Chinese Machine Reading Comprehension and Beyond

Yiming Cui

Research Center for Social Computing and Information Retrieval (SCIR), Harbin Institute of Technology, China
Joint Laboratory of HIT and iFLYTEK Research (HFL), Beijing, China

The 3rd Workshop on Machine Reading for Question Answering (MRQA 2021)
Nov 10, 2021


STAY SAFE, STAY WELL

OUTLINE

- Quick Revisit: Chinese Machine Reading Comprehension
  - CMRC Dataset Series, Chinese Pre-trained Language Models
- Multilingual & Cross-lingual Machine Reading Comprehension
  - Dual BERT, WEAM
- Explainable Machine Reading Comprehension
  - RDG, ExpMRC, Attention in MRC
- Summary
- References & Useful Resources

Chinese Machine Reading Comprehension

INTRODUCTION

- To comprehend human language is essential in A.I.
- **M**achine **R**eading **C**omprehension has been a trending topic in NLP research

Computing Intelligence

Perceptual Intelligence

Cognitive Intelligence

INTRODUCTION

- **Machine Reading Comprehension (MRC)**
  - Read and comprehend passage(s) and answer relevant questions
- **Type of MRC Datasets**
  - Cloze-style: CNN/DailyMail (Hermann et al., 2015), CBT (Hill et al., 2015), PD&CFT (Cui et al., 2016)
  - Span-extraction: SQuAD (Rajpurkar et al., 2016), CMRC 2018 (Cui et al., 2019)
  - Multi-choice: MCTest (Richardson et al., 2013), RACE (Lai et al., 2017), C^3 (Sun et al., 2020)
  - Conversational: CoQA (Reddy et al., 2018), QuAC (Choi et al., 2018)
  - Multi-hop: HotpotQA (Yang et al., 2018)
  - Multi-modal: VCR (Zellers et al., 2019)
  - ……

CHINESE MRC

- **Our efforts in Chinese MRC**
  - Cloze-style MRC
    - PD&CFT (Cui et al., COLING 2016), CMRC 2017 (Cui et al., LREC 2018)
  - Span-Extraction MRC
    - CMRC 2018 (Cui et al., EMNLP 2019)
  - Sentence-cloze MRC
    - CMRC 2019 (Cui et al., COLING 2020)

PD&CFT / CMRC 2017

- **Two cloze-style Chinese MRC datasets**
  - PD&CFT: First Chinese cloze-style MRC dataset
  - We also created a new dataset for the first evaluation workshop on Chinese MRC (CMRC 2017)

1 ||| People Daily (Jan 1). According to report of "New York Times", the Wall Street stock market continued to rise as the global stock market in the last day of 2013, ending with the highest record or near record of this year.
2 ||| "New York times" reported that the S&P 500 index rose 29.6% this year, which is the largest increase since 1997.
3 ||| Dow Jones industrial average index rose 26.5%, which is the largest increase since 1996.
4 ||| NASDAQ rose 38.3%.
5 ||| In terms of December 31, due to the prospects in employment and possible acceleration of economy next year, there is a rising confidence in consumers.
6 ||| As reported by Business Association report, consumer confidence rose to 78.1 in December, significantly higher than 72 in November.
7 ||| Also as "Wall Street journal" reported that 2013 is the best U.S. stock market since 1995.
8 ||| In this year, to chase the "silly money" is the most wise way to invest in U.S. stock.
9 ||| The so-called "silly money" XXXXX is that, to buy and hold the common combination of U.S. stock.
10 ||| This strategy is better than other complex investment methods, such as hedge funds and the methods adopted by other professional investors.

Passage

The so-called "silly money" XXXXX is that, to buy and hold the common combination of U.S. stock.

Question

Answer

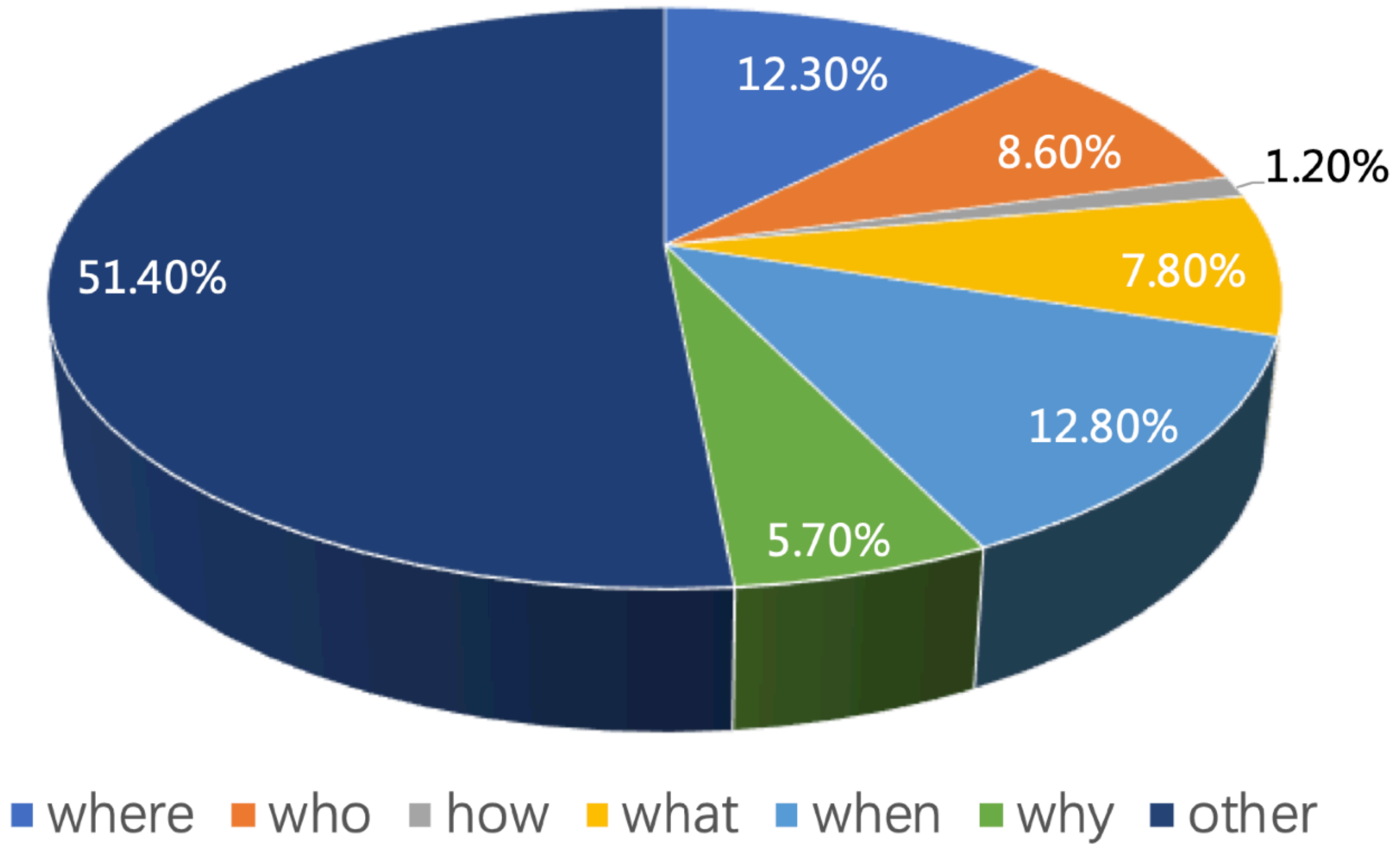
strategy

[Cui et al., COLING 2016] Consensus Attention-based Neural Networks for Chinese Reading Comprehension
[Cui et al., LREC 2018] Dataset for the First Evaluation on Chinese Machine Reading Comprehension

CMRC 2018

- **A Span-Extraction Dataset for Chinese MRC**
    - Similar to SQuAD, CMRC 2018 is a span-extraction Chinese MRC dataset (~18K questions)
    - We also propose a challenging set that is composed of hard questions, which need comprehensive reasoning over multiple sentences

[Passage] 《黄色脸孔》是柯南·道尔所著的福尔摩斯探案的56个短篇故事之一，收录于《福尔摩斯回忆录》。孟罗先生素来与妻子恩爱，但自从最近邻居新入伙后，孟罗太太则变得很奇怪，曾经凌晨时份外出，又藉丈夫不在家时偷偷走到邻居家中。于是孟罗先生向福尔摩斯求助，福尔摩斯听毕孟罗先生的故事后，认为孟罗太太被来自美国的前夫勒索，所以不敢向孟罗先生说出真相，所以吩咐孟罗先生，如果太太再次走到邻居家时，即时联络他，他会第一时间赶到。孟罗太太又走到邻居家，福尔摩斯陪同孟罗先生冲入，却发现**邻居家中的人是孟罗太太与前夫生的女儿，因为孟罗太太的前夫是黑人，她怕孟罗先生嫌弃混血儿，所以不敢说出真相。**	**[Passage]** ”The Adventure of the Yellow Face”, one of the 56 short Sherlock Holmes stories written by Sir Arthur Conan Doyle, is the third tale from The Memoirs of Sherlock Holmes. Mr. Munro has always been loved by his wife, but since the new neighbors recently joined, Mrs. Munro has become very strange. She used to go out in the early hours of the morning and secretly went to her neighbors when her husband was not at home. ... Mrs. Munro went to the neighbor’s house again, and Holmes accompanied Mr. Munro to rush in, only to find that the neighbor’s family was the daughter of Mrs. Munro and her ex-husband, **because Mrs. Munro’s ex-husband was black, and she was afraid of Mr. Munro hate the mixed-race, so she did not dare to tell the truth.**
[Question] 孟罗太太为什么在邻居新入伙后变得很奇怪? **[Answer 1]** 邻居家中的人是孟罗太太与前夫生的女儿，因为孟罗太太的前夫是黑人，她怕孟罗先生嫌弃混血儿 **[Answer 2]** 邻居家中的人是孟罗太太与前夫生的女儿，因为孟罗太太的前夫是黑人，她怕孟罗先生嫌弃混血儿，所以不敢说出真相。 **[Answer 3]** 邻居家中的人是孟罗太太与前夫生的女儿，因为孟罗太太的前夫是黑人，她怕孟罗先生嫌弃混血儿，所以不敢说出真相。	**[Question]** Why Mrs. Munro became strange after the new neighbors moved in? **[Answer 1]** because Mrs. Munro’s ex-husband was black, and she was afraid of Mr. Munro hate the mixed-race **[Answer 2]** because Mrs. Munro’s ex-husband was black, and she was afraid of Mr. Munro hate the mixed-race **[Answer 3]** because Mrs. Munro’s ex-husband was black, and she was afraid of Mr. Munro hate the mixed-race, so she did not dare to tell the truth.



[Cui et al., EMNLP 2019] A Span-Extraction Dataset for Chinese Machine Reading Comprehension

CMRC 2018

- **Latest Submissions**
  - Most recent submissions are based on powerful pre-trained language models, such as MacBERT
  - Top systems are about to reach the human performance on the normal test set
  - However, there is still a large gap (~30%) to human on the challenge set

CMRC 2018

A Span-Extraction Dataset for Chinese Machine Reading Comprehension

Rank	Model	Test		Challenge	
		EM	F1	EM	F1
	Human Performance *Joint Laboratory of HIT and iFLYTEK Research* [Cui et al., EMNLP 2019]	92.400	97.914	90.382	95.248
1 Dec 8, 2020	MacBERT-large-extData-v2 (single model) *AISpeech*	**80.409**	**93.768**	36.706	66.905
2 Nov 12, 2020	MacBERT-large-extData (single model) *AISpeech*	77.998	92.882	**38.492**	**67.109**
3 Nov 3, 2020	RoBERTa-wwm-ext-large-extData (single model) *AISpeech*	76.997	92.171	32.540	63.597
4 May 1, 2020	MacBERT-large (single model) *Joint Laboratory of HIT and iFLYTEK Research* [Cui et al., Findings of EMNLP 2020]	74.786	90.693	31.923	60.177

[Cui et al., EMNLP 2019] A Span-Extraction Dataset for Chinese Machine Reading Comprehension

CMRC 2019

- **A Sentence Cloze Dataset for Chinese MRC**
  - We propose sentence cloze task for MRC
    - A natural extension to cloze-style machine reading comprehension
    - Instead of filling a word or an entity in the blank, we require the machine to fill in the sentence
  - Test the ability of sentence-level inference in MRC
  - Release a challenging Chinese dataset **CMRC 2019**, which consists of 100K blanks
  - State-of-the-art PLMs still lag behind human performance on this dataset

[Passage]
A long time ago, there was a queen. **[BLANK1]** Soon after the child was born, the Queen died. **[BLANK2]** The stepmother didn't like her very much. She made Snow White do the housework all day and all night. A wizard had given this Queen a glass. The glass could speak. It was on the wall in the Queen's room. Every day the Queen looked in the glass to see how beautiful she was. As she looked in the glass, she asked: "Tell me, glass upon the wall, who is most beautiful of all?" And the glass said: "The Queen is most beautiful of all.". Years went by. Snow-white grew up and became a little girl. Every day the Queen looked in the glass and said, "Tell me, glass upon the wall, **[BLANK3]**" And the glass said, "Snow-white is most beautiful of all.". When the Queen heard this, **[BLANK4]**. She said, "Snow-white is not more beautiful than I am. There is no one who is more beautiful than I am.". So she called a hunter and said, "Take Snow-white into the forest and kill her.". The hunter took Snow-white to the forest, but he did not kill her, because she was so beautiful and so lovely. He put Snow White in the forest and went away.

[Candidates]
0: The king married another queen
1: She had a pretty daughter named Snow White
2: The king was also passed away
3: who is most beautiful of all?
4: she was very happy
5: she was very angry

Fake candidates

[Answers]
1, 0, 3, 5

Correct order of sentence ID

[Cui et al., COLING 2020] A Sentence Cloze Dataset for Chinese Machine Reading Comprehension

CMRC 2019

- **Results**
  - Human: ~95% on QAC (Question-level accuracy) and 75~81% on PAC (Passage-level)
  - PLM-based baseline systems achieve high scores on QAC but not on PAC
  - Top submissions adopt data augmentation, ensemble, etc. PAC is still far from human.

