Teaching Machines to Ask and Answer Questions

Knowledge Assessment and Information Acquisition in Reading Comprehension

Yifan Gao

PhD Oral Defense

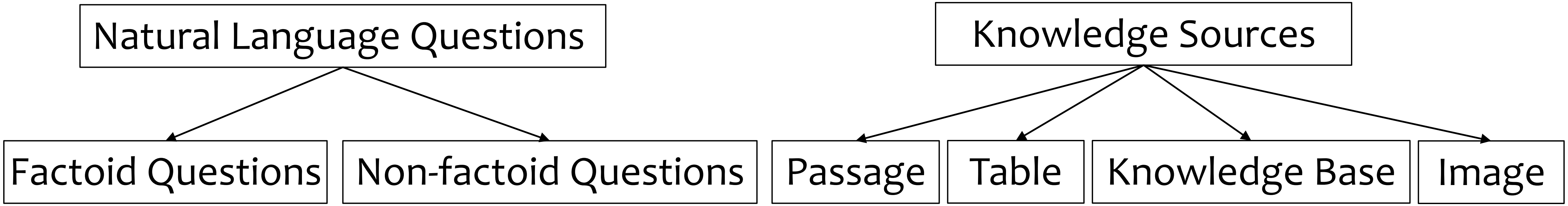
Supervisors: Prof. Irwin King, Prof. Michael R. Lyu

Committee: Prof. Yufei Tao, Prof. Jeffrey Xu Yu, Prof. Hsin-Hsi Chen

香港中文大學
The Chinese University of Hong Kong

Question Answering

Goal: answer natural language questions using certain knowledge sources



"Since question can be devised to query any aspect of text comprehension, the ability to answer questions is the strongest possible demonstration of understanding"

Wendy Lehnert, 1977

Question ~~Answering~~ Asking (Generation)

Goal: Generate natural language questions using certain knowledge sources

Why machines need to ask questions?

In our day-to-day communications, we are not always passively answering questions, especially when

- Knowledge assessment: We want to assess someone's understanding towards a passage
- Information acquisition: We need to gather enough information before answering it

Moving Beyond Merely Answering Questions

In our day-to-day communications, we are not always passively answering questions, especially when

we want to assess someone's understanding towards a passage

Passage: Incumbent Democratic President Bill Clinton was ineligible to serve a third term due to term limitations in the 22nd Amendment of the Constitution, and Vice President Gore was able to secure the Democratic nomination with relative ease.

⇨

Asking Questions

Q1: What political party is Clinton a member of?

A1: Democratic

Q2: What was he ineligible to serve?

A2: third term

Q3: Why?

A3: term limitations

Moving Beyond Merely Answering Questions

In our day-to-day communications, we are not always passively answering questions, especially when

we need to gather enough information before answering it

Knowledge: 7(a) loans provides business loans to American small businesses. The loan program is designed to assist for-profit businesses that are not able to get other financing from other resources.

Question: I am a 34-year-old man from the United States. I am the owner of an American small business. Is the 7(a) Loan Program for me?

⇨

1. Understand the text
2. Ask questions to gather more personal information
3. Assess the eligibility
4. Answer

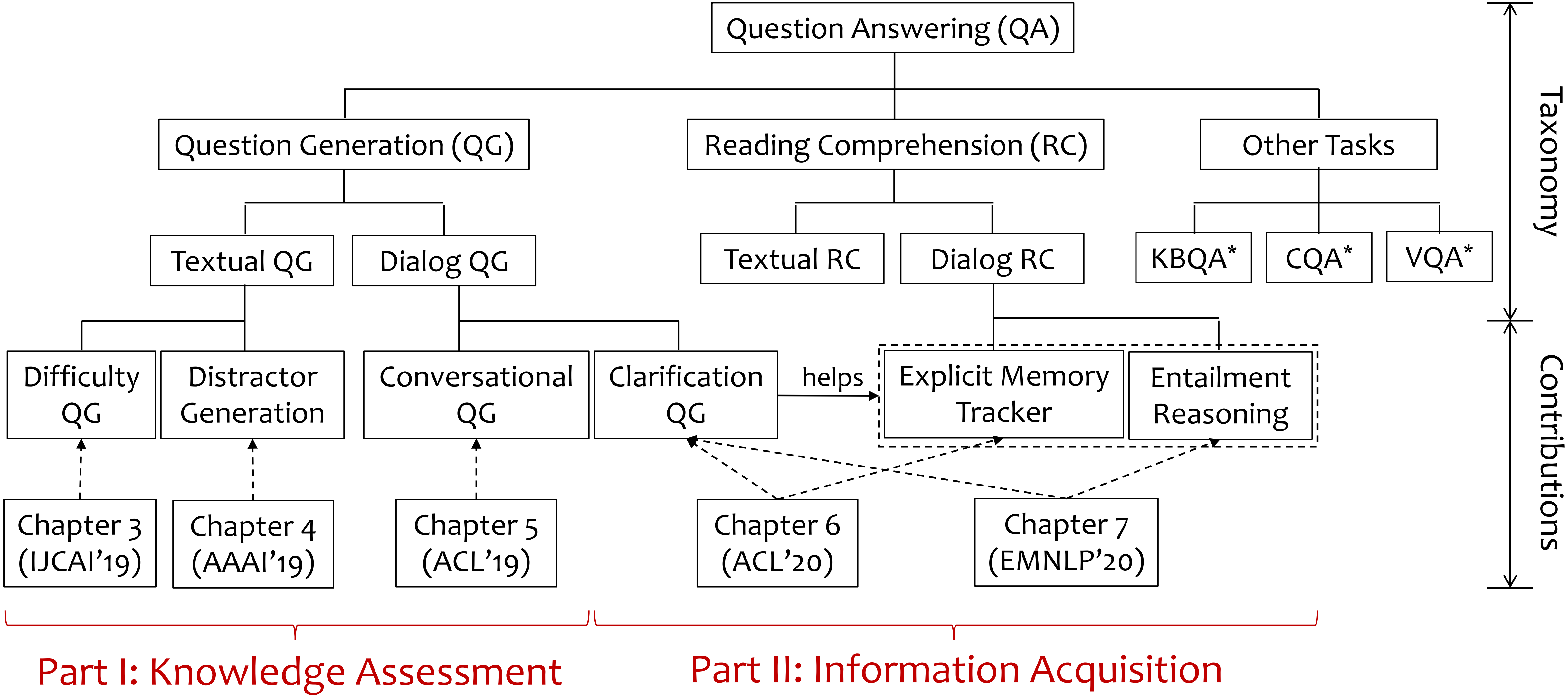
Both Question Asking and Answering are Important!

Ask Questions: Explore Unknown

Answer Questions: Fill Information Gap

Research Question: Can machines be engaged in question asking and answering in our life?

Taxonomy & Contributions



* KBQA: Knowledge Base QA; CQA: Community QA; VQA: Visual QA

Part I: Knowledge Assessment

- Goal: We want to assess someone's understanding towards a passage through Question Generation

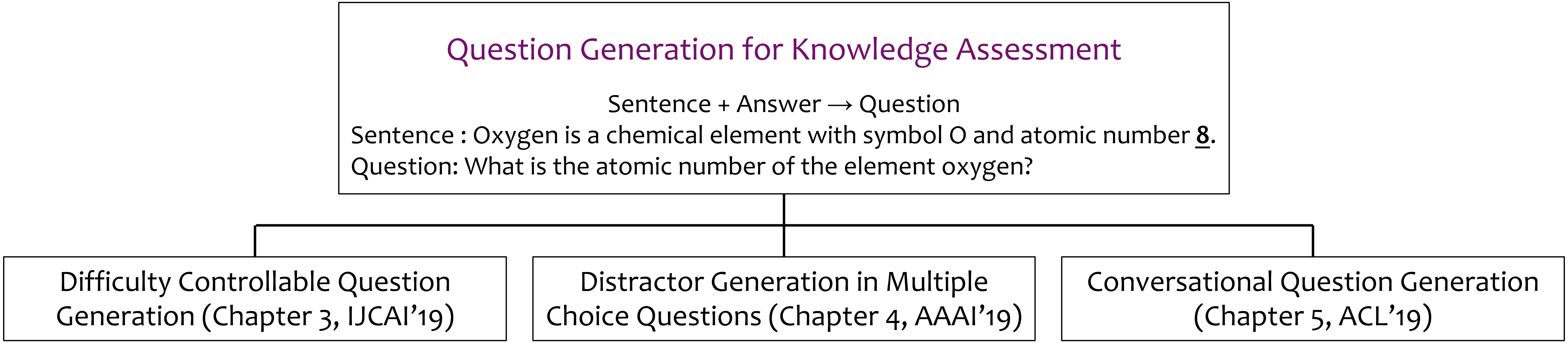
- Question Generation: Given the context and the answer, generate a natural language question

- Example
  - **Sentence:** Oxygen is a chemical element with symbol O and atomic number 8.
  - **Question:** What is the atomic number of the element oxygen?

answer

- Applications
  - Intelligence tutor systems
  - Syntactic QA pair generation

Part I: Knowledge Assessment



Question Generation for Knowledge Assessment

Sentence + Answer → Question

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

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

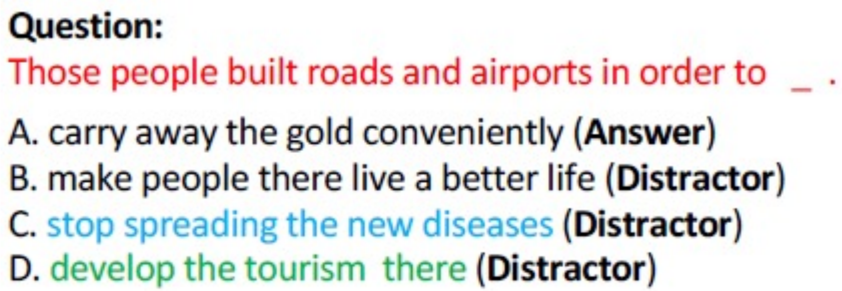
Difficulty Controllable Question Generation (Chapter 3, IJCAI'19)

Distractor Generation in Multiple Choice Questions (Chapter 4, AAAI'19)

Conversational Question Generation (Chapter 5, ACL'19)

Sentence + Answer + "Easy" → Easy Question

Sentence + Answer + "Hard" → Hard Question



Question:

Those people built roads and airports in order to _ .

A. carry away the gold conveniently (**Answer**)

B. make people there live a better life (**Distractor**)

C. stop spreading the new diseases (**Distractor**)

D. develop the tourism there (**Distractor**)

Q1: What political party is Clinton a member of?

A1: Democratic

Q2: What was he ineligible to serve?

A2: third term

Q3: Why?

A3: term limitations

Difficulty Controllable QG: Motivation

- Questions have different difficulty levels

- For humans, we want to balance the number of hard questions and easy questions in knowledge testing

- For machines, we want to test how a QA system works for questions with diverse difficulty levels

Can we control the difficulty of generated questions?

Difficulty Controllable QG: A New Task

- Input: a sentence, a text fragment (answer) in the sentence, and a **difficulty level**
- Output: a question that is asked about the fragment and satisfy the difficulty level

S_1 : Oxygen is a chemical element with symbol O and atomic number 8.
D_1: Easy
Q_1: What is the atomic number of the element oxygen?

answer

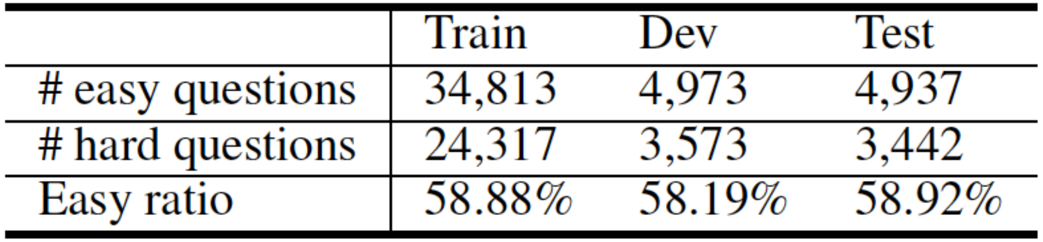
S_2 : The electric guitar is often emphasized, used with distortion and other effects, both as a rhythm instrument using repetitive riffs with a varying degree of complexity, and as a solo lead instrument.
D_2: Hard
Q_2: What instrument is usually at the center of a hard rock sound?

Difficulty Controllable QG: Data Preparation

- Challenges
  1. No existing QA dataset has difficulty labels for questions
  2. "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
  1. Simplify the problem
     - Two difficulty levels: Easy and Hard
     - Limit the difficulty of questions to how they are asked, not from the understanding of the passage of text
  2. Develop an automatic labelling protocol (next slide)
  3. Study the correlation between automatically labelled difficulty with human difficulty (next slide)

Difficulty Controllable QG: Data Preparation

- Automatic labelling protocol
  - Employ two reading comprehension systems, namely R-Net[1]and BiDAF[2]
  - 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



	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%

- Correlation with human difficulty
  - Human rating on 100 sampled Easy & Hard questions
  - 1-3 scale, 3 for the most difficult → Easy: 1.90; Hard: 2.52

[1] Wang, Wenhui, et al. "Gated Self-Matching Networks for Reading Comprehension and Question Answering." ACL, 2017.

[2] Seo, Minjoon, et al. "Bidirectional Attention Flow for Machine Comprehension." ICLR, 2017.

Difficulty Controllable QG: Exploring a few intuitions…

1. 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?

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

Difficulty Controllable QG: 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

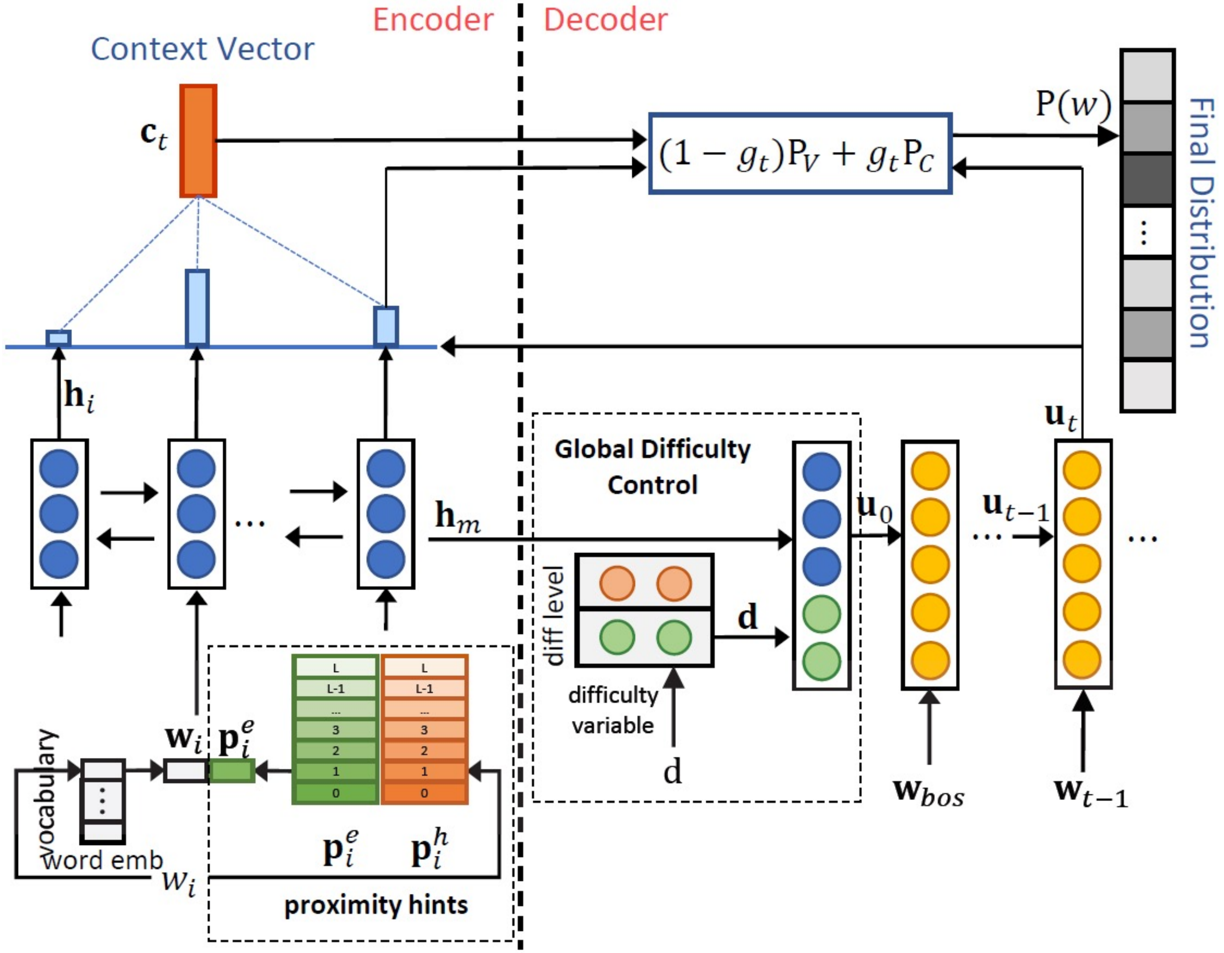
- Finding: the distance for hard questions is significantly larger than that for easy questions (9.71 vs. 7.67)

- Difficulty Level Proximity Hints (DLPH)
  - 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: $(\mathbf{p}_0^h, \mathbf{p}_1^h, \mathbf{p}_2^h, \dots \mathbf{p}_L^h)$
  - $\mathbf{p}_L$ means the word has L absolute distance with the answer span

Difficulty Controllable QG: Proposed Framework

- Characteristic-rich Encoder
  - Concatenate word embeddings and embeddings for **proximity hints**
    $$\mathbf{x} = [\mathbf{w}; \mathbf{p}]$$
  - Bidirectional LSTMs encode the sequence

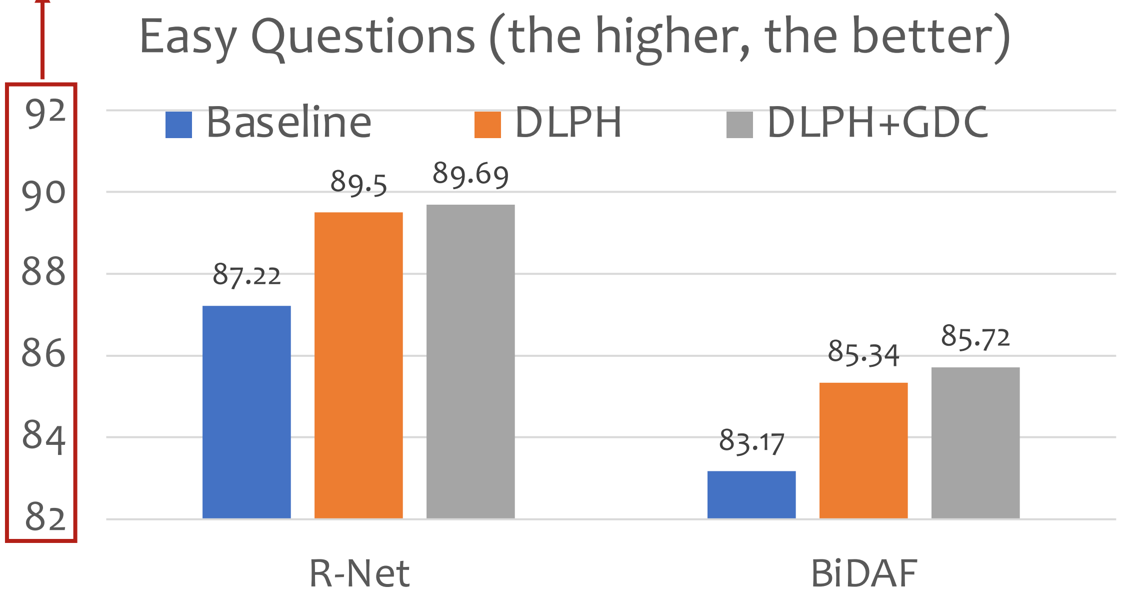
- Difficulty-controllable Decoder
  - **Global Difficulty Control**: use style variable to initialize decoder state
    $$\mathbf{u}_0 = [\mathbf{h}_m; \mathbf{d}]$$
  - Decoder with Attention & Copy

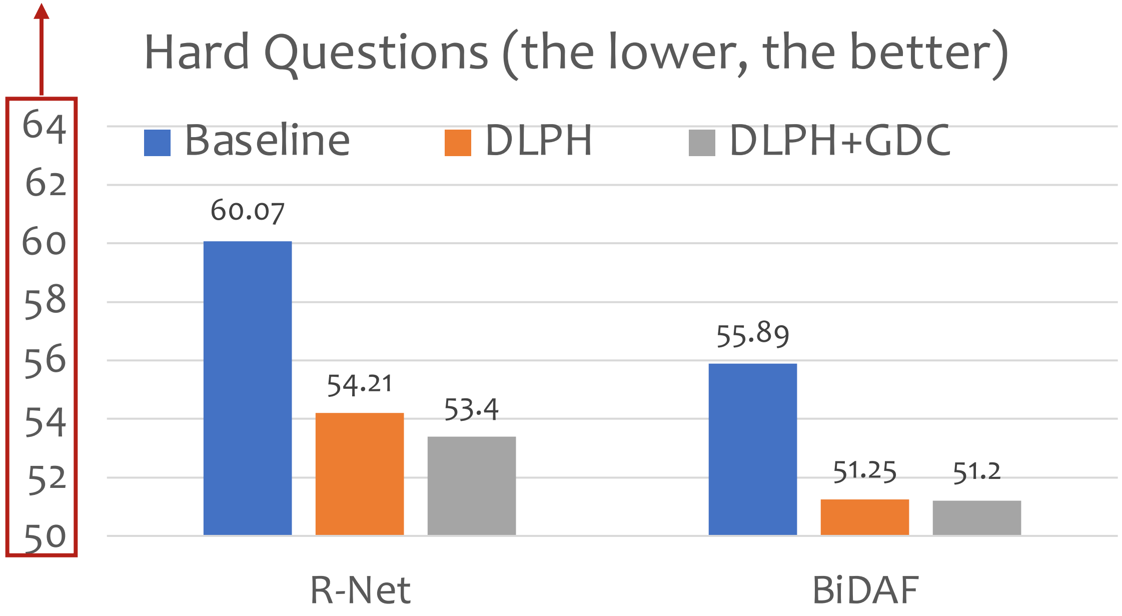


Difficulty Controllable QG: Difficulty Control Results

- **Metric:** Employ reading comprehension systems (R-Net, BiDAF) to evaluate the difficulty of generated questions
  - Baseline: Answer-Aware Neural Question Generation[1]
  - DLPH: Difficulty Level Proximity Hints
  - DLPH+GDC: Difficulty Level Proximity Hints + Global Difficulty Control

Percent of questions that can be correctly answered by reading comprehension systems

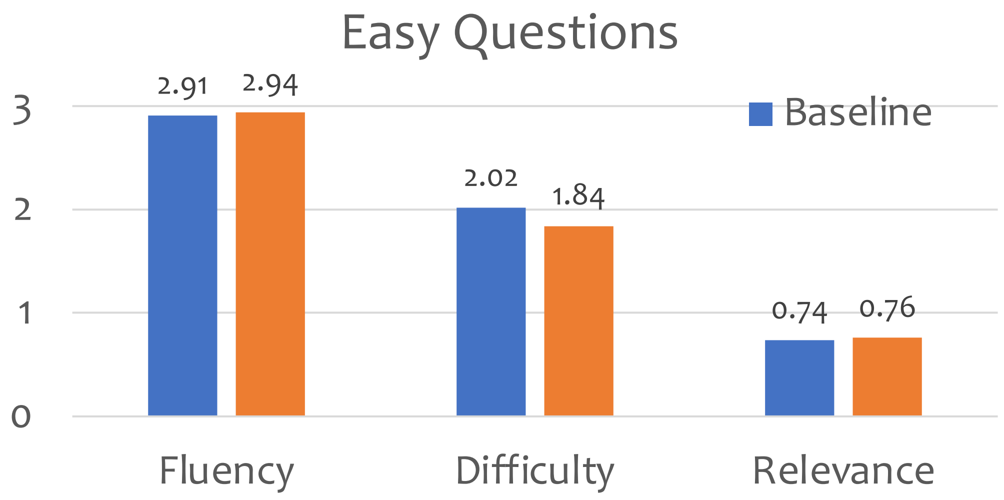

Easy Questions (the higher, the better)

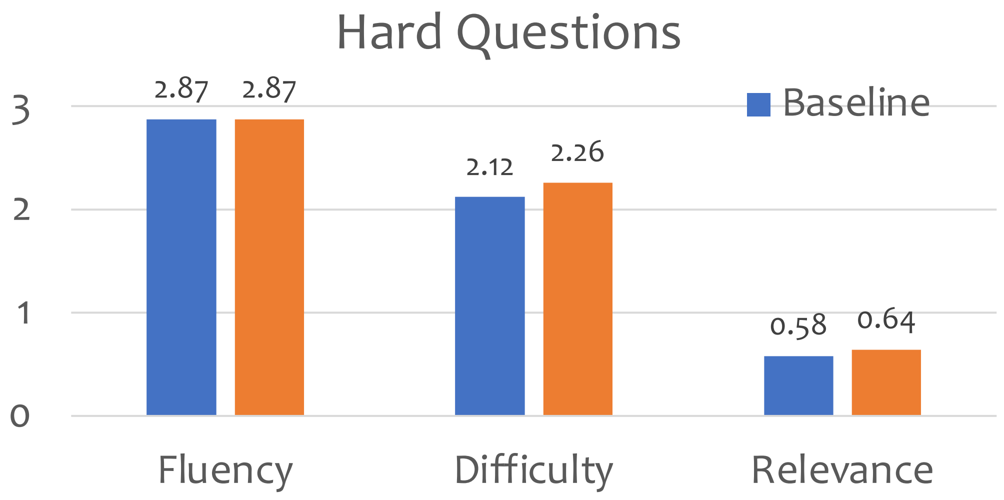

Hard Questions (the lower, the better)

[1] Zhou, Qingyu, et al. "Neural question generation from text: A preliminary study." NLPCC. Springer, Cham, 2017.

Difficulty Controllable QG: Human Evaluation

- 3 annotators rate the 100 easy questions and 100 hard questions
- Metrics
  - Fluency: grammatical correctness and fluency, 1-3 scale, 3 for best
  - Difficulty: difficulty of generated questions, 1-3 scale, 3 for best
  - Relevance: if the question is ask about the answer, 0-1 scale, 1 for best

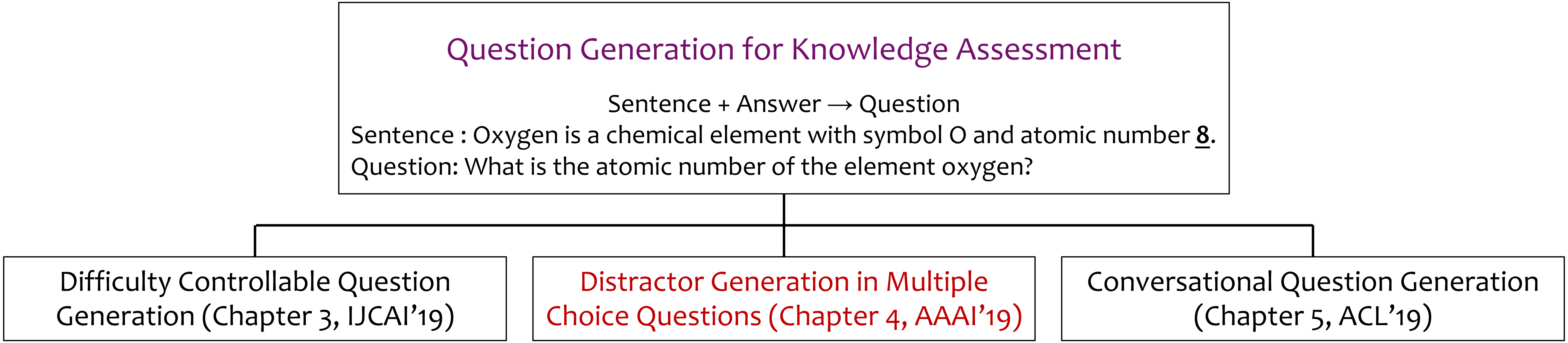




Difficulty Controllable QG: Conclusion

- 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

Part I: Knowledge Assessment



Question Generation for Knowledge Assessment

Sentence + Answer → Question

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

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

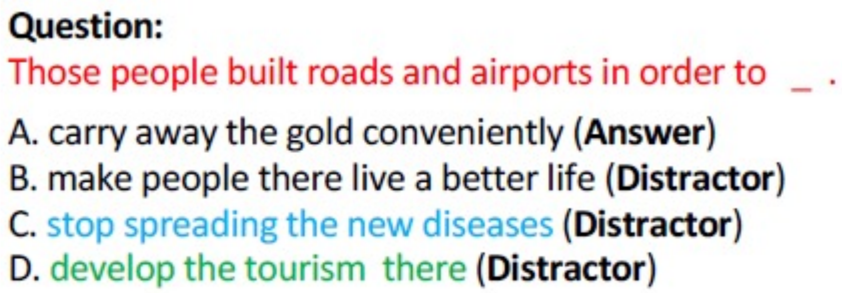
Difficulty Controllable Question Generation (Chapter 3, IJCAI'19)

Distractor Generation in Multiple Choice Questions (Chapter 4, AAAI'19)

Conversational Question Generation (Chapter 5, ACL'19)

Sentence + Answer + "Easy" → Easy Question

Sentence + Answer + "Hard" → Hard Question



Question:

Those people built roads and airports in order to _ .

A. carry away the gold conveniently (**Answer**)

B. make people there live a better life (**Distractor**)

C. stop spreading the new diseases (**Distractor**)

D. develop the tourism there (**Distractor**)

Q1: What political party is Clinton a member of?

