Heliyon 8 (2022) e09290

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

ExpMRC: explainability evaluation for machine reading comprehension

Yiming Cui [a,b,*], Ting Liu [a], Wanxiang Che [a], Zhigang Chen [b], Shijin Wang [b,c]

[a] *Research Center for SCIR, Harbin Institute of Technology, Harbin 150001, China*
[b] *State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, Beijing 100010, China*
[c] *iFLYTEK AI Research (Central China), Wuhan 430000, China*

ARTICLE INFO

Dataset link: https://github.com/ymcui/expmrc

Keywords:
Machine reading comprehension
Explainable artificial intelligence
Natural language processing

ABSTRACT

Achieving human-level performance on some Machine Reading Comprehension (MRC) datasets is no longer challenging with the help of powerful Pre-trained Language Models (PLMs). However, it is necessary to provide both answer prediction and its explanation to further improve the MRC system's reliability, especially for real-life applications. In this paper, we propose a new benchmark called ExpMRC for evaluating the textual explainability of the MRC systems. ExpMRC contains four subsets, including SQuAD, CMRC 2018, RACE$^+$, and C^3, with additional annotations of the answer's evidence. The MRC systems are required to give not only the correct answer but also its explanation. We use state-of-the-art PLMs to build baseline systems and adopt various unsupervised approaches to extract both answer and evidence spans without human-annotated evidence spans. The experimental results show that these models are still far from human performance, suggesting that the ExpMRC is challenging. Resources (data and baselines) are available through https://github.com/ymcui/expmrc.

1. Introduction

Machine Reading Comprehension is a task that requires machines to read and comprehend given passages and answer questions. The MRC-related study has received wide attention over the past few years. We have seen tremendous efforts to create challenging datasets [1, 2, 3, 4, 5, 6] and design effective models [7, 8, 9].

However, although the state-of-the-art systems can achieve better performance than the average human on some MRC datasets with the help of pre-trained language models [10, 11, 12], the explainability of these systems remains uncertain, such as the internal mechanism in neural models and giving text explanations. This raises concerns about utilizing these models in real-world applications. In a realistic view, question answering (QA) or MRC systems that only give final predictions cannot convince the users since these results lack explainability. In this context, Explainable Artificial Intelligence (XAI) [13] has received much more attention in recent years. XAI aims to produce more explainable machine learning models while preserving high model output accuracy and allowing humans to understand its intrinsic mechanism.

Understanding the intrinsic mechanism of the neural network is a challenging issue. In natural language processing field, there are several intense discussions on the relevant topics, such as *whether attention can be explanations* [14, 15, 16, 17]. However, the community has not come to a consensus on this question. Nonetheless, we could seek post-hoc explainability approaches, which target models that are not readily interpretable by design. Post-hoc approaches resort to diverse means to enhance the model's interpretability [18]. One of the post-hoc approaches for NLP is to generate text explanations, which is a practical method for alleviating the absence of the neural network's explainability [19]. Although the text explanation does not necessarily interpret the model's intrinsic mechanism, it is informative to know both the predicted answer and its text explanation, especially for real-life applications.

To better evaluate the MRC model's explainability, in this paper, we propose a comprehensive benchmark ExpMRC for the machine reading comprehension in a multilingual and multitask way, which evaluates the accuracy of both answers and their explanations. The proposed ExpMRC contains four subsets, including SQuAD [3], CMRC 2018 [5], RACE$^+$ (similar to RACE [4]), and C^3 [6], with additional annotations of the evidence spans, covering span-extraction MRC and multi-choice MRC in both English and Chinese. The MRC model should not only extract an answer span or select an answer choice for the question but also extract a passage span as evidence, which creates more challenges to the existing MRC tasks. The resulting dataset contains 11K human-annotated evidence spans over 4K questions. The contributions of our paper are as follows.

[*] Corresponding author at: Research Center for SCIR, Harbin Institute of Technology, Harbin 150001, China.
E-mail addresses: ymcui@ir.hit.edu.cn, ymcui@iflytek.com (Y. Cui).

https://doi.org/10.1016/j.heliyon.2022.e09290
Received 1 December 2021; Received in revised form 27 December 2021; Accepted 13 April 2022

Table 1. Examples in ExpMRC. The evidence is marked with underline. The answer is in blue.