System	Dev QAC	Dev PAC	Test QAC	Test PAC
Human Performance	*95.9*	*81.0*	*95.3*	*75.0*
Random Selection	7.6	0.0	7.5	0.0
Baseline Systems				
BERT	71.2	10.0	71.0	8.8
BERT-multilingual	66.8	6.67	66.0	5.0
BERT-wwm	72.4	9.3	71.4	7.6
BERT-wwm-ext	75.0	12.7	73.7	9.2
RoBERTa-wwm-ext	75.9	11.0	75.8	12.4
RoBERTa-wwm-ext-large	82.6	23.3	81.7	23.0

System	Dev QAC	Dev PAC	Test QAC	Test PAC
Human Performance	*95.9*	*81.0*	*95.3*	*75.0*
Random Selection	7.6	0.0	7.5	0.0
Top Submissions from CMRC 2019				
bert_scp_spm†	90.9	60.0	90.8	57.6
mojito†	88.2	48.0	86.0	41.8
DA-BERT†	86.3	34.3	84.4	27.6

[Cui et al., COLING 2020] A Sentence Cloze Dataset for Chinese Machine Reading Comprehension

CHINESE PLMS

- **Chinese PLM Series**
  - Pre-trained language models (PLMs) have become a new
  - To accelerate the Chinese NLP research, we create and open-source a series of Chinese PLMs
    - Including BERT, XLNet, RoBERTa, ELECTRA, MacBERT, etc.
  - With these PLMs, there is a significant boost in MRC performances

	CMRC 2018						DRCD			
	Dev		Test		Challenge		Dev		Test	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
BERT	65.5 (64.4)	84.5 (84.0)	70.0 (68.7)	87.0 (86.3)	18.6 (17.0)	43.3 (41.3)	83.1 (82.7)	89.9 (89.6)	82.2 (81.6)	89.2 (88.8)
BERT-wwm	66.3 (65.0)	85.6 (84.7)	70.5 (69.1)	87.4 (86.7)	21.0 (19.3)	47.0 (43.9)	84.3 (83.4)	90.5 (90.2)	82.8 (81.8)	89.7 (89.0)
BERT-wwm-ext	67.1 (65.6)	85.7 (85.0)	71.4 (70.0)	87.7 (87.0)	24.0 (20.0)	47.3 (44.6)	85.0 (84.5)	91.2 (90.9)	83.6 (83.0)	90.4 (89.9)
RoBERTa-wwm-ext	67.4 (66.5)	87.2 (86.5)	72.6 (71.4)	89.4 (88.8)	26.2 (24.6)	51.0 (49.1)	86.6 (85.9)	92.5 (92.2)	85.6 (85.2)	92.0 (91.7)
ELECTRA-base	68.4 **(68.0)**	84.8 (84.6)	73.1 **(72.7)**	87.1 (86.9)	22.6 (21.7)	45.0 (43.8)	87.5 (87.0)	92.5 (92.3)	86.9 (86.6)	91.8 (91.7)
MacBERT-base	**68.5** (67.3)	**87.9 (87.1)**	**73.2** (72.4)	**89.5 (89.2)**	**30.2 (26.4)**	**54.0 (52.2)**	**89.4 (89.2)**	**94.3 (94.1)**	**89.5 (88.7)**	**93.8 (93.5)**
ELECTRA-large	69.1 (68.2)	85.2 (84.5)	73.9 (72.8)	87.1 (86.6)	23.0 (21.6)	44.2 (43.2)	88.8 (88.7)	93.3 (93.2)	88.8 (88.2)	93.6 (93.2)
RoBERTa-wwm-ext-large	68.5 (67.6)	88.4 (87.9)	74.2 (72.4)	90.6 (90.0)	31.5 **(30.1)**	60.1 (57.5)	89.6 (89.1)	94.8 (94.4)	89.6 (88.9)	94.5 (94.1)
MacBERT-large	**70.7 (68.6)**	**88.9 (88.2)**	**74.8 (73.2)**	**90.7 (90.1)**	**31.9** (29.6)	**60.2 (57.6)**	**91.2 (90.8)**	**95.6 (95.3)**	**91.7 (90.9)**	**95.6 (95.3)**

▲ Results on CMRC 2018 (Simplified Chinese) and DRCD (Traditional Chinese)

[Cui et al., IEEE/ACM TASLP] Pre-training with Whole Word Masking for Chinese BERT
[Cui et al., Findings of EMNLP 2020] Revisiting Pre-trained Models for Chinese Natural Language Processing

Multilingual & Cross-lingual MRC

BACKGROUND

- **Problem 1: Most of the MRC research is mainly for the English dataset**
  - Languages other than English are not well-addressed due to the lack of data

TriviaQA
… NaturalQuestions HotpotQA
…
NarrativeQA CNN / DailyMail
MultiRC CLOTH
SQuAD
DuoRC ARC
MCTest
… QuAC RACE
DROP
MS MARCO DREAM
CBT
SCT
NewsQA CoQA …
SearchQA
RecipeQA

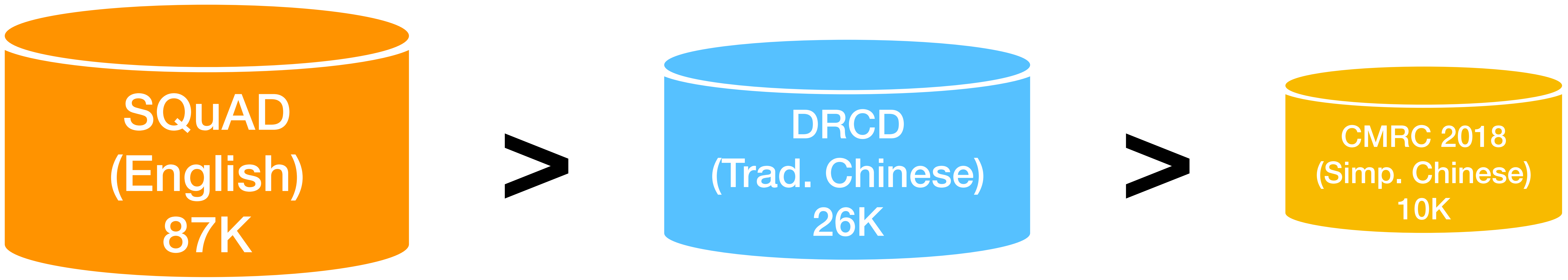
▲ English MRC Datasets

C^3 …
WebQA
PD&CFT CMRC 2018
DRCD
CJRC
DuReader
CMRC 2017
CMRC 2019
ChID
…

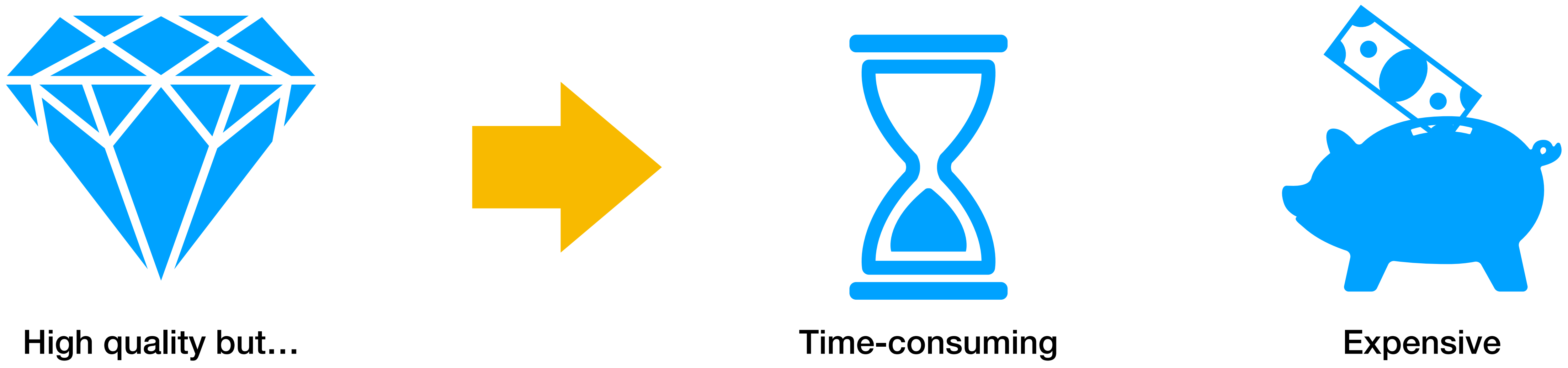
▲ Chinese MRC Datasets

Background

- Problem 2: Existing Chinese MRC datasets are relatively small



- Problem 3: Annotating training data is time-consuming and expensive



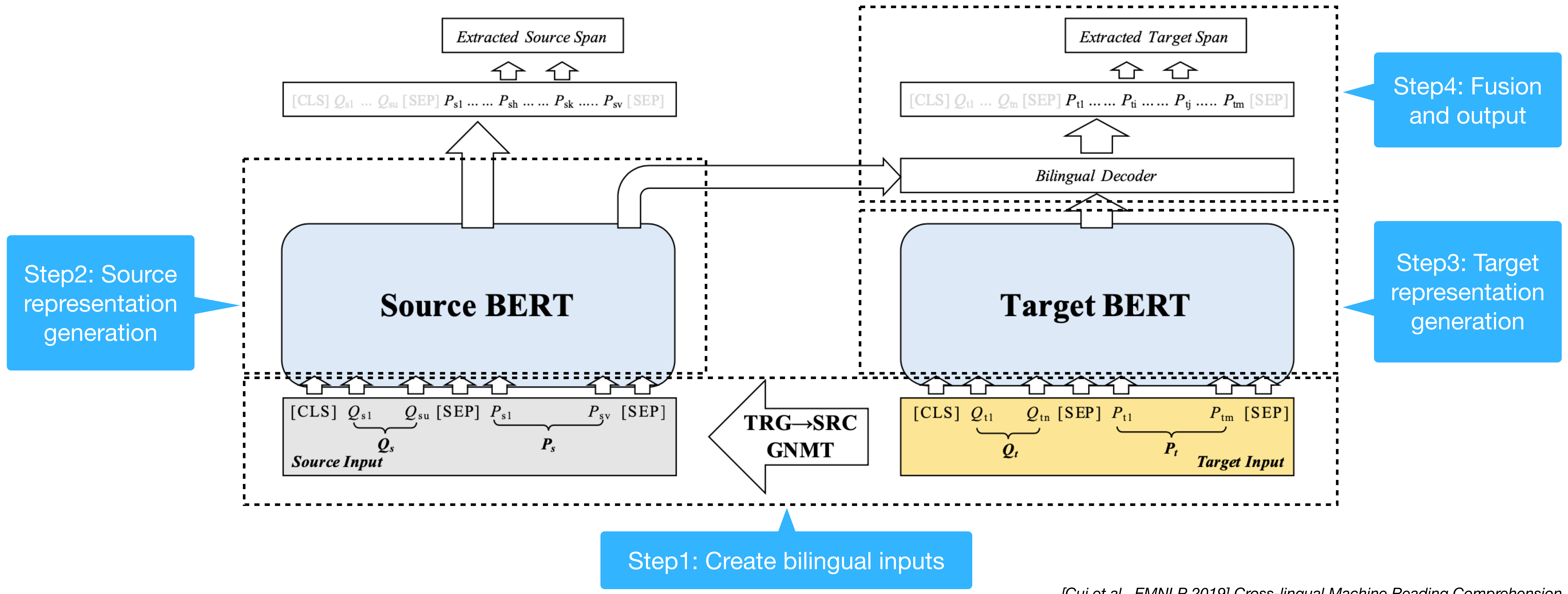
MULTILINGUAL MRC

- **Question**
  - *Can we use English data to help improve MRC performance in other languages?*
- **Solutions**
  - Dual BERT (Cui et al., EMNLP 2019)
    - Simultaneously model <Passage, Question> in both source and target language.
    - Promising results on two public Chinese MRC datasets and set new state-of-the-art performances, indicating the potentials in CLMRC research
  - WEAM (Word-Exchange Aligning Model) (Yang et al., MRQA 2021)
    - Use statistical alignment matrix to help word aligning in multilingual PLMs
    - Achieves better performance than TLM on MLQA and XNLI

[Cui et al., EMNLP 2019] Cross-lingual Machine Reading Comprehension

Dual BERT

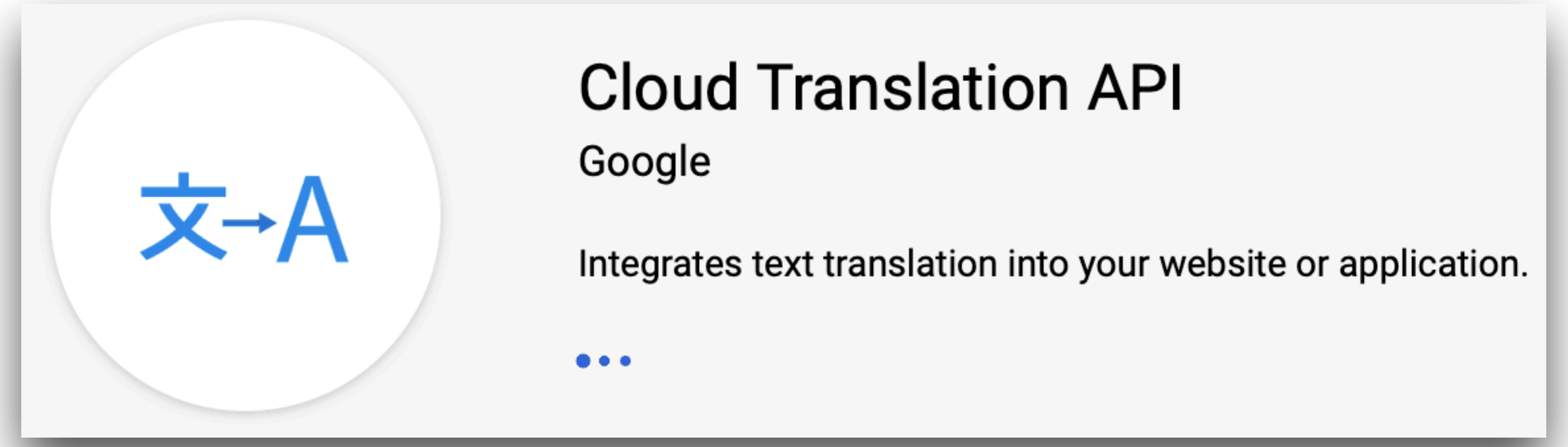
- Overview of Dual BERT



[Cui et al., EMNLP 2019] Cross-lingual Machine Reading Comprehension

DUAL BERT

- **Step 1: Creating bilingual corpus**
  - Use Google Neural Machine Translation (GNMT) to translate <P, Q, A> to the source language
  - Recover translated answer $\boldsymbol{A_{trans}}$ to an EXACT passage span as the answer in the source language
  - Choose an arbitrary passage span that has the highest F1-score to $\boldsymbol{A_{trans}}$



	MT02	MT03	MT04	MT05	MT06	MT08	Average
$\text{AST}_{feature}$ (Cheng et al., 2018)	46.10	**44.07**	**45.61**	44.06	**44.44**	34.94	43.20
GNMT (March 25, 2019)	**46.26**	43.40	44.17	**44.14**	43.86	**37.61**	**43.24**

▲ GNMT performance on NIST MT 02~08 datasets

- **Step 2 & 3: Modeling passage and question in both source/target spaces**
  - We use multilingual BERT for modeling input passage and question

[Cui et al., EMNLP 2019] Cross-lingual Machine Reading Comprehension

Dual BERT

- **Step 4: Fusion and output**
  - We use **Self-Adaptive Attention (SAA)** to create a fused representation

$$A_T = \mathbf{softmax}(B_T \cdot B_T^\top)$$
$$A_S = \mathbf{softmax}(B_S \cdot B_S^\top)$$

$$\tilde{A}_{TS} = A_T \cdot A_{TS} \cdot {A_S}^\top, \tilde{A}_{TS} \in \mathbb{R}^{L_T * L_S}$$
$$R' = \mathbf{softmax}(\tilde{A}_{TS}) \cdot B_S$$

  - An additional dense layer with residual connection

$$R = W_r R' + b_r, \quad W_r \in \mathbb{R}^{h*h}$$
$$H_T = concat[B_T, \mathbf{LayerNorm}(B_T + R)]$$

  - Output start/end probabilities and training

Loss for target prediction ↓

$$\mathcal{L} = \mathcal{L}_T + \lambda \mathcal{L}_{aux}$$

↑ *Loss for source prediction*

Dynamically determined by the similarity between source and target span representation

$$\lambda = \max\{0, \cos < \tilde{H}_S, \tilde{H}_T >\}$$

[Cui et al., EMNLP 2019] Cross-lingual Machine Reading Comprehension

Dual BERT

- **Results**
  - Back-Translation Approaches
    - SimpleMatch → Aligner → Verifier: The more information we use, the better performance we get
  - Without SQuAD Weights
    - Modeling input in bilingual space could substantially improve performance
  - With SQuAD Weights
    - Mixed Training > Cascade Training
  - Dual BERT outperforms all baselines