A1: Democratic

Q2: What was **he** ineligible to serve?

A2: third term

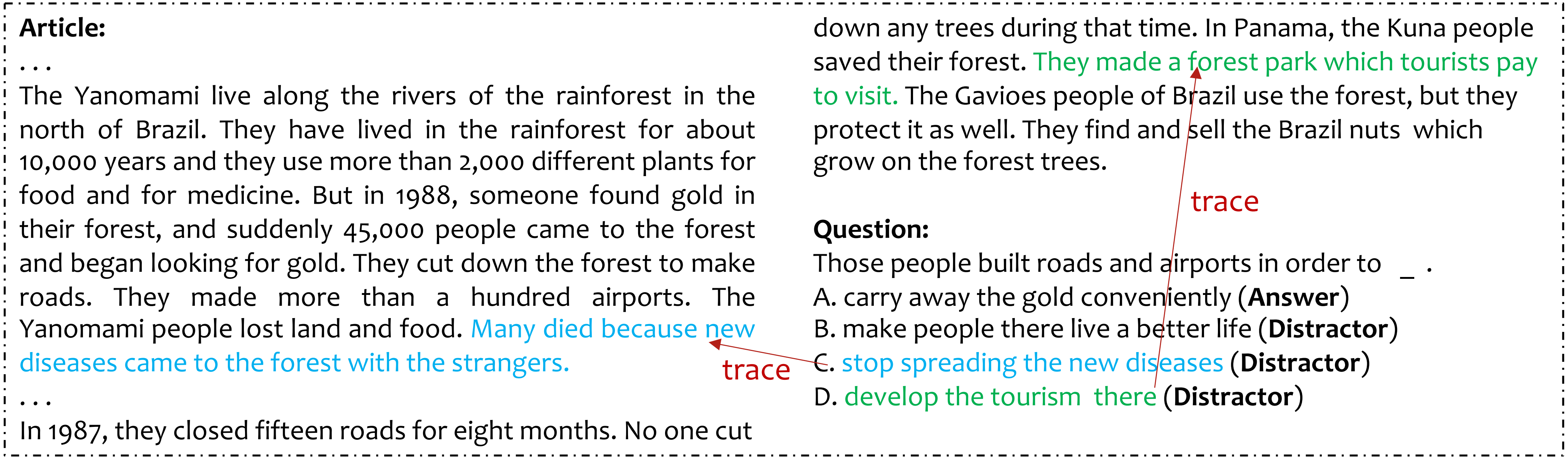
Q3: **Why**?

A3: term limitations

Difficulty in how questions are asked → Understand RC passages and design challenging distractors!

Distractor Generation: Motivation

- In multiple choice reading comprehension questions, distractors (wrong options) are difficulty to design because
  - Poor distractors can make the question almost trivial to solve.
  - Reasonable distractors should have some trace in the article.



Article:

. . .

The Yanomami live along the rivers of the rainforest in the north of Brazil. They have lived in the rainforest for about 10,000 years and they use more than 2,000 different plants for food and for medicine. But in 1988, someone found gold in their forest, and suddenly 45,000 people came to the forest and began looking for gold. They cut down the forest to make roads. They made more than a hundred airports. The Yanomami people lost land and food. Many died because new diseases came to the forest with the strangers.

. . .

In 1987, they closed fifteen roads for eight months. No one cut down any trees during that time. In Panama, the Kuna people saved their forest. They made a forest park which tourists pay to visit. The Gavioes people of Brazil use the forest, but they protect it as well. They find and sell the Brazil nuts which grow on the forest trees.

Question:

Those people built roads and airports in order to _ .

A. carry away the gold conveniently (**Answer**)

B. make people there live a better life (**Distractor**)

C. stop spreading the new diseases (**Distractor**)

D. develop the tourism there (**Distractor**)

Distractor Generation: Task Definition

- Input: an article, a pair of question and its correct option
- Output: context and question related, grammatically consistent distractors
  - Not share the same semantic meaning as the correct answer
  - Consistent with the semantic of the question
  - Have some trace in article

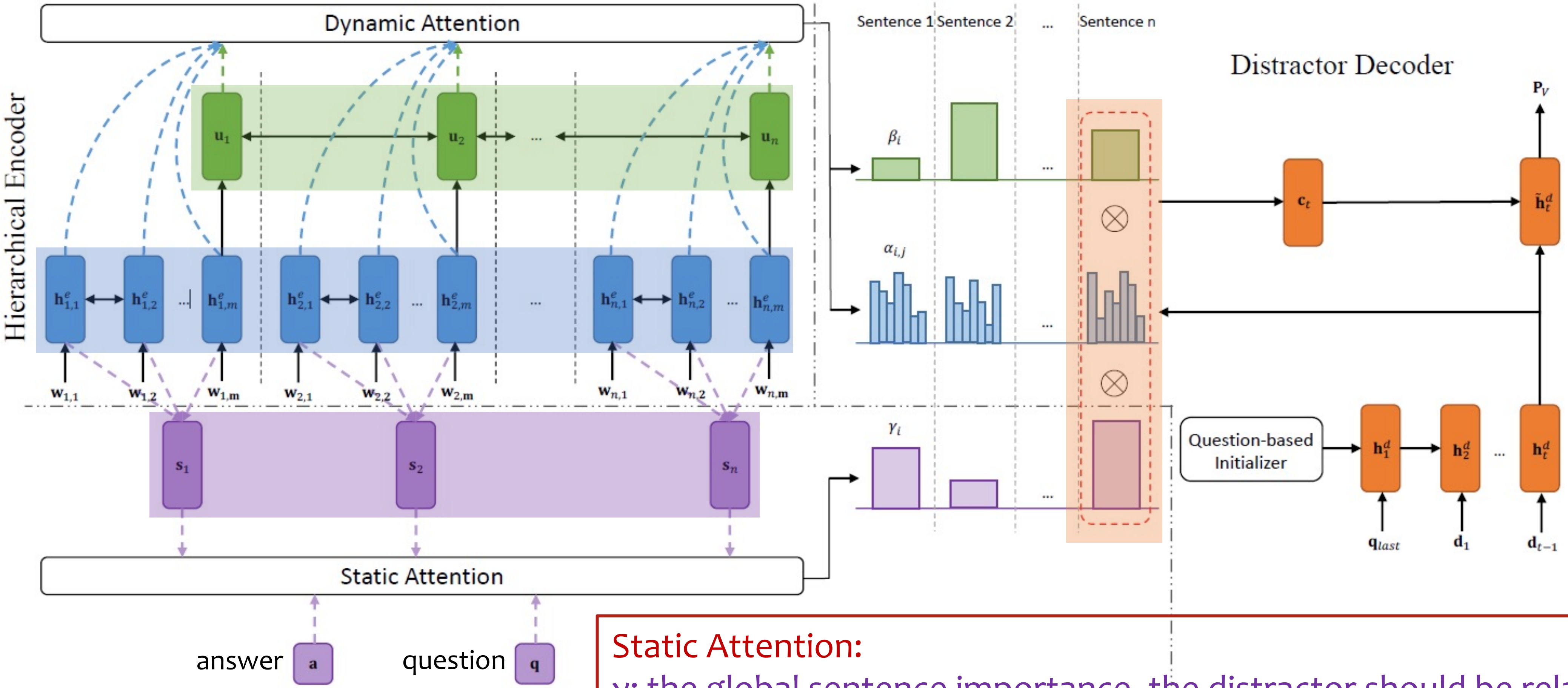
- Applications
  - Aid the preparation of MCQ reading comprehension datasets
  - Helpful to alleviate instructors' workload in designing MCQs

Distractor Generation: **Hierarchical** Encoder-Decoder Framework

Dynamic Attention:

β: emphasize the sentence that is important for generating the current distractor

α: word importance in each sentence



Combine Static and Dynamic Attention:

$$\tilde{\alpha}_{i,j} = \frac{\alpha_{i,j}\beta_i\gamma_i}{\sum_{ij}\alpha_{i,j}\beta_i\gamma_i}$$

Static Attention:

γ: the global sentence importance, the distractor should be related to question and should not share the same semantic meaning with the correct answer.

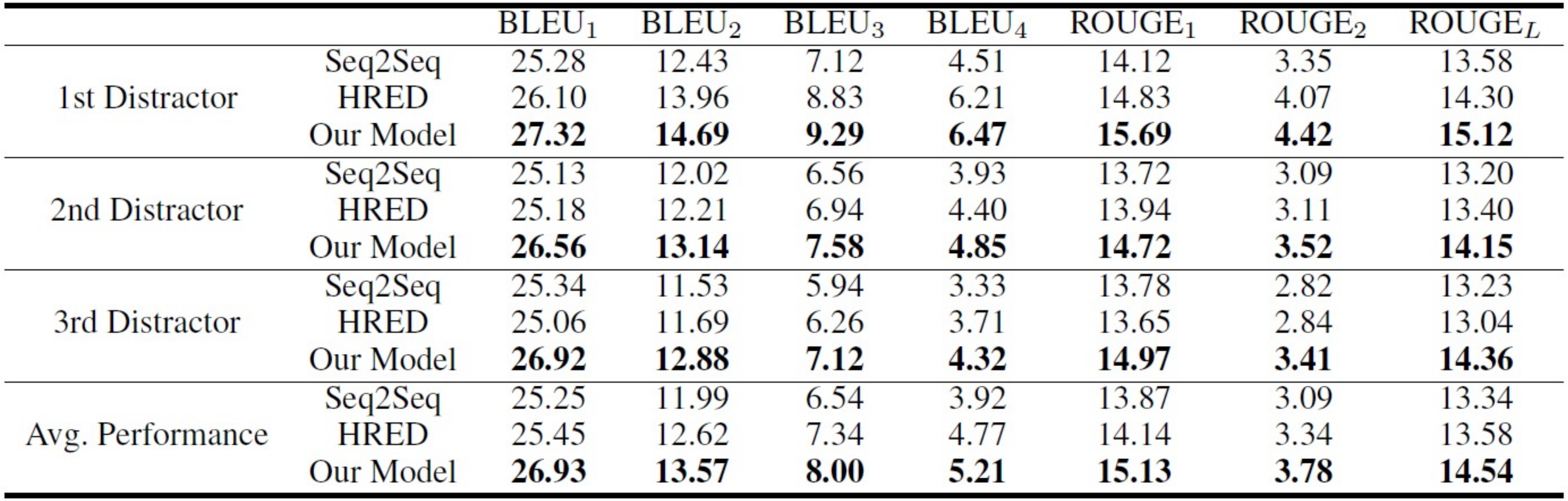
Distractor Generation: Dataset & Evaluation Metrics

- RACE Dataset[1]
  - 27,933 articles from English examinations of Chinese students from grade 7 to 12
  - Exclude trivial distractors, having no semantic relevance with the article
  - On average, each question receives 2.1 distractors
- Automatic Evaluation Metrics
  - BLEU (**Bi**Lingual **E**valuation **U**nderstudy)[2]
    - Measure the average n-gram **precision** on a set of reference sentences, with a penalty for overly short sentences.
    - BLEU-n (n=1/2/34) is BLEU score that uses up to n-grams for counting co-occurrences.
  - ROUGE (**R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation)[3]
    - Evaluate n-grams **recall** of the summaries with gold standard sentences as references
    - ROUGE-n (n=1/2): Overlap of n-grams between the system and reference summaries
    - ROUGE-L: Longest Common Subsequence (LCS) based statistics

[1] Lai, Guokun, et al. "RACE: Large-scale ReAding Comprehension Dataset From Examinations." EMNLP, 2017.
[2] Papineni, Kishore, et al. "BLEU: a Method for Automatic Evaluation of Machine Translation." ACL 2002.
[3] Lin, Chin-Yew. "ROUGE: a Package for Automatic Evaluation of Summaries." WAS 2004.

Distractor Generation: Automatic Evaluation Results

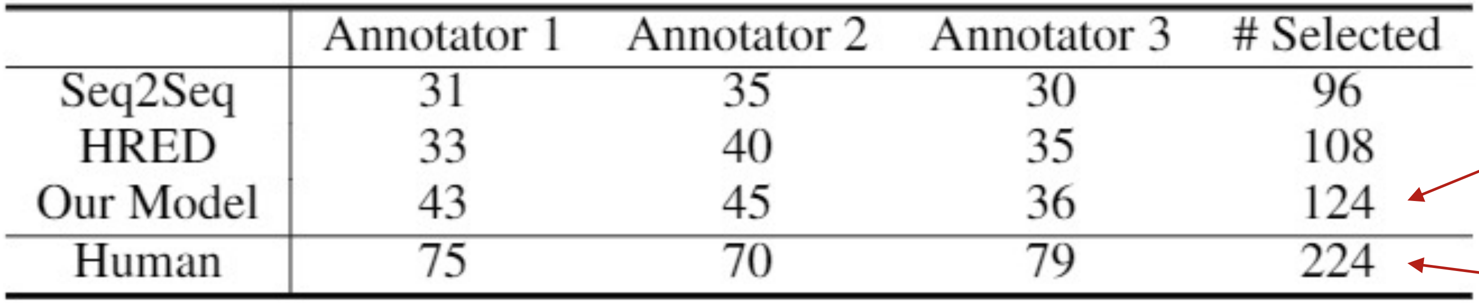
- Generation:
  - We generate 3 distractors using beam search with window size 50
  - Select first three outputs with Jaccard Distance > 0.5
- Models:
  - Seq2Seq: Sequence-to-Sequence model with attention and copy mechanism
  - HRED: HieRarchical Encoder-Decoder model
  - Our Model: hierarchical encoder-decoder model + static and dynamic attention mechanism



		BLEU_1	BLEU_2	BLEU_3	BLEU_4	ROUGE_1	ROUGE_2	ROUGE_L
1st Distractor	Seq2Seq	25.28	12.43	7.12	4.51	14.12	3.35	13.58
	HRED	26.10	13.96	8.83	6.21	14.83	4.07	14.30
	Our Model	**27.32**	**14.69**	**9.29**	**6.47**	**15.69**	**4.42**	**15.12**
2nd Distractor	Seq2Seq	25.13	12.02	6.56	3.93	13.72	3.09	13.20
	HRED	25.18	12.21	6.94	4.40	13.94	3.11	13.40
	Our Model	**26.56**	**13.14**	**7.58**	**4.85**	**14.72**	**3.52**	**14.15**
3rd Distractor	Seq2Seq	25.34	11.53	5.94	3.33	13.78	2.82	13.23
	HRED	25.06	11.69	6.26	3.71	13.65	2.84	13.04
	Our Model	**26.92**	**12.88**	**7.12**	**4.32**	**14.97**	**3.41**	**14.36**
Avg. Performance	Seq2Seq	25.25	11.99	6.54	3.92	13.87	3.09	13.34
	HRED	25.45	12.62	7.34	4.77	14.14	3.34	13.58
	Our Model	**26.93**	**13.57**	**8.00**	**5.21**	**15.13**	**3.78**	**14.54**

Distractor Generation: Human Evaluation

- We hire 3 annotators to answer 540 MCQs with five options
  - correct answer
  - the human designed distractors by high-school teachers
  - distractors generated from different methods (Seq2Seq, HRED, Our Model)
- We count how many times of individual pipelines are successful in *confusing the annotators*
  - "Confusing" means the annotator select the distractor when they do reading comprehension tests



	Annotator 1	Annotator 2	Annotator 3	# Selected
Seq2Seq	31	35	30	96
HRED	33	40	35	108
Our Model	43	45	36	124
Human	75	70	79	224

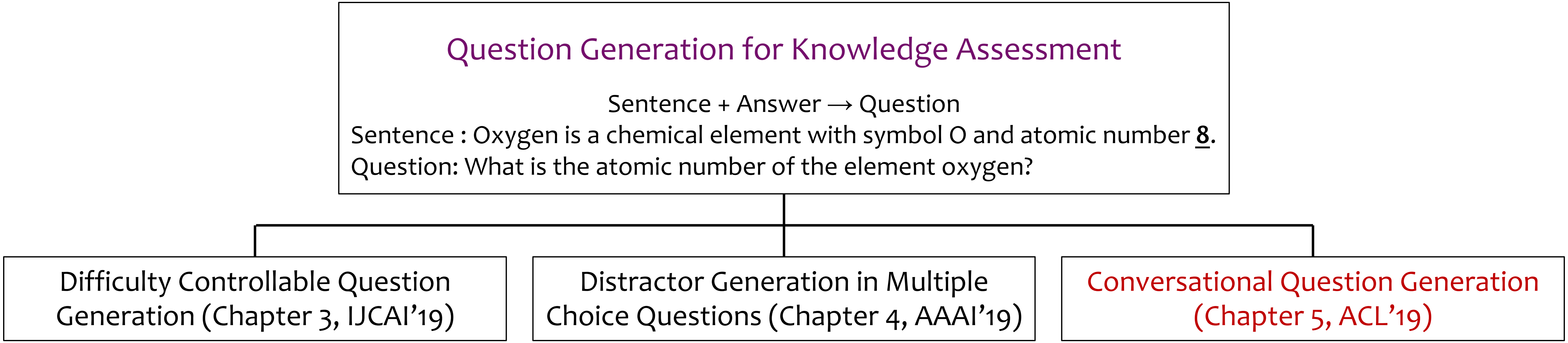
Our distractors are the best in model-generated ones

Human designed distractors are still much better than model generated ones

Distractor Generation: Conclusion

- We propose distractor generation for multiple choice reading comprehension questions
- The proposed hierarchical encoder-decoder approach with static and dynamic attention outperforms all previous baselines
- Human evaluation shows our model can really generate confusing distractors
- Code & Dataset: https://github.com/Yifan-Gao/Distractor-Generation-RACE

Part I: Knowledge Assessment



Question Generation for Knowledge Assessment

Sentence + Answer → Question

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

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

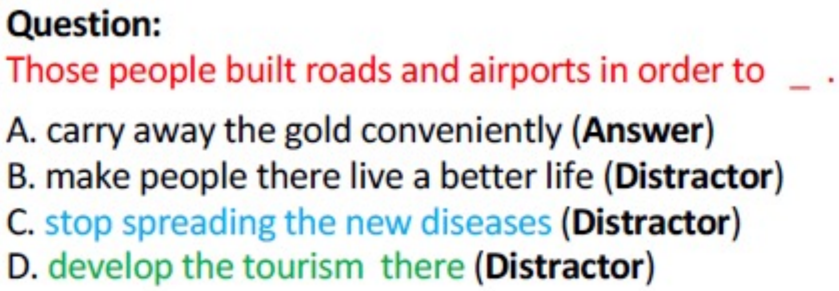
Difficulty Controllable Question Generation (Chapter 3, IJCAI'19)

Distractor Generation in Multiple Choice Questions (Chapter 4, AAAI'19)

Conversational Question Generation (Chapter 5, ACL'19)

Sentence + Answer + "Easy" → Easy Question

Sentence + Answer + "Hard" → Hard Question



Question:

Those people built roads and airports in order to _ .

A. carry away the gold conveniently (**Answer**)

B. make people there live a better life (**Distractor**)

C. stop spreading the new diseases (**Distractor**)

D. develop the tourism there (**Distractor**)

Q1: What political party is Clinton a member of?
A1: Democratic
Q2: What was he ineligible to serve?
A2: third term
Q3: Why?
A3: term limitations

Difficulty in how questions are asked → Understand RC passages and design challenging distractors → Ask questions in a better way!

Conversational QG: Motivation

So far, all questions are asked in a standalone manner

Incumbent Democratic President Bill Clinton was ineligible to serve a third term due to term limitations in the 22nd Amendment of the Constitution, and Vice President Gore was able to secure the Democratic nomination with relative ease.

Q: What political party is Clinton a member of?
A: Democratic
Q: What was Clinton ineligible to serve?
A: third term
Q: Why was he ineligible to serve a third term?
A: term limitations

Standalone Interaction ☹

Conversational QG: Question Generation + Conversation

Human test knowledge through conversations involving interconnected questions

Machines?

Standalone Interaction

Q: What political party is Clinton a member of?

A: Democratic

Q: What was Clinton ineligible to serve?

A: third term

Q: Why was he ineligible to serve a third term?

A: term limitations

Conversational Questions

Q1: What political party is Clinton a member of?

A1: Democratic

Q2: What was **he** ineligible to serve?

A2: third term

Q3: **Why**?

A3: term limitations

Conversation is the ultimate way for human-machine interactions

Conversational QG: Task Definition

- A New Setting:
  - A system needs to ask a series of interconnected questions grounded in a passage through a question-answering style conversation
  - Every question after the first turn might be dependent on the conversation history

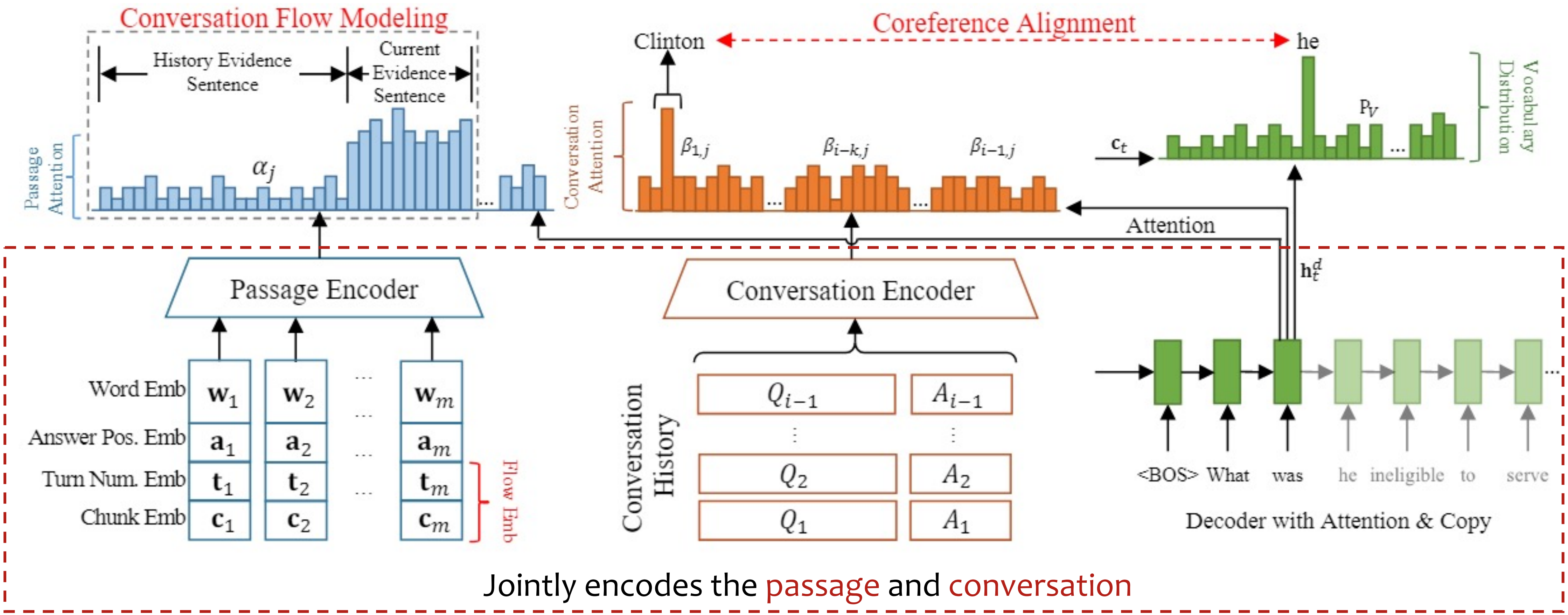
Incumbent Democratic President Bill Clinton was ineligible to serve a third term due to term limitations in the 22nd Amendment of the Constitution, and Vice President Gore was able to secure the Democratic nomination with relative ease.

Q1: What political party is Clinton a member of?
A1: Democratic
Q2: What was **he** ineligible to serve?
A2: third term
Q3: Why?
A3: term limitations

Challenge 1: Conversational questions depend on the conversation so far (refer back to the conversation history using coreference)

Challenge 2: A coherent conversation must have smooth transitions between turns (We expect the narrative structure of passages can influence the conversation flow of our interconnected questions)

Conversational QG: Proposed Encoder-Decoder Framework



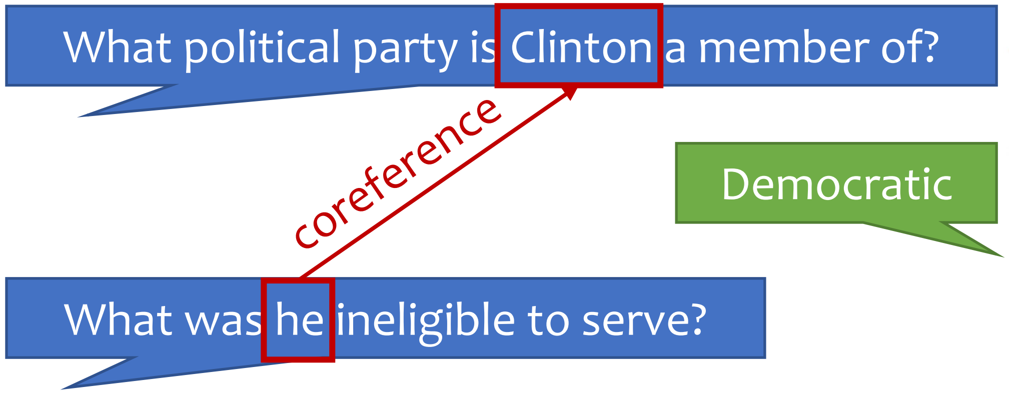
Jointly encodes the passage and conversation

Conversational QG: Coreference Alignment

Generate conversational interconnected questions depending on the conversation so far

Explicitly **align** coreferent mentions in conversation history with corresponding pronominal references in generated questions

- Preprocessing Stage

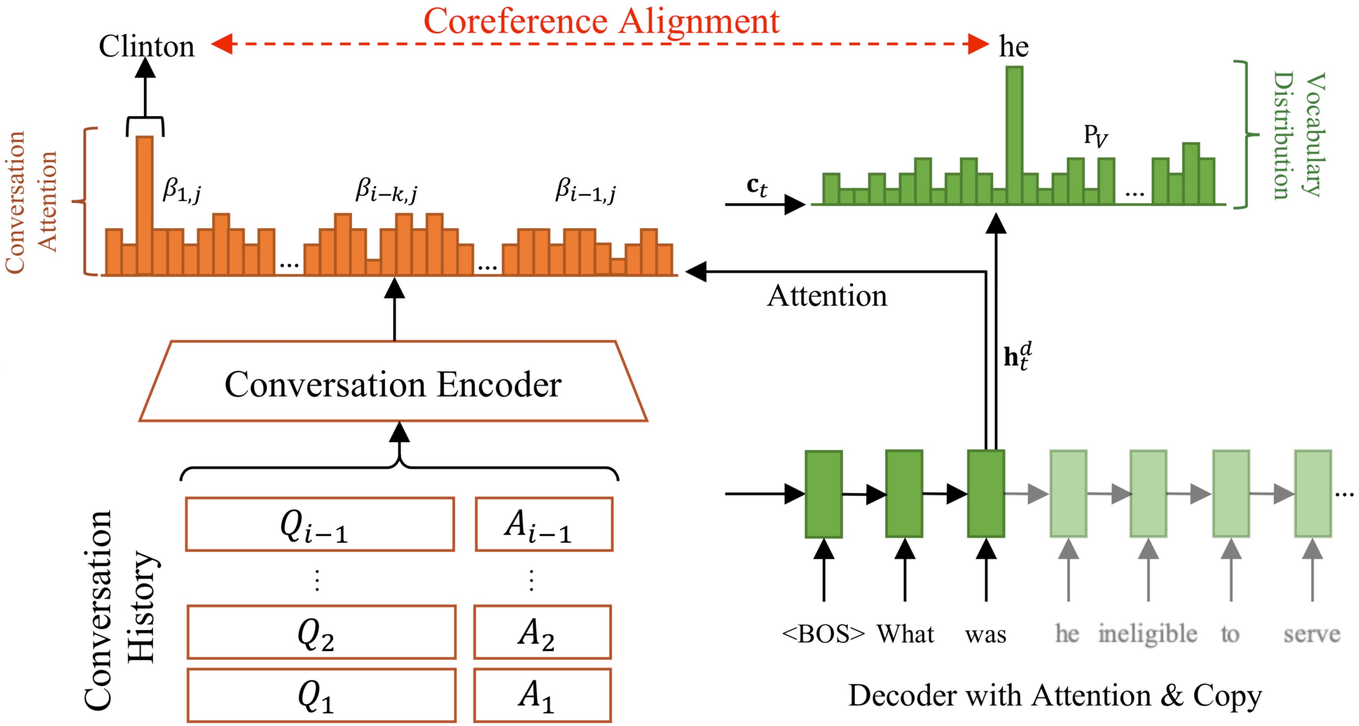

What political party is Clinton a member of?
coreference
Democratic
What was he ineligible to serve?