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.
C^3	...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

- We release a new MRC benchmark called ExpMRC, which aims to evaluate the accuracy of the final answer as well as its explanation, encouraging the community to build explainable MRC systems.
- We propose several baseline systems that adopt pseudo-training approaches for ExpMRC that do not use any evidence span annotations.
- The experimental results on ExpMRC show that the current competitive pre-trained language models are still far from satisfactory in providing explanations for the predicted answer, suggesting that the proposed ExpMRC is challenging.

2. Related work

Machine reading comprehension has been regarded as an important task to test how well the machine comprehends human languages. In the earlier stage, as most of the models [7, 8, 20] are solely trained on the training data of each dataset without much prior knowledge, their performances are not very impressive. However, as the pre-trained language models emerged during these years, such as BERT [10], RoBERTa [11], and ELECTRA [12], many systems achieved better performances than average humans on several MRC datasets, such as SQuAD 1.1 [3] and SQuAD 2.0 [21] datasets.

After reaching the ‘over-human’ performance, there is another issue to be addressed. The decision process and the explanation of these artifacts remain unclear, raising concerns about their reliability and usability in real-life applications. In this context, XAI becomes more important than ever, not only in NLP but also in various directions in AI. However, most cutting-edge systems have been developed on neural networks, and investigating the explainability of these approaches is nontrivial.

In NLP, some researchers conducted analyses to better understand the internal mechanism of BERT-based architecture. For example, [22] discovered that there are repetitive attention patterns across different heads in the multi-head attention mechanism indicating its over-parametrization. However, perhaps the most popular discussion is *whether the attention can be explanations*. Some researchers argue that attention cannot be used as explanations, such as [15], who verified that using completely different attention weights can also achieve the same prediction. In contrast, some works hold positive attitudes about this topic [16, 17]. These works have brought us different views of attention-based models, but there is still no consensus about this important topic.

In MRC, a multi-hop explainable QA dataset called HotpotQA [23] was proposed. HotpotQA requires the machine to retrieve relevant documents and extract a passage span as the answer along with its evidence sentences. Various models [24, 25] have been proposed to address this task using supervised learning approaches with labeled training data. However, unfortunately, most works focus on achieving higher scores on the benchmark without specifically caring about the explainability.

For the explainability studies in MRC, [26] propose a method to extract evidence sentences from multi-choice MRC tasks. [27] propose to use system performance rather than visualizing attention score to better reveal the model’s explainability. [28] investigate a few black-box attacks at the character, word, and sentence level for MRC systems. [19] propose an unsupervised approach to extract rationale in the passage for MRC systems.

Although various efforts have been made, we argue that explainability is a universal demand for all MRC tasks and different languages but is not restricted to English multi-hop QA. Another issue is that annotating evidence for each task is not feasible. We should also seek unsupervised or semi-supervised approaches that do not rely on additional annotated evidence to minimize costs. In this context, we propose ExpMRC to specifically focus on evaluating explainability on four tasks, covering span-extraction and multi-choice MRC in both English and Chinese. ExpMRC does not provide any newly annotated *training data*. We encourage our community to focus on designing unsupervised approaches to improve the explainability with generalizable approaches for different MRC tasks and even different languages. To the best of our knowledge, this is the first MRC benchmark in a multitask and multilingual setting, which can be used in not only explainability evaluation but also in various directions, such as cross-lingual studies.

3. ExpMRC

3.1. Subset selection

The motivation for our dataset is to provide a comprehensive MRC benchmark for evaluating not only the answer prediction accuracy but also how well it gives for its explanation. Therefore, our dataset is not completely composed of new data. We adopt several well-designed MRC datasets and newly annotated data to form ExpMRC to minimize the repetitive annotations and place our work well in line with previous works.

Specifically, ExpMRC contains the following four subsets, including two span-extraction MRC datasets and two multi-choice MRC datasets. Examples in ExpMRC are depicted in Table 1. SQuAD, CMRC 2018, C^3 are partly developed from the respective original dataset. RACE$^+$ is a newly annotated subset, where we do not adopt the original RACE dataset.

- **SQuAD** [3] is a well-known dataset for span-extraction MRC. Given a Wikipedia passage, the system should extract a passage span as the answer to the question.
- **CMRC 2018** [5] is also a span-extraction MRC dataset but in Chinese. In addition to the traditional train/dev/test split, a challenge

set was also released that requires multi-sentence inference while keeping the original span-extraction setting.