#	System	CMRC 2018 Dev EM	CMRC 2018 Dev F1	CMRC 2018 Test EM	CMRC 2018 Test F1	CMRC 2018 Challenge EM	CMRC 2018 Challenge F1	DRCD Dev EM	DRCD Dev F1	DRCD Test EM	DRCD Test F1
	Human Performance	*91.1*	*97.3*	*92.4*	*97.9*	*90.4*	*95.2*	-	-	*80.4*	*93.3*
	P-Reader (single model)†	59.9	81.5	65.2	84.4	15.1	39.6	-	-	-	-
	Z-Reader (single model)†	79.8	92.7	74.2	88.1	13.9	37.4	-	-	-	-
	MCA-Reader (ensemble)†	66.7	85.5	71.2	88.1	15.5	37.1	-	-	-	-
	RCEN (ensemble)†	76.3	91.4	68.7	85.8	15.3	34.5	-	-	-	-
	r-net (single model)†	-	-	-	-	-	-	-	-	29.1	44.4
	DA (Yang et al., 2019)	49.2	65.4	-	-	-	-	55.4	67.7	-	-
1	GNMT+BERT$_{SQ-B_{en}}$ ♠	15.9	40.3	20.8	45.4	4.2	20.2	28.1	50.0	26.6	48.9
2	GNMT+BERT$_{SQ-L_{en}}$ ♠	16.8	42.1	21.7	47.3	5.2	22.0	28.9	52.0	28.7	52.1
3	GNMT+BERT$_{SQ-L_{en}}$+SimpleMatch♠	26.7	56.9	31.3	61.6	9.1	35.5	36.9	60.6	37.0	61.2
4	GNMT+BERT$_{SQ-L_{en}}$+Aligner	46.1	66.4	49.8	69.3	16.5	40.9	60.1	70.5	59.5	70.7
5	GNMT+BERT$_{SQ-L_{en}}$+Verifier	64.7	84.7	68.9	86.8	20.0	45.6	83.5	90.1	82.6	89.6
6	BERT$_{B_{cn}}$	63.6	83.9	67.8	86.0	18.4	42.1	83.4	90.1	81.9	89.0
7	BERT$_{B_{mul}}$	64.1	84.4	68.6	86.8	18.6	43.8	83.2	89.9	82.4	89.5
8	**Dual BERT**	65.8	86.3	70.4	88.1	23.8	47.9	84.5	90.8	83.7	90.3
9	BERT$_{SQ-B_{mul}}$ ♠	56.5	77.5	59.7	79.9	18.6	41.4	66.7	81.0	65.4	80.1
10	BERT$_{SQ-B_{mul}}$ + Cascade Training	66.6	87.3	71.8	89.4	25.6	52.3	85.2	91.4	84.4	90.8
11	BERT$_{B_{mul}}$ + Mixed Training	66.8	87.5	72.6	89.8	26.7	53.4	85.3	91.6	84.7	91.2
12	**Dual BERT (w/ SQuAD)**	68.0	88.1	73.6	90.2	27.8	55.2	86.0	92.1	85.4	91.6

[Cui et al., EMNLP 2019] Cross-lingual Machine Reading Comprehension

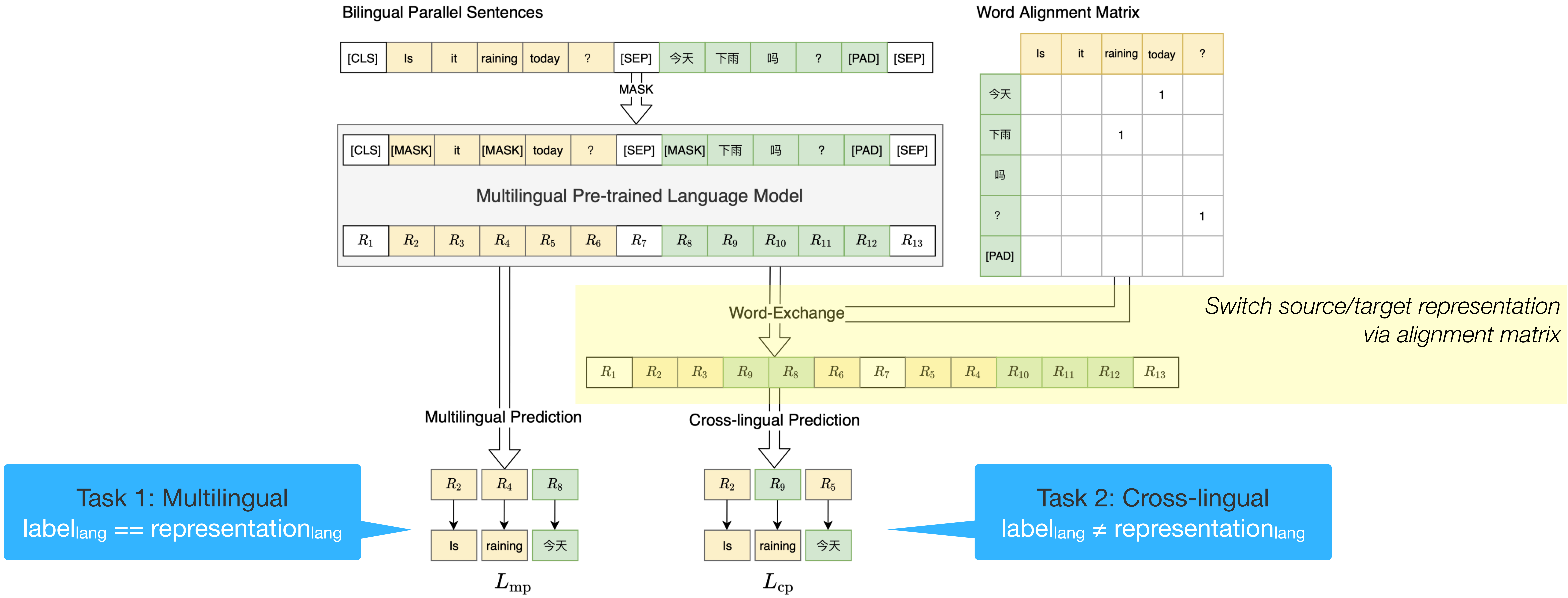
WEAM

- **Bilingual Alignment Pre-training for Zero-shot Cross-lingual Transfer**
  - The pre-training tasks of the multilingual LMs can be divided into two groups
    - Training on monolingual data from multiple languages, like Multilingual Masked LM (MMLM)
    - Or on bilingual parallel data, like Translation Language Model (TLM)
  - We propose the **Word-Exchange Aligning Model (WEAM)** to incorporate word alignment info
    - WEAM consists of a multilingual and a cross-lingual prediction task, trained on parallel corpora.
    - The multilingual prediction task predicts the original masked word in a standard way, while the cross-lingual task predicts the corresponding word from the representations in the other language.
    - WEAM uses statistical alignment information as prior knowledge to guide the cross-lingual prediction.

[Yang et al., MRQA 2021] Bilingual Alignment Pre-training for Zero-shot Cross-lingual Transfer

WEAM

- Overview of Word-Exchange Aligning Model (WEAM)



[Yang et al., MRQA 2021] Bilingual Alignment Pre-training for Zero-shot Cross-lingual Transfer

WEAM

- **Experiments**
  - Pre-training: train from mBERT with Europarl en-es (1.8M), en-de (1.9M), and en-zh (5.1M) data
  - Results
    - Training with bilingual data improves zero-shot performance on es/de/zh
    - Incorporating alignment information could give further improvements to TLM

Model	en	es	de	zh	AVG(all)	AVG(zero-shot)
Translate-Train						
mBERT	82.1	77.8	75.9	75.7	77.9	76.5
Zero-Shot						
mBERT	82.1	74.3	71.1	69.3	74.2	71.6
Word-aligned BERT	80.1	75.5	73.1	-	-	-
mBERT+TLM	82.0	75.0	73.5	73.1	75.9	73.9
mBERT+WEAM	**82.6**	**76.4**	**74.5**	**74.4**	**77.0**	**75.1**

▲ Results on MLQA

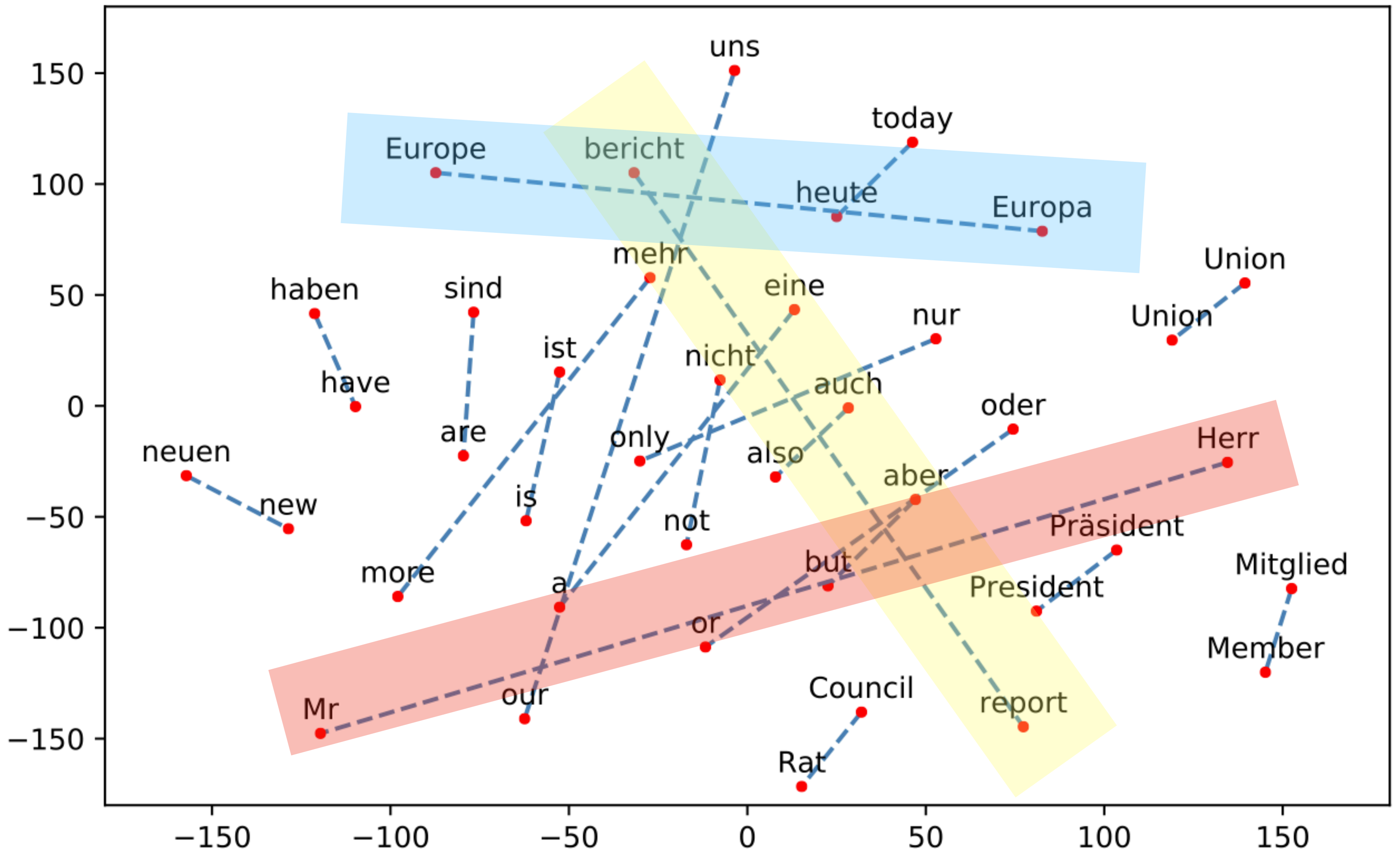
Model	en	es	de	zh	AVG(all)	AVG(zero-shot)
Translate-Train						
mBERT	77.7	53.9	62.0	61.4	63.8	60.3
mBERT (ours)	80.3	67.1	63.5	63.6	68.6	65.7
Zero-Shot						
mBERT	77.7	64.3	57.9	57.5	64.4	61.0
mBERT+TLM	**80.0**	65.7	63.1	62.0	67.7	64.6
mBERT+WEAM	79.7	**67.8**	**64.3**	**63.7**	**68.9**	**66.2**

▲ Results on XNLI

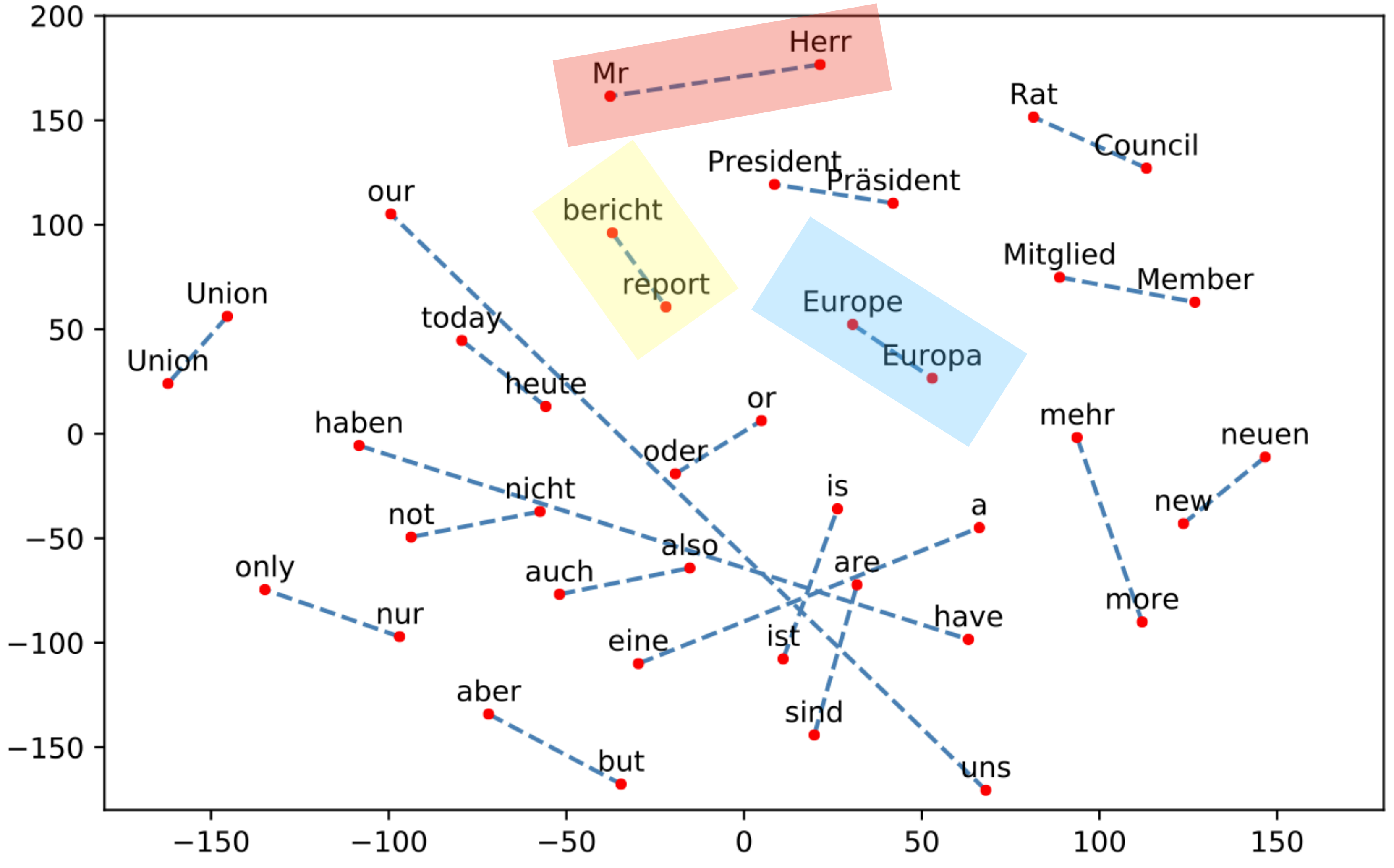
[Yang et al., MRQA 2021] Bilingual Alignment Pre-training for Zero-shot Cross-lingual Transfer

WEAM

- **Visualization of Embeddings**
  - Word vectors from mBERT word embeddings layer before and after WEAM pre-training
  - Word pairs are identified by *FastAlign*
  - After pre-training, most of the word pairs are getting closer



▲ **Before** WEAM Pre-training



▲ **After** WEAM Pre-training

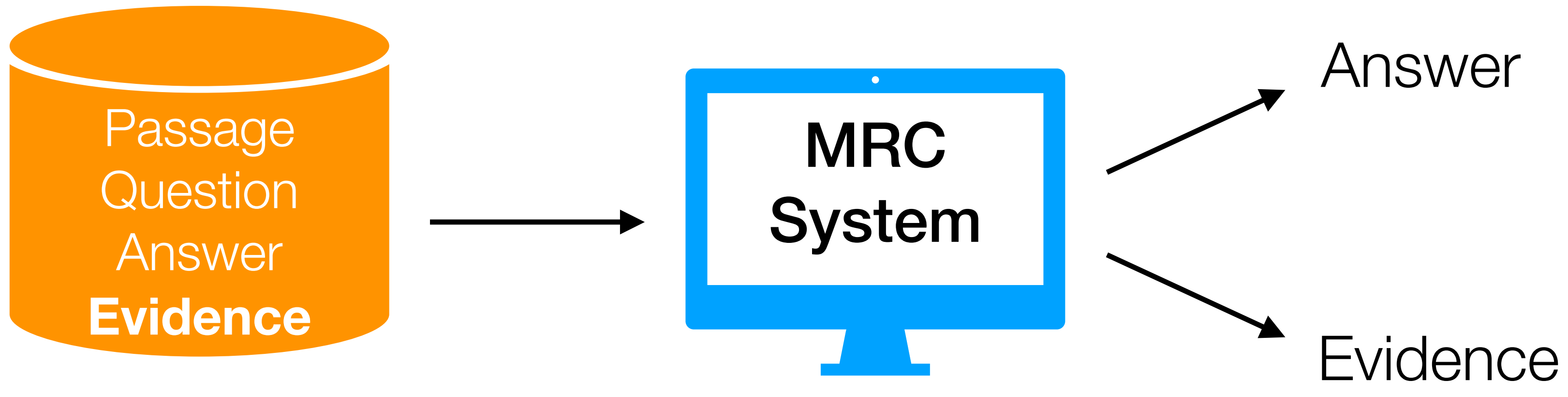
[Yang et al., MRQA 2021] Bilingual Alignment Pre-training for Zero-shot Cross-lingual Transfer

Towards Explainable MRC

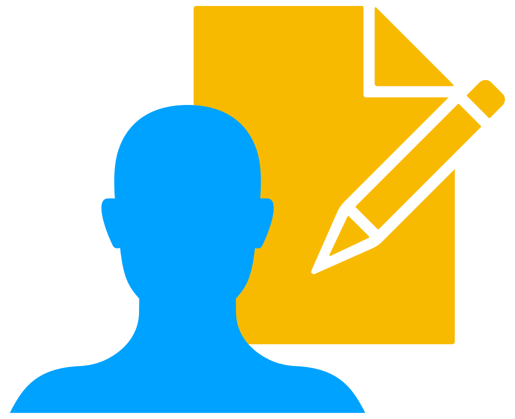
Explainable MRC

Post-hoc

Can we generate useful explanations in an unsupervised way?