- Training Stage

e.g., Clinton

e.g., he

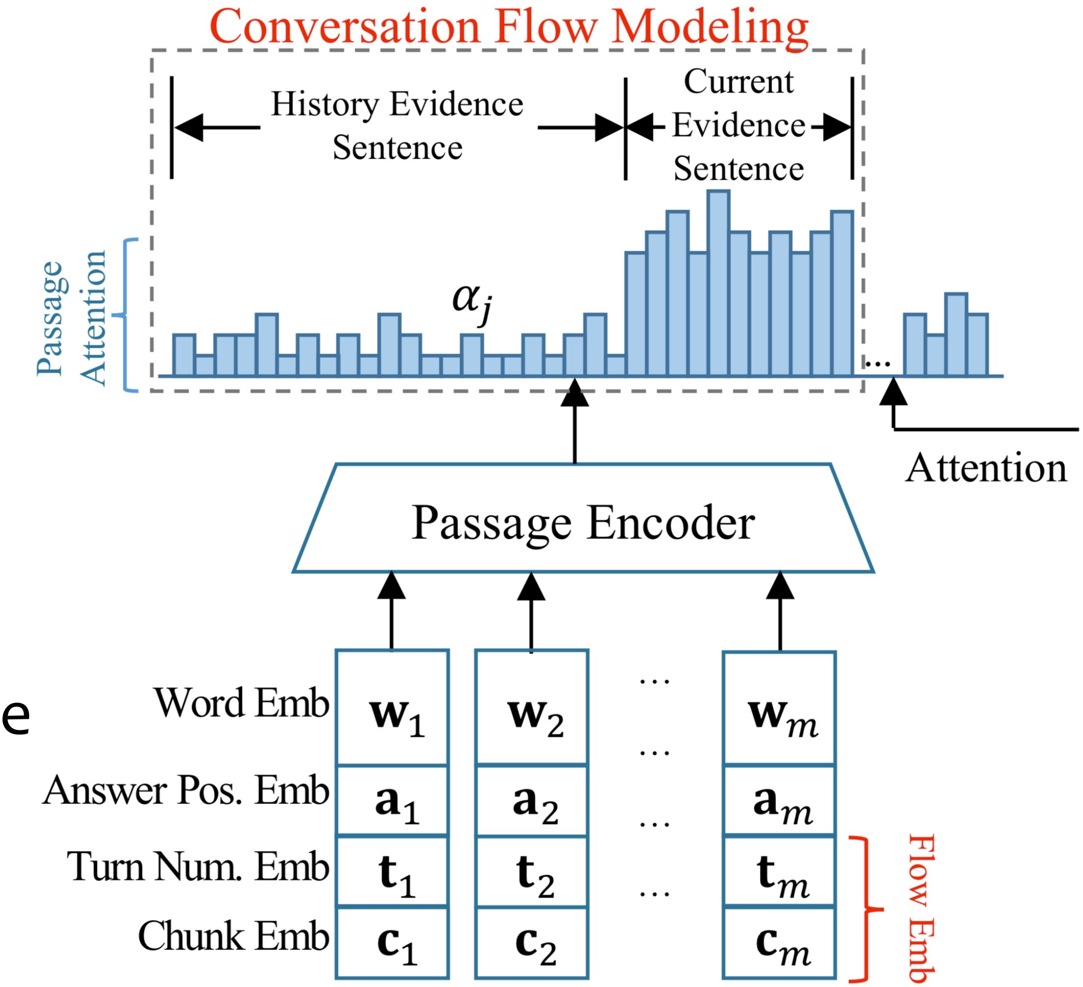
$$\mathcal{L}_{\text{coref}} = -(\lambda_1 \log \frac{\Sigma_j \beta^c_j}{\Sigma_{k,j} \beta_{i-k,j}} + \lambda_2 \log p_{\text{coref}}) * s_c$$



Conversational QG: Conversation Flow Modeling

Model the conversation flow to transit focus inside the passage smoothly across turns

1. Focus on sentences contain key information to generate the current turn question Current Evidence Sentence
2. Ignore sentences questioned several turns ago History Evidence Sentence



Conversation Flow Modeling
History Evidence Sentence
Current Evidence Sentence
Passage Attention
α_j
...
Attention
Passage Encoder
Word Emb
Answer Pos. Emb
Turn Num. Emb
Chunk Emb
$\mathbf{w}_1$ $\mathbf{w}_2$ ... $\mathbf{w}_m$
$\mathbf{a}_1$ $\mathbf{a}_2$ ... $\mathbf{a}_m$
$\mathbf{t}_1$ $\mathbf{t}_2$ ... $\mathbf{t}_m$
$\mathbf{c}_1$ $\mathbf{c}_2$ ... $\mathbf{c}_m$
Flow Emb

Focus on Current Evidence Sentence (CES): and ignore History Evidence Sentence (HES): via a flow loss:

$$\mathcal{L}_{\text{flow}} = -\lambda_3 \log \frac{\Sigma_{j:w_j \in \text{CES}} \alpha_j}{\Sigma_j \alpha_j} + \lambda_4 \frac{\Sigma_{j:w_j \in \text{HES}} \alpha_j}{\Sigma_j \alpha_j}$$

Conversational QG: Dataset & Main Results

- CoQA Dataset[1]
  - A large-scale conversational question answering dataset
  - 8k conversations, 127k QA pairs
  - Short question length: 5.5 tokens (previous QG dataset: 10.1 tokens)
- Main Results

Baselines:

PGNet: Pointer-Generator Network

NQG: Neural Question Generation[2]

Model Ablations:

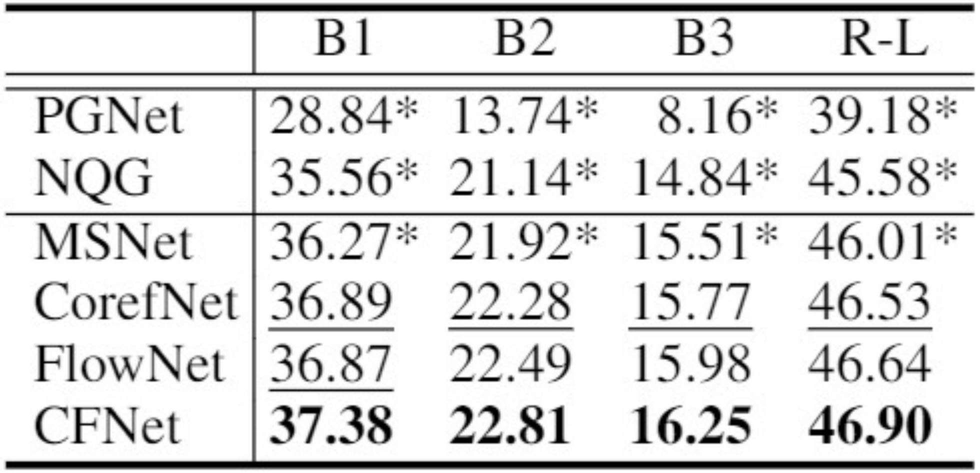
MSNet: Multi-Source EncDec

CorefNet: Coreference Alignment

FlowNet: Conversation Flow Modeling

Our Full Model:

CFNet: Multi-Source EncDec +
Coreference Alignment +
Conversational Flow Modeling



	B1	B2	B3	R-L
PGNet	28.84*	13.74*	8.16*	39.18*
NQG	35.56*	21.14*	14.84*	45.58*
MSNet	36.27*	21.92*	15.51*	46.01*
CorefNet	36.89	22.28	15.77	46.53
FlowNet	36.87	22.49	15.98	46.64
CFNet	**37.38**	**22.81**	**16.25**	**46.90**

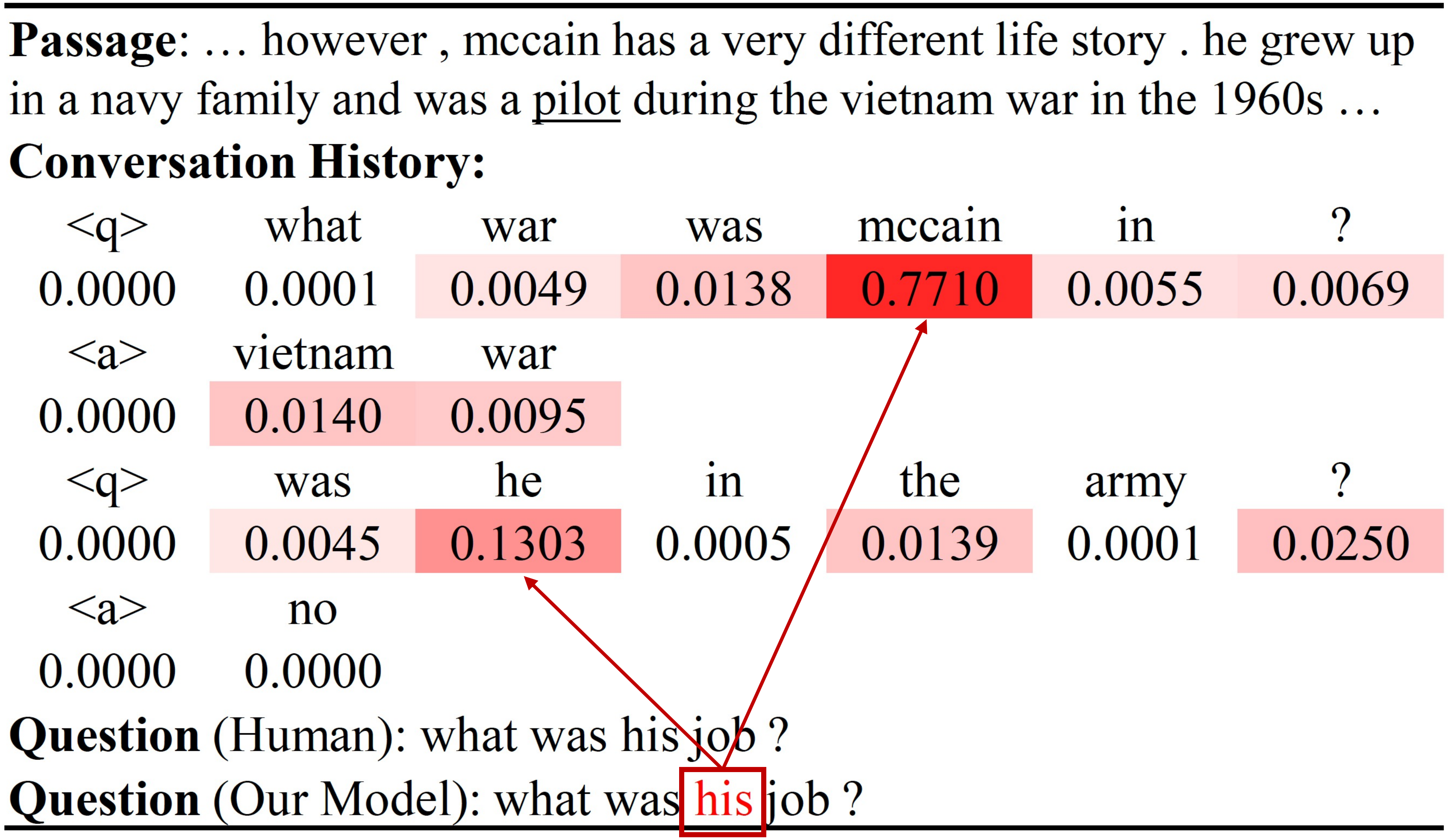
underline: p-value<0.05, *: p-value<0.01

[1] Reddy, Siva, et al. "CoQA: A Conversational Question Answering Challenge." Transactions of ACL, 2019.

[2] Du, Xinya, and Claire Cardie. "Harvesting Paragraph-Level Question-Answer Pairs from Wikipedia." ACL 2018

Conversational QG: Coreference Alignment Analysis

The attention probability when the model predicts a pronoun



Passage: … however , mccain has a very different life story . he grew up in a navy family and was a pilot during the vietnam war in the 1960s …

Conversation History:

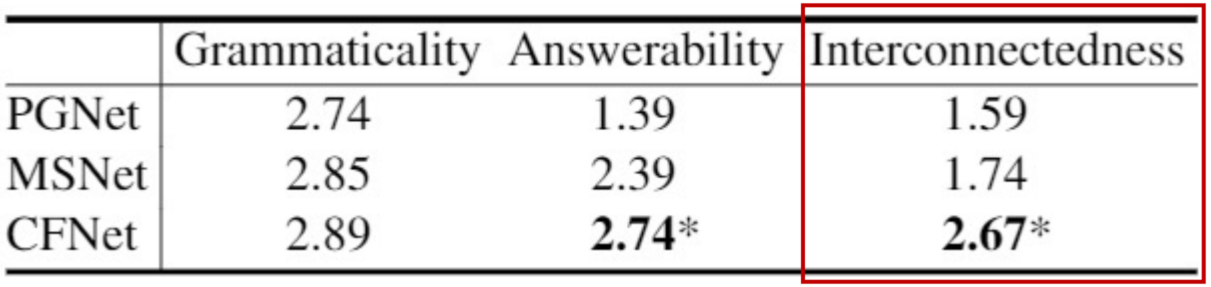
<q>	what	war	was	mccain	in	?
0.0000	0.0001	0.0049	0.0138	0.7710	0.0055	0.0069
<a>	vietnam	war				
0.0000	0.0140	0.0095				
<q>	was	he	in	the	army	?
0.0000	0.0045	0.1303	0.0005	0.0139	0.0001	0.0250
<a>	no					
0.0000	0.0000					

Question (Human): what was his job ?

Question (Our Model): what was his job ?

Conversational QG: Human Evaluation

- We hire 5 annotators to rate 93 questions
- Rating criteria (1-3 scale, 3 for the best)
  - Grammaticality: the grammatical correctness and fluency
  - Answerability: whether the generated question can be answered by the current answer
  - **Interconnectedness**: whether the generated questions are *conversational* or not



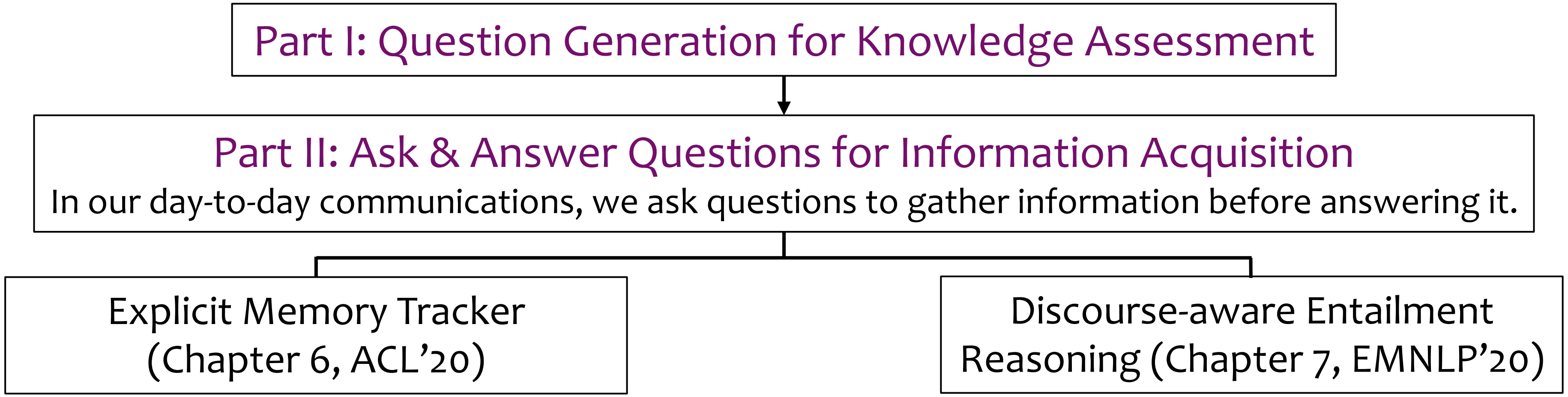
	Grammaticality	Answerability	Interconnectedness
PGNet	2.74	1.39	1.59
MSNet	2.85	2.39	1.74
CFNet	2.89	**2.74***	**2.67***

*: p-value<0.01

Conversational QG: Conclusion

- A new setting: Conversational Question Generation
- Coreference Alignment
- Conversation Flow Modeling
- Code & Models: https://github.com/Yifan-Gao/conversational-QG

Part II: Information Acquisition



Part I: Question Generation for Knowledge Assessment

Part II: Ask & Answer Questions for Information Acquisition

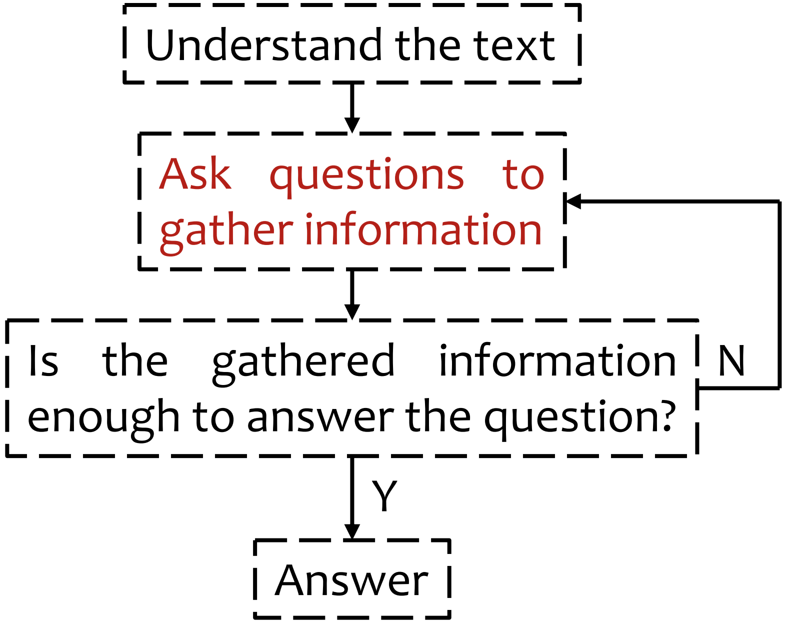
In our day-to-day communications, we ask questions to gather information before answering it.

Explicit Memory Tracker (Chapter 6, ACL'20)

Discourse-aware Entailment Reasoning (Chapter 7, EMNLP'20)

Knowledge Article: 7(a) loans provides business loans to American small businesses. The loan program is designed to assist for-profit businesses that are not able to get other financing from other resources.

Question: I am a 34-year-old man from the United States. I am the owner of an American small business. Is the 7(a) Loan Program for me?



Understand the text

Ask questions to gather information

Is the gathered information enough to answer the question?

Answer

Part II: Information Acquisition -- Background

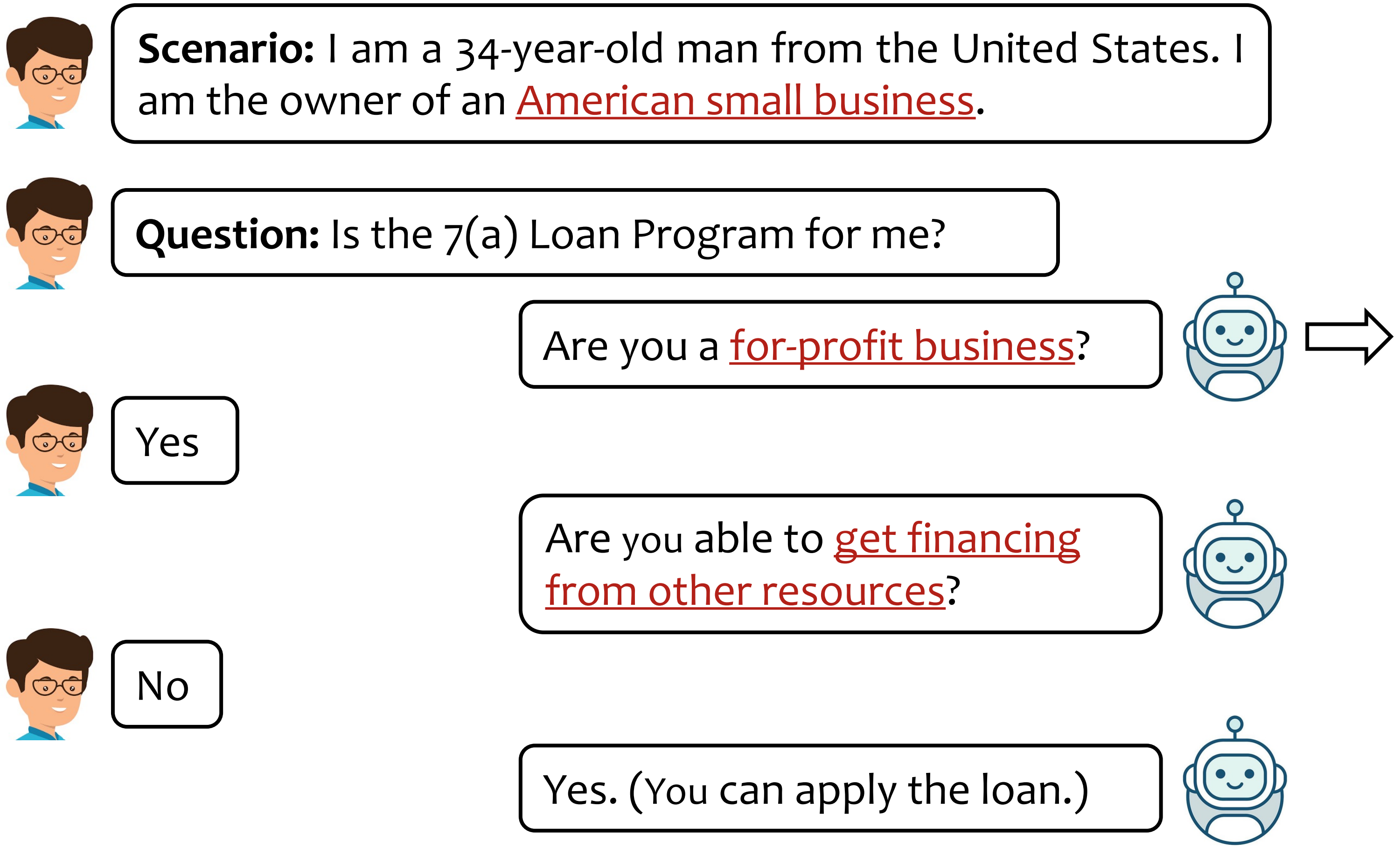
Scenario: I am a 34-year-old man from the United States. I am the owner of an American small business.

Question: Is the 7(a) Loan Program for me?

⇨

Rule Text:
7(a) loans provides business loans to American small businesses. The loan program is designed to assist for-profit businesses that are not able to get other financing from other resources.

Part II: Information Acquisition -- Background



Scenario: I am a 34-year-old man from the United States. I am the owner of an American small business.

Question: Is the 7(a) Loan Program for me?

Are you a for-profit business?

Yes

Are you able to get financing from other resources?

No

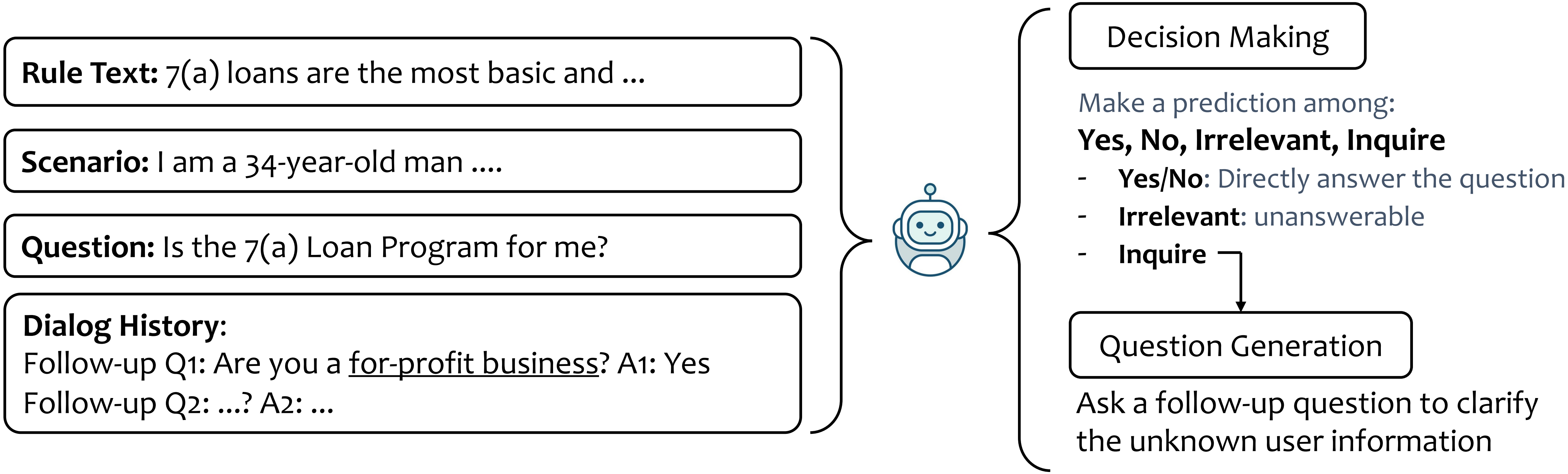
Yes. (You can apply the loan.)

Rule Text:

7(a) loans provides business loans to ① American small businesses. The loan program is designed to assist ② for-profit businesses that are ③ not able to get other financing from other resources.

Part II: Information Acquisition -- Task Definition

ShARC[1]: Shaping Answers with Rules through Conversation



Rule Text: 7(a) loans are the most basic and ...

Scenario: I am a 34-year-old man ....

Question: Is the 7(a) Loan Program for me?

Dialog History:
Follow-up Q1: Are you a for-profit business? A1: Yes
Follow-up Q2: ...? A2: ...

Decision Making

Make a prediction among:
Yes, No, Irrelevant, Inquire

- **Yes/No**: Directly answer the question
- **Irrelevant**: unanswerable
- **Inquire**

Question Generation

Ask a follow-up question to clarify the unknown user information

[1] Saeidi, Marzieh, et al. "Interpretation of Natural Language Rules in Conversational Machine Reading." EMNLP, 2018.

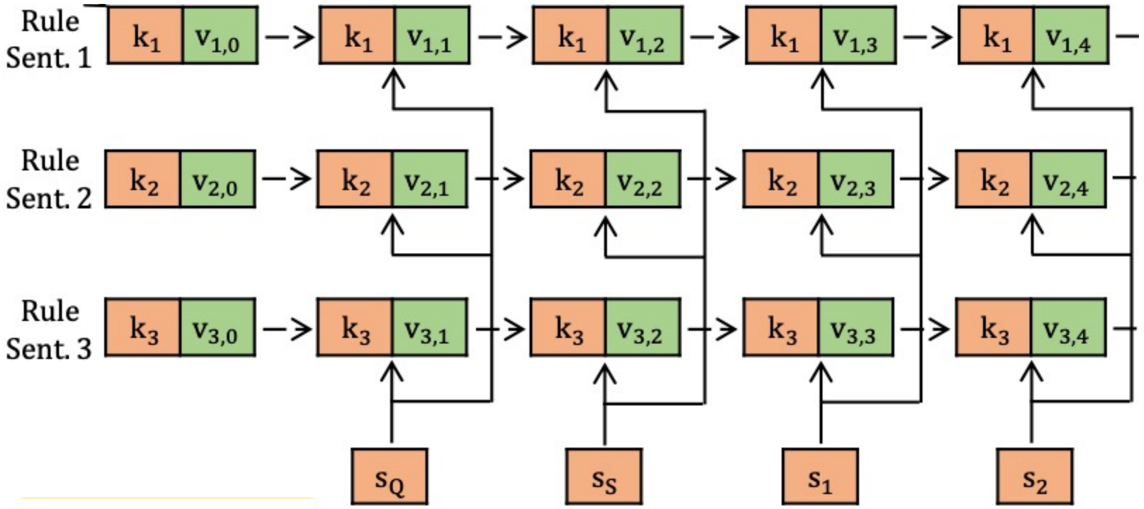
Part II: Information Acquisition

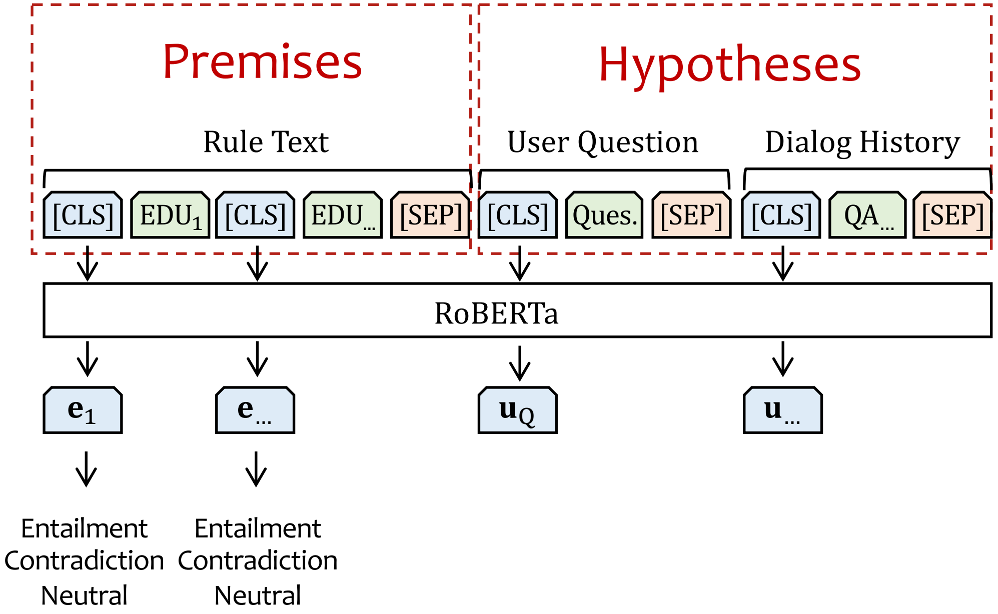
Part II: Ask & Answer Questions for Information Acquisition

In our day-to-day communications, we ask questions to gather information before answering it.