- **RACE$^+$** is a new subset that is similar to RACE [4]. While we can use RACE as the C^3 counterpart, we decided not to adopt it. We had some in-house collected multi-choice MRC data, which is similar to RACE and is also designed for middle and high school students in China. More importantly, these data contain additional hints on the answering process, which are very helpful for evidence annotation. Thus, we decided to use our data instead of RACE.
- **C^3** [6] is a Chinese multi-choice MRC dataset. The system should choose the correct option as the answer after reading the passage and question. To ensure domain consistency, we only use non-dialogue subsets C$^3_{\mathrm{M}}$.

As the test set of SQuAD is not publicly available, we cannot adopt it directly.[1] Instead, we follow the original dataset construction steps to replicate the subset for testing purposes, where the subset is annotated from English Wikipedia passages. Note that we select the passages that do not appear in the original SQuAD training and development set.

At this point, we have four subsets (SQuAD, CMRC 2018, RACE$^+$, and C^3) to be annotated, containing both span-extraction and multi-choice MRC tasks in both English and Chinese. As SQuAD, CMRC 2018, C^3 datasets are well-defined datasets with careful annotation procedures, we did not perform additional pre-processing. Regarding our RACE$^+$, we follow the pre-processing steps as in RACE [4], as they share similar characteristics. Note that to preserve the integrity of the test set results, following previous works [3, 5, 21], we do not release the test sets to the public. To get the test set results, the participants should submit their system and get tested under the online platform (without direct access to the hidden test set).

3.2. Annotation process

All four subsets contain passages, questions, candidates (if applicable), and answers. We only need to annotate their evidence span on top of them. Before evidence annotation, the annotators are required to consider whether a question is appropriate for annotation. After removing sensitive content, we skipped some questions based on the following criteria.

- The evidence is a simple combination of the question and answer without much syntactical or semantical variance, such as the evidence span being the same or similar to the question text, where the question word is replaced by the answer.
- The questions require external knowledge to be solved and cannot only be inferred from the passage. That is, the evidence should not be formed by passage span.
- The conclusive questions of the whole passage, such as 'what is the best title or main idea for this passage?', etc. In this situation, the evidence span might be very long.

After the initial check, first, the annotators are asked to read the question and the correct answer (passage span or option text). Because, as the ground truth answer already exists in the original dataset, it is unnecessary to require the annotators to answer the questions again, which increases their burden when they recommend the wrong answer, and they will eventually consult the ground truth answer to find the correct evidence. Then, the annotators select (copy-and-paste) a span from the passage that can be evidence of the answer. The evidence should be a minimal passage span that can support the answer and does not always need to be a complete sentence or clause. We encourage the annotators to select the evidence that needs reasoning skills, although this is not a usual case in these datasets, especially in span-extraction MRC, where most of the questions do not need reasoning.

Selecting a single contiguous span makes the task much easier for the model, or it will become a sequence labeling task. During the annotation, if a redundant span is included to form a single span, we instructed our annotator that the length of the redundant span should not exceed 30% of the valid span length. However, in most cases (over 90%), a single contiguous span is enough for our selected datasets. It could be problematic for other datasets that require long-range inference, but this does not often happen in our ExpMRC.

The annotators are paid approximately $0.50 per evidence for all types of MRC data. Depending on the dataset language, the annotators are either English-majored or Chinese-majored graduate students from China.[2]

Following previous works, we also adopt multiple evidence references for each question to maximize the inter-agreement between the annotators. During annotation, we do not reveal the annotated evidence span of the other annotators to the current annotator to increase the diversity and avoid copy-and-paste behavior. After the preliminary annotation, all evidence spans are checked one by one to ensure a high-quality dataset. Finally, the annotations are verified that the correct answer can be selected by only reading the evidence and question to ensure that the annotation is valid.

3.3. Data statistics

The statistics of the proposed ExpMRC are listed in Table 2. Note that the 'token' in Table 2 represents the character for Chinese and the word for English.

For all subsets, we provide $2 \sim 4$ referential evidence spans for each question. It should be noted that ExpMRC does not provide any newly annotated training data. We believe there will be a significant improvement in the performance when there is a proper amount of labeled training data for evidence.[3] However, we believe that the explainability is within the model but does not largely depend on the labeled training set. We expect our community to develop a self-explainable system and evaluate its generalizability in a multilingual or multitask setting. If these systems generalize well in ExpMRC, they can also be applied to other MRC systems with a different task form or language. Also, developing an unsupervised or semi-supervised system significantly saves the cost of annotating evidence text, which is a promising way to develop generalizable and explainable MRC systems. However, if this is in a supervised setting (similar to what we do in HotpotQA), it will be hard to generalize to other settings.