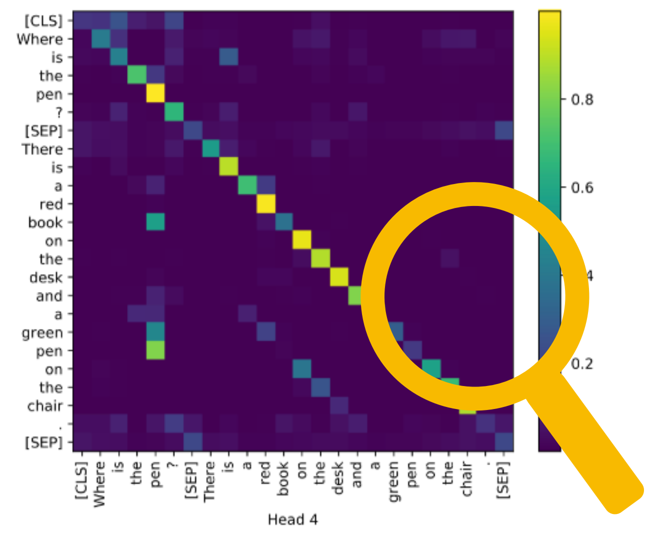
How to evaluate the quality of explanations?



Human Evaluation

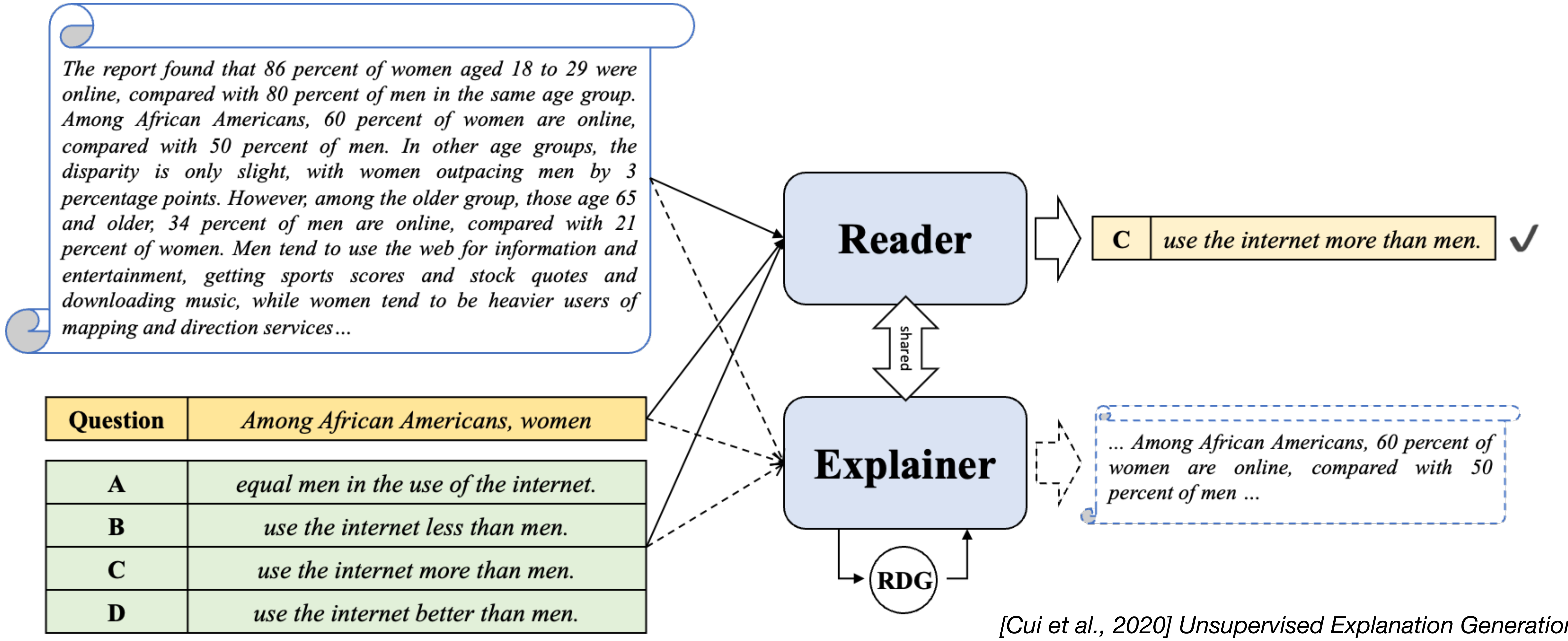
Intrinsic

What's the differences in attention map for MRC models?



RDG

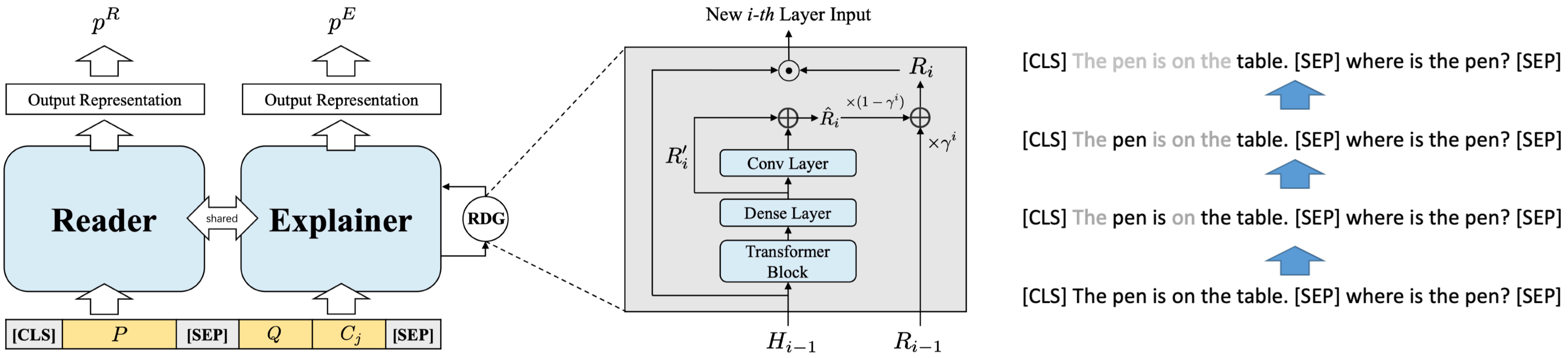
- **Unsupervised Explanation Generation for Machine Reading Comprehension**
  - Deep-learning based MRC systems lack explainability
  - Annotating rationale for MRC data is time-consuming and expensive
  - We propose a self-explainable MRC model: **Recursive Dynamic Gating (RDG)**



[Cui et al., 2020] Unsupervised Explanation Generation for Machine Reading Comprehension

RDG

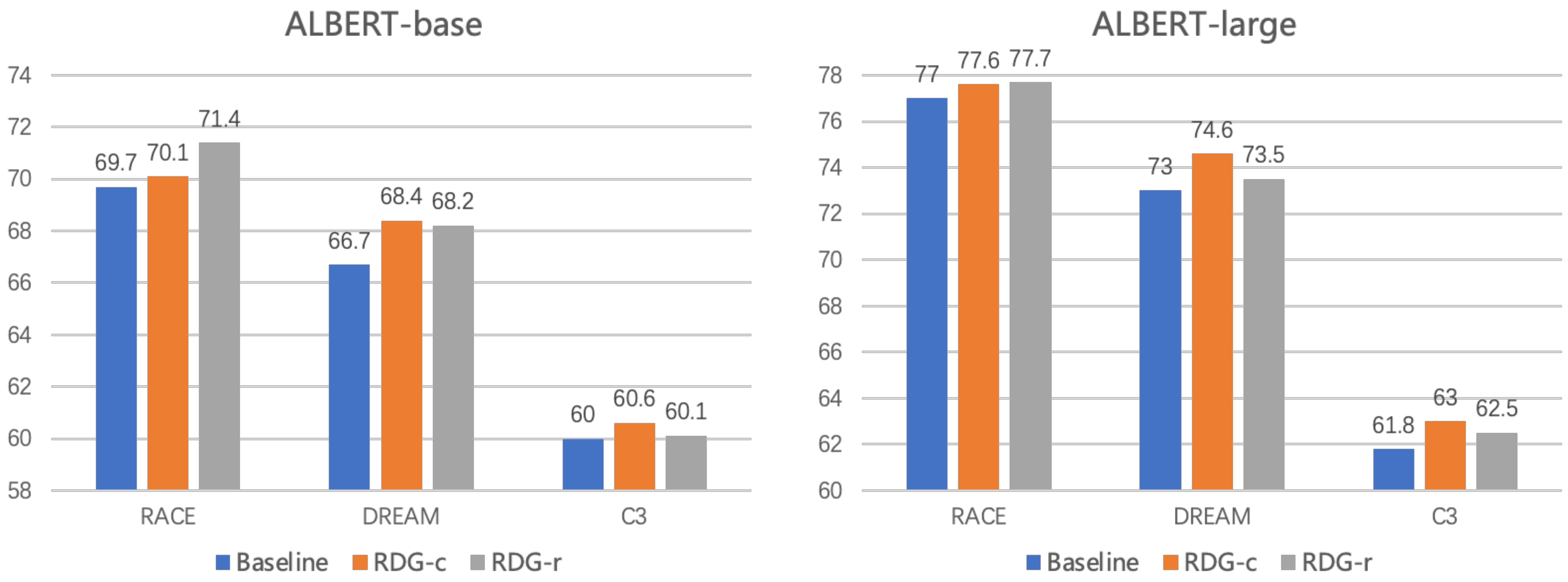
- **Recursive Dynamic Gating (RDG)**
  - Reader: Normal MRC system that learns to identify the correct answer
  - Explainer: Learn from Reader and try to find the most important words in the passage
  - Approach: (Soft-)filtering the passage information in each transformer layer



[Cui et al., 2020] Unsupervised Explanation Generation for Machine Reading Comprehension

RDG

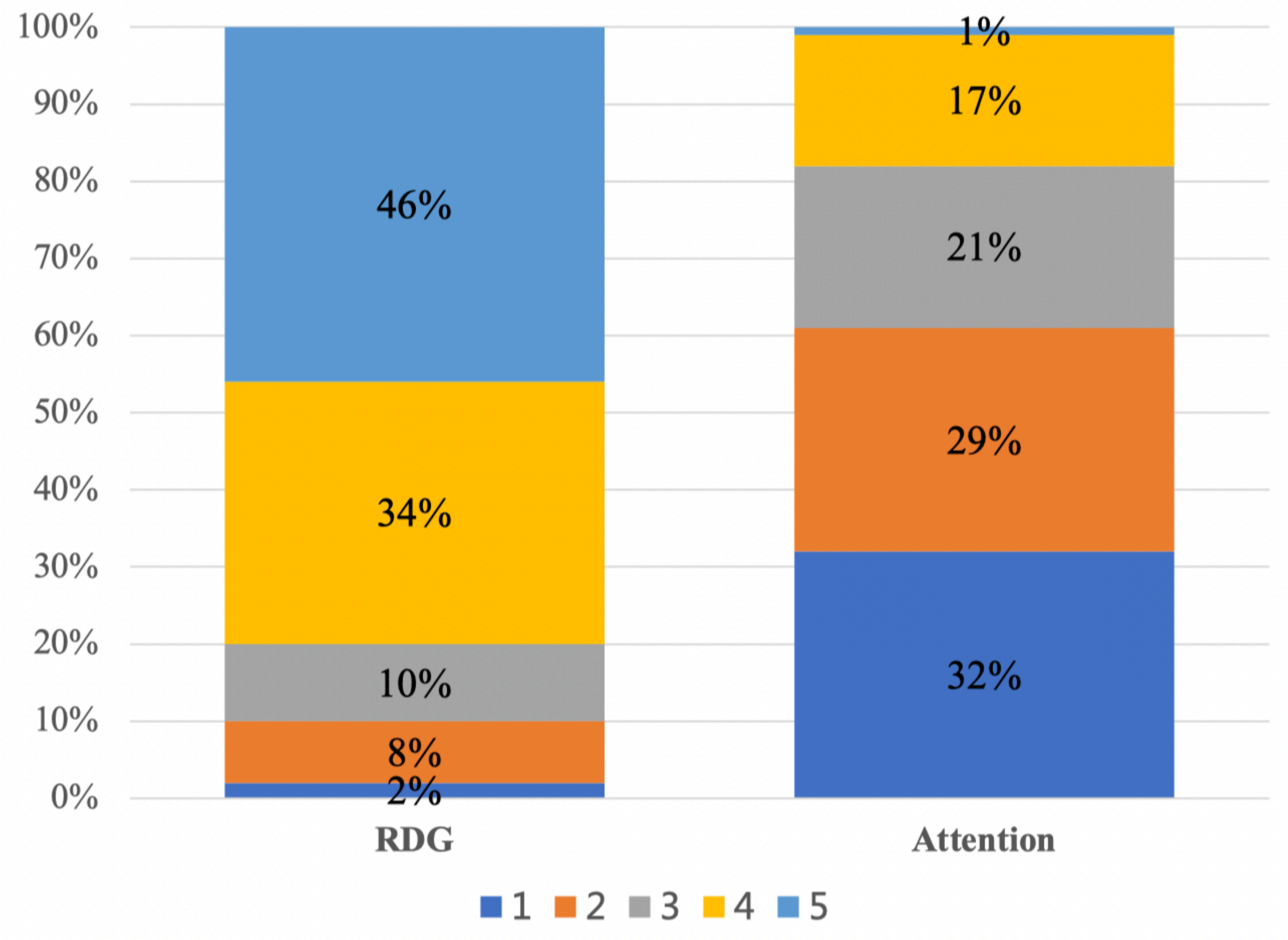
- **Experimental Results on RACE, DREAM, and C³ (zh)**
  - Applying RDG achieves better performance than non-explainable MRC systems
  - Explainability comes with no performance cost — Better answer prediction and explainability



[Cui et al., 2020] Unsupervised Explanation Generation for Machine Reading Comprehension