Explicit Memory Tracker (Chapter 6, ACL’20)

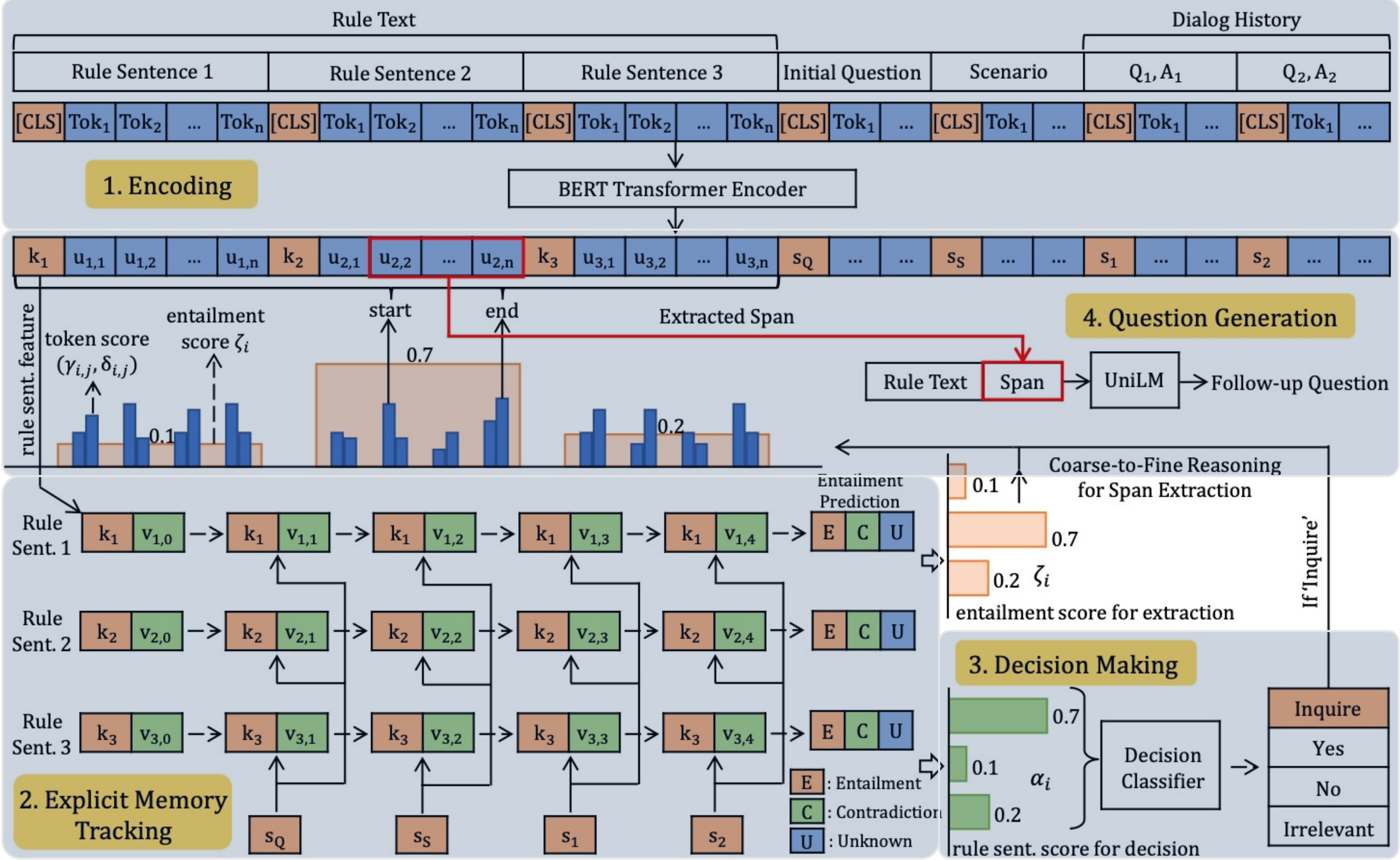
Discourse-aware Entailment Reasoning (Chapter 7, EMNLP’20)


Rule Sent. 1
Rule Sent. 2
Rule Sent. 3
k_1 $v_{1,0}$ $v_{1,1}$ $v_{1,2}$ $v_{1,3}$ $v_{1,4}$
k_2 $v_{2,0}$ $v_{2,1}$ $v_{2,2}$ $v_{2,3}$ $v_{2,4}$
k_3 $v_{3,0}$ $v_{3,1}$ $v_{3,2}$ $v_{3,3}$ $v_{3,4}$
s_Q
s_S
s_1
s_2

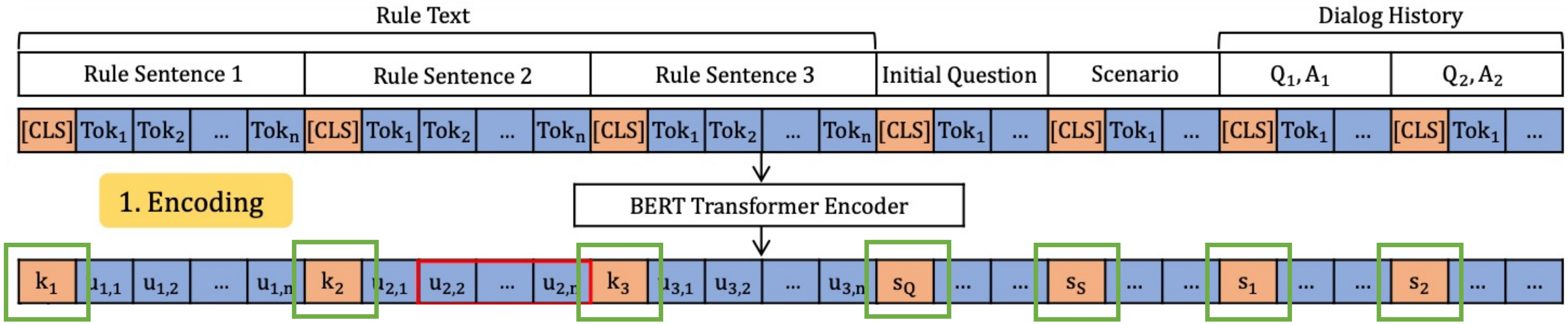


Explicit Memory Tracker: Framework

Overall Process: 1. Encoding ⟶ 2. Explicit Memory Tracking ⟶ 3. Decision Making ⟶ 4. Question Generation


Rule Text
Dialog History
Rule Sentence 1
Rule Sentence 2
Rule Sentence 3
Initial Question
Scenario
Q_1, A_1
Q_2, A_2
[CLS] Tok_1 Tok_2 ... Tok_n
1. Encoding
BERT Transformer Encoder
k_1 $u_{1,1}$ $u_{1,2}$ ... $u_{1,n}$ k_2 $u_{2,1}$ $u_{2,2}$ ... $u_{2,n}$ k_3 $u_{3,1}$ $u_{3,2}$ ... $u_{3,n}$ s_Q ... s_S ... s_1 ... s_2 ...
rule sent. feature
token score $(\gamma_{i,j}, \delta_{i,j})$
entailment score ζ_i
start
end
Extracted Span
0.1
0.7
0.2
4. Question Generation
Rule Text
Span
UniLM
Follow-up Question
Coarse-to-Fine Reasoning for Span Extraction
Entailment Prediction
0.1
0.7
0.2 ζ_i
entailment score for extraction
If 'Inquire'
Rule Sent. 1
Rule Sent. 2
Rule Sent. 3
k_1 $v_{1,0}$ k_1 $v_{1,1}$ k_1 $v_{1,2}$ k_1 $v_{1,3}$ k_1 $v_{1,4}$
k_2 $v_{2,0}$ k_2 $v_{2,1}$ k_2 $v_{2,2}$ k_2 $v_{2,3}$ k_2 $v_{2,4}$
k_3 $v_{3,0}$ k_3 $v_{3,1}$ k_3 $v_{3,2}$ k_3 $v_{3,3}$ k_3 $v_{3,4}$
E C U
s_Q s_S s_1 s_2
2. Explicit Memory Tracking
E : Entailment
C : Contradiction
U : Unknown
3. Decision Making
0.7
0.1
0.2
α_i
rule sent. score for decision
Decision Classifier
Inquire
Yes
No
Irrelevant

Explicit Memory Tracker: Encoding


Rule Text
Dialog History
Rule Sentence 1
Rule Sentence 2
Rule Sentence 3
Initial Question
Scenario
Q_1, A_1
Q_2, A_2
[CLS] Tok_1 Tok_2 ... Tok_n
1. Encoding
BERT Transformer Encoder
k_1 $u_{1,1}$ $u_{1,2}$ ... $u_{1,n}$
k_2 $u_{2,1}$ $u_{2,2}$ ... $u_{2,n}$
k_3 $u_{3,1}$ $u_{3,2}$ ... $u_{3,n}$
s_Q ... s_S ... s_1 ... s_2 ...

1. Parse the rule text into multiple rule sentences according to rules
2. Insert **[CLS]** token at the start of each rule sentence, initial question, scenario, and question-answer pairs in the dialog history
3. Concatenate all information and feed to BERT for encoding
4. **[CLS]** symbol is treated as the feature representation of the sentence that follows it

Explicit Memory Tracker: Implication Finding

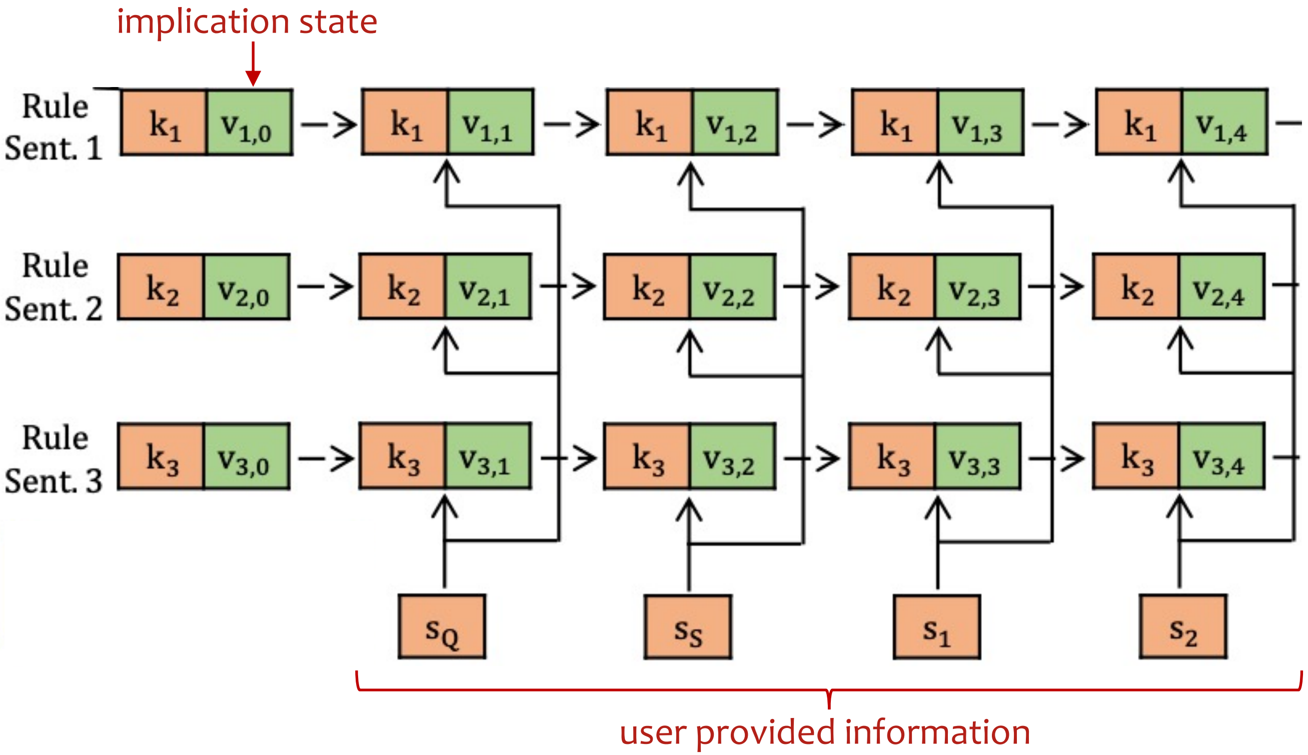
Rule sentences $\mathbf{k}_1, \mathbf{k}_2, \dots, \mathbf{k}_M$

Find Implications

User provided information:

- Initial question $\mathbf{s}_Q$
- Scenario $\mathbf{s}_S$
- Dialog history $\mathbf{s}_1, \dots, \mathbf{s}_P$

- We propose Explicit Memory Tracker (EMT), a gated recurrent memory-augmented neural network
- EMT explicitly tracks the states ($\mathbf{v}_i$) of rule sentences ($\mathbf{k}_i$) by sequentially reading the user provided information ($\mathbf{s}_i$)

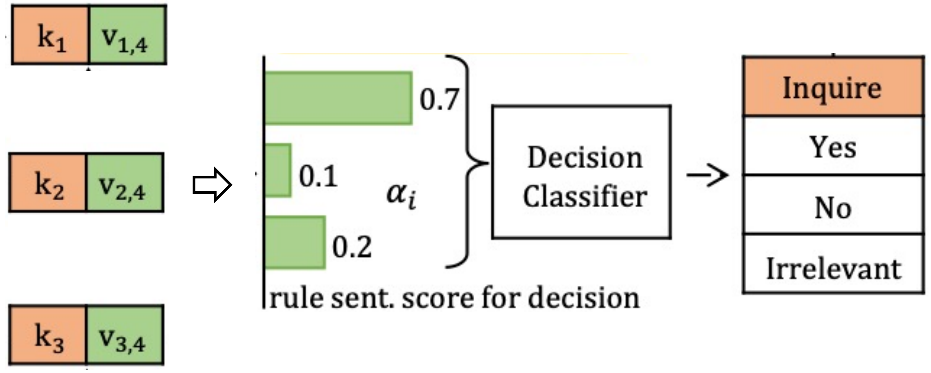


Explicit Memory Tracker: Decision Making

After P turns of explicit memory tracking, keys and final states of rule sentences are denoted as $(\mathbf{k}_1, \mathbf{v}_1), \ldots, (\mathbf{k}_M, \mathbf{v}_M)$:

- Decision making
- Question generation

Based on the most up-to-date key-value states of rule sentences $(\mathbf{k}_1, \mathbf{v}_1), \ldots, (\mathbf{k}_M, \mathbf{v}_M)$, EMT makes a decision among *Yes, No, Irrelevant, Inquire:*



$$\alpha_i = \mathbf{w}_\alpha^\top [\mathbf{k}_i; \mathbf{v}_i] + b_\alpha \in \mathbb{R}^1$$

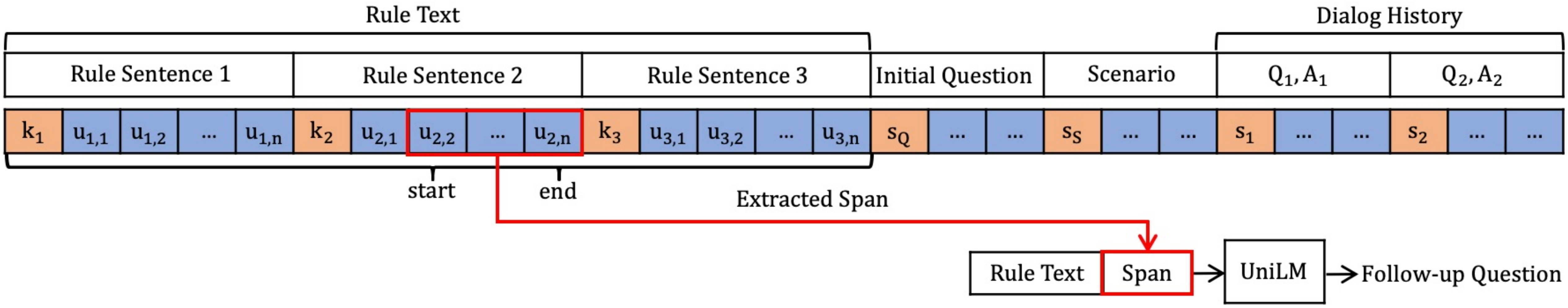
$$\tilde{\alpha}_i = \mathrm{softmax}(\alpha)_i \in [0, 1]$$

$$\mathbf{c} = \sum_i \tilde{\alpha}_i [\mathbf{k}_i; \mathbf{v}_i] \in \mathbb{R}^d$$

$$\mathbf{z} = \mathbf{W}_z \mathbf{c} + \mathbf{b}_z \in \mathbb{R}^4$$

Explicit Memory Tracker: Question Generation

- When the decision is ‘Inquire’, a follow-up question is required for further clarification
  1) Extract a span inside the rule text which contains the underspecified user information
  2) Finetune UniLM[1] for question generation, a pretrained language model
  3) [CLS] rule text [SEP] span [SEP] → follow-up question



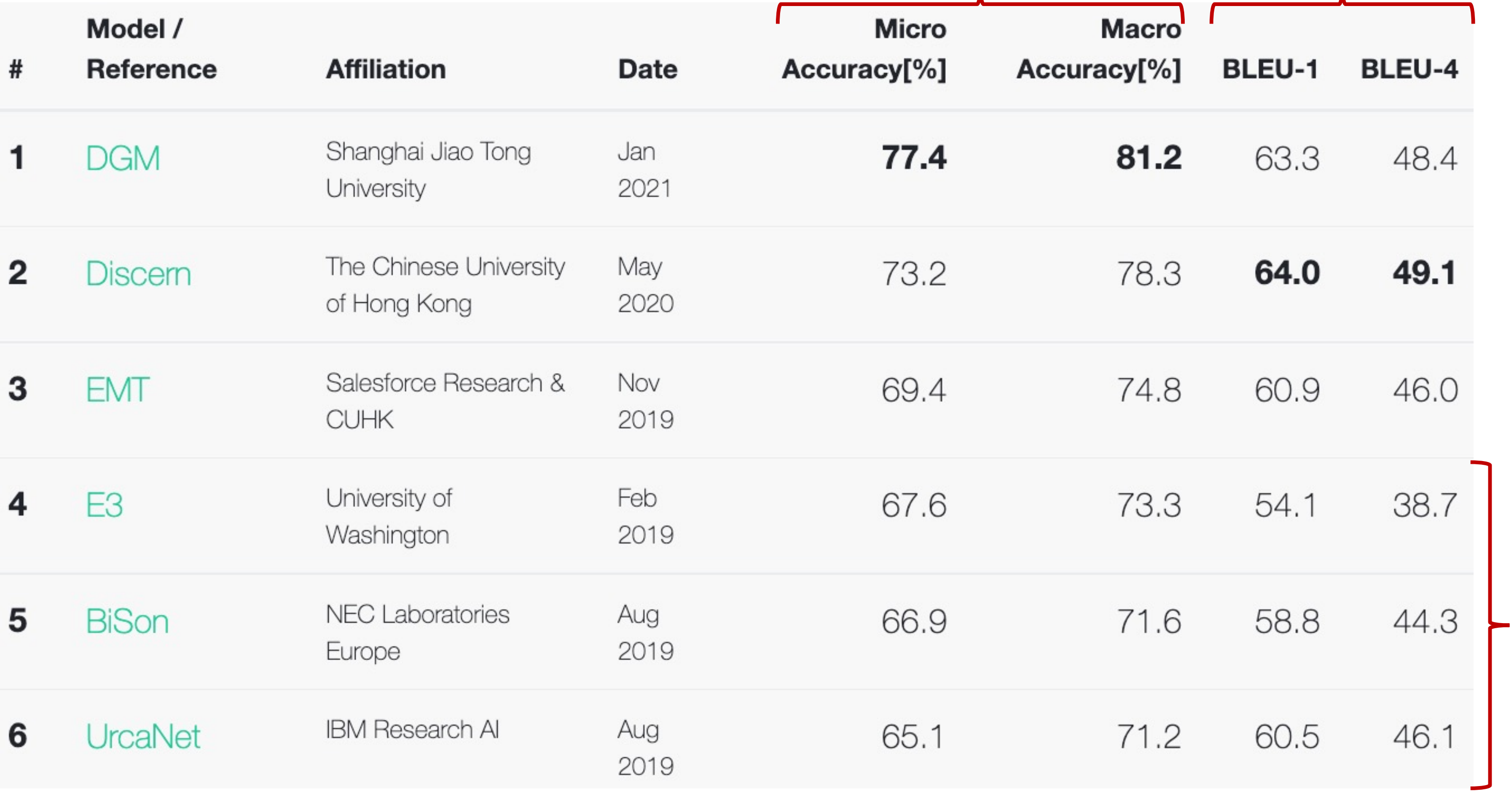
[1] Dong, Li, et al. “Unified Language Model Pre-training for Natural Language Understanding and Generation.” NeurIPS, 2019.

Explicit Memory Tracker: Dataset & Main Results

- Dataset:
  - ShARC conversational machine reading dataset[1]
  - Test set is not public, leaderboard: https://sharc-data.github.io/leaderboard.html

Decision Making

Question Generation



#	Model / Reference	Affiliation	Date	Micro Accuracy[%]	Macro Accuracy[%]	BLEU-1	BLEU-4
1	DGM	Shanghai Jiao Tong University	Jan 2021	**77.4**	**81.2**	63.3	48.4
2	Discern	The Chinese University of Hong Kong	May 2020	73.2	78.3	**64.0**	**49.1**
3	EMT	Salesforce Research & CUHK	Nov 2019	69.4	74.8	60.9	46.0
4	E3	University of Washington	Feb 2019	67.6	73.3	54.1	38.7
5	BiSon	NEC Laboratories Europe	Aug 2019	66.9	71.6	58.8	44.3
6	UrcaNet	IBM Research AI	Aug 2019	65.1	71.2	60.5	46.1

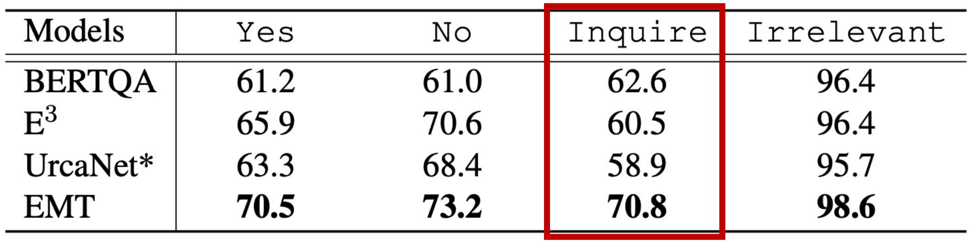
Explicit Memory Tracker

outperform all previous models (by Nov. 2019)

[1] Saeidi, Marzieh, et al. "Interpretation of Natural Language Rules in Conversational Machine Reading." EMNLP, 2018.

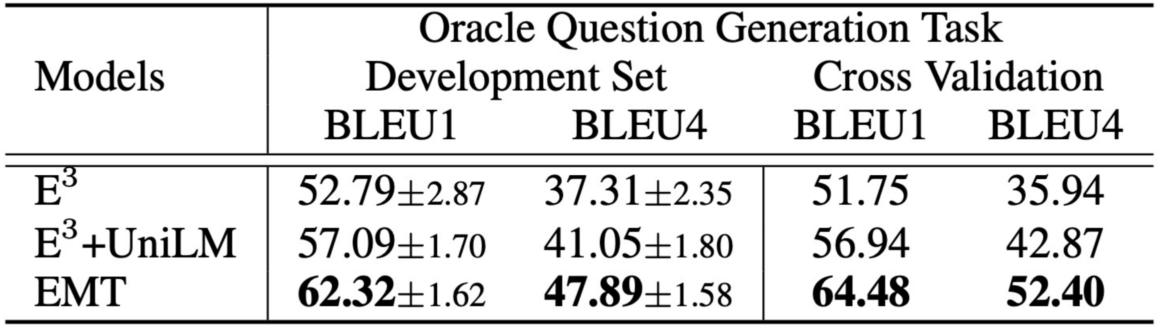
Explicit Memory Tracker: Detailed Analysis

- Class-wise Decision Accuracy
  - EMT makes better "Inquire" decision when necessary



Models	Yes	No	Inquire	Irrelevant
BERTQA	61.2	61.0	62.6	96.4
E^3	65.9	70.6	60.5	96.4
UrcaNet*	63.3	68.4	58.9	95.7
EMT	**70.5**	**73.2**	**70.8**	**98.6**

- Oracle Question Generation Results
  - If two models have different *Inquire* predictions, the follow-up questions for evaluation will be different, making the comparison unfair
  - We propose oracle QG setting: models are required to generate follow-up questions *whenever* the ground truth decision is *Inquire*

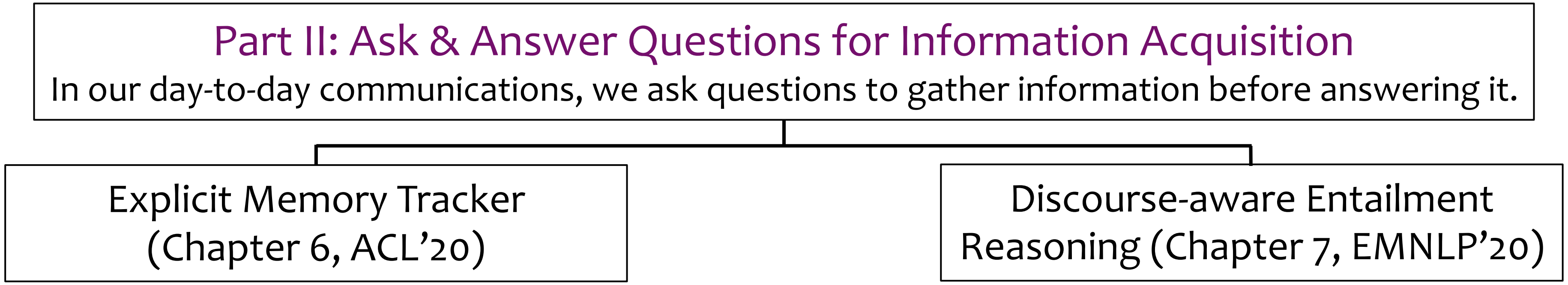


Models	Oracle Question Generation Task			
	Development Set		Cross Validation	
	BLEU1	BLEU4	BLEU1	BLEU4
E^3	52.79±2.87	37.31±2.35	51.75	35.94
E^3+UniLM	57.09±1.70	41.05±1.80	56.94	42.87
EMT	**62.32**±1.62	**47.89**±1.58	**64.48**	**52.40**

Explicit Memory Tracker: Conclusion

- We propose a new approach called Explicit Memory Tracker (EMT) for conversational machine reading
- EMT achieved a new state-of-the-art result on the ShARC CMR challenge
- Code & Models: https://github.com/Yifan-Gao/explicit_memory_tracker

Part II: Information Acquisition

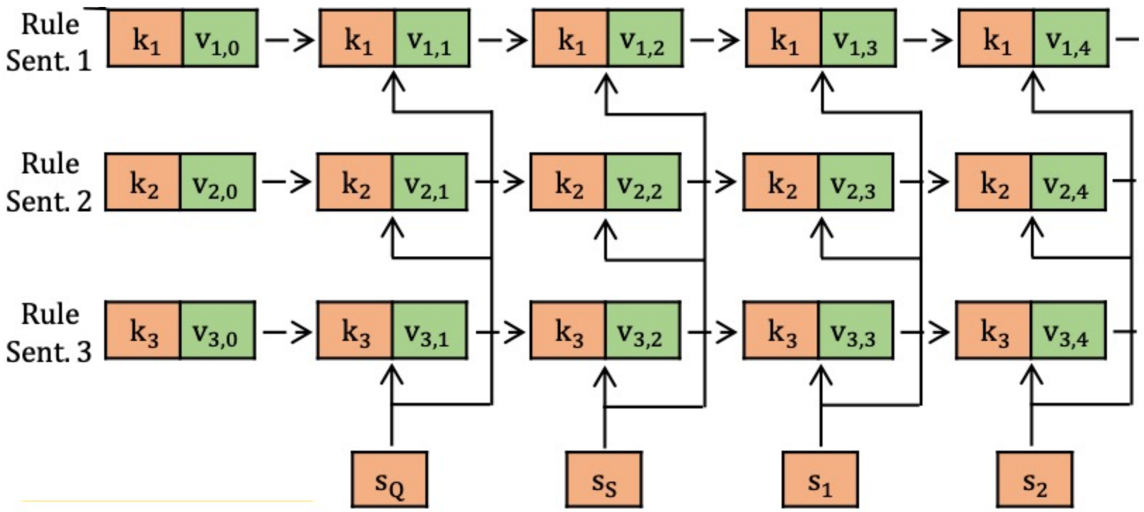


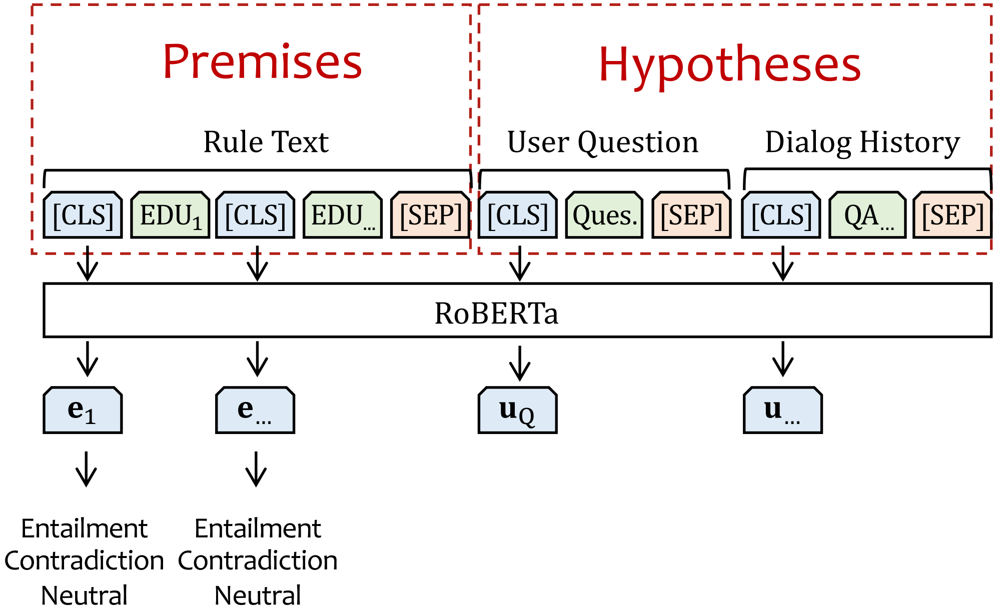
Part II: Ask & Answer Questions for Information Acquisition

In our day-to-day communications, we ask questions to gather information before answering it.