We also provide statistics to see what skills are needed when we find evidence text in multi-choice MRC. We can also see that the subsets of span-extraction MRC tasks exhibit more types of 'surface matching' (simple word matching) and 'semantic matching' (such as 'man' and 'male') to find evidence. While, for multi-choice MRC tasks, there are more evidences that require complex reasoning, which demonstrates that it is harder to extract evidence for these subsets.

The distribution of the question type in each task's development set is depicted in Fig. 1. There are fewer questions of *'who, when, and where'* in RACE$^+$ and C^3, suggesting that these subsets are much more difficult, which is in line with the statistics above.

4. Baselines

Given that the proposed ExpMRC is designed to evaluate the explainability in terms of the system's explanation text, we mainly focus

[1] As CMRC 2018 is our previous work, although the test set is not publicly available, we can still use it for annotation.

[2] The annotators are full-intern students. The cost is only used for estimating the total cost of the project.

[3] Specifically, it refers to the ground truth evidence, as the answers are available in each original training set.

Table 2. Statistics of the proposed ExpMRC.

	SQuAD		CMRC 2018		RACE$^+$		C^3	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Language	English		Chinese		English		Chinese	
Answer Type	passage span		passage span		multi-choice		multi-choice	
Domain	Wikipedia		Wikipedia		exams		exams	
Passage Num.	319	313	369	399	167	168	273	244
Question Num.	501	502	515	500	561	564	505	500
Max Answer Num.	3	3	3	3	1	1	1	1
Max Evidence Num.	2	2	3	3	2	2	4	4
Avg/Max Passage Tokens	146/369	157/352	467/961	468/930	311/514	324/603	426/1096	413/1011
Avg/Max Question Tokens	12/28	11/28	15/37	15/37	15/39	16/55	14/28	14/31
Avg/Max Answer Tokens	3/25	3/27	6/64	5/33	6/20	6/27	7/25	7/35
Avg/Max Evidence Tokens	26/62	28/76	43/175	52/313	23/162	23/82	37/199	41/180
Surface Matching	-	-	-	-	61%	58%	63%	62%
Semantic Matching	-	-	-	-	14%	16%	20%	18%
Complex Reasoning	-	-	-	-	25%	26%	17%	20%

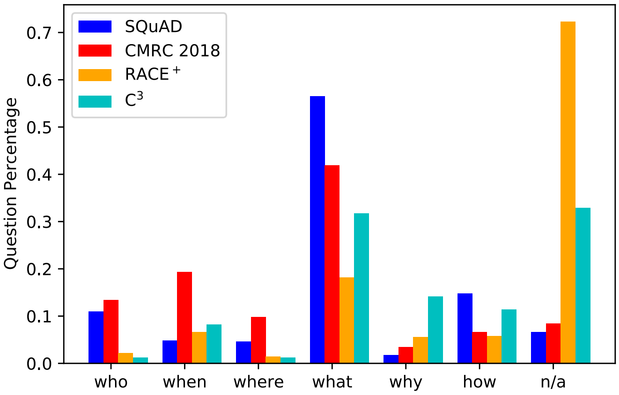


Fig. 1. Distribution of question types.

on the *unsupervised approaches* for our baseline systems, where ground truth evidence spans are not provided in the respective original training set.[4] We use pre-trained language models as the backbones to generate answers to the questions. Then we apply several methods to generate evidence spans, where we classify them into non-learning and machine learning baselines.

4.1. Non-learning baselines

For non-learning baselines, we mainly use the prediction and question as the clues for finding evidence. For simplicity, we only consider extracting sentence-level evidence in these baselines, although the ground truth evidence may not always be a complete sentence. We first split the passage into several sentences using '.!?' as delimiters. Then we select one of the passage sentences as the evidence prediction. As a preliminary, we should train a normal MRC system using the respective original training set that contains <passage, question, answer> to get *predicted answer*. In order to find more accurate evidence sentences, we adopt three approaches.

- **Most Similar Sentence**: We calculate the token-level F1 score between the predicted answer span (or choice text) and each passage sentence. Then we select the sentence that has the highest F1 as the evidence prediction. In span-extraction MRC tasks, the extracted evidence is the sentence that contains the prediction span in most cases.
- **Most Similar Sentence with Question**: Similar to the 'Most Similar Sentence', but we use both the question text and predicted answer span as the key to finding the most similar passage sentence.
- **Answer Sentence**: In span-extraction MRC tasks, we can directly extract the sentence that contains the answer prediction as evidence.

These approaches largely rely on the accuracy of answer prediction, as an incorrect prediction will directly affect the evidence finding process.