RDG

- **Quality of Explanation**
  - Human evaluation
    - RDG achieves an average score of **4.14** (out of 5), while attention achieves **2.26**
  - Quantitative evaluation
    - Hypothesis: Good explanation helps humans in question answering process
    - Setups: Input generated explanations and the question to the model, and compare which system could give higher answer accuracy
    - Results: The explanation generated by RDG has better accuracy in prediction, suggesting that it has much meaningful information

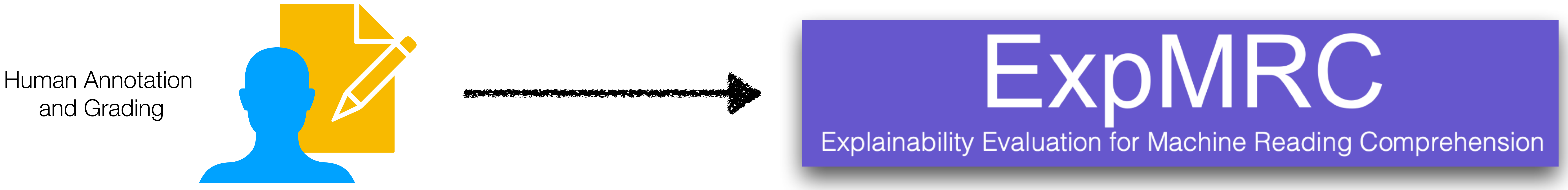


System	RACE Dev	RACE Test	DREAM Dev	DREAM Test	C^3 Dev	C^3 Test
Baseline	72.3	71.4	68.1	68.2	60.5	60.1
Att-GT	59.5	57.8	50.2	49.0	48.7	48.8
RDG-GT	71.3	69.1	69.5	67.5	64.6	64.4
Att-Pred	55.9	53.7	46.0	45.6	46.5	46.0
RDG-Pred	64.4	62.3	62.9	61.9	56.9	56.6

[Cui et al., 2020] Unsupervised Explanation Generation for Machine Reading Comprehension

ExpMRC

- **ExpMRC: Explainability Evaluation for Machine Reading Comprehension**
    - Explainability and interpretability is not well-studied in MRC
    - A new comprehensive benchmark for explainable MRC
    - Propose several unsupervised baselines for ExpMRC



[Cui et al., 2021] ExpMRC: Explainability Evaluation for Machine Reading Comprehension

ExpMRC

- **Dataset Annotation**
  - Including four subsets, featuring multilingual and multitask settings
  - Annotate a span in the passage as the evidence text
  - Principles
    - Not a simple combination of the question and answer
    - Encourage multi-sentence reasoning

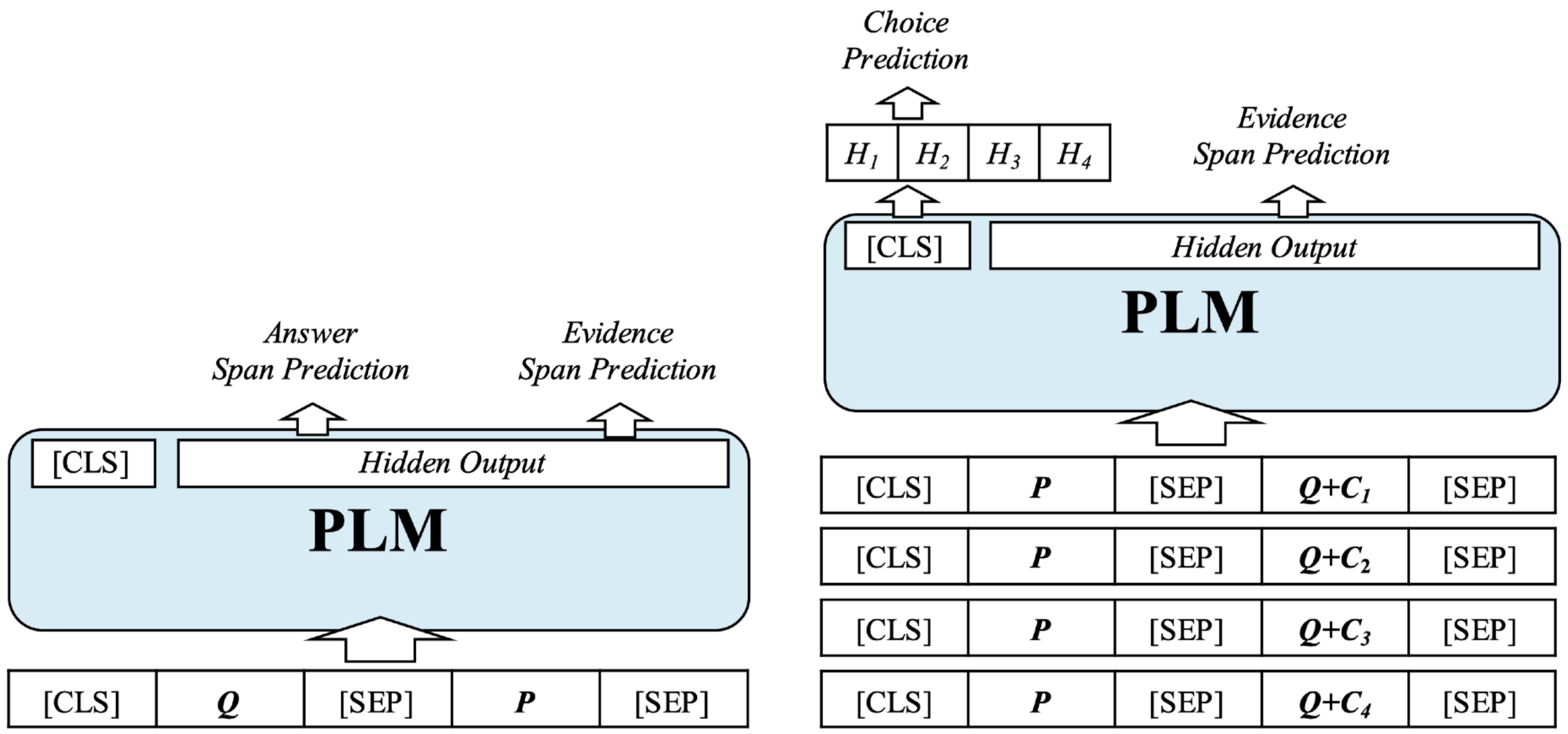
	English	Chinese
Span-Extraction	SQuAD	CMRC 2018
Multi-Choice	RACE+	C3

Subset	Passage	Question & Answer
SQuAD	...Competition amongst employers tends to drive up wages due to the nature of the job, since there is a relative shortage of workers for the particular position. Professional and labor organizations may limit the supply of workers which results in higher demand and greater incomes for members. Members may also receive higher wages through collective bargaining ...	Q: Who works to get workers higher compensation? A: Professional and labor organizations
CMRC 2018	...钩盲蛇（学名：“Ramphotyphlops braminus”）是蛇亚目盲蛇科下的一种无毒蛇种，主要分布在非洲及亚洲，不过现在钩盲蛇的分布已推广至世界各地。钩盲蛇是栖息于地洞的蛇种，由于体型细小，加上善于掘洞...	Q: 钩盲蛇一般生活在什么地形中? A: 地洞
RACE+	...My biology teacher, Mr. Clark, divided us into three groups and asked us to play a game about natural selection and how birds find food. He gave the first group one spoon to every student, the second group forks and my group knives. ...When I almost picked a bean, it dropped back to the ground. When I finally picked up several beans, one of my friends ran into me. I fell over. All my beans dropped to the ground! Just at that moment, Mr. Clark called us back. ...	Q: How many beans did the writer get at last? A: None. B: One. C: Several. D: Many.
C3	...大学生活是走上社会的预演，可以说，大学里的处世态度和人际关系的成功与否，直接决定着将来在社会上的成败。人是社会性的动物，生活中的每个人都离不开别人的帮助，同时也在帮助着别人。不管是学习、生活、工作，都要求自己要有良好的处理人际关系的能力。一个人要想有良好的人际关系，就要遵循以下几个原则：一是“主动”。要主动和别人交往，主动帮助别人。二是“诚信”。...	Q: 说话人认为什么因素决定在社会上的成败? A: 工作的态度 B: 朋友的数量 C: 大学里的学习成绩 D: 大学里的人际关系

[Cui et al., 2021] ExpMRC: Explainability Evaluation for Machine Reading Comprehension

ExpMRC

- **Unsupervised Baselines**
  - Non-learning baselines: Most Similar Sentence (w/ Question), Answer Sentence (SE-MRC only)
  - Machine Learning baselines: Pseudo-training data approach



(a) Span-Extraction MRC (b) Multi-Choice MRC

[Cui et al., 2021] ExpMRC: Explainability Evaluation for Machine Reading Comprehension

ExpMRC

- **Experimental Results**
  - Evaluation Metrics
    - Answer/Evidence/Overall F1
  - Finding evidences for span-extraction MRC is easier than multi-choice MRC
  - Using pseudo evidence data for training can also improve the accuracy of answer prediction
  - Overall, there is still a large gap between baselines and human performance, especially for multi-choice MRC settings

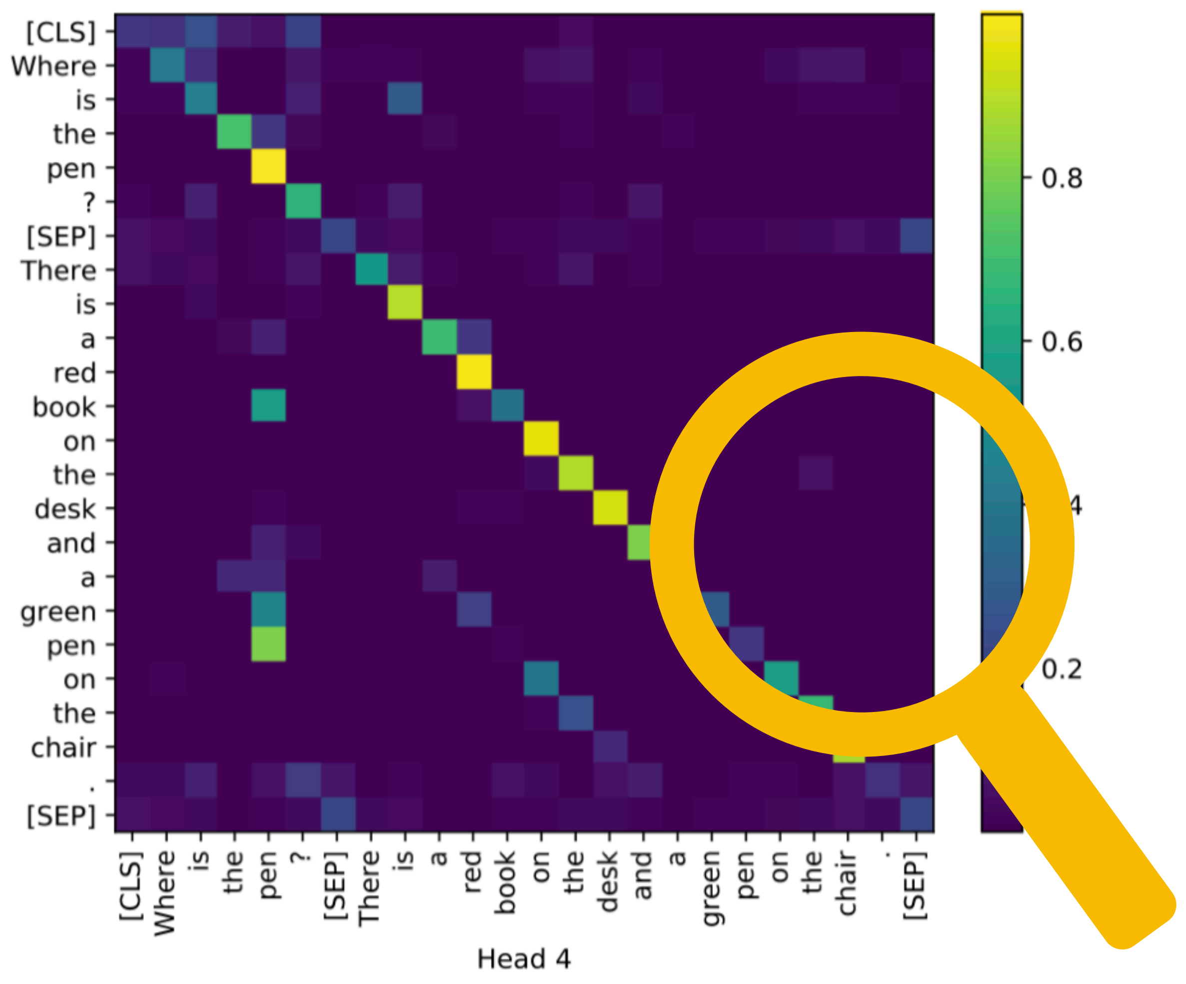
System	SQuAD (dev)			SQuAD (test)			CMRC 2018 (dev)			CMRC 2018 (test)		
	Ans.	Evi.	All	Ans.	Evi.	All	Ans.	Evi.	All	Ans.	Evi.	All
Human Performance	*90.8*	*92.1*	*83.6*	*91.3*	*92.9*	*84.7*	*97.7*	*94.6*	*92.4*	*97.9*	*94.6*	*92.6*
PLM Base-level Baselines												
Most Similar Sent.	**87.4**	81.8	74.5	87.1	85.4	76.1	**82.3**	71.9	60.1	84.4	62.2	52.9
Most Similar Sent. w/ Ques.	**87.4**	81.0	72.9	87.1	84.8	75.6	**82.3**	76.9	63.9	84.4	**69.8**	**59.9**
Predicted Answer Sent.	**87.4**	**84.1**	**76.4**	87.1	**89.1**	**79.6**	**82.3**	**78.0**	**66.8**	84.4	69.1	59.8
Pseudo-data Training	87.0	79.5	70.6	**88.0**	78.6	69.8	81.5	73.2	60.4	**85.9**	61.3	52.4
PLM Large-level Baselines												
Most Similar Sent.	**93.0**	83.9	79.3	92.3	85.7	80.4	82.8	71.6	60.3	88.6	63.0	55.9
Most Similar Sent. w/ Ques.	**93.0**	81.9	77.4	92.3	85.1	79.8	82.8	76.3	63.6	88.6	**71.0**	63.2
Predicted Answer Sent.	**93.0**	**85.4**	**81.8**	92.3	**89.6**	**83.6**	82.8	**77.7**	**66.9**	88.6	70.6	**63.3**
Pseudo-data Training	92.9	80.7	75.6	**93.9**	80.1	74.8	**83.8**	73.1	62.7	**89.6**	62.9	55.3