Explicit Memory Tracker (Chapter 6, ACL’20)

Discourse-aware Entailment Reasoning (Chapter 7, EMNLP’20)


Rule Sent. 1
Rule Sent. 2
Rule Sent. 3
k_1 $v_{1,0}$ k_1 $v_{1,1}$ k_1 $v_{1,2}$ k_1 $v_{1,3}$ k_1 $v_{1,4}$
k_2 $v_{2,0}$ k_2 $v_{2,1}$ k_2 $v_{2,2}$ k_2 $v_{2,3}$ k_2 $v_{2,4}$
k_3 $v_{3,0}$ k_3 $v_{3,1}$ k_3 $v_{3,2}$ k_3 $v_{3,3}$ k_3 $v_{3,4}$
s_Q
s_S
s_1
s_2



DISCERN: Challenges in Document Interpretation

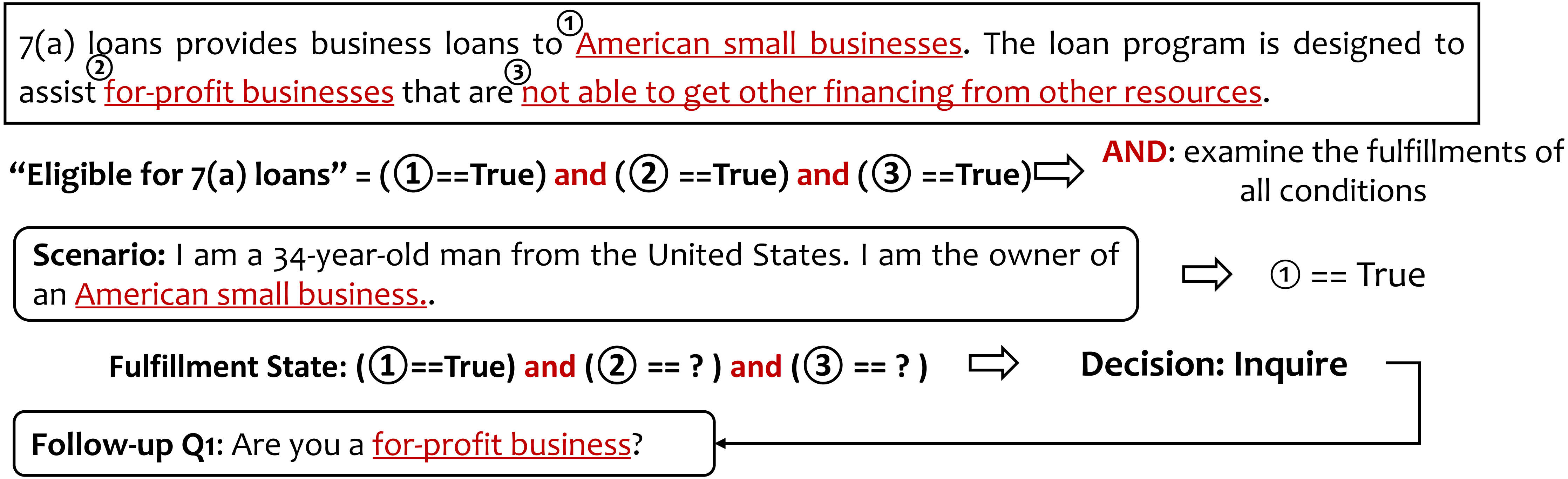
- Document Interpretation
  - Identification of Conditions
  - Determination of Logical Structures

7(a) loans provides business loans to ①American small businesses. The loan program is designed to assist ②for-profit businesses that are ③not able to get other financing from other resources.

"Eligible for 7(a) loans" = (①==True) and (② ==True) and (③ ==True)

DISCERN: Challenges in Dialog Understanding

- Dialog Understanding
  - Track the user’s fulfillment over the conditions
  - Jointly consider the fulfillment states and the logical structure of rules



7(a) loans provides business loans to ① American small businesses. The loan program is designed to assist ② for-profit businesses that are ③ not able to get other financing from other resources.

“Eligible for 7(a) loans” = (①==True) and (② ==True) and (③ ==True) ⇨ **AND**: examine the fulfillments of all conditions

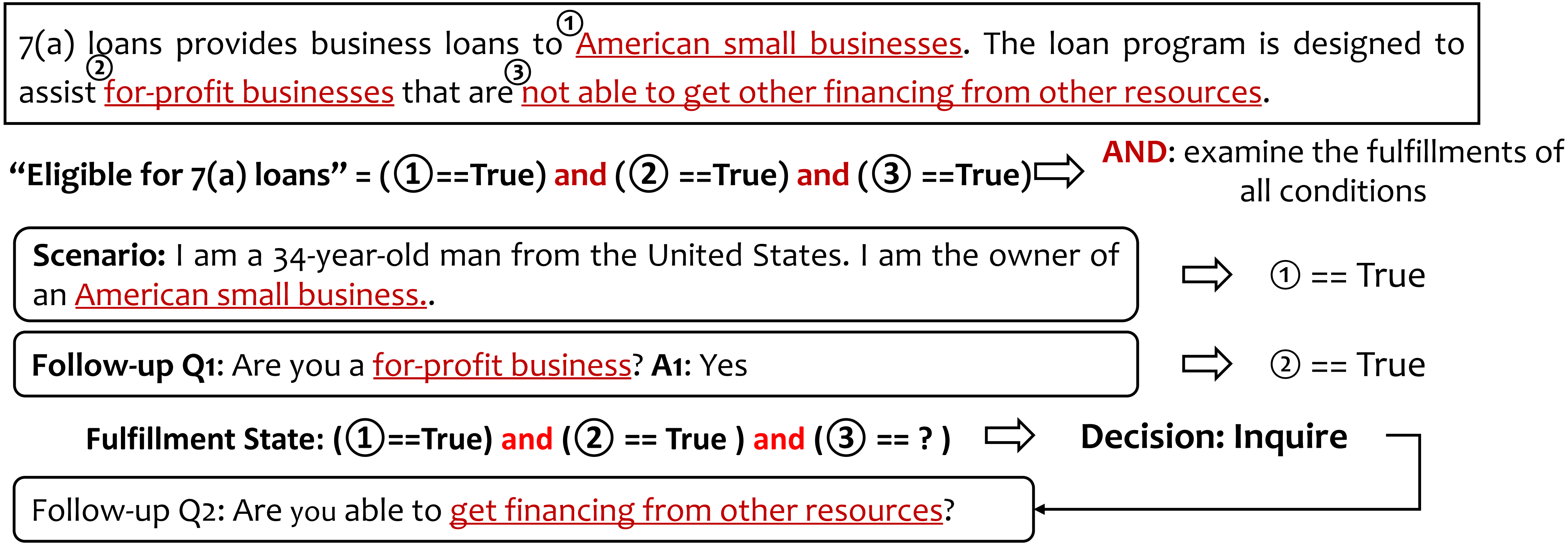
Scenario: I am a 34-year-old man from the United States. I am the owner of an American small business.. ⇨ ① == True

Fulfillment State: (①==True) and (② == ?) and (③ == ?) ⇨ **Decision: Inquire**

Follow-up Q1: Are you a for-profit business?

DISCERN: Challenges in Dialog Understanding

- Dialog Understanding
  - Track the user's fulfillment over the conditions
  - Jointly consider the fulfillment states and the logical structure of rules



7(a) loans provides business loans to ① American small businesses. The loan program is designed to assist ② for-profit businesses that are ③ not able to get other financing from other resources.

"Eligible for 7(a) loans" = (①==True) and (② ==True) and (③ ==True) ⇨ **AND**: examine the fulfillments of all conditions

Scenario: I am a 34-year-old man from the United States. I am the owner of an American small business.. ⇨ ① == True

Follow-up Q1: Are you a for-profit business? **A1:** Yes ⇨ ② == True

Fulfillment State: (①==True) and (② == True) and (③ == ?) ⇨ **Decision: Inquire**

Follow-up Q2: Are you able to get financing from other resources?

DISCERN: Challenges in Dialog Understanding

- Dialog Understanding
  - Track the user’s fulfillment over the conditions
  - Jointly consider the fulfillment states and the logical structure of rules

7(a) loans provides business loans to ①American small businesses. The loan program is designed to assist ②for-profit businesses that are ③not able to get other financing from other resources.

“Eligible for 7(a) loans” = (①==True) and (② ==True) and (③ ==True) ⇨ **AND**: examine the fulfillments of all conditions

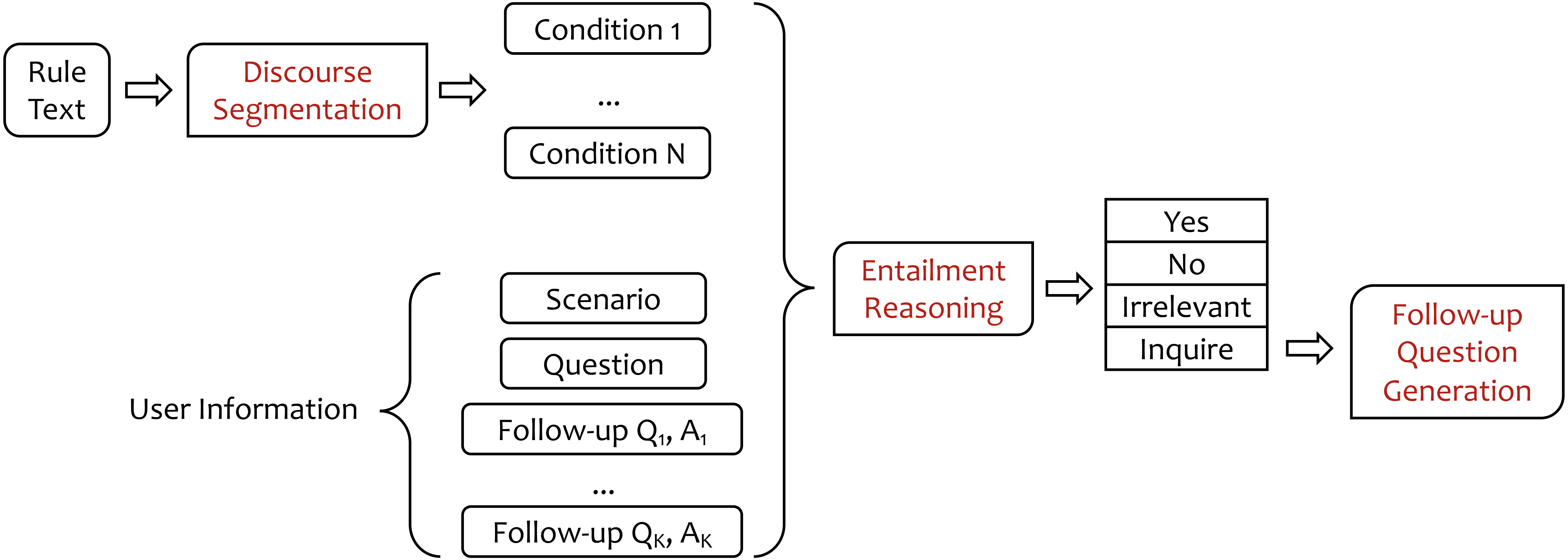
Scenario: I am a 34-year-old man from the United States. I am the owner of an American small business.. ⇨ ① == True

Follow-up Q1: Are you a for-profit business? **A1:** Yes ⇨ ② == True

Follow-up Q2: Are you able to get financing from other resources? **A2:** No ⇨ ③ == True

Fulfillment State: (①==True) and (② == True) and (③ == True) ⇨ **Decision: Yes** (You can apply the loan.)

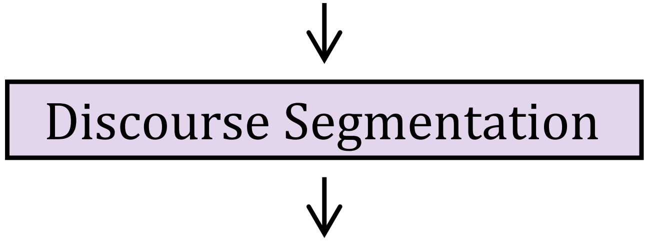
DISCERN: Discourse-Aware Entailment Reasoning Network for Conversational Machine Reading



DISCERN: Rule Segmentation

- Goal: Parse the rule into individual conditions for entailment reasoning
- Challenges: One rule sentence may contain several in-line conditions
- Discourse Segmentation
  - In the rhetorical structure theory of discourse parsing (Mann and Thompson, 1988), texts are split into clause-like units called Elementary Discourse Units (EDUs)

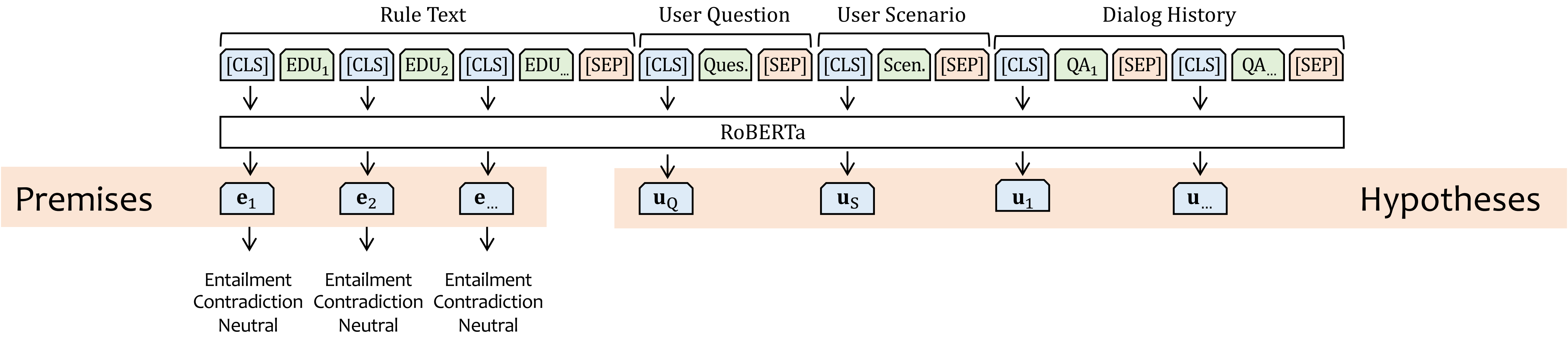
Rule Text: If a worker has taken more leave than they're entitled to, their employer must not take money from their final pay unless it's been agreed beforehand in writing.



Discourse Segmentation

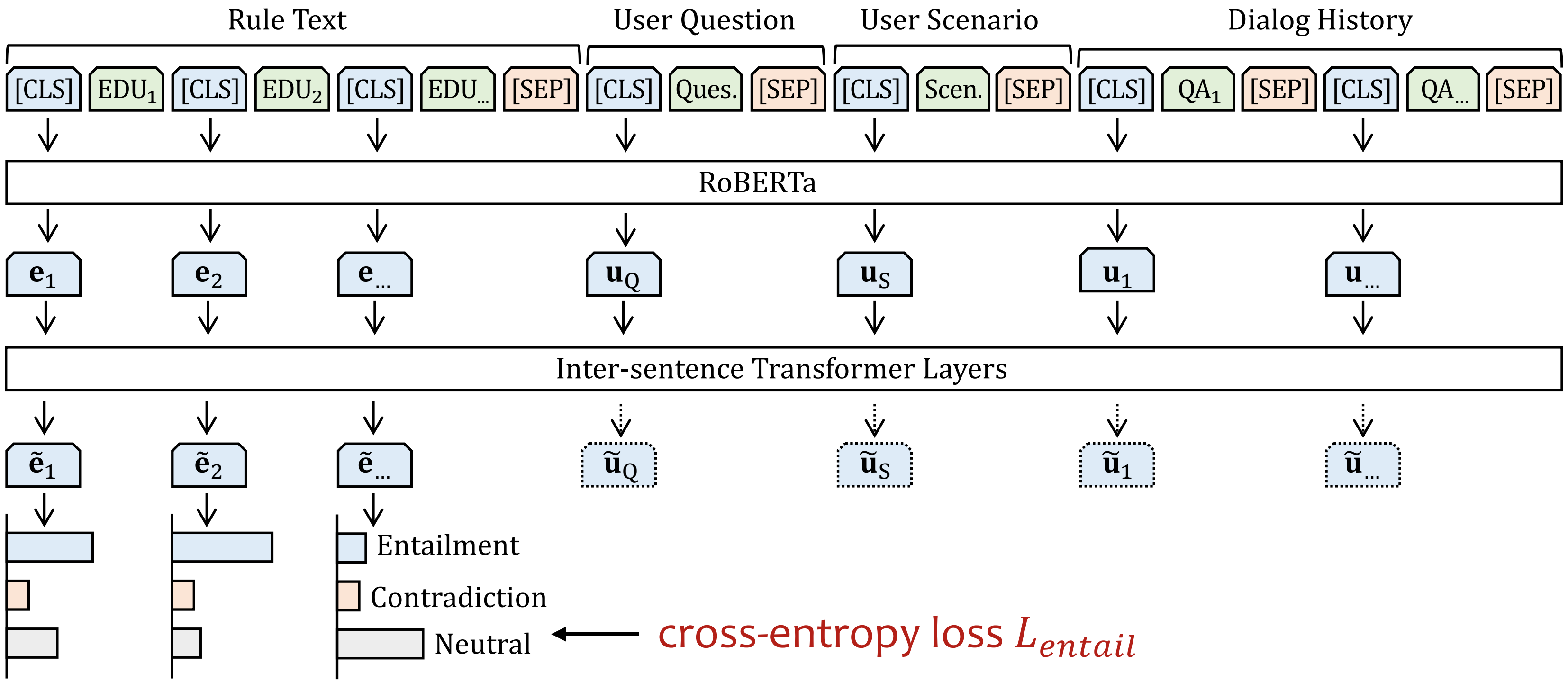
[If a worker has taken more leave than they're entitled to,]$_{\text{EDU1}}$ [their employer must not take money from their final pay]$_{\text{EDU2}}$ [unless it's been agreed beforehand in writing.]$_{\text{EDU3}}$

DISCERN: Entailment Prediction



Multi-Sentence Entailment Prediction

DISCERN: Entailment Prediction

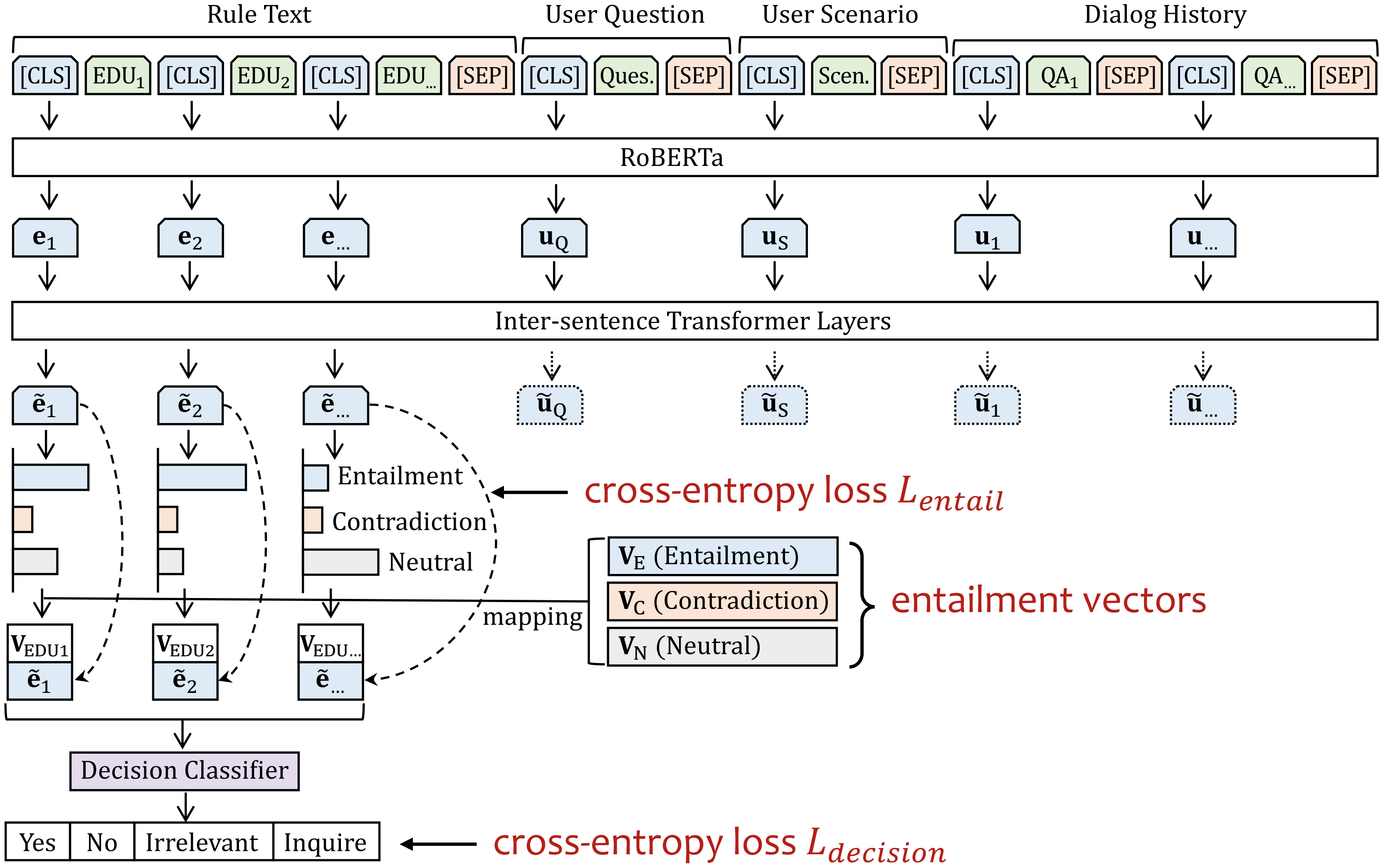


cross-entropy loss L_{entail}

User Info. → Rule

Rule → Logical Structure

DISCERN: Decision Making

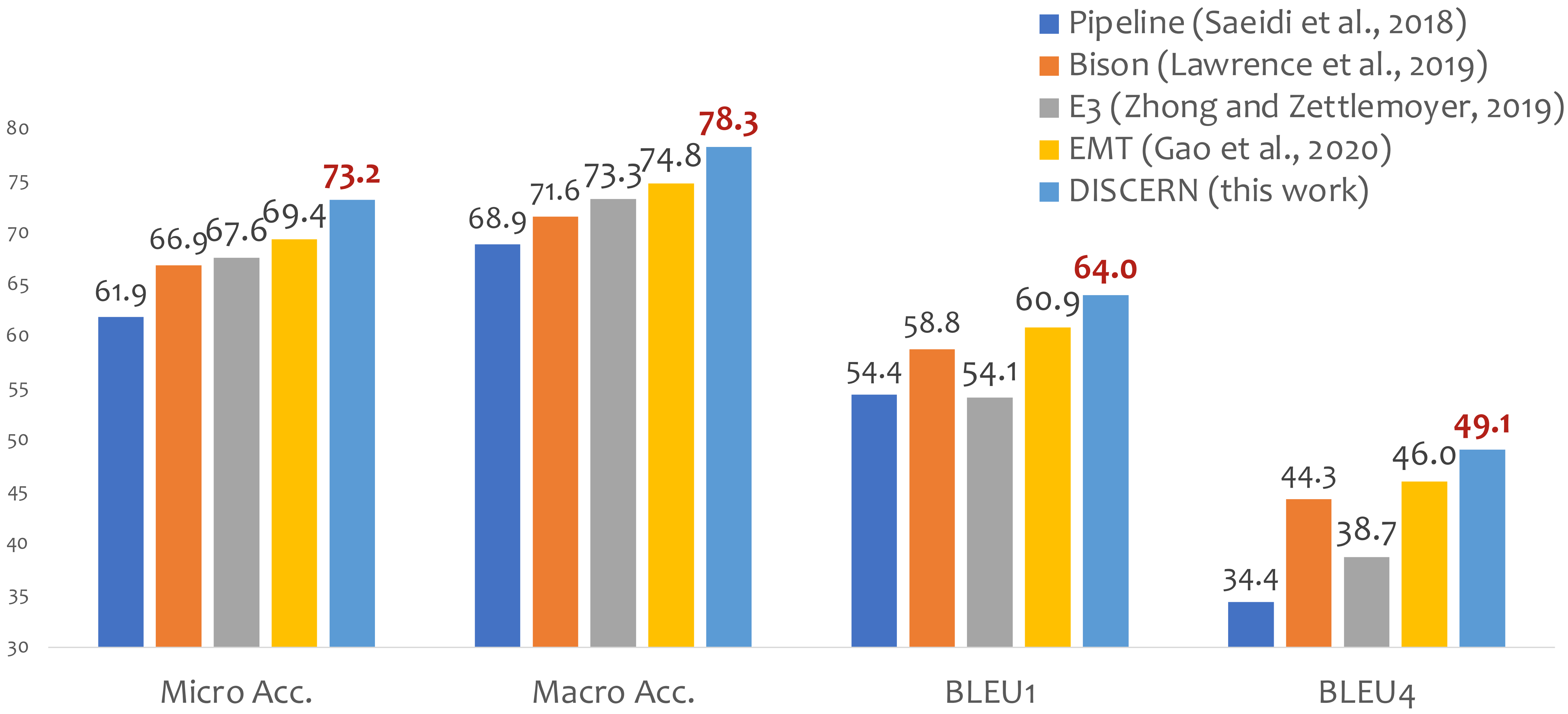


cross-entropy loss L_{entail}

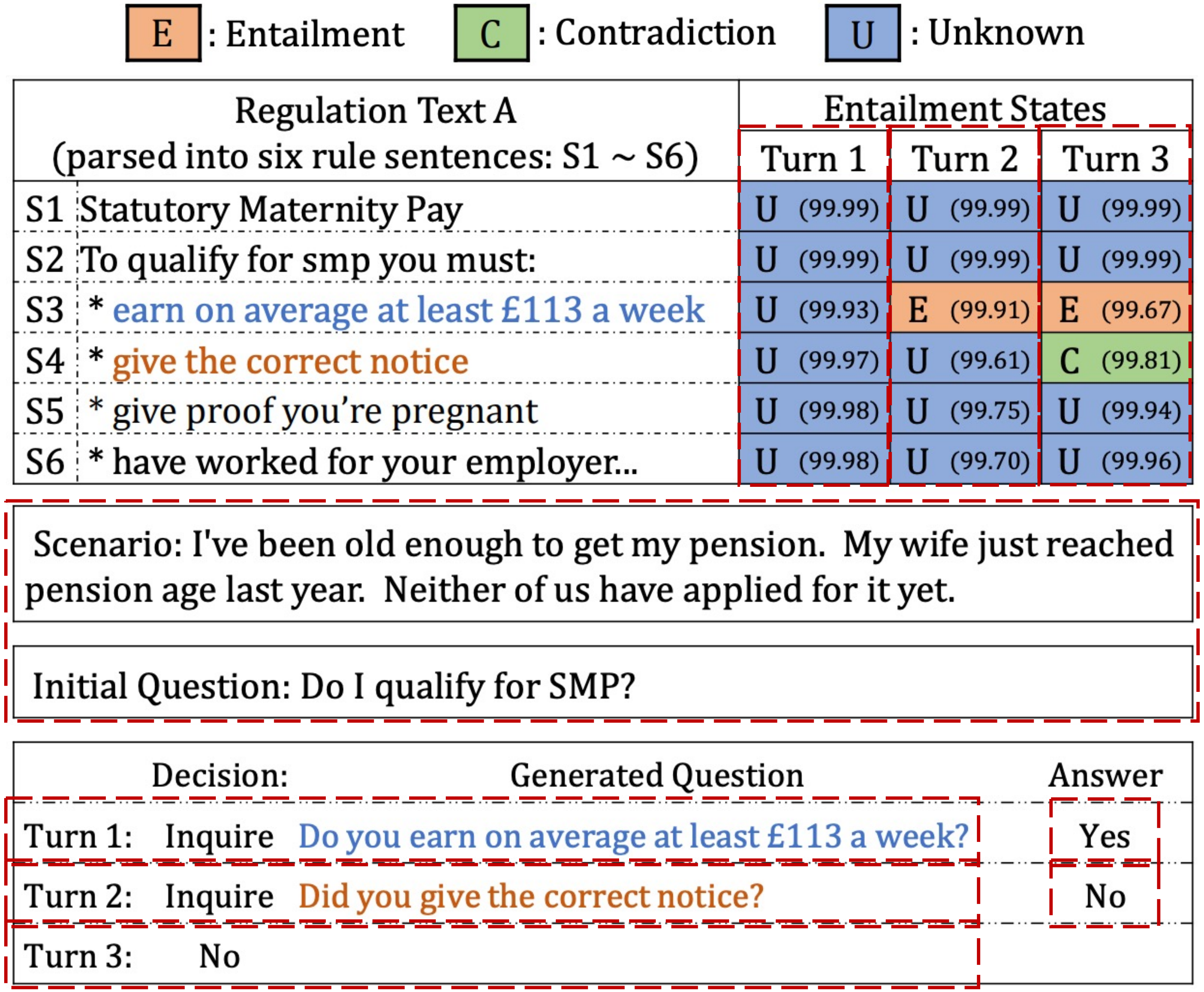
entailment vectors

cross-entropy loss $L_{decision}$

DISCERN: Main Results



DISCERN: Interpretability



E : Entailment   C : Contradiction   U : Unknown

Regulation Text A (parsed into six rule sentences: S1 ~ S6)		Entailment States Turn 1	 Turn 2	 Turn 3
S1	Statutory Maternity Pay	U (99.99)	U (99.99)	U (99.99)
S2	To qualify for smp you must:	U (99.99)	U (99.99)	U (99.99)
S3	* earn on average at least £113 a week	U (99.93)	E (99.91)	E (99.67)
S4	* give the correct notice	U (99.97)	U (99.61)	C (99.81)
S5	* give proof you're pregnant	U (99.98)	U (99.75)	U (99.94)
S6	* have worked for your employer...	U (99.98)	U (99.70)	U (99.96)

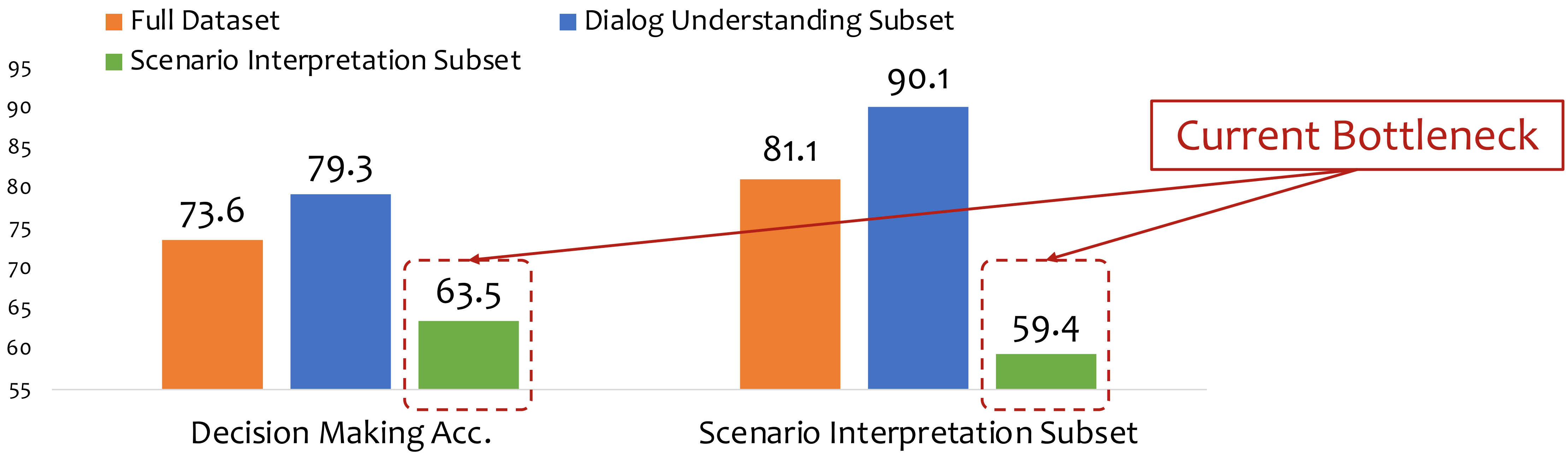
Scenario: I've been old enough to get my pension. My wife just reached pension age last year. Neither of us have applied for it yet.