4.2. Machine learning baselines

As no training data are provided in ExpMRC, we seek a pseudo-training approach to accomplish a machine learning baseline system. First, we generate pseudo-evidence for each sample in the respective training set, which has no evidence annotation. We use the ground truth answer and question text to find the most similar passage sentence as the pseudo-evidence to form pseudo-training data. Then we use the pseudo-training data and PLM to train a model that outputs both answer and evidence. Specifically, we add an additional task head on top of the PLM's final hidden representation, alongside its original answer prediction task, as shown in Fig. 2.

- **Span-Extraction MRC**: The concatenation of the question Q and passage P are fed into PLM, and we use the final hidden representation with two fully-connected layers to predict the start and end positions of the answer span. The input sequence forms as in Fig. 2, where `[CLS]` is the special starting token and `[SEP]` is the special token for separation.
- **Multi-Choice MRC**: The concatenation of the passage P, question Q, and choice C_i are fed into the PLM to obtain four pooled representations (assuming we have four candidates). Then we use a fully-connected layer with softmax activation to predict the final choice.

The evidence prediction is identical to the answer prediction in span-extraction MRC, where we project the final hidden representation $\boldsymbol{h} \in \mathbb{R}^{n \times h}$ into the start and end probabilities $p^{\mathrm{s}}, p^{\mathrm{e}} \in \mathbb{R}^{n}$, as shown in Equation (1). We calculate the standard cross-entropy loss of the start and end positions for evidence span prediction, as shown in Equation (2).

$$p^{\star} = \mathbf{softmax}(\boldsymbol{h}\boldsymbol{w}^{\star} + b^{\star})\ ,\ \star \in \{s, e\} \tag{1}$$

$$\mathcal{L}_E = -\frac{1}{2N} \sum_{i=1}^{N} (y_i^{\mathrm{s}} \log p^{\mathrm{s}} + y_i^{\mathrm{e}} \log p^{\mathrm{e}}) \tag{2}$$

[4] The term 'unsupervised' specifically refers that we do not utilize additional annotated evidence spans, but we can still use the original training data that contains annotated answer spans.

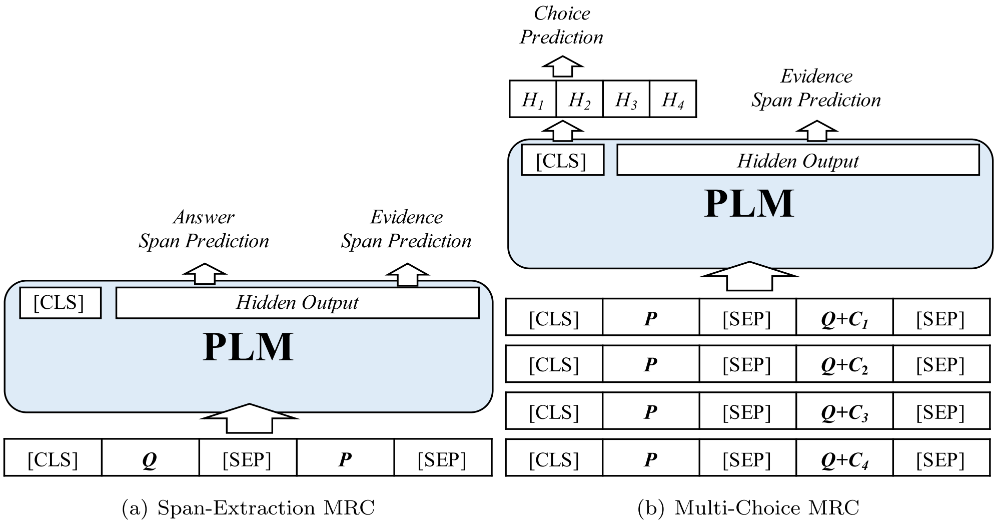


Fig. 2. Neural network architecture of the baselines.

The final training loss $\mathcal{L}$ is the sum of answer prediction loss $\mathcal{L}_A$ and the evidence prediction loss $\lambda\mathcal{L}_E$ ($\lambda \in [0,1]$, as the pseudo-training data are not quite accurate), as shown in Equation (3).

$$\mathcal{L} = \mathcal{L}_A + \lambda\mathcal{L}_E \tag{3}$$

5. Evaluation

5.1. Evaluation metrics

To evaluate how well the MRC model can generate explanations for the answers, we use the following metrics, which are divided into answer evaluation and evidence evaluation.

