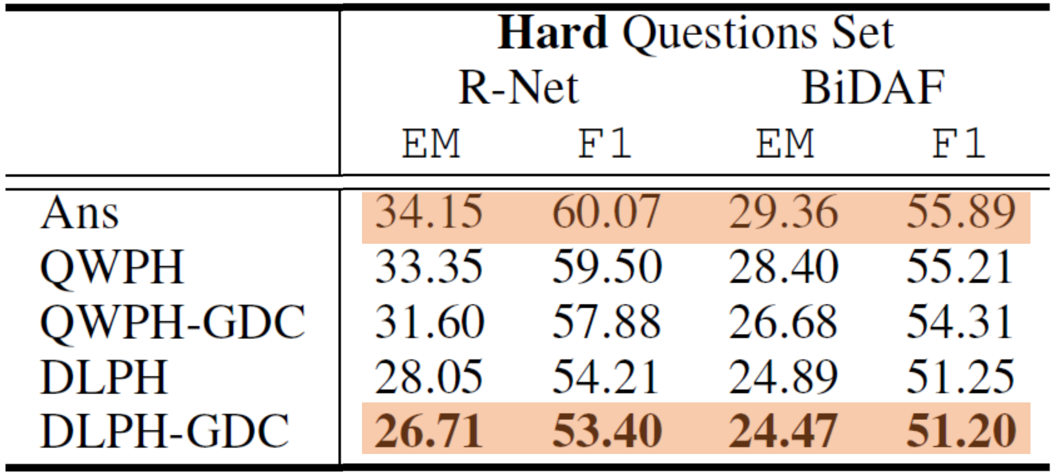
	Easy Questions Set				**Hard** Questions Set			
	R-Net		BiDAF		R-Net		BiDAF	
	EM	F1	EM	F1	EM	F1	EM	F1
Ans	82.16	87.22	75.43	83.17	34.15	60.07	29.36	55.89
QWPH	82.66	87.37	76.10	83.90	33.35	59.50	28.40	55.21
QWPH-GDC	84.35	88.86	77.23	84.78	31.60	57.88	26.68	54.31
DLPH	85.49	89.50	78.35	85.34	28.05	54.21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

For hard questions, questions irrelevant to the answer can also yield lower scores. We will discuss it now!

Human Evaluation



	Hard Question Set		
	F	D	R
Ans	2.87	2.12	0.58
DLPH-GDC	2.87	2.26	0.64

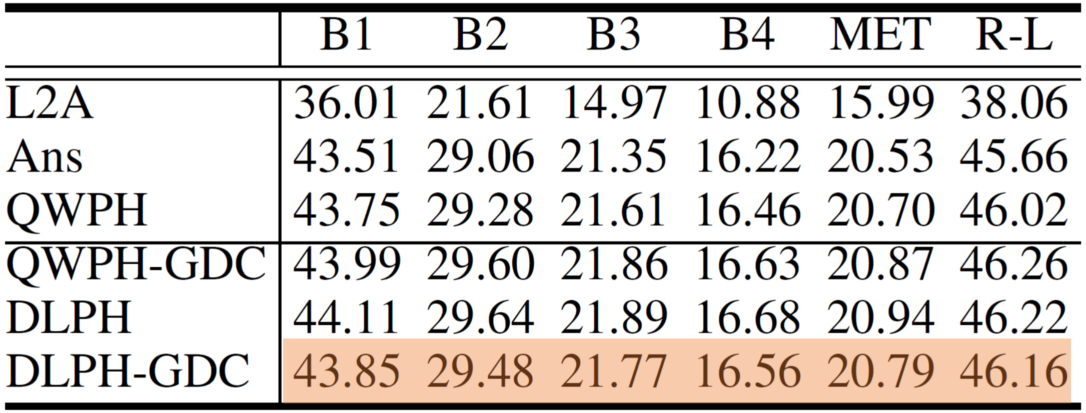


	Hard Questions Set			
	R-Net		BiDAF	
	EM	F1	EM	F1
Ans	34.15	60.07	29.36	55.89
QWPH	33.35	59.50	28.40	55.21
QWPH-GDC	31.60	57.88	26.68	54.31
DLPH	28.05	54.21	24.89	51.25
DLPH-GDC	**26.71**	**53.40**	**24.47**	**51.20**

By comparing the **Relevance** scores in human evaluation results and **EM/F1** scores in difficulty control results for Hard Question Set, we find that the questions generated by DLPH-GDC are *more relevant* and *more difficult* than those generated by the Ans baseline.

Automatic Evaluation of Question Quality

- We evaluate the similarity of generated questions with the ground truth questions by feeding the ground truth difficulty labels
- Metrics: BLEU (B), METEOR (MET), ROUGE-L (R-L)



	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

DLPH-GDC sacrifices a little in N-gram based performance here while achieving the best difficulty control capability