Initial Question: Do I qualify for SMP?

	Decision:	Generated Question	Answer
Turn 1:	Inquire	Do you earn on average at least £113 a week?	Yes
Turn 2:	Inquire	Did you give the correct notice?	No
Turn 3:	No		

DISCERN: How Far Has the Problem Been Solved?

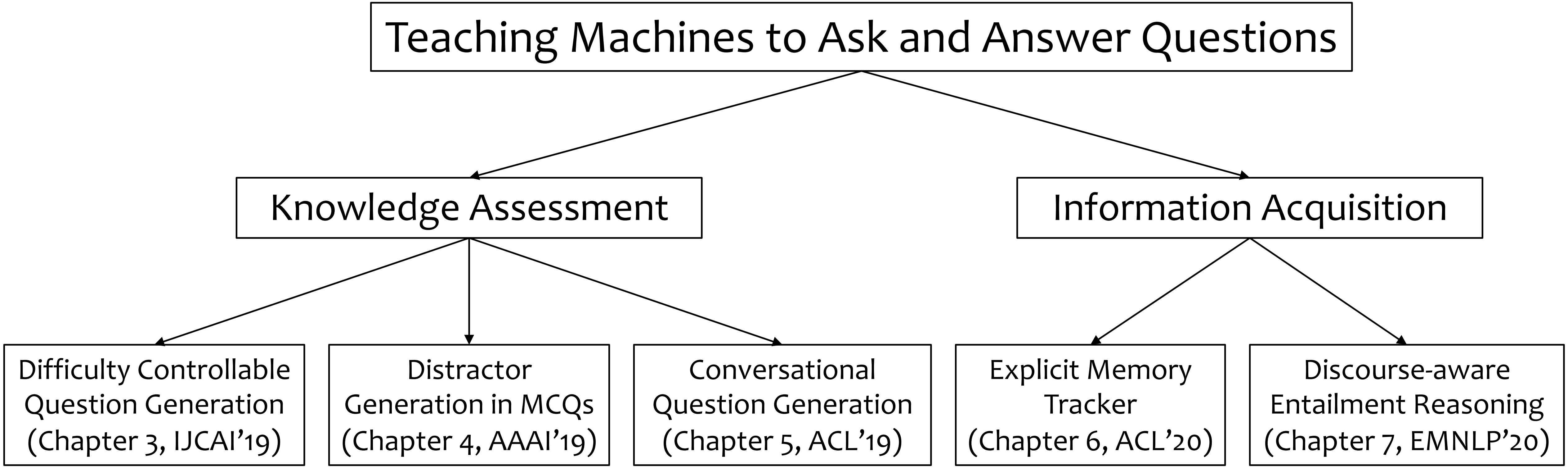
- Disentangle the challenge between scenario interpretation and dialog understanding
  - Full Dataset: Full development set of ShARC
  - Dialog Understanding Subset: User information only contains dialog history
  - Scenario Interpretation Subset: User information only contains user scenario



DISCERN: Conclusion

- We present DISCERN, a system that does discourse-aware entailment reasoning for conversational machine reading
- Results on the ShARC benchmark shows that DISCERN outperforms existing methods by a large margin
- We also conduct comprehensive analyses to unveil the limitations of DISCERN and challenges for ShARC
- Code & Models: https://github.com/Yifan-Gao/DISCERN

Can Machine Be Engaged in Question Asking and Answering?



Teaching Machines to Ask and Answer Questions

Knowledge Assessment

Information Acquisition

Difficulty Controllable Question Generation (Chapter 3, IJCAI'19)

Distractor Generation in MCQs (Chapter 4, AAAI'19)

Conversational Question Generation (Chapter 5, ACL'19)

Explicit Memory Tracker (Chapter 6, ACL'20)

Discourse-aware Entailment Reasoning (Chapter 7, EMNLP'20)

Yes! Machines can be engaged in question asking and answering!

Future Work (1)

- Open-Domain Information Acquisition

(Known before answering questions)

Knowledge: 7(a) loans provides business loans to American small businesses. The loan program is designed to assist for-profit businesses that are not able to get other financing from other resources.

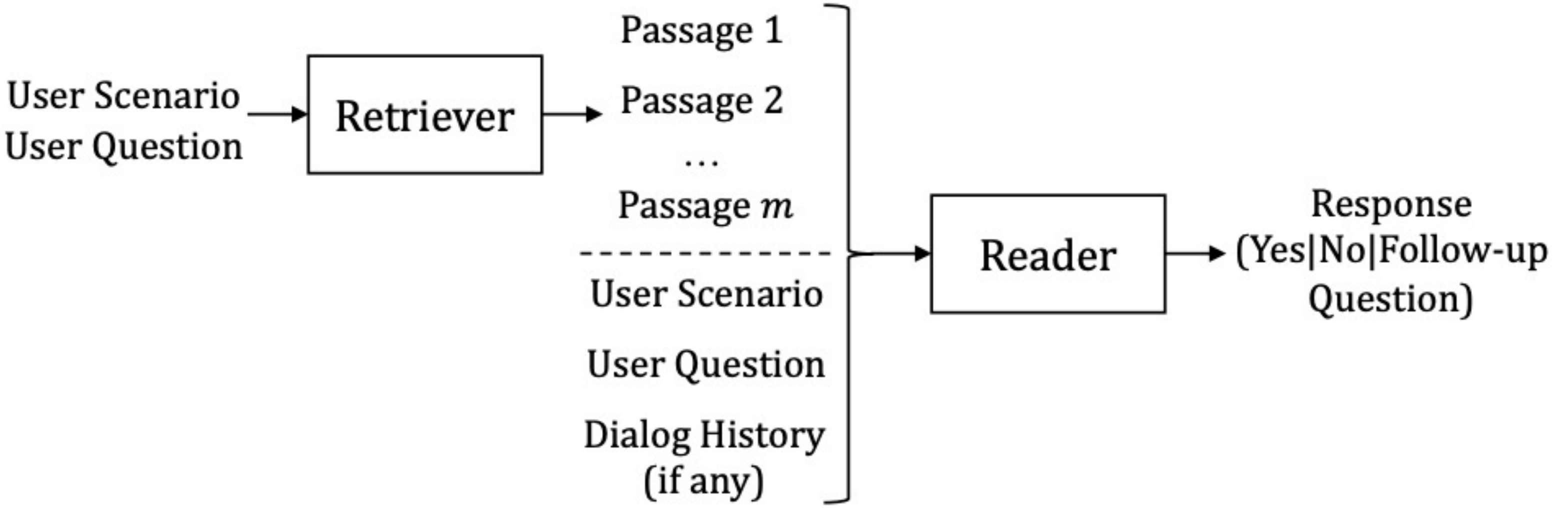
Question: I am a 34-year-old man from the United States. I am the owner of an American small business. Is the 7(a) Loan Program for me?

1. Understand the text
2. Ask questions to gather more personal information
3. Assess the eligibility
4. Answer

Future Work (1)

- Open-Domain Information Acquisition

- Ideally, a system should retrieve relevant rule text, and conduct the machine reading to answer questions



Future Work (2)

In our day-to-day communications, we are not always passively answering questions, especially when

(Appear in >50% Google search queries!)

questions are ambiguous so we need to make clarification

Question: What's the most points scored in an NBA game?

You may want to ask:

- What's the most points scored in an NBA game by combined team?
- What's the most points scored in an NBA game by a single team?
- What's the most points scored in an NBA game by an individual?

Publications (as first author)

Knowledge Assessment

- Difficulty Controllable Question Generation for Reading Comprehension. **IJCAI 2019.**
- Generating Distractors for Reading Comprehension Questions from Real Examinations. **AAAI 2019.** *Code & Dataset: https://github.com/Yifan-Gao/Distractor-Generation-RACE*
- Interconnected Question Generation with Coreference Alignment and Conversation Flow Modeling. **ACL 2019**. *Code & Models: https://github.com/Yifan-Gao/conversational-QG*

Information Acquisition

- Explicit Memory Tracker with Coarse-to-Fine Reasoning for Conversational Machine Reading. **ACL 2020**. *Code & Models: https://github.com/Yifan-Gao/explicit_memory_tracker*
- DISCERN: Discourse-Aware Entailment Reasoning Network for Conversational Machine Reading. **EMNLP 2020**. *Code & Models: https://github.com/Yifan-Gao/DISCERN*

Others

- Dialogue Generation on Infrequent Sentence Functions via Structured Meta-Learning. **Findings of ACL: EMNLP 2020.**

Reference

[1] Wang, Wenhui, et al. “Gated Self-Matching Networks for Reading Comprehension and Question Answering.” ACL, 2017.

[2] Seo, Minjoon, et al. “Bidirectional Attention Flow for Machine Comprehension.” ICLR, 2017.

[3] Zhou, Qingyu, et al. “Neural question generation from text: A preliminary study.” NLPCC. Springer, Cham, 2017.

[4] Lai, Guokun, et al. “RACE: Large-scale ReAding Comprehension Dataset From Examinations.” EMNLP, 2017.

[5] Papineni, Kishore, et al. “BLEU: a Method for Automatic Evaluation of Machine Translation.” ACL 2002.

[6] Lin, Chin-Yew. “ROUGE: a Package for Automatic Evaluation of Summaries.” WAS 2004.

[7] Reddy, Siva, et al. “CoQA: A Conversational Question Answering Challenge.” Transactions of ACL, 2019.

[8] Du, Xinya, and Claire Cardie. “Harvesting Paragraph-Level Question-Answer Pairs from Wikipedia.” ACL 2018.

[9] Saeidi, Marzieh, et al. “Interpretation of Natural Language Rules in Conversational Machine Reading.” EMNLP, 2018.

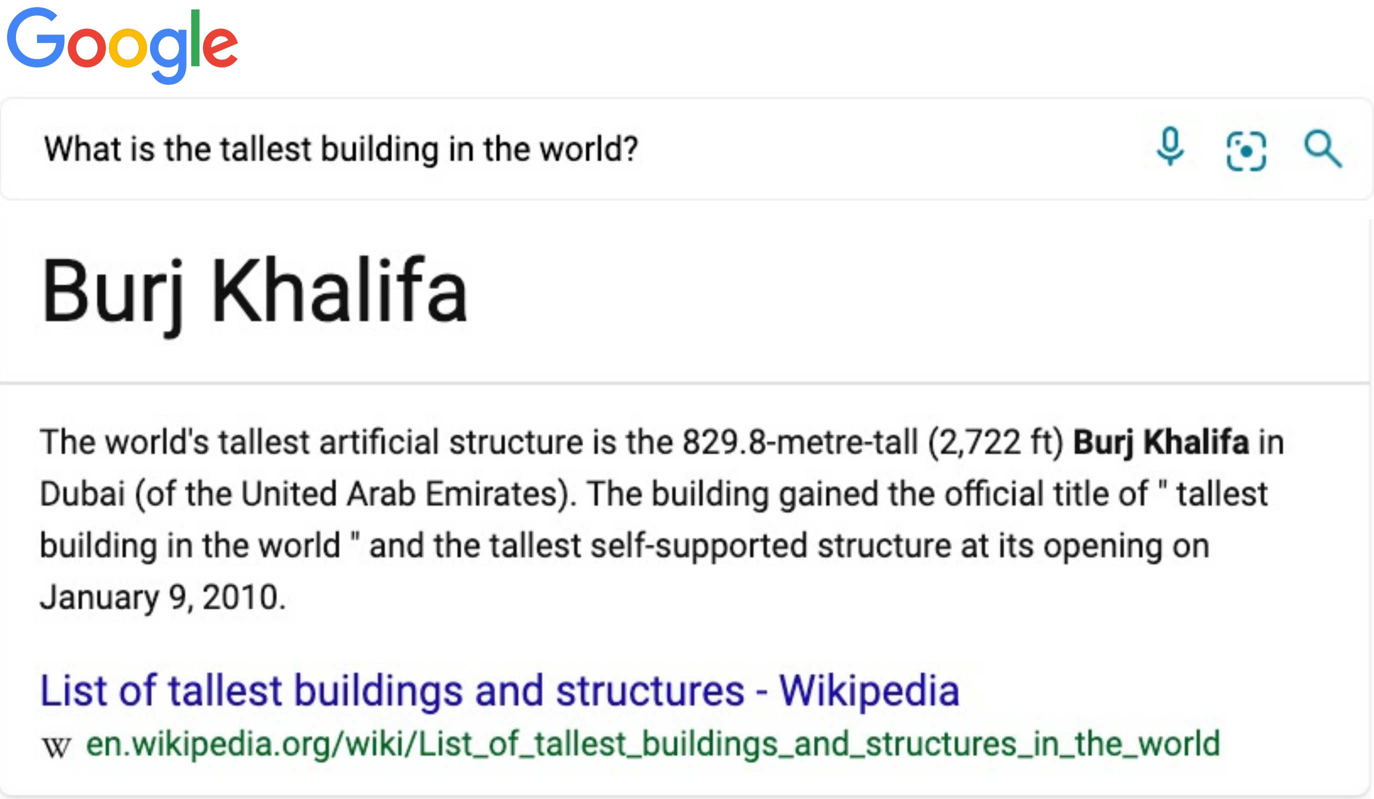
[10] Dong, Li, et al. “Unified Language Model Pre-training for Natural Language Understanding and Generation.” NeurIPS, 2019.

Thank you!

Backup Slides

Introduction: Question Answering Applications

- Personal Assistants
- Smart Home and Entertainment
- Search engines



Google

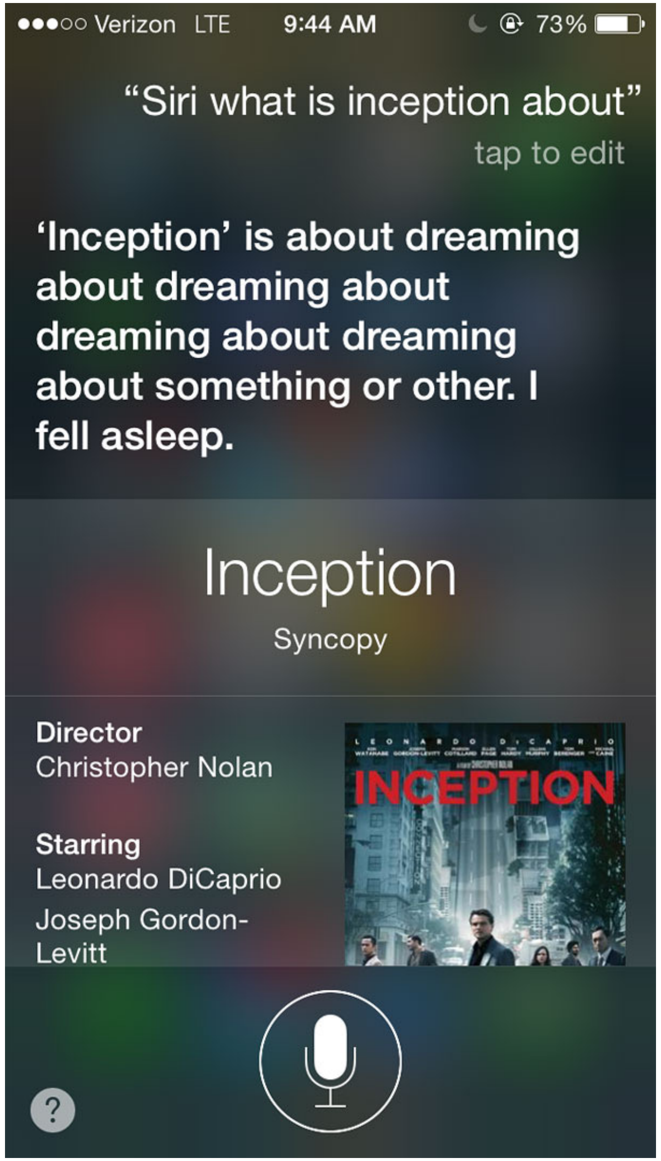
What is the tallest building in the world?

Burj Khalifa

The world's tallest artificial structure is the 829.8-metre-tall (2,722 ft) **Burj Khalifa** in Dubai (of the United Arab Emirates). The building gained the official title of " tallest building in the world " and the tallest self-supported structure at its opening on January 9, 2010.

List of tallest buildings and structures - Wikipedia

W en.wikipedia.org/wiki/List_of_tallest_buildings_and_structures_in_the_world



"Siri what is inception about"

tap to edit

'Inception' is about dreaming about dreaming about dreaming about dreaming about something or other. I fell asleep.

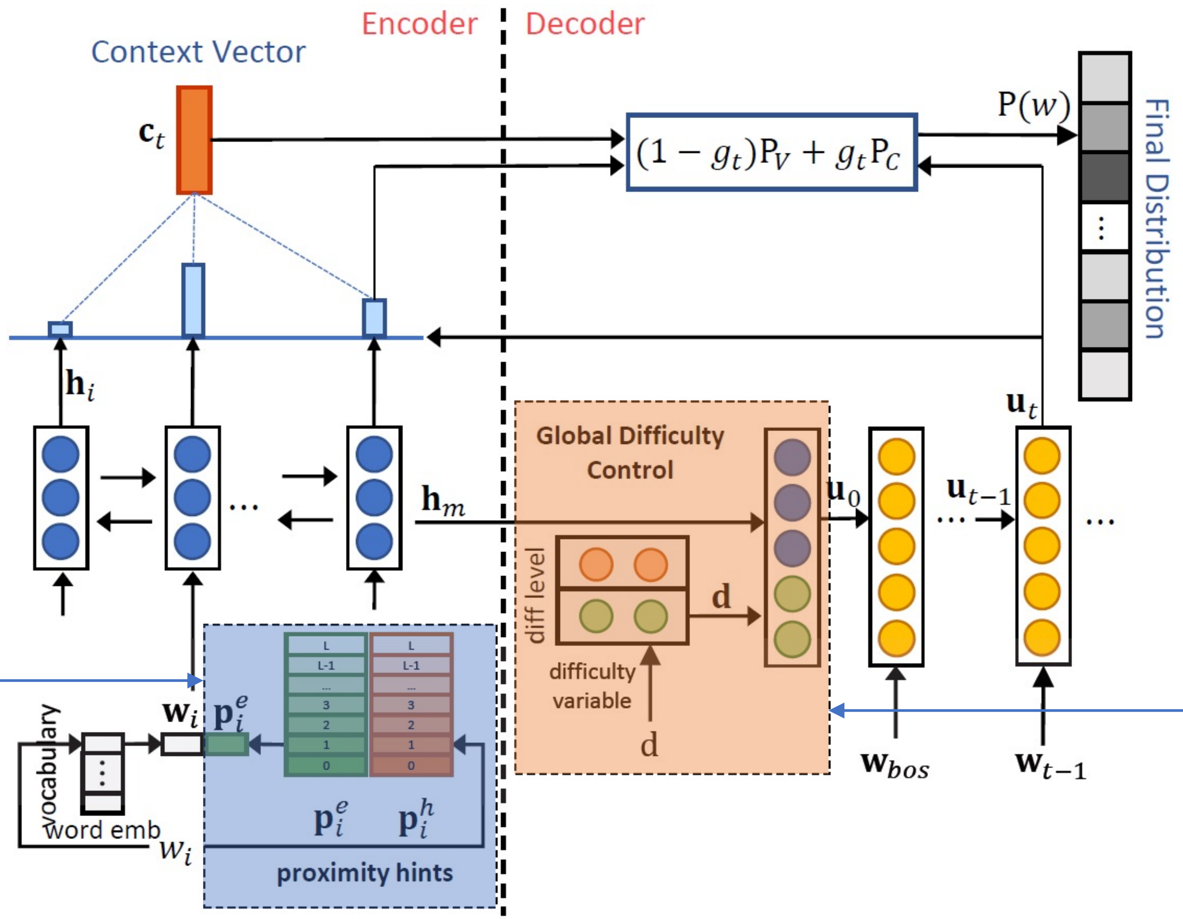
Inception

Syncopy

Director
Christopher Nolan

Starring
Leonardo DiCaprio
Joseph Gordon-Levitt

Difficulty Controllable QG: Proposed Framework



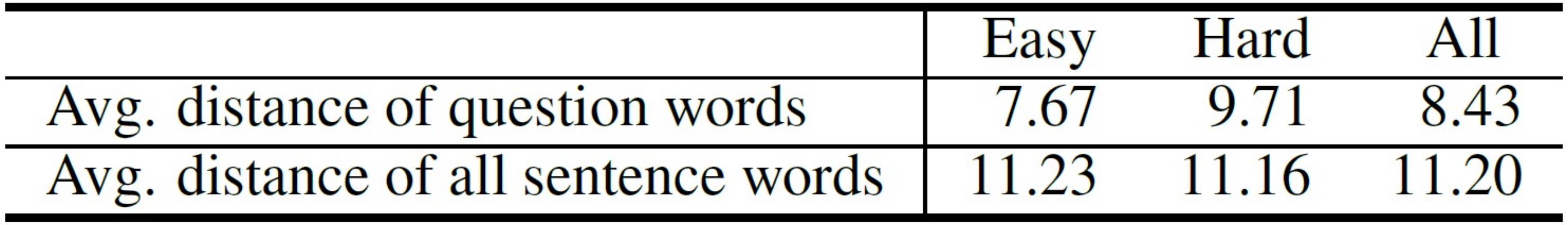
First Intuition

Second Intuition

Difficulty Controllable QG: Exploring Proximity Hints

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

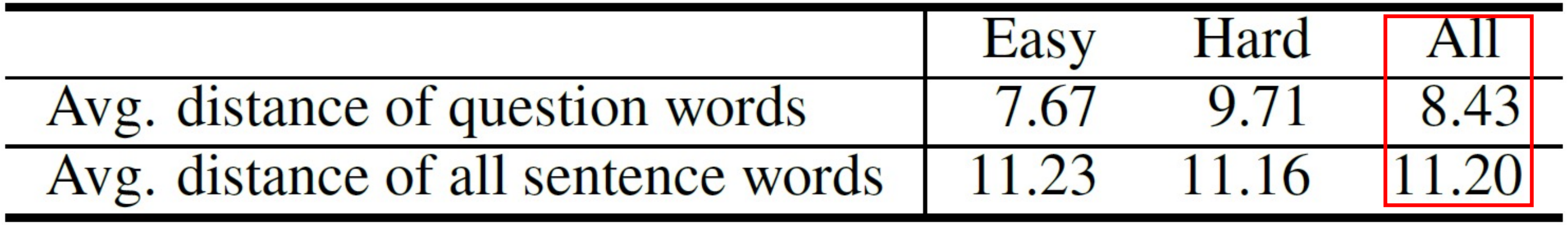
Question: What is the **atomic number** of the **element oxygen**?
Sentence: **Oxygen** is a chemical **element** with symbol O and **atomic number** 8.
Distance: 11 7 2 1



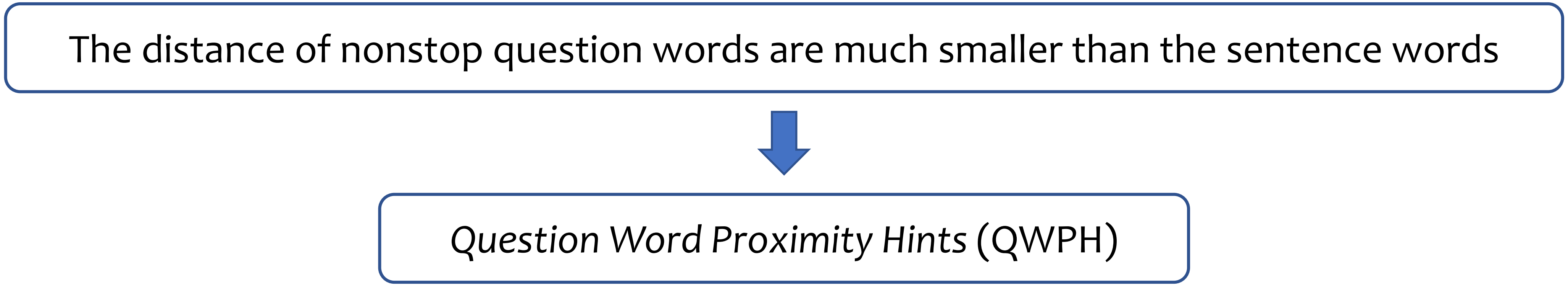
	Easy	Hard	All
Avg. distance of question words	7.67	9.71	8.43
Avg. distance of all sentence words	11.23	11.16	11.20

Difficulty Controllable QG: 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

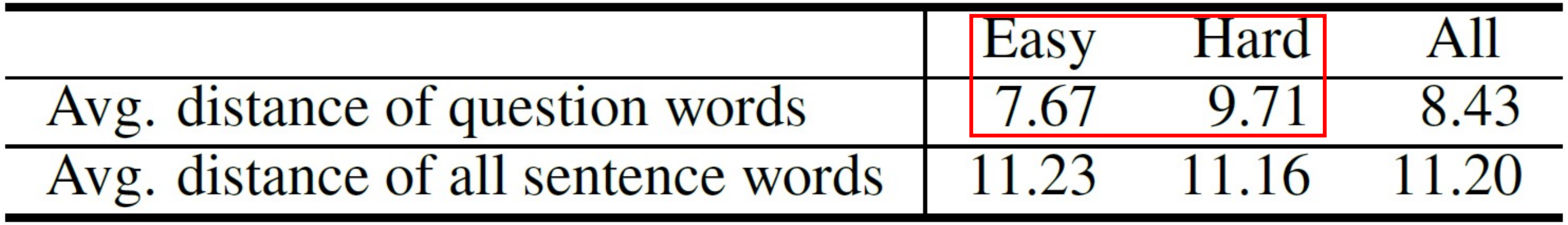


The distance of nonstop question words are much smaller than the sentence words

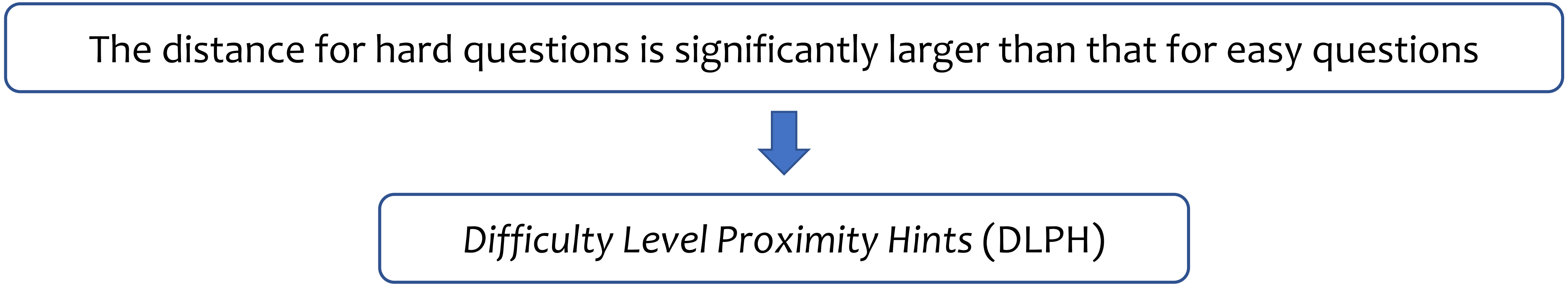
Question Word Proximity Hints (QWPH)

Difficulty Controllable QG: 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)

Difficulty Controllable QG: Evaluation Metrics

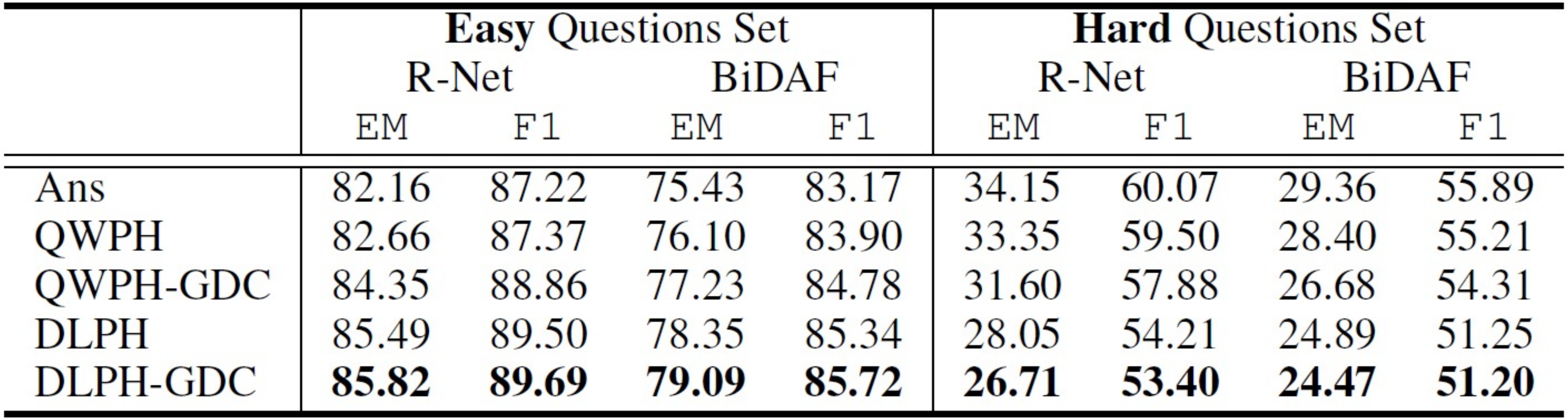
- Automatic Evaluation
  - Employ reading comprehension systems to evaluate the difficulty of generated questions
  - N-gram based similarity: BLEU, ROUGE, METEOR
- Human Evaluation
  - Fluency, Difficulty, Relevance

Difficulty Controllable QG: 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 Controllable QG: Difficulty Control Results

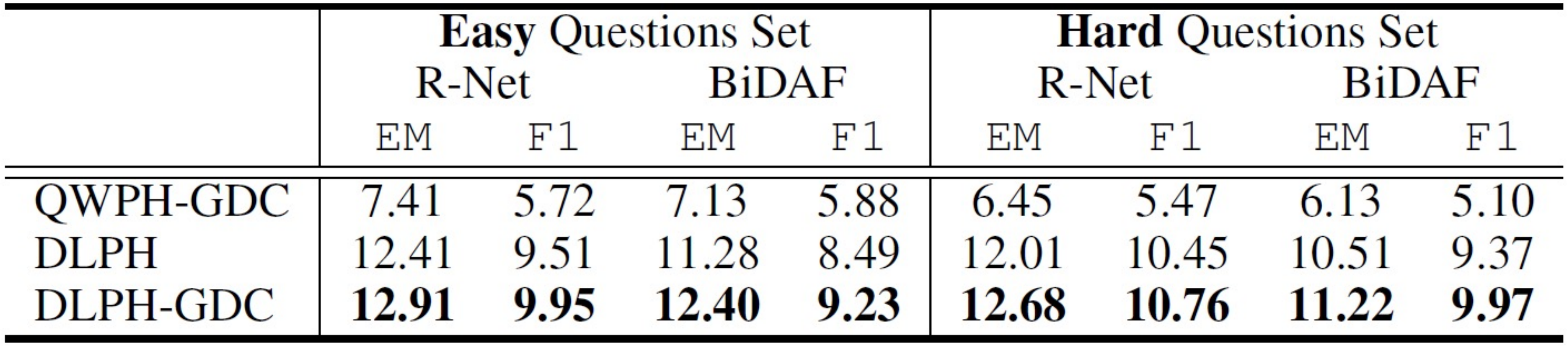
- **Metric:** Employ reading comprehension systems (R-Net, BiDAF) to evaluate the difficulty of generated questions
- For easy questions, higher score indicates better difficulty-control, while for hard questions, lower indicates better.