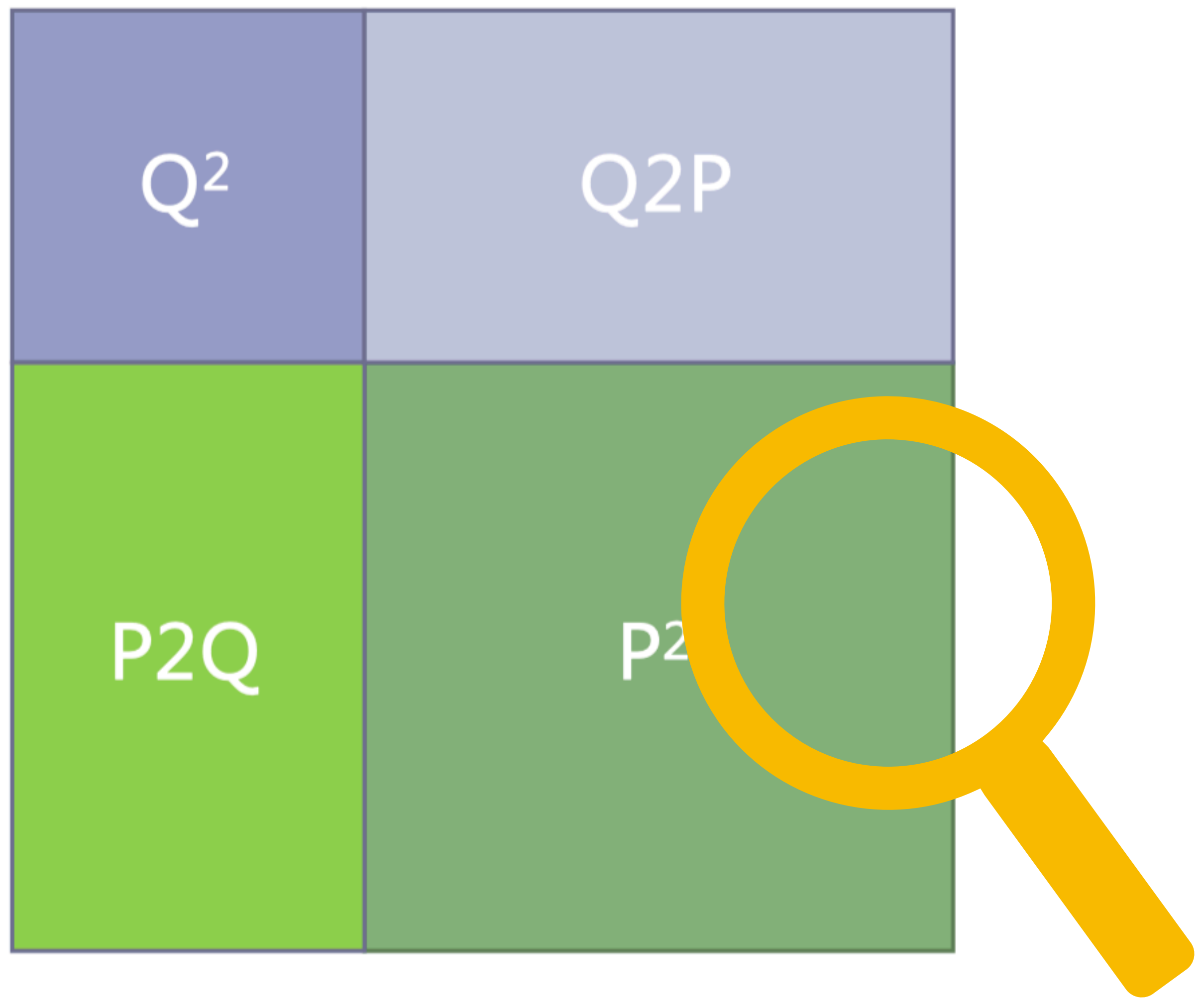
System	RACE$^+$ (dev)			RACE$^+$ (test)			C^3 (dev)			C^3 (test)		
	Ans.	Evi.	All	Ans.	Evi.	All	Ans.	Evi.	All	Ans.	Evi.	All
Human Performance	*92.0*	*92.4*	*85.4*	*93.6*	*90.5*	*84.4*	*95.3*	*95.7*	*91.1*	*94.3*	*97.7*	*90.0*
PLM Base-level Baselines												
Most Similar Sent.	62.4	36.6	28.2	59.8	34.4	26.3	68.7	57.7	**47.7**	66.8	52.2	41.2
Most Similar Sent. w/ Ques.	62.4	44.5	31.5	59.8	41.8	**27.3**	68.7	**62.3**	47.3	66.8	57.4	**42.3**
Pseudo-data Training	**63.6**	**45.7**	**31.7**	**60.1**	**43.5**	27.1	**70.9**	59.9	43.5	**69.0**	**57.5**	40.6
PLM Large-level Baselines												
Most Similar Sent.	**69.0**	37.6	29.9	68.1	36.8	28.9	73.1	59.4	49.9	72.0	52.7	43.9
Most Similar Sent. w/ Ques.	**69.0**	**48.0**	**36.8**	68.1	**42.5**	**31.3**	73.1	63.2	**50.9**	72.0	58.4	46.0
Pseudo-data Training	**69.0**	45.9	32.6	**70.4**	41.3	30.8	**76.4**	**64.3**	50.7	**74.4**	**59.9**	**47.3**

[Cui et al., 2021] ExpMRC: Explainability Evaluation for Machine Reading Comprehension

Attention in MRC

- **Understanding Attention in Machine Reading Comprehension**
  - Should we analyze the attention map as a whole?
  - What's the differences in attention map for MRC models?

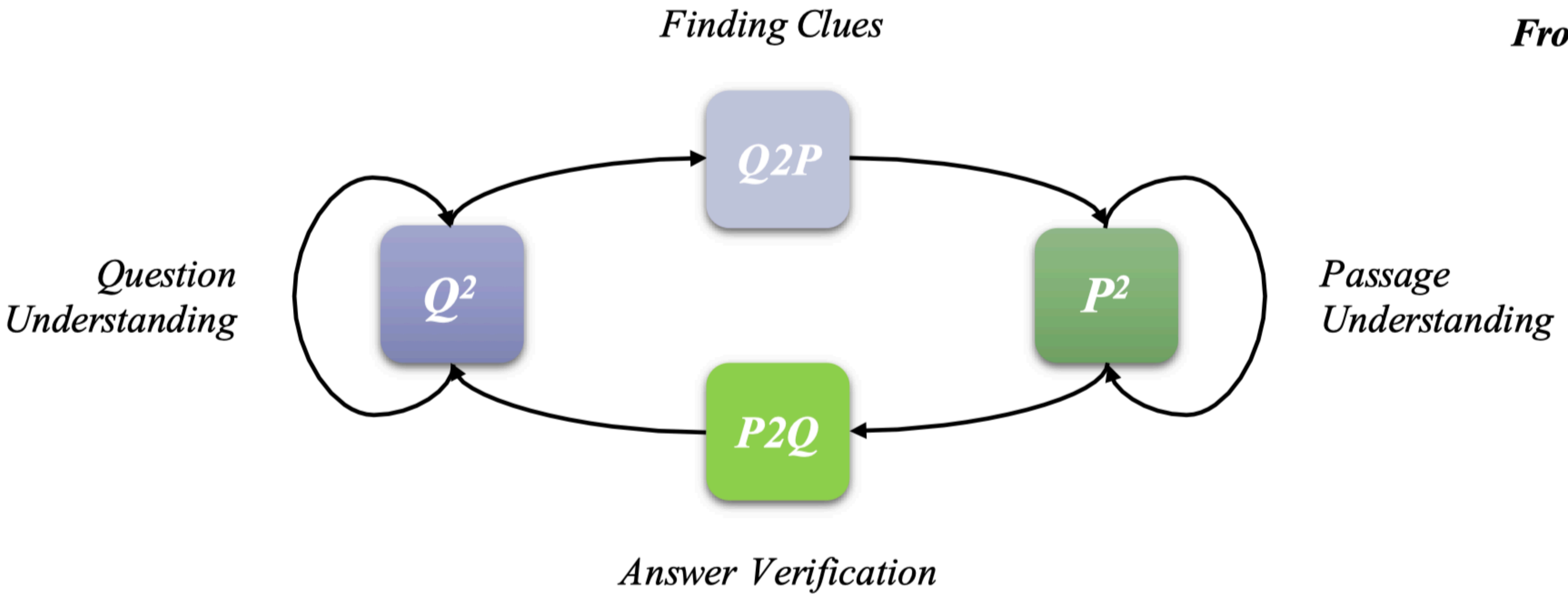


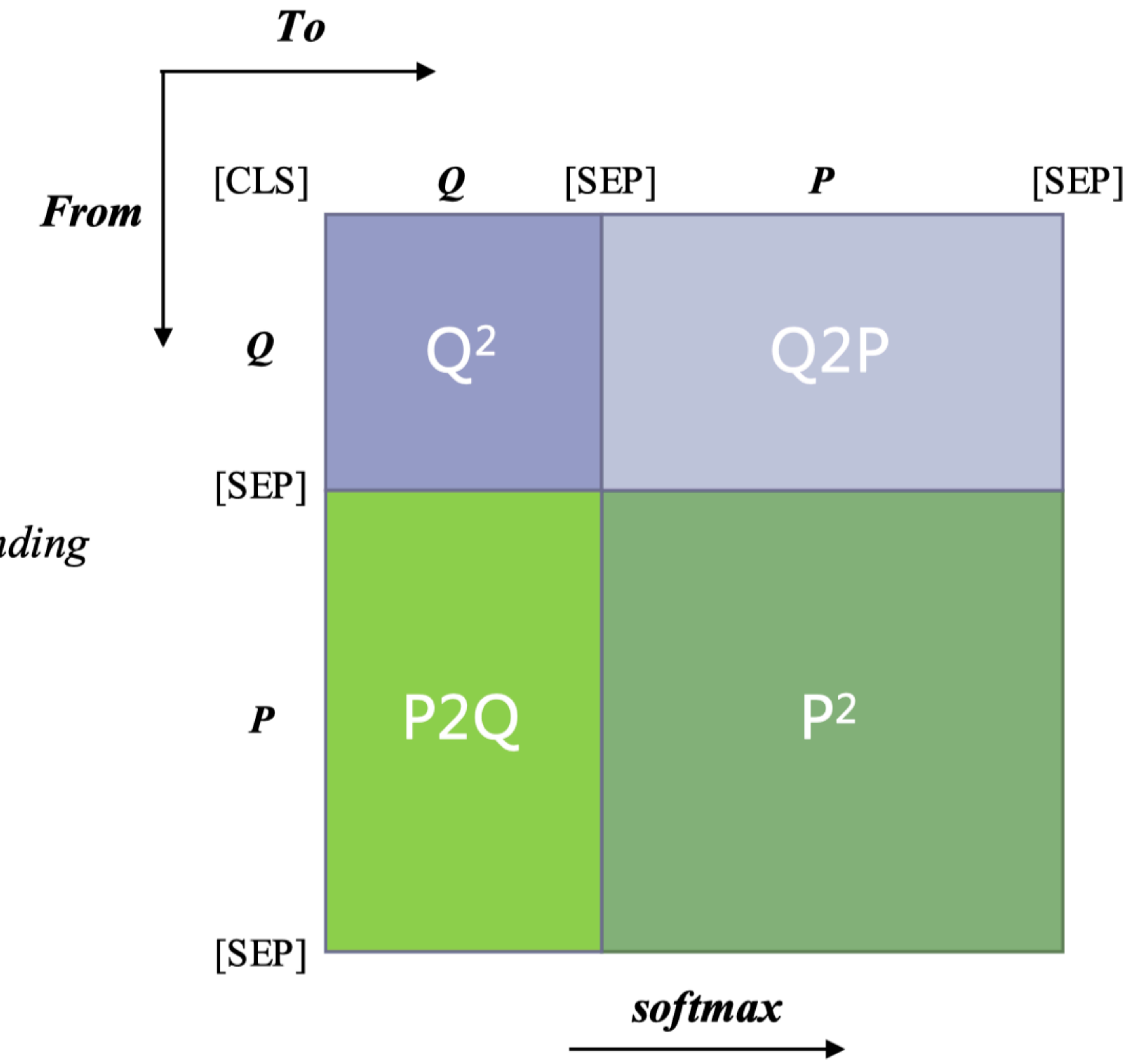


[Cui et al., 2021] *Understanding Attention in Machine Reading Comprehension*

Attention in MRC

- **Attention Zones for MRC**
  - Typical input format: [CLS] Question [SEP] Passage [SEP]
  - Divide attention matrix into four zones: Q^2, Q2P, P2Q, P^2