	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 Controllable QG: 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.

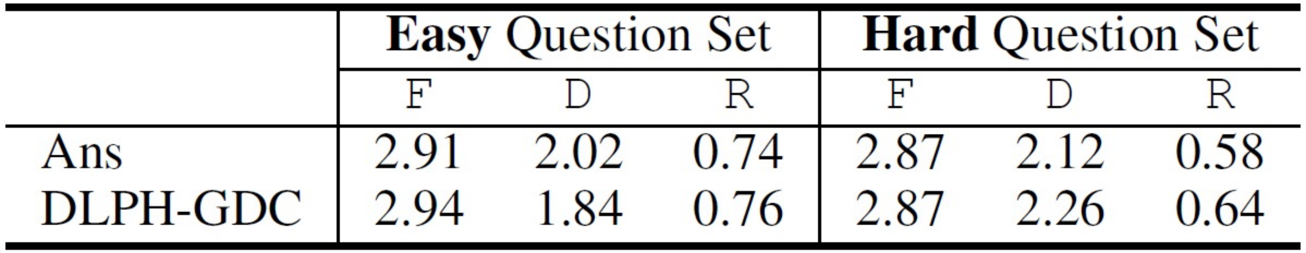


	Easy Questions Set				**Hard** Questions Set			
	R-Net		BiDAF		R-Net		BiDAF	
	EM	F1	EM	F1	EM	F1	EM	F1
QWPH-GDC	7.41	5.72	7.13	5.88	6.45	5.47	6.13	5.10
DLPH	12.41	9.51	11.28	8.49	12.01	10.45	10.51	9.37
DLPH-GDC	**12.91**	**9.95**	**12.40**	**9.23**	**12.68**	**10.76**	**11.22**	**9.97**

(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.

Difficulty Controllable QG: Human Evaluation

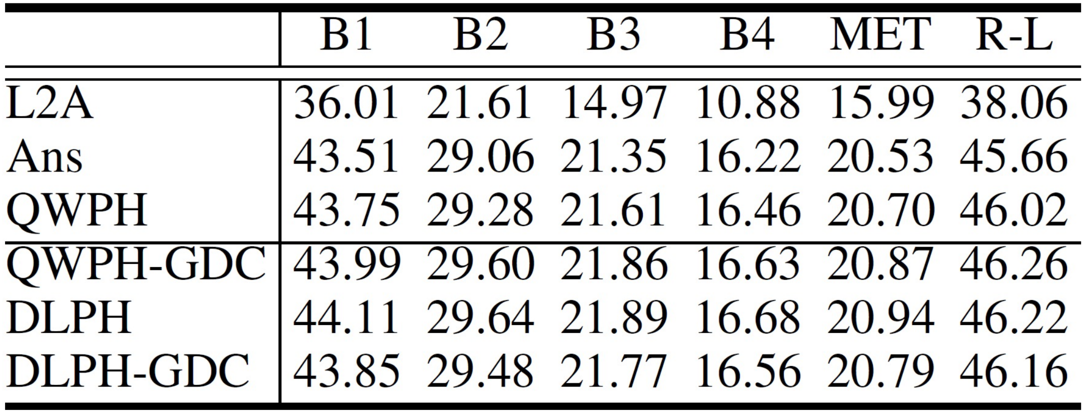


	Easy Question Set			**Hard** Question Set		
	F	D	R	F	D	R
Ans	2.91	2.02	0.74	2.87	2.12	0.58
DLPH-GDC	2.94	1.84	0.76	2.87	2.26	0.64

- **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

Difficulty Controllable QG: Automatic Evaluation of Question Quality

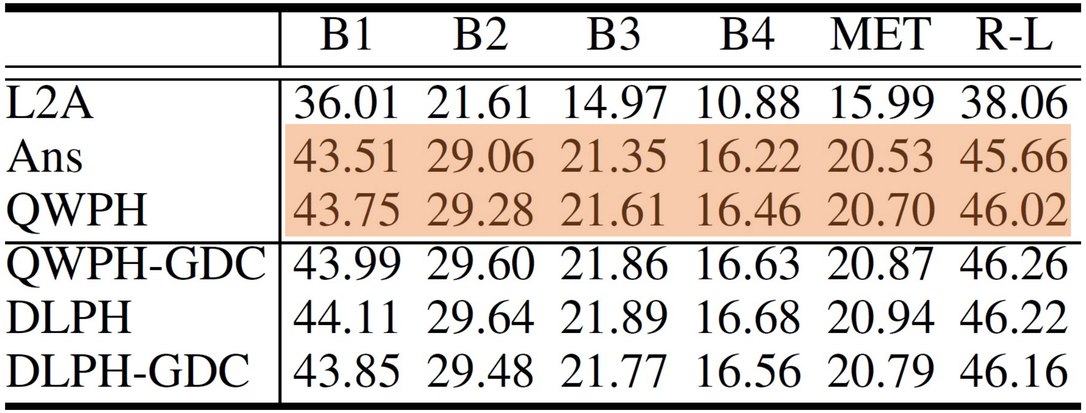
- We evaluate the similarity of generated questions with the ground truth questions by feeding the ground truth difficulty labels
- Metrics: BLEU (B), METEOR (MET), ROUGE-L (R-L)



	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

Difficulty Controllable QG: 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)

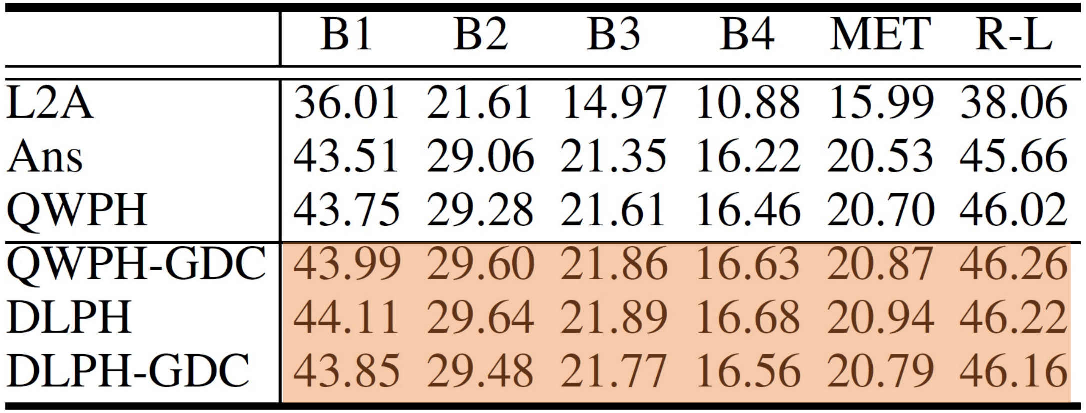


	B1	B2	B3	B4	MET	R-L
L2A	36.01	21.61	14.97	10.88	15.99	38.06
Ans	43.51	29.06	21.35	16.22	20.53	45.66
QWPH	43.75	29.28	21.61	16.46	20.70	46.02
QWPH-GDC	43.99	29.60	21.86	16.63	20.87	46.26
DLPH	44.11	29.64	21.89	16.68	20.94	46.22
DLPH-GDC	43.85	29.48	21.77	16.56	20.79	46.16

Further distinguish the different distance help generate better questions

Difficulty Controllable QG: 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)

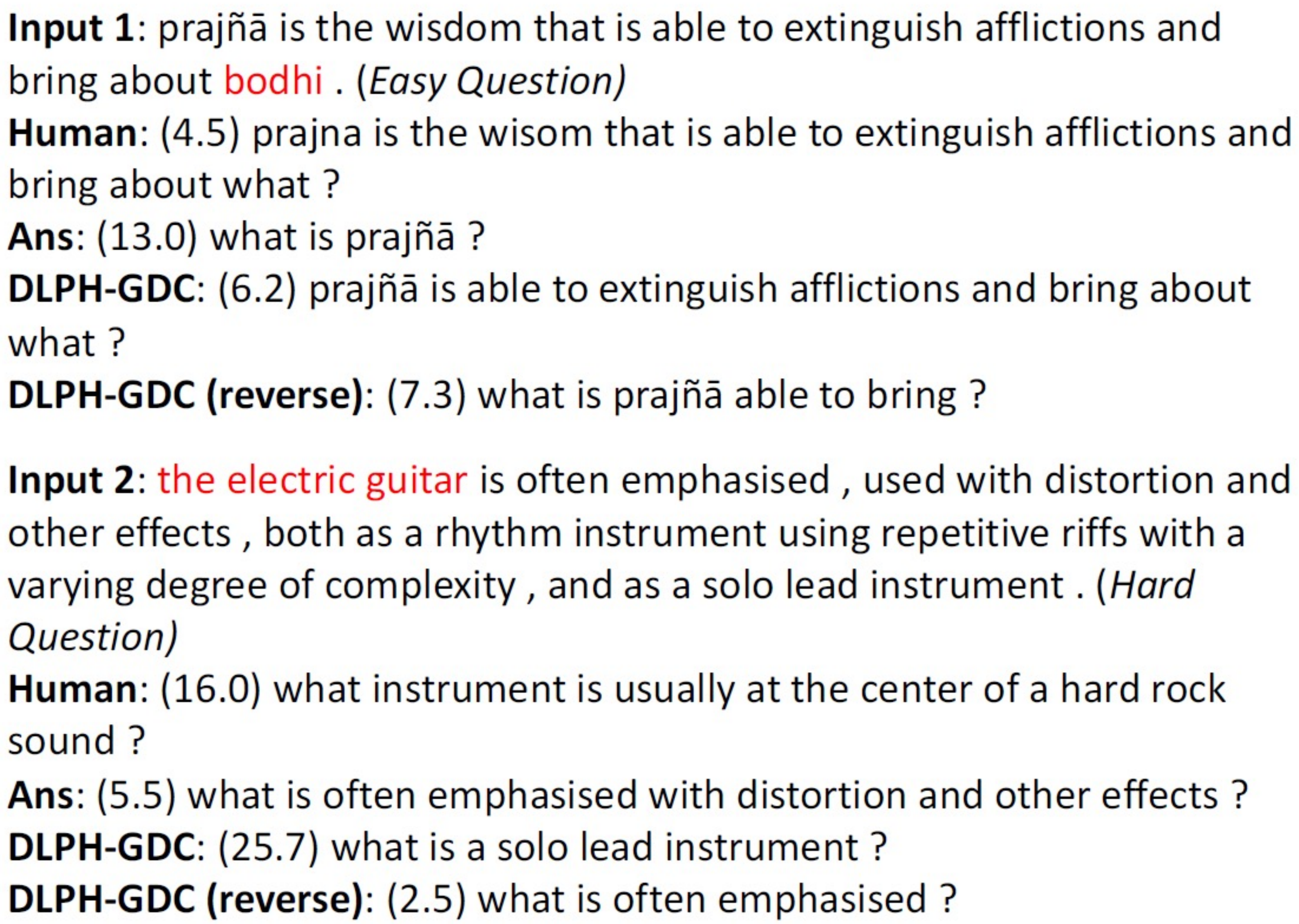


	B1	B2	B3	B4	MET	R-L
L2A	36.01	21.61	14.97	10.88	15.99	38.06
Ans	43.51	29.06	21.35	16.22	20.53	45.66
QWPH	43.75	29.28	21.61	16.46	20.70	46.02
QWPH-GDC	43.99	29.60	21.86	16.63	20.87	46.26
DLPH	44.11	29.64	21.89	16.68	20.94	46.22
DLPH-GDC	43.85	29.48	21.77	16.56	20.79	46.16

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

Difficulty Controllable QG: Case Study

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



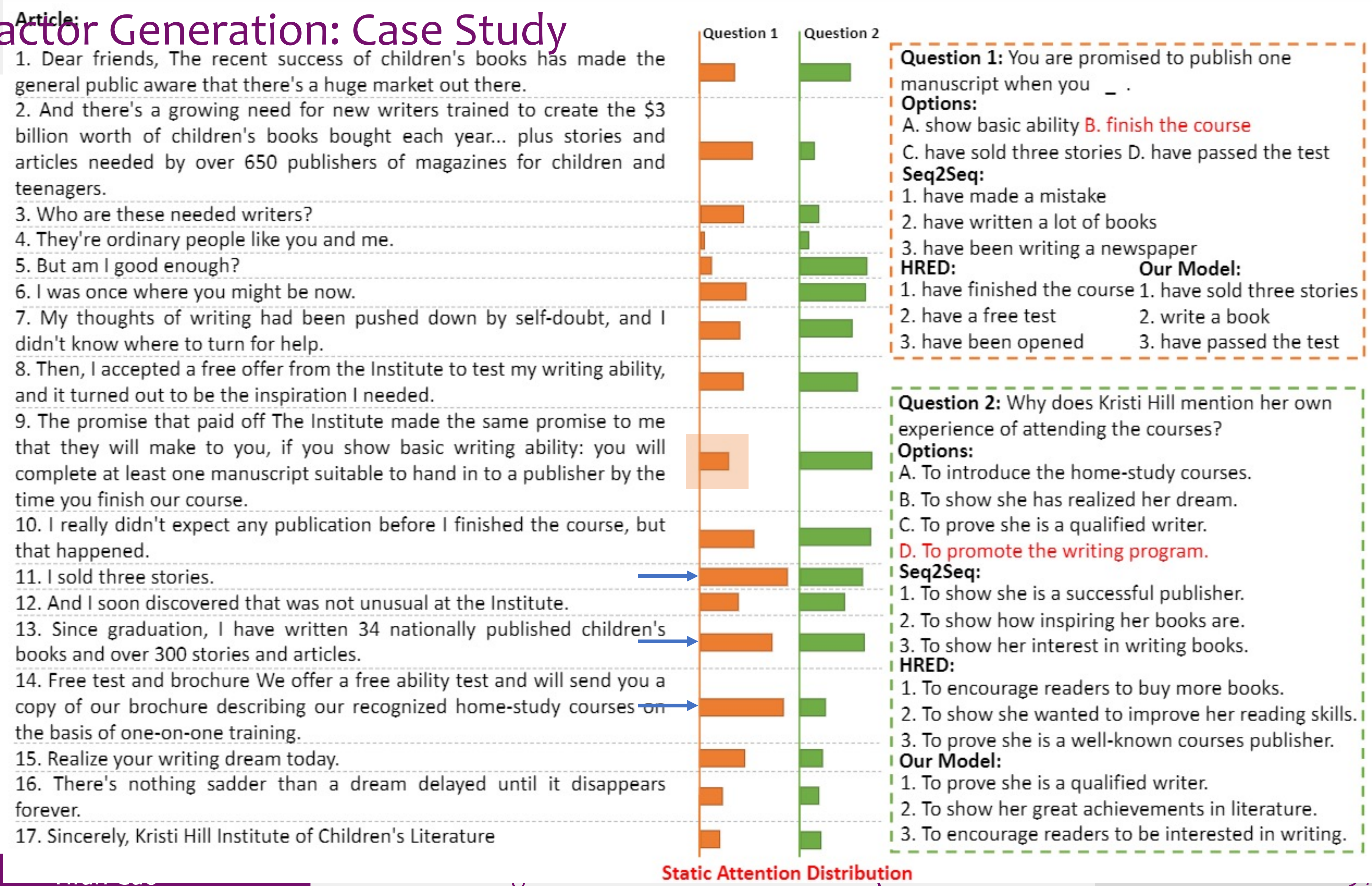
Input 1: prajñā is the wisdom that is able to extinguish afflictions and bring about bodhi . (*Easy Question)*
Human: (4.5) prajna is the wisom that is able to extinguish afflictions and bring about what ?
Ans: (13.0) what is prajñā ?
DLPH-GDC: (6.2) prajñā is able to extinguish afflictions and bring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

Difficulty Controllable QG: 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

Distractor Generation: Case Study



Article:

1. Dear friends, The recent success of children's books has made the general public aware that there's a huge market out there.

2. And there's a growing need for new writers trained to create the $3 billion worth of children's books bought each year... plus stories and articles needed by over 650 publishers of magazines for children and teenagers.

3. Who are these needed writers?

4. They're ordinary people like you and me.

5. But am I good enough?

6. I was once where you might be now.

7. My thoughts of writing had been pushed down by self-doubt, and I didn't know where to turn for help.

8. Then, I accepted a free offer from the Institute to test my writing ability, and it turned out to be the inspiration I needed.

9. The promise that paid off The Institute made the same promise to me that they will make to you, if you show basic writing ability: you will complete at least one manuscript suitable to hand in to a publisher by the time you finish our course.

10. I really didn't expect any publication before I finished the course, but that happened.

11. I sold three stories.

12. And I soon discovered that was not unusual at the Institute.

13. Since graduation, I have written 34 nationally published children's books and over 300 stories and articles.

14. Free test and brochure We offer a free ability test and will send you a copy of our brochure describing our recognized home-study courses on the basis of one-on-one training.

15. Realize your writing dream today.

16. There's nothing sadder than a dream delayed until it disappears forever.

17. Sincerely, Kristi Hill Institute of Children's Literature

Question 1: You are promised to publish one manuscript when you _ .

Options:

A. show basic ability B. finish the course

C. have sold three stories D. have passed the test

Seq2Seq:

1. have made a mistake
2. have written a lot of books
3. have been writing a newspaper

HRED:	**Our Model:**
1. have finished the course	1. have sold three stories
2. have a free test	2. write a book
3. have been opened	3. have passed the test

Question 2: Why does Kristi Hill mention her own experience of attending the courses?

Options:

A. To introduce the home-study courses.

B. To show she has realized her dream.

C. To prove she is a qualified writer.

D. To promote the writing program.

Seq2Seq:

1. To show she is a successful publisher.
2. To show how inspiring her books are.
3. To show her interest in writing books.

HRED:

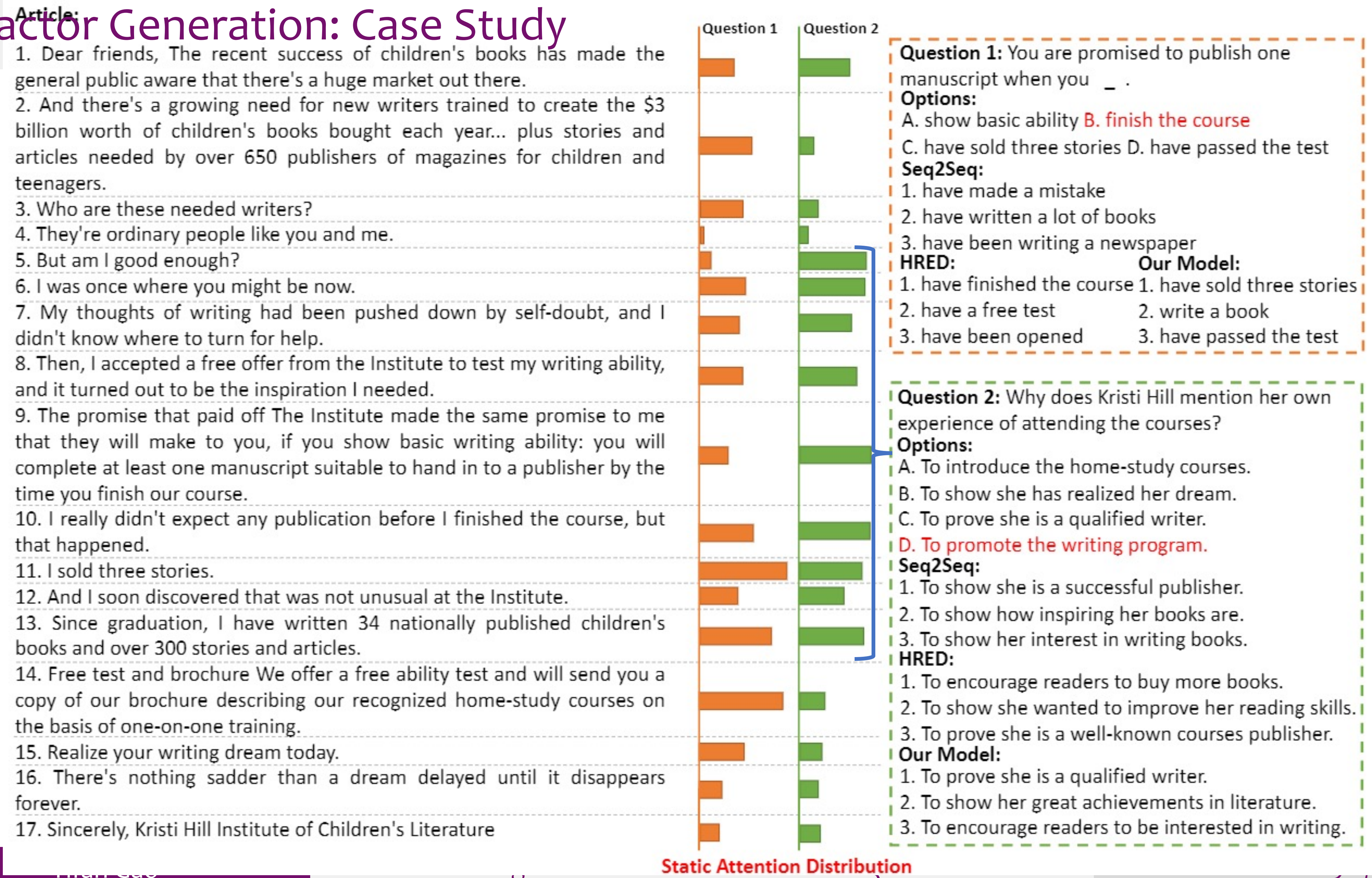
1. To encourage readers to buy more books.
2. To show she wanted to improve her reading skills.
3. To prove she is a well-known courses publisher.

Our Model:

1. To prove she is a qualified writer.
2. To show her great achievements in literature.
3. To encourage readers to be interested in writing.

Static Attention Distribution

Distractor Generation: Case Study



Article:

1. Dear friends, The recent success of children's books has made the general public aware that there's a huge market out there.
2. And there's a growing need for new writers trained to create the $3 billion worth of children's books bought each year... plus stories and articles needed by over 650 publishers of magazines for children and teenagers.
3. Who are these needed writers?
4. They're ordinary people like you and me.
5. But am I good enough?
6. I was once where you might be now.
7. My thoughts of writing had been pushed down by self-doubt, and I didn't know where to turn for help.
8. Then, I accepted a free offer from the Institute to test my writing ability, and it turned out to be the inspiration I needed.
9. The promise that paid off The Institute made the same promise to me that they will make to you, if you show basic writing ability: you will complete at least one manuscript suitable to hand in to a publisher by the time you finish our course.
10. I really didn't expect any publication before I finished the course, but that happened.
11. I sold three stories.
12. And I soon discovered that was not unusual at the Institute.
13. Since graduation, I have written 34 nationally published children's books and over 300 stories and articles.
14. Free test and brochure We offer a free ability test and will send you a copy of our brochure describing our recognized home-study courses on the basis of one-on-one training.
15. Realize your writing dream today.
16. There's nothing sadder than a dream delayed until it disappears forever.
17. Sincerely, Kristi Hill Institute of Children's Literature

Question 1: You are promised to publish one manuscript when you _ .

Options:

A. show basic ability B. finish the course

C. have sold three stories D. have passed the test

Seq2Seq:

1. have made a mistake
2. have written a lot of books
3. have been writing a newspaper

HRED:

1. have finished the course
2. have a free test
3. have been opened

Our Model:

1. have sold three stories
2. write a book
3. have passed the test

Question 2: Why does Kristi Hill mention her own experience of attending the courses?

Options:

A. To introduce the home-study courses.

B. To show she has realized her dream.

C. To prove she is a qualified writer.

D. To promote the writing program.

Seq2Seq:

1. To show she is a successful publisher.
2. To show how inspiring her books are.
3. To show her interest in writing books.

HRED:

1. To encourage readers to buy more books.
2. To show she wanted to improve her reading skills.
3. To prove she is a well-known courses publisher.

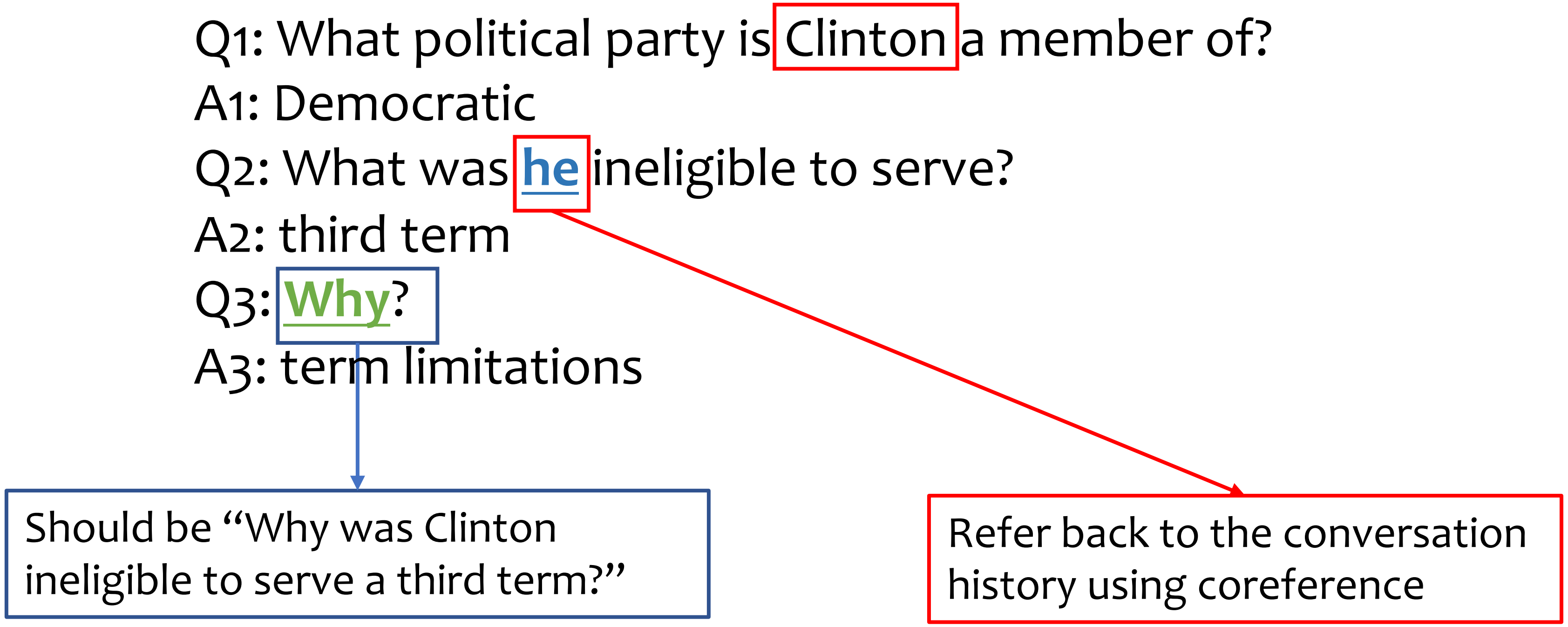
Our Model:

1. To prove she is a qualified writer.
2. To show her great achievements in literature.
3. To encourage readers to be interested in writing.

Static Attention Distribution

Conversational QG: Challenges

1. Generate conversational interconnected questions depending on the conversation so far



Q1: What political party is Clinton a member of?

A1: Democratic

Q2: What was **he** ineligible to serve?

A2: third term

Q3: **Why**?

A3: term limitations

Should be "Why was Clinton ineligible to serve a third term?"

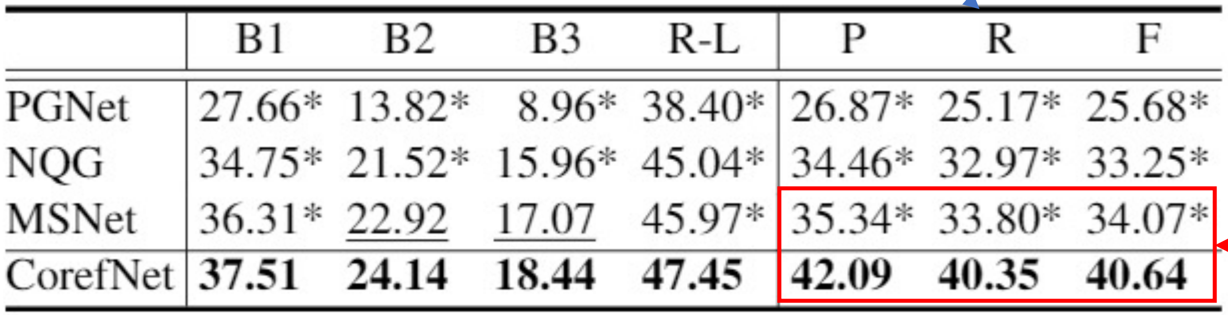
Refer back to the conversation history using coreference

Conversational QG: Coreference Alignment Analysis

<u>Coreference Set</u>

- Each sample in the coreference set requires a pronoun resolution

Precision, Recall, F-score of pronouns
in generated questions



	B1	B2	B3	R-L	P	R	F
PGNet	27.66*	13.82*	8.96*	38.40*	26.87*	25.17*	25.68*
NQG	34.75*	21.52*	15.96*	45.04*	34.46*	32.97*	33.25*
MSNet	36.31*	<u>22.92</u>	<u>17.07</u>	45.97*	35.34*	33.80*	34.07*
CorefNet	**37.51**	**24.14**	**18.44**	**47.45**	**42.09**	**40.35**	**40.64**

A large margin!

<u>underline</u>: p-value<0.05, *: p-value<0.01

Explicit Memory Tracker: Tracking

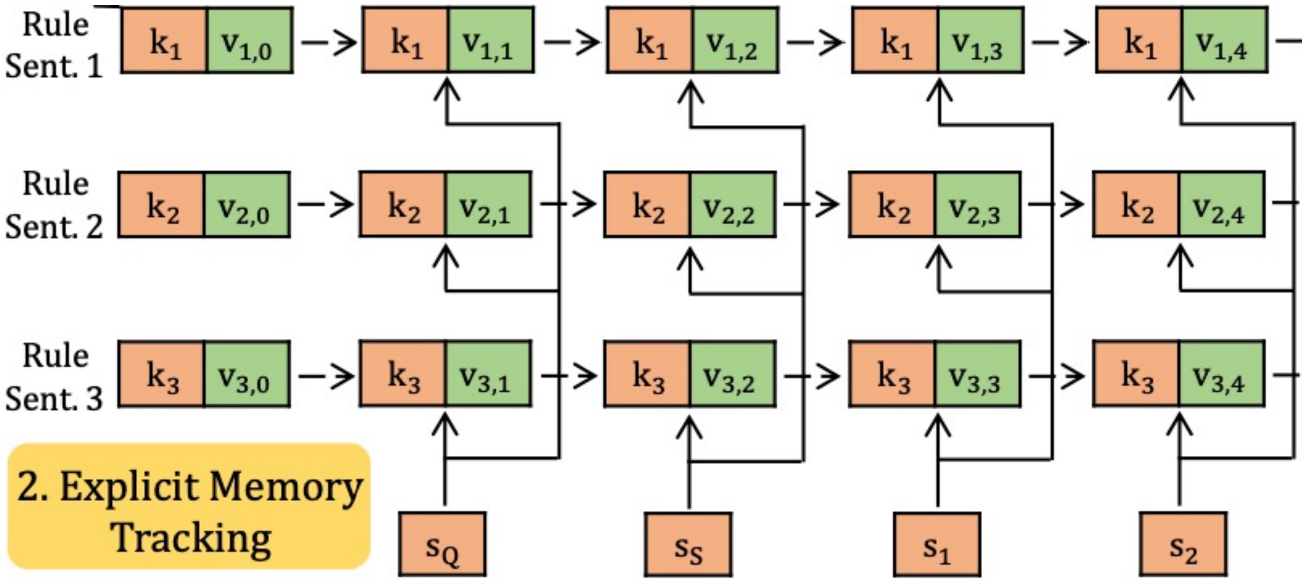
EMT assigns a state $\mathbf{v}_i$ to each key $\mathbf{k}_i$, and sequentially reads user information

At time step t:

$$\tilde{\mathbf{v}}_{i,t} = \mathrm{ReLU}(\mathbf{W}_k\mathbf{k}_i + \mathbf{W}_v\mathbf{v}_{i,t} + \mathbf{W}_s\mathbf{s}_t),$$

$$g_i = \sigma(\mathbf{s}_t^\top \mathbf{k}_i + \mathbf{s}_t^\top \mathbf{v}_{i,t}) \in [0,1],$$

$$\mathbf{v}_{i,t} = \mathbf{v}_{i,t} + g_i \odot \tilde{\mathbf{v}}_{i,t} \in \mathbb{R}^d, \mathbf{v}_{i,t} = \frac{\mathbf{v}_{i,t}}{\|\mathbf{v}_{i,t}\|}$$


Rule Sent. 1
Rule Sent. 2
Rule Sent. 3
k_1 $v_{1,0}$ k_1 $v_{1,1}$ k_1 $v_{1,2}$ k_1 $v_{1,3}$ k_1 $v_{1,4}$
k_2 $v_{2,0}$ k_2 $v_{2,1}$ k_2 $v_{2,2}$ k_2 $v_{2,3}$ k_2 $v_{2,4}$
k_3 $v_{3,0}$ k_3 $v_{3,1}$ k_3 $v_{3,2}$ k_3 $v_{3,3}$ k_3 $v_{3,4}$
s_Q
s_S
s_1
s_2
2. Explicit Memory Tracking

Keys and final states of rule sentences are denoted as $(\mathbf{k}_1, \mathbf{v}_1), \ldots, (\mathbf{k}_M, \mathbf{v}_M)$

- ❖ Decision Making Module
- ❖ Question Generation Module

Explicit Memory Tracker: Decision Making

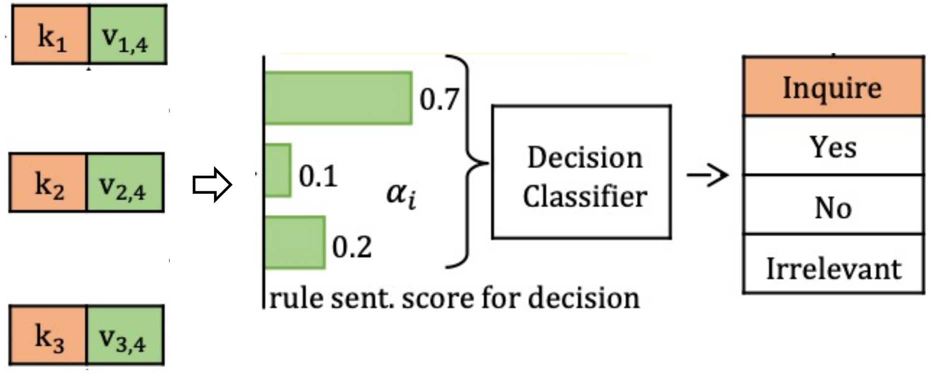
Based on the most up-to-date key-value states of rule sentences $(\mathbf{k}_1, \mathbf{v}_1), \ldots, (\mathbf{k}_M, \mathbf{v}_M)$, EMT makes a decision among *Yes, No, Irrelevant, Inquire*

$$\alpha_i = \mathbf{w}_\alpha^\top [\mathbf{k}_i; \mathbf{v}_i] + b_\alpha \in \mathbb{R}^1$$

$$\tilde{\alpha}_i = \mathrm{softmax}(\alpha)_i \in [0, 1]$$

$$\mathbf{c} = \sum_i \tilde{\alpha}_i [\mathbf{k}_i; \mathbf{v}_i] \in \mathbb{R}^d$$

$$\mathbf{z} = \mathbf{W}_z \mathbf{c} + \mathbf{b}_z \in \mathbb{R}^4$$



The decision making module is trained with the following cross entropy loss:

$$\mathcal{L}_{\text{dec}} = -\log\ \mathrm{softmax}(\mathbf{z})_l$$

Explicit Memory Tracker: Question Generation

When the decision is ‘Inquire’, a follow-up question is required for further clarification.

We adopt a two-step approach:

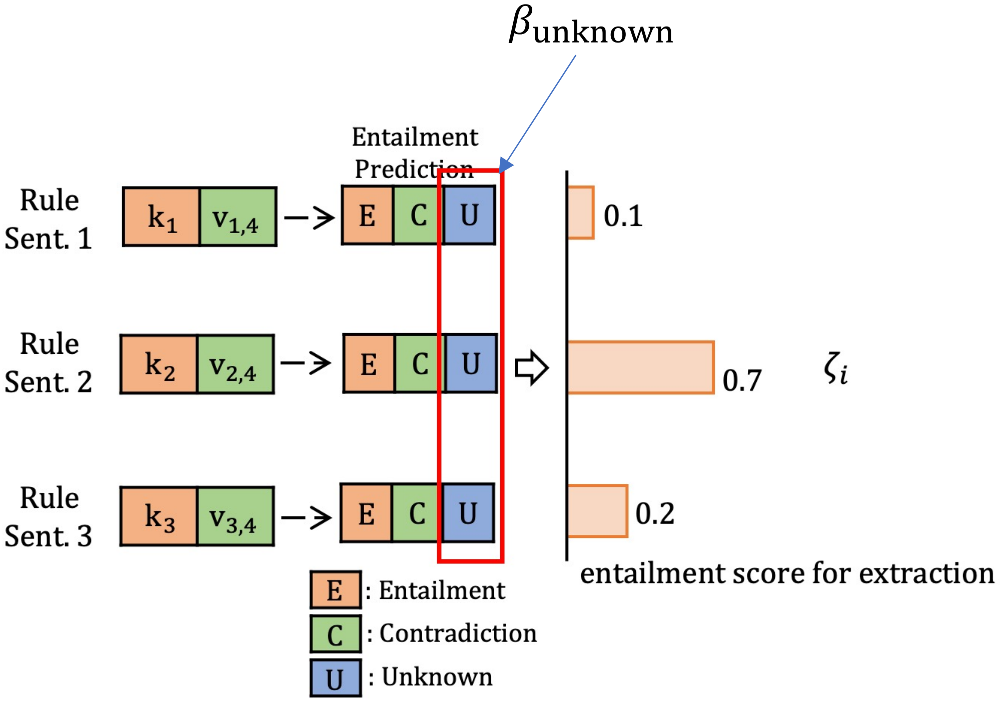
1. Extract a span inside the rule text which contains the underspecified user information
2. Rephrase the extracted span into a follow-up question

ACL 2020, Explicit Memory Tracker with Coarse-to-Fine Reasoning for Conversational Machine Reading

Explicit Memory Tracker: Span Extraction for Question Generation

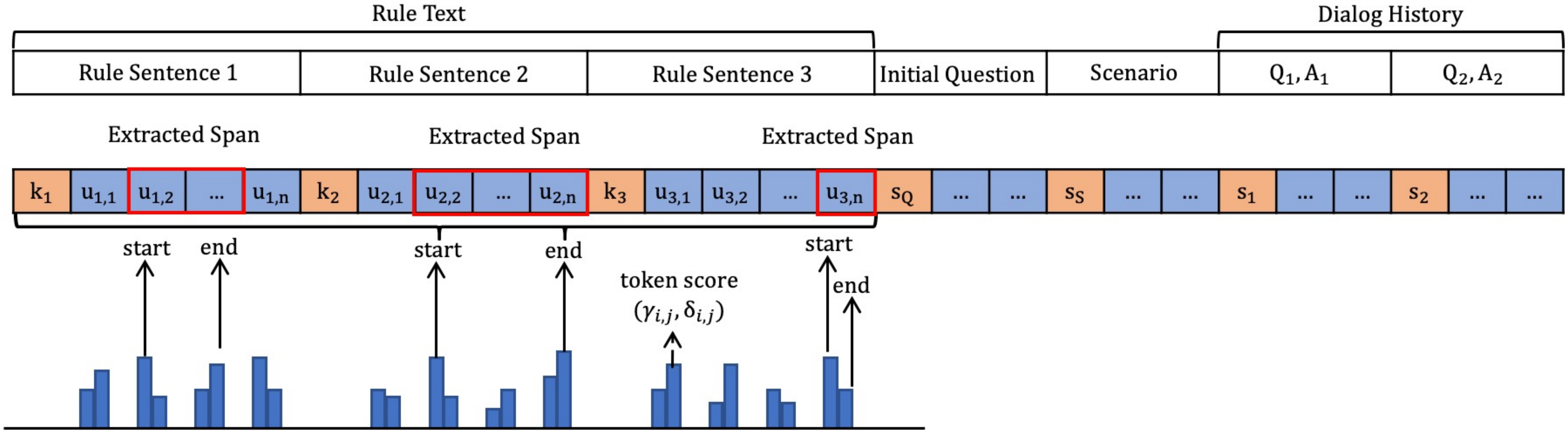
1. Coarse-to-fine Underspecified Span Extraction
   1) Identify underspecified rule sentence ζ_i

$$\zeta_i = \text{softmax}(\beta_{\text{unknown}})_i \in [0, 1]$$



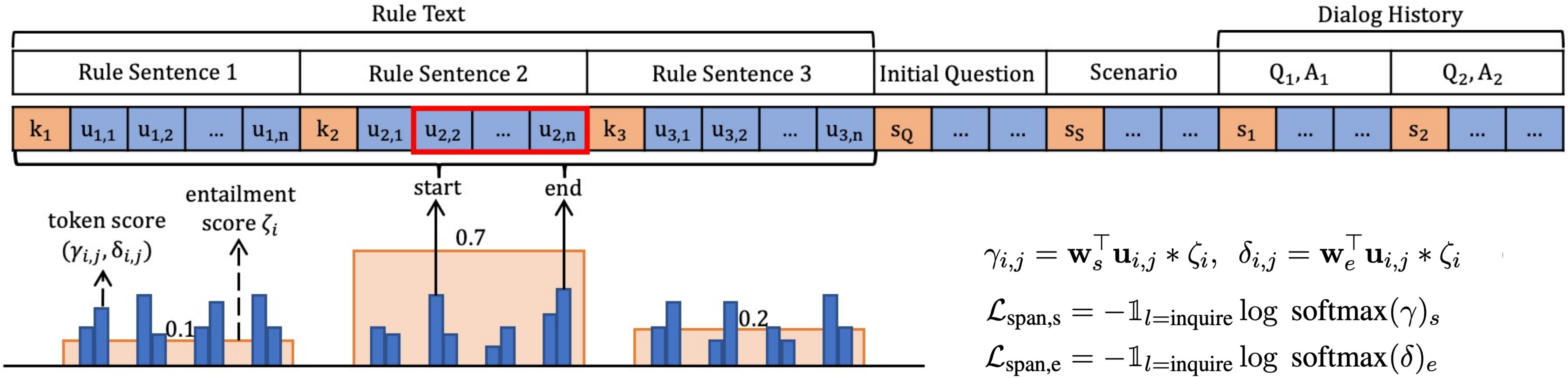
Explicit Memory Tracker: Span Extraction for Question Generation

1. Coarse-to-fine Underspecified Span Extraction
   1) Identify underspecified rule sentence ζ_i
   2) Extract a span within each rule sentence $(\gamma_{i,j}, \delta_{i,j})$



Explicit Memory Tracker: Span Extraction for Question Generation

1. Coarse-to-fine Underspecified Span Extraction
   1) Identify underspecified rule sentence ζ_i
   2) Extract a span within each rule sentence $(\gamma_{i,j}, \delta_{i,j})$
   3) Select the span with the highest span score ζ_i * $(\gamma_{i,j}, \delta_{i,j})$



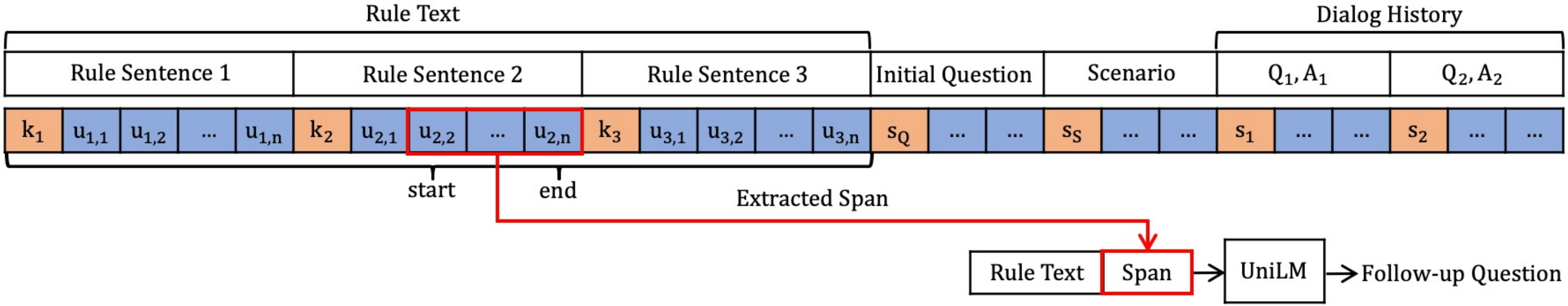
$$\gamma_{i,j} = \mathbf{w}_s^\top \mathbf{u}_{i,j} * \zeta_i, \quad \delta_{i,j} = \mathbf{w}_e^\top \mathbf{u}_{i,j} * \zeta_i$$

$$\mathcal{L}_{\text{span,s}} = -\mathbb{1}_{l=\text{inquire}} \log \ \text{softmax}(\gamma)_s$$

$$\mathcal{L}_{\text{span,e}} = -\mathbb{1}_{l=\text{inquire}} \log \ \text{softmax}(\delta)_e$$

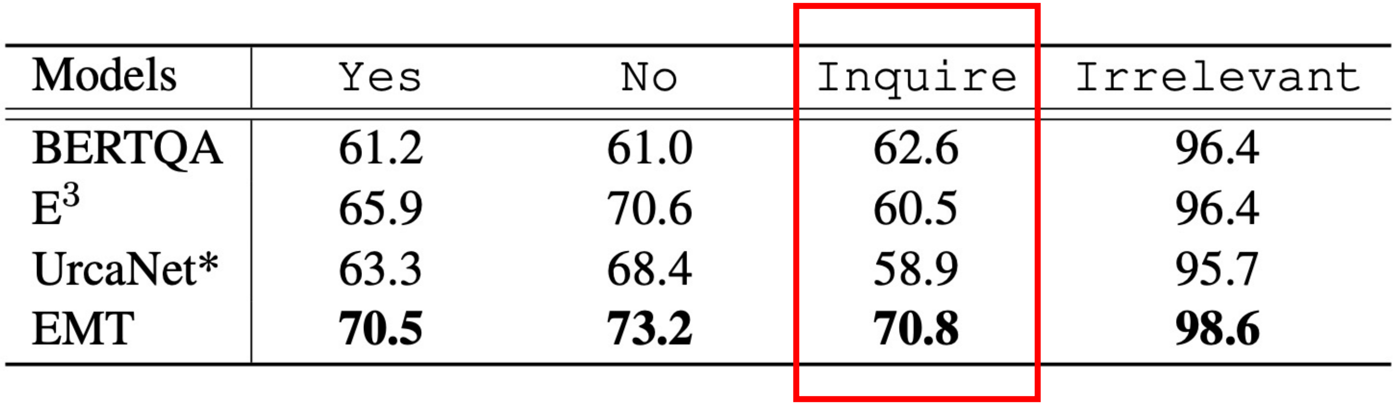
Explicit Memory Tracker: Span Extraction for Question Generation

2. Question Rephrasing
   1) Finetune UniLM (Dong et al, 2019), a pretrained language model
   2) [CLS] rule text [SEP] span [SEP]



Rule Text
Dialog History
Rule Sentence 1
Rule Sentence 2
Rule Sentence 3
Initial Question
Scenario
Q_1, A_1
Q_2, A_2
k_1 $u_{1,1}$ $u_{1,2}$ ... $u_{1,n}$ k_2 $u_{2,1}$ $u_{2,2}$ ... $u_{2,n}$ k_3 $u_{3,1}$ $u_{3,2}$ ... $u_{3,n}$ s_Q ... ... s_S ... ... s_1 ... ... s_2 ... ...
start
end
Extracted Span
Rule Text
Span
UniLM
Follow-up Question

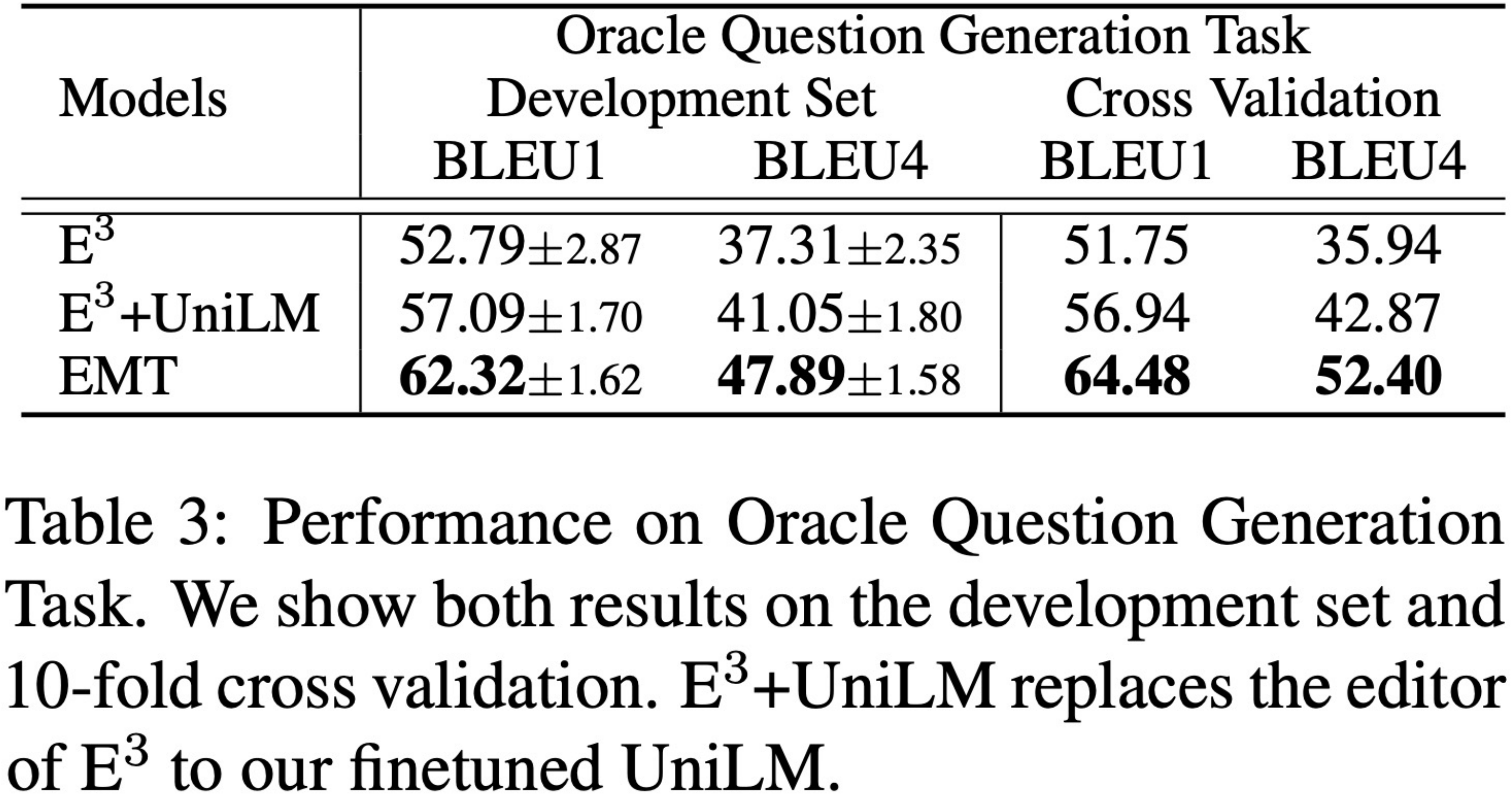
Explicit Memory Tracker: Class-wise Decision Prediction Accuracy



Models	Yes	No	Inquire	Irrelevant
BERTQA	61.2	61.0	62.6	96.4
E^3	65.9	70.6	60.5	96.4
UrcaNet*	63.3	68.4	58.9	95.7
EMT	**70.5**	**73.2**	**70.8**	**98.6**

Table 2: Class-wise decision prediction accuracy on the development set (*: reported in the paper).

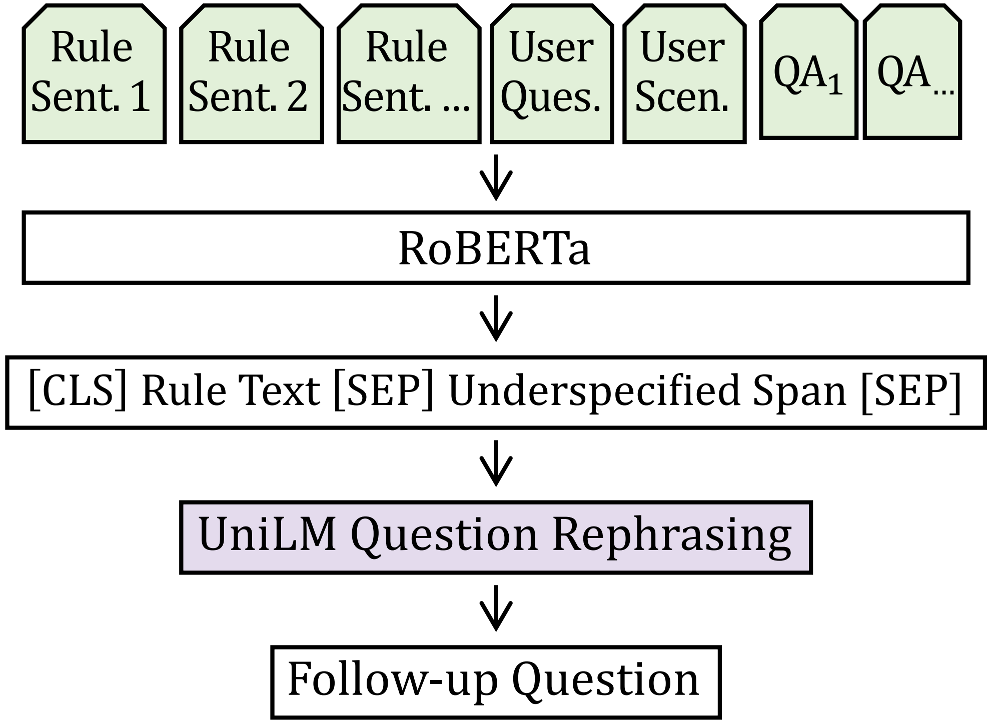
Explicit Memory Tracker: Oracle Question Generation Task



Models	Oracle Question Generation Task			
	Development Set		Cross Validation	
	BLEU1	BLEU4	BLEU1	BLEU4
E^3	52.79±2.87	37.31±2.35	51.75	35.94
E^3+UniLM	57.09±1.70	41.05±1.80	56.94	42.87
EMT	**62.32**±1.62	**47.89**±1.58	**64.48**	**52.40**

Table 3: Performance on Oracle Question Generation Task. We show both results on the development set and 10-fold cross validation. E^3+UniLM replaces the editor of E^3 to our finetuned UniLM.

DISCERN: Follow-up Question Generation


Rule Sent. 1
Rule Sent. 2
Rule Sent. ...
User Ques.
User Scen.
QA$_1$
QA$_{...}$
RoBERTa
[CLS] Rule Text [SEP] Underspecified Span [SEP]
UniLM Question Rephrasing
Follow-up Question