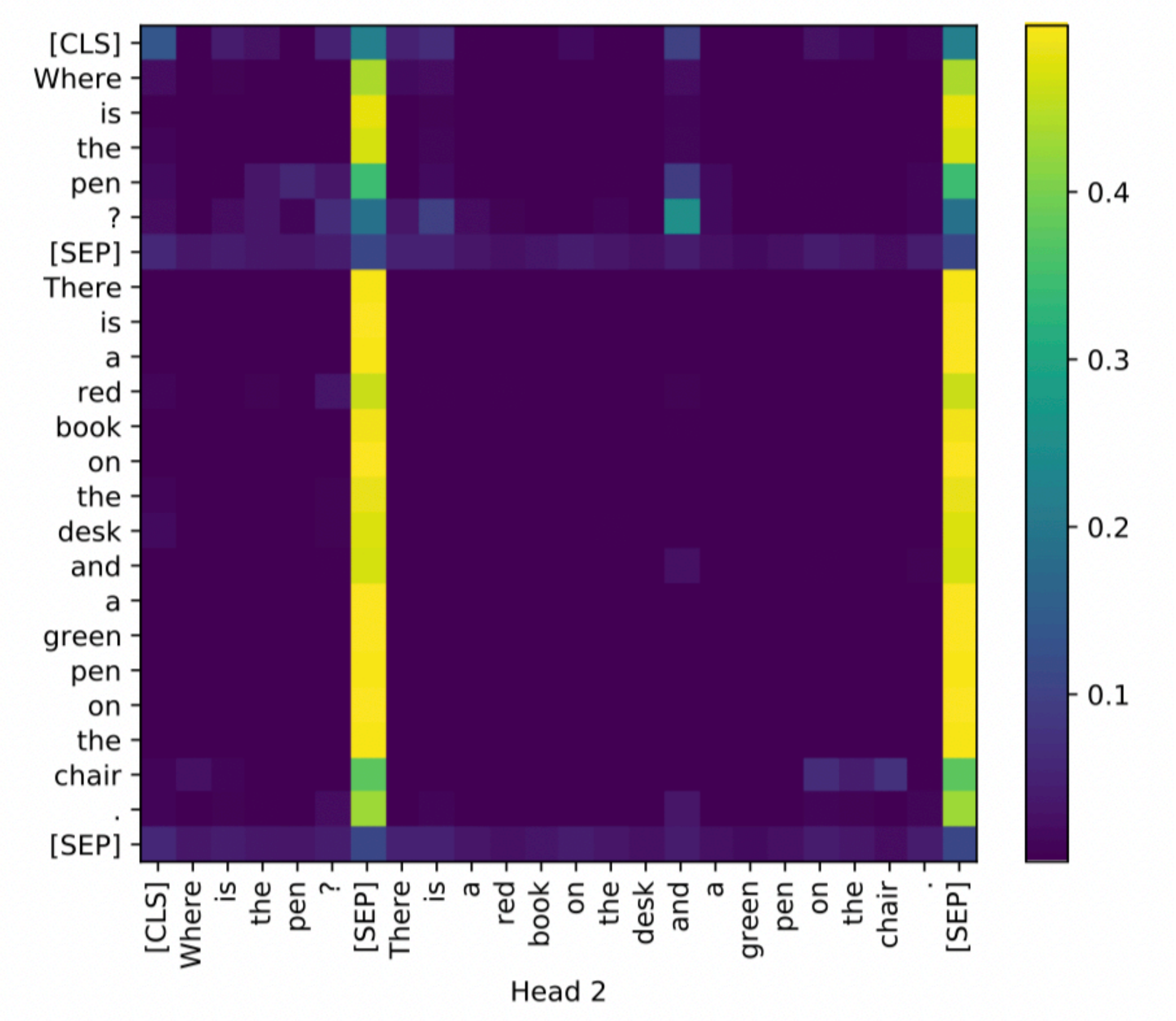


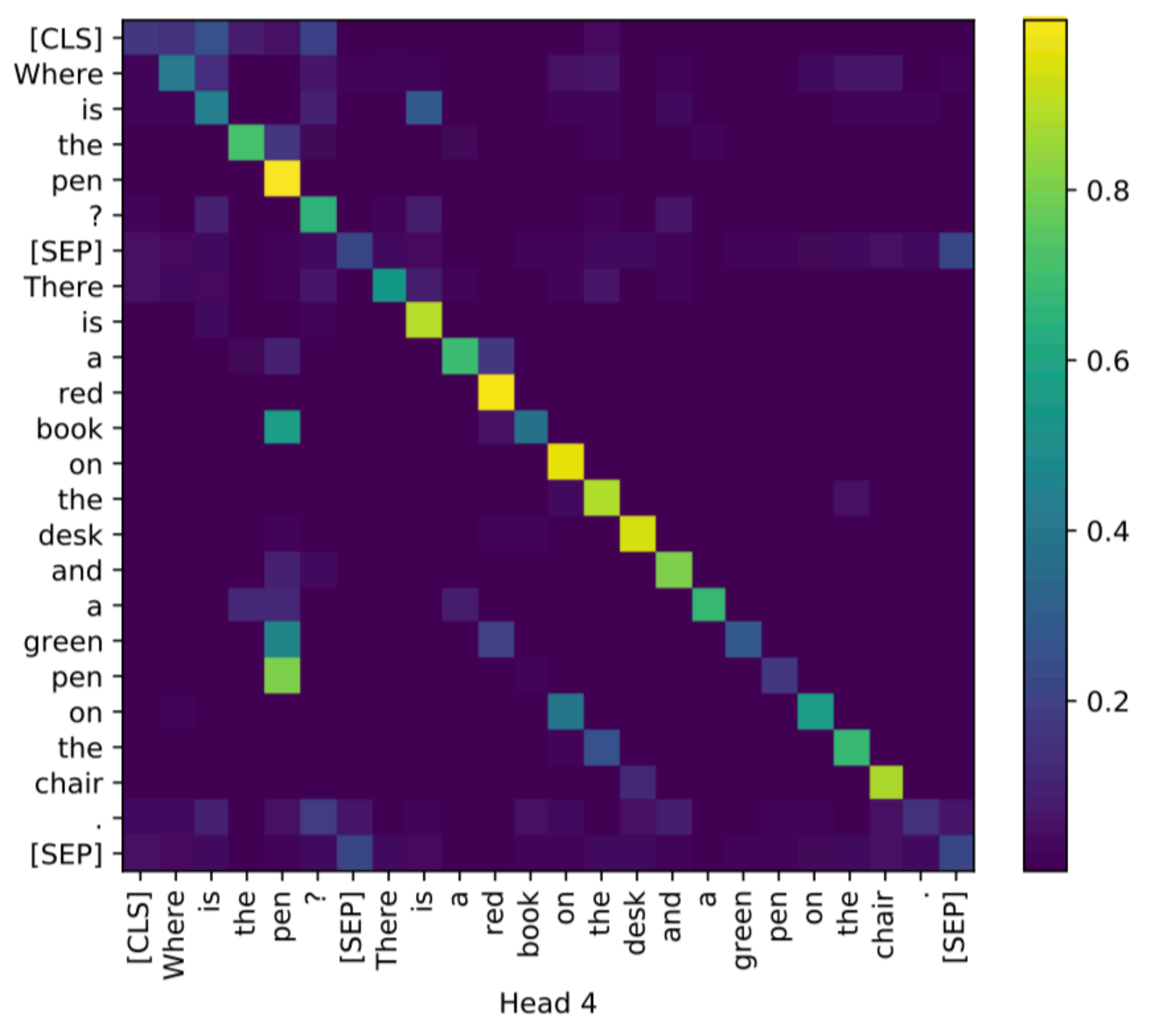


[Cui et al., 2021] Understanding Attention in Machine Reading Comprehension

ATTENTION IN MRC

- **Higher Attention Value ≠ Higher Impact on Performance**
  - Kovaleva et al. (2019): Higher attention values for special tokens and diagonal elements
  - Let's remove (mask) those tokens to see if they are important to answer prediction
  - Observation: Not all these tokens are critical to performance



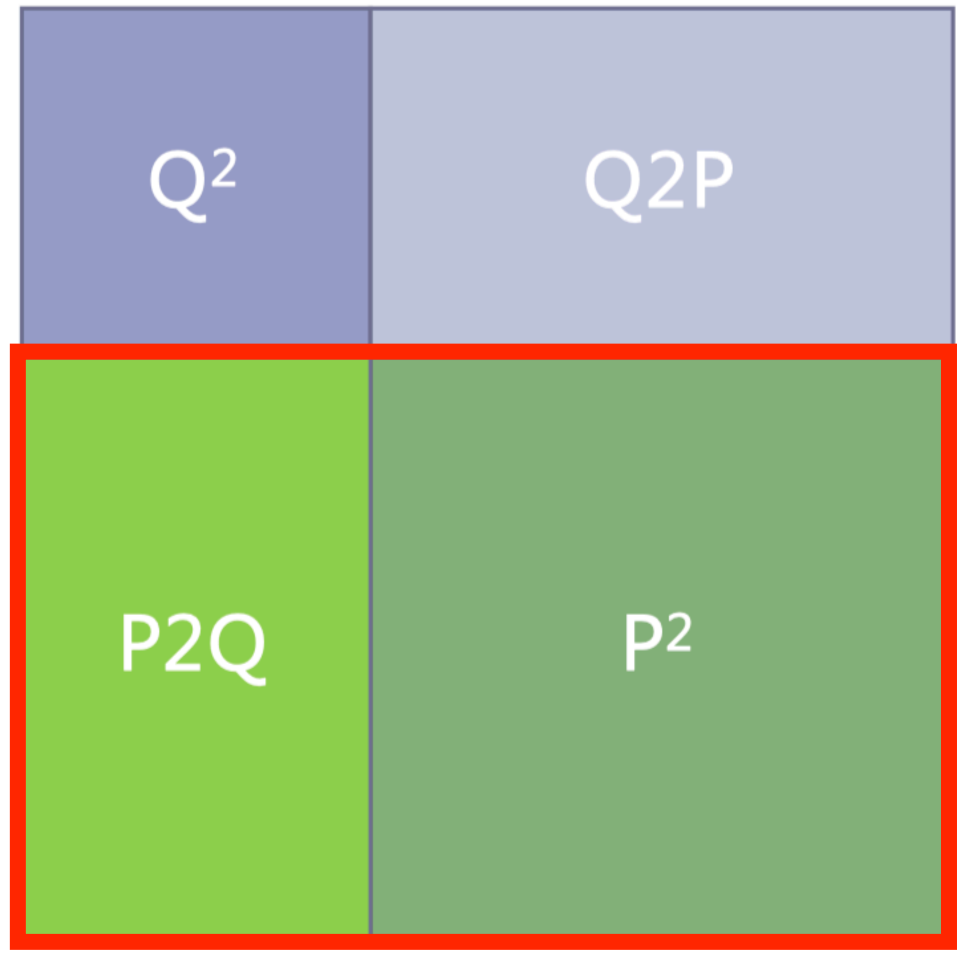


	SQuAD EM	SQuAD F1	CMRC 2018 EM	CMRC 2018 F1
Baseline	80.687	88.129	63.796	84.789
No [CLS]	80.802	88.276	64.119	84.858
No Mid [SEP]	80.689	88.082	63.896	84.626
No End [SEP]	80.522	87.959	64.299	84.866
No All	78.956	86.414	63.659	83.945
No Diagonal	80.645	88.241	64.548	84.908
No Q^2	76.395	84.195	60.100	80.625
No Q2P	79.941	87.352	64.517	84.592
No P2Q	12.763	16.355	15.070	18.466
No P^2	34.441	51.792	16.278	42.906

[Cui et al., 2021] Understanding Attention in Machine Reading Comprehension

Attention in MRC

- **High correlation in P2 and P2Q**
  - Experiment 1: Removing top-10 attention values in each attention zone
  - Experiment 2: Correlation of masking top-k^{th} attention value and its rank (k)
  - Overall, P2Q and P^2 seems to highly correlate with answer prediction



	SQuAD (en)	CMRC 2018 (zh)
Q^2	65.272	58.652
Q2P	79.743	63.324
P2Q	45.790	43.939
P^2	78.412	63.175

▲ Exp1: Removing Top-10 attention values

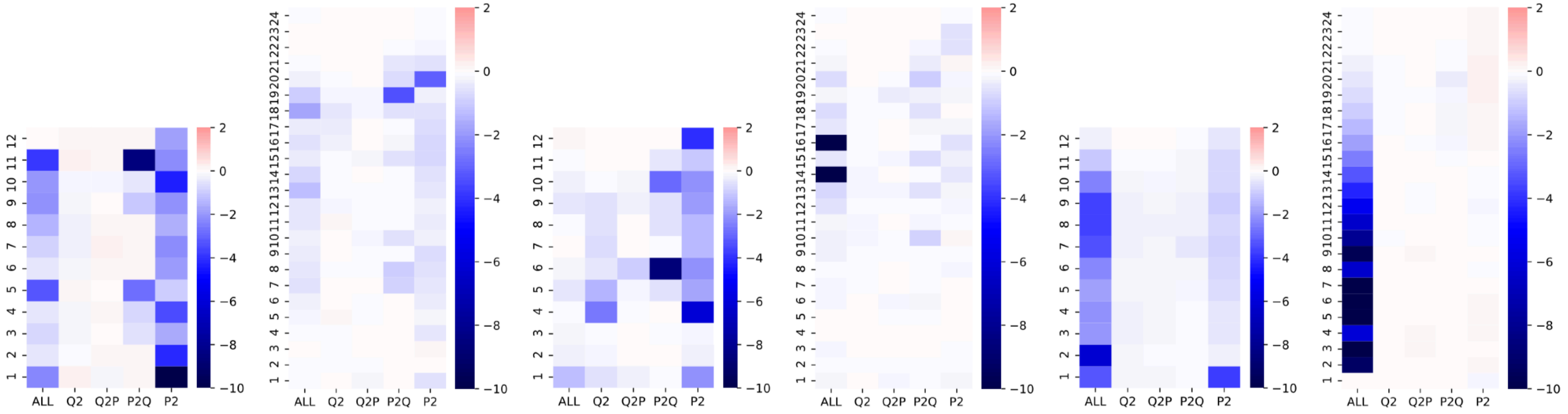
	SQuAD (en)	CMRC 2018 (zh)
Q^2	0.624 ± 0.083	-0.316 ± 0.370
Q2P	0.159 ± 0.435	0.134 ± 0.531
P2Q	0.765 ± 0.017	0.778 ± 0.118
P^2	0.534 ± 0.216	0.291 ± 0.299

▲ Exp2: Correlation of masking top-k^{th} attention value and its rank

[Cui et al., 2021] Understanding Attention in Machine Reading Comprehension

Attention in MRC

- **Different Patterns for Different PLMs**
  - Investigating behaviors of different attention zones in different PLMs (BERT, ELECTRA, ALBERT)
  - P2Q and P^2 are the most important attention zones to the performance
  - Large models are more robust than base models (knowledge distributions)
  - Cross-layer parameter sharing (ALBERT) makes it a unique pattern to other PLMs



(a) $\text{BERT}_{\text{base}}$ (b) $\text{BERT}_{\text{large}}$ (c) $\text{ELECTRA}_{\text{base}}$ (d) $\text{ELECTRA}_{\text{large}}$ (e) $\text{ALBERT}_{\text{base}}$ (f) $\text{ALBERT}_{\text{large}}$

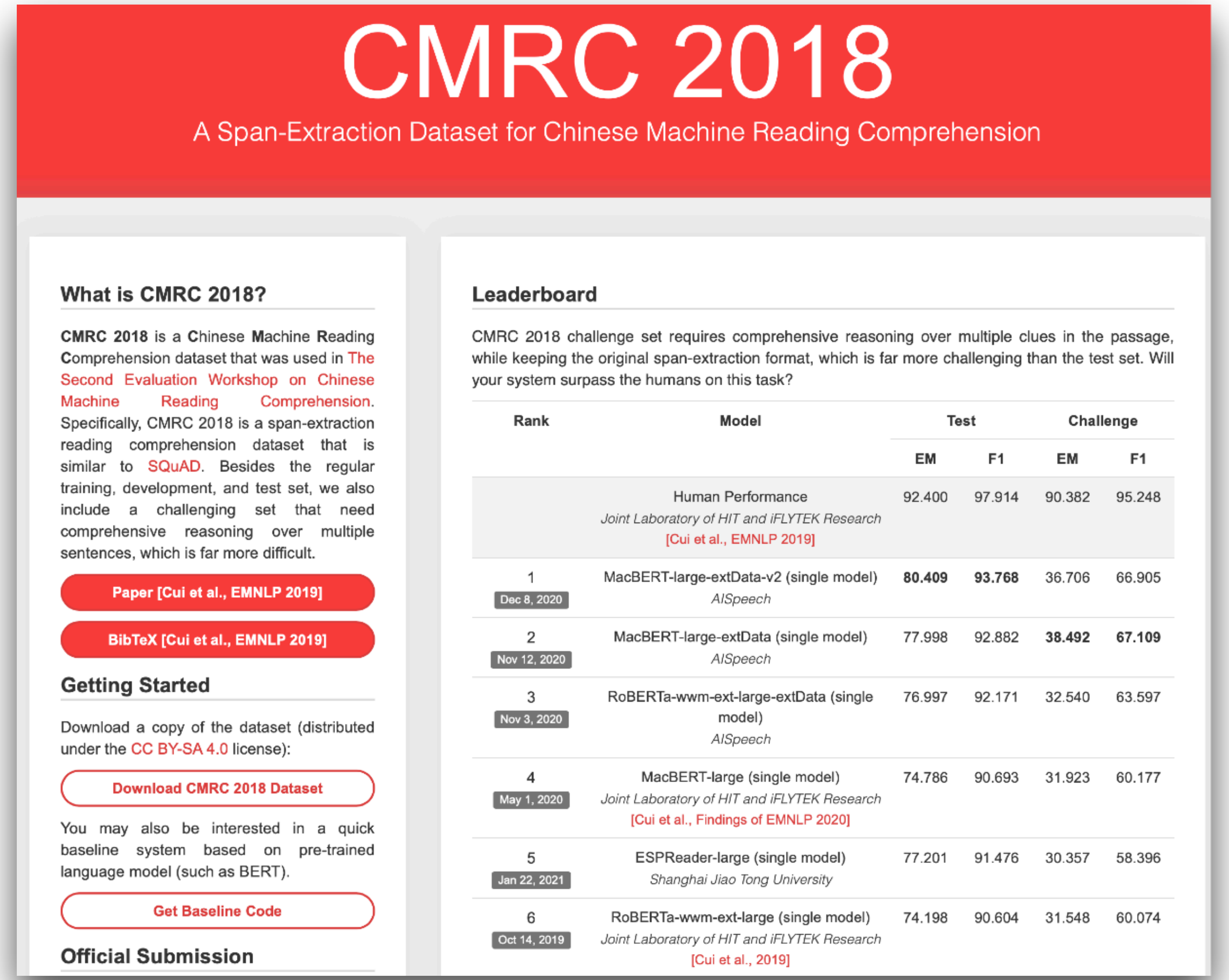
[Cui et al., 2021] Understanding Attention in Machine Reading Comprehension

SUMMARY

- **Chinese MRC**
  - A series of Chinese MRC datasets and pre-trained language models
- **Multilingual & Cross-lingual MRC**
  - **DualBERT**: Enhance Chinese MRC performance by utilizing English data
  - **WEAM**: Enhance cross-lingual ability with the knowledge from the alignment matrix
- **Explainable MRC**
  - **RDG**: Extend MRC system with explainable post-hoc explanations
  - **ExpMRC**: Evaluating explanation extraction for MRC systems
  - **Attention in MRC**: analyzing attention behavior specifically for MRC tasks

SUBMISSIONS

- **Participate in Our MRC Evaluations**
  - Open submissions for CMRC 2018 (zh), CMRC 2019 (zh), and ExpMRC (zh/en)



CMRC 2018

A Span-Extraction Dataset for Chinese Machine Reading Comprehension

What is CMRC 2018?

CMRC 2018 is a **C**hinese **M**achine **R**eading **C**omprehension dataset that was used in The Second Evaluation Workshop on Chinese Machine Reading Comprehension. Specifically, CMRC 2018 is a span-extraction reading comprehension dataset that is similar to SQuAD. Besides the regular training, development, and test set, we also include a challenging set that need comprehensive reasoning over multiple sentences, which is far more difficult.

Paper [Cui et al., EMNLP 2019]

BibTeX [Cui et al., EMNLP 2019]

Getting Started

Download a copy of the dataset (distributed under the CC BY-SA 4.0 license):

Download CMRC 2018 Dataset

You may also be interested in a quick baseline system based on pre-trained language model (such as BERT).

Get Baseline Code

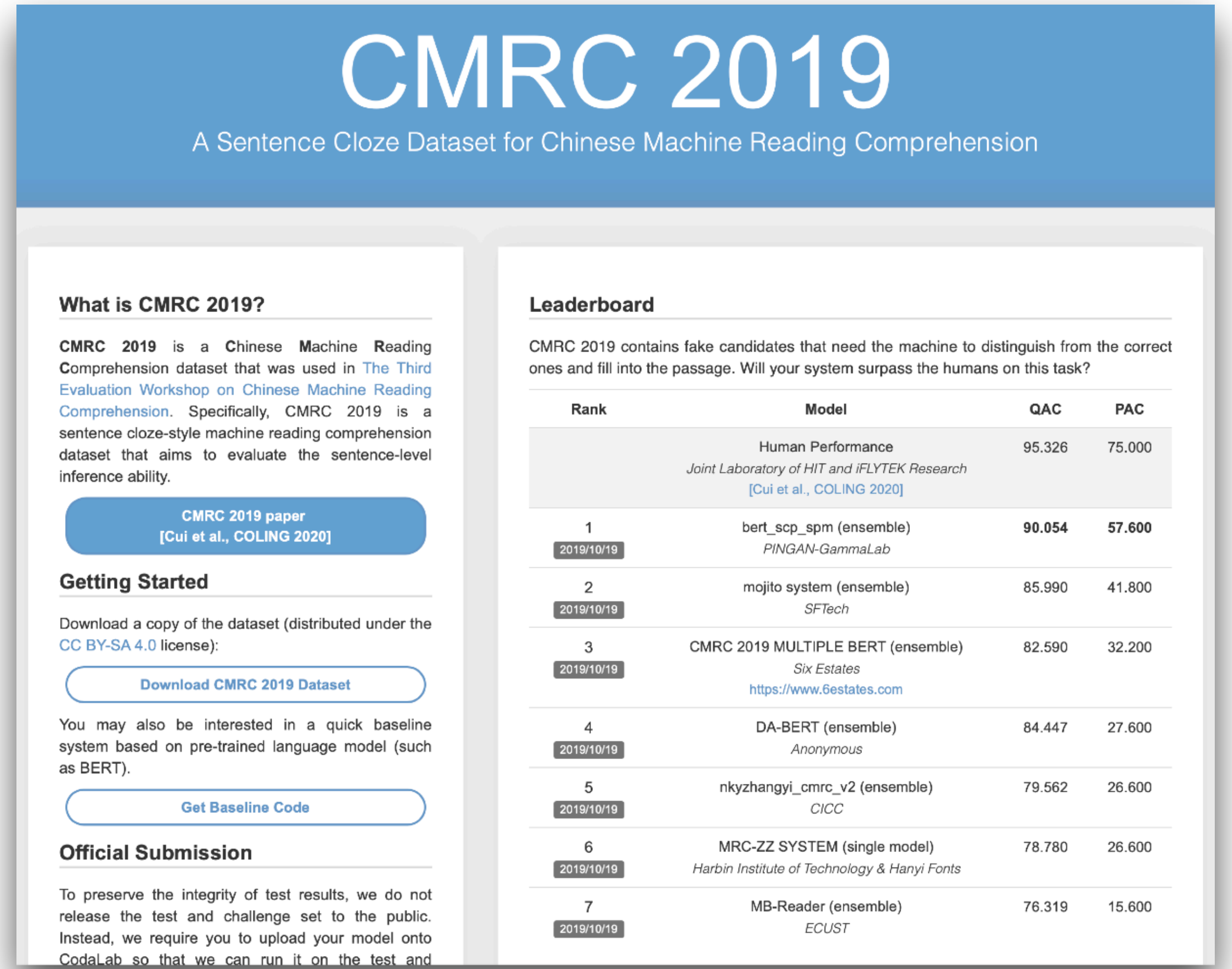
Official Submission

Leaderboard

CMRC 2018 challenge set requires comprehensive reasoning over multiple clues in the passage, while keeping the original span-extraction format, which is far more challenging than the test set. Will your system surpass the humans on this task?

Rank	Model	Test		Challenge	
		EM	F1	EM	F1
	Human Performance Joint Laboratory of HIT and iFLYTEK Research [Cui et al., EMNLP 2019]	92.400	97.914	90.382	95.248
1 Dec 8, 2020	MacBERT-large-extData-v2 (single model) AISpeech	**80.409**	**93.768**	36.706	66.905
2 Nov 12, 2020	MacBERT-large-extData (single model) AISpeech	77.998	92.882	**38.492**	**67.109**
3 Nov 3, 2020	RoBERTa-wwm-ext-large-extData (single model) AISpeech	76.997	92.171	32.540	63.597
4 May 1, 2020	MacBERT-large (single model) Joint Laboratory of HIT and iFLYTEK Research [Cui et al., Findings of EMNLP 2020]	74.786	90.693	31.923	60.177
5 Jan 22, 2021	ESPReader-large (single model) Shanghai Jiao Tong University	77.201	91.476	30.357	58.396
6 Oct 14, 2019	RoBERTa-wwm-ext-large (single model) Joint Laboratory of HIT and iFLYTEK Research [Cui et al., 2019]	74.198	90.604	31.548	60.074

https://ymcui.com/cmrc2018/



CMRC 2019

A Sentence Cloze Dataset for Chinese Machine Reading Comprehension

What is CMRC 2019?

CMRC 2019 is a **C**hinese **M**achine **R**eading **C**omprehension dataset that was used in The Third Evaluation Workshop on Chinese Machine Reading Comprehension. Specifically, CMRC 2019 is a sentence cloze-style machine reading comprehension dataset that aims to evaluate the sentence-level inference ability.

CMRC 2019 paper [Cui et al., COLING 2020]

Getting Started

Download a copy of the dataset (distributed under the CC BY-SA 4.0 license):

Download CMRC 2019 Dataset

You may also be interested in a quick baseline system based on pre-trained language model (such as BERT).

Get Baseline Code

Official Submission

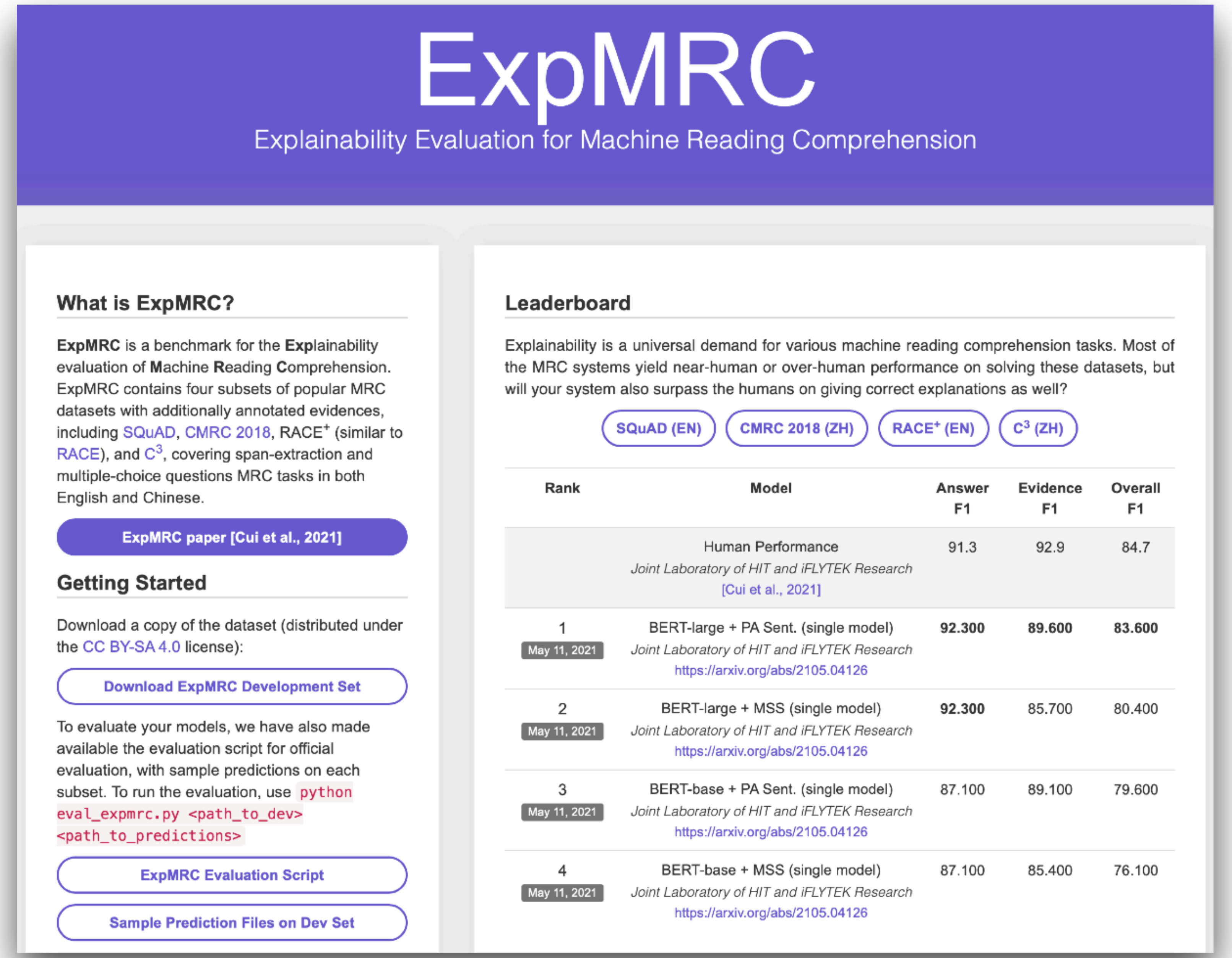
To preserve the integrity of test results, we do not release the test and challenge set to the public. Instead, we require you to upload your model onto CodaLab so that we can run it on the test and

Leaderboard

CMRC 2019 contains fake candidates that need the machine to distinguish from the correct ones and fill into the passage. Will your system surpass the humans on this task?

Rank	Model	QAC	PAC
	Human Performance Joint Laboratory of HIT and iFLYTEK Research [Cui et al., COLING 2020]	95.326	75.000
1 2019/10/19	bert_scp_spm (ensemble) PINGAN-GammaLab	**90.054**	**57.600**
2 2019/10/19	mojito system (ensemble) SFTech	85.990	41.800
3 2019/10/19	CMRC 2019 MULTIPLE BERT (ensemble) Six Estates https://www.6estates.com	82.590	32.200
4 2019/10/19	DA-BERT (ensemble) Anonymous	84.447	27.600
5 2019/10/19	nkyzhangyi_cmrc_v2 (ensemble) CICC	79.562	26.600
6 2019/10/19	MRC-ZZ SYSTEM (single model) Harbin Institute of Technology & Hanyi Fonts	78.780	26.600
7 2019/10/19	MB-Reader (ensemble) ECUST	76.319	15.600

https://ymcui.com/cmrc2019/



ExpMRC

Explainability Evaluation for Machine Reading Comprehension

What is ExpMRC?

ExpMRC is a benchmark for the **Exp**lainability evaluation of **M**achine **R**eading **C**omprehension. ExpMRC contains four subsets of popular MRC datasets with additionally annotated evidences, including SQuAD, CMRC 2018, RACE+ (similar to RACE), and C3, covering span-extraction and multiple-choice questions MRC tasks in both English and Chinese.

ExpMRC paper [Cui et al., 2021]

Getting Started

Download a copy of the dataset (distributed under the CC BY-SA 4.0 license):

Download ExpMRC Development Set

To evaluate your models, we have also made available the evaluation script for official evaluation, with sample predictions on each subset. To run the evaluation, use `python eval_expmrc.py <path_to_dev> <path_to_predictions>`

ExpMRC Evaluation Script

Sample Prediction Files on Dev Set

Leaderboard

Explainability is a universal demand for various machine reading comprehension tasks. Most of the MRC systems yield near-human or over-human performance on solving these datasets, but will your system also surpass the humans on giving correct explanations as well?

SQuAD (EN) | CMRC 2018 (ZH) | RACE+ (EN) | C3 (ZH)

Rank	Model	Answer F1	Evidence F1	Overall F1
	Human Performance Joint Laboratory of HIT and iFLYTEK Research [Cui et al., 2021]	91.3	92.9	84.7
1 May 11, 2021	BERT-large + PA Sent. (single model) Joint Laboratory of HIT and iFLYTEK Research https://arxiv.org/abs/2105.04126	**92.300**	**89.600**	**83.600**
2 May 11, 2021	BERT-large + MSS (single model) Joint Laboratory of HIT and iFLYTEK Research https://arxiv.org/abs/2105.04126	**92.300**	85.700	80.400
3 May 11, 2021	BERT-base + PA Sent. (single model) Joint Laboratory of HIT and iFLYTEK Research https://arxiv.org/abs/2105.04126	87.100	89.100	79.600
4 May 11, 2021	BERT-base + MSS (single model) Joint Laboratory of HIT and iFLYTEK Research https://arxiv.org/abs/2105.04126	87.100	85.400	76.100

https://ymcui.com/expmrc/

USEFUL RESOURCES

- CMRC 2017 (Cui et al., LREC 2018)
  - https://github.com/ymcui/cmrc2017
- CMRC 2018 (Cui et al., EMNLP 2019)
  - https://github.com/ymcui/cmrc2018
- CMRC 2019 (Cui et al., COLING 2020)
  - https://github.com/ymcui/cmrc2019
- ExpMRC (Cui et al., 2021)
  - https://github.com/ymcui/expmrc
- Chinese PLMs: BERT-wwm, RoBERTa, XLNet, ELECTRA, MacBERT (Cui et al., IEEE/ACM TASLP, Findings of EMNLP 2020)
  - https://github.com/ymcui/Chinese-BERT-wwm
  - https://github.com/ymcui/Chinese-XLNet
  - https://github.com/ymcui/Chinese-ELECTRA
  - https://github.com/ymcui/MacBERT

References

- Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In NeurIPS 2015.
- Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. arXiv preprint arXiv:1511.02301.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In EMNLP 2016.
- Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for Chinese machine reading comprehension. In EMNLP-IJCNLP 2019.
- Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In EMNLP 2013.
- Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In EMNLP 2017.
- Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2020. Investigating Prior Knowledge for Challenging Chinese Machine Reading Comprehension. In TACL.
- Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. Transactions of the Association for Computational Linguistics, 7:249–266.
- Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2018. DRCD: a Chinese machine reading comprehension dataset. arXiv preprint arXiv:1806.00920.
- Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In EMNLP 2018.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In EMNLP 2018.
- Kovaleva, O.; Romanov, A.; Rogers, A.; and Rumshisky, A. 2019. Revealing the Dark Secrets of BERT. In EMNLP 2019.

References

- Rowan Zellers, Yonatan Bisk, Ali Farhadi, Yejin Choi. 2019. From Recognition to Cognition: Visual Commonsense Reasoning. In CVPR 2019.
- Yiming Cui, Ting Liu, Zhipeng Chen, Shijin Wang, and Guoping Hu. 2016. Consensus attention-based neural networks for Chinese reading comprehension. In Proceedings of COLING 2016.
- Yiming Cui, Ting Liu, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2018. Dataset for the first evaluation on Chinese machine reading comprehension. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pages 2721–2725, Paris, France, may. European Language Resources Association (ELRA).
- Yiming Cui, Ting Liu, Ziqing Yang, Zhipeng Chen, Wentao Ma, Wanxiang Che, Shijin Wang, Guoping Hu. 2020. A Sentence Cloze Dataset for Chinese Machine Reading Comprehension. In COLING 2020.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang. 2019. Pre-training with whole word masking for Chinese BERT. In IEEE/ACM TASLP.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. Findings of EMNLP 2020.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, Guoping Hu. 2019. Cross-Lingual Machine Reading Comprehension. In EMNLP 2019.
- Ziqing Yang, Wentao Ma, Yiming Cui, Jiani Ye, Wanxiang Che, Shijin Wang. Bilingual Alignment Pre-training for Zero-shot Cross-lingual Transfer. In MRQA 2021.
- Yiming Cui, Ting Liu, Shijin Wang, Guoping Hu. 2020. Unsupervised Explanation Generation for Machine Reading Comprehension. In arXiv pre-print: 2011.06737.
- Yiming Cui, Wei-Nan Zhang, Wanxiang Che, Ting Liu, Zhigang Chen. 2021. Understanding Attention in Machine Reading Comprehension. In arXiv pre-print: 2108.11574.
- Yiming Cui, Ting Liu, Wanxiang Che, Zhigang Chen, Shijin Wang. 2021. ExpMRC: Explainability Evaluation for Machine Reading Comprehension. In arXiv pre-print: 2105.04126.

Thank you !

https://github.com/ymcui

https://ymcui.com

me@ymcui.com