For answer evaluation, we strictly follow the original evaluation script for each subset. Specifically, we use the F1-score (F1) to evaluate SQuAD and CMRC 2018. We discard Exact Match (EM) and only evaluated F1 for simplicity. Note that, as these datasets are in different languages, the evaluation details are slightly different. For RACE$^+$ and C^3, we use accuracy for evaluation.

For evidence evaluation, we use F1 metrics, as most of the evidence spans are quite long, and it is difficult for the machine to extract the evidence spans exactly and thus we do not adopt EM. Also, the central idea of the evidence is to provide enough information to support the answer, so it is proper to adopt F1. Note that we only evaluate the correctness of evidence in this metric, regardless of the correctness of the answer. Altogether, we also use an overall F1 metric to provide a comprehensive evaluation of the system. For each instance, we calculate the score of the answer metric and evidence metric. The overall F1 of each instance is obtained by multiplying both terms, as shown in Equation (4).

$$\texttt{F1}_{\texttt{overall}} = \texttt{F1}_{\texttt{answer}} \times \texttt{F1}_{\texttt{evidence}} \tag{4}$$

Finally, the overall F1 of all instances is obtained by averaging all instance-level F1. The overall F1 reflects the correctness of both the answer and its evidence.

5.2. Human performance

Following previous works [3, 4, 5], we also report human performance to estimate how well humans perform on this dataset. Following [5], we use a *cross-validation approach* that regards one of the candidates as the prediction and treats the rest of the candidates as ground truths. Final scores are obtained by averaging all possible combinations.

- **SQuAD, CMRC 2018**: In these datasets, there are multiple references for both answer and evidence, and thus we use the cross-validation approach for both and obtain their products as instance-level human performance.
- **RACE$^+$, C^3**: As these datasets have only one reference answer, we invite three annotators to answer a random set of 100 questions in each set to obtain the averaged human answer performance. For the evidence, we directly use the cross-validation approach for the selected random set. Similarly, the instance-level human performance is obtained by the product of the answer and evidence score.

Note that as the evidence spans are annotated by referring to either the answers or additional hints, the actual human performance can be lower, and thus, these results should be regarded as *ceiling* human performance roughly. Finally, we average the scores in all instances to obtain the final overall human performance. Note that the answers and the evidences are not annotated by the same annotator, where the former is from the original dataset, and the latter is ours.

6. Experiments

6.1. Setups

We use pre-trained language models as the baseline system backbones. Specifically, we use BERT-base and BERT-large-wwm [10] for English tasks, and MacBERT-base/large [29] for Chinese tasks. We use a universal initial learning rate of 3e-5 and iterate two training epochs for all tasks. The maximum sequence length is set to 512, and the QA length is 128 in all experiments. We use Adam [30] with weight decay optimizer for training. All experiments are performed on a single Cloud TPU v2 for base-level PLMs and v3 for large-level PLMs. We set $\lambda = 0.01$ for span-extraction tasks and $\lambda = 0.1$ for multi-choice tasks in the final loss function to penalize the evidence pseudo-data training, which we found to be effective. Further investigation is discussed in Section 6.3.

6.2. Baseline results

The results are in Table 3, where 5-run maximum scores are reported.

Overall, the best-performing baselines are still far behind the human performance, indicating that the proposed dataset is challenging. Additionally, the gaps in multi-choice MRC subsets are larger than those in span-extraction MRC. For all subsets, adding question text for similarity calculation is more effective than only using the predicted answer.

Table 3. Baseline results on SQuAD, CMRC 2018, RACE$^+$, and C^3. B: base, L: large. 'Sent.' for 'sentence', 'Ques.' for 'question'. 'Ans.', 'Evi.', and 'All' denote the answer/evidence/overall score, respectively.

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*
Most Similar Sent. (B)	**87.4**	81.8	74.5	87.1	85.4	76.1	**82.3**	71.9	60.1	84.4	62.2	52.9
MSS. w/ Ques. (B)	**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. (B)	**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 (B)	87.0	79.5	70.6	**88.0**	78.6	69.8	81.5	73.2	60.4	**85.9**	61.3	52.4
Most Similar Sent. (L)	**93.0**	83.9	79.3	92.3	85.7	80.4	82.8	71.6	60.3	88.6	63.0	55.9
MSS. w/ Ques. (L)	**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. (L)	**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 (L)	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*
Most Similar Sent. (B)	62.4	36.6	28.2	59.8	34.4	26.3	68.7	57.7	**47.7**	66.8	52.2	41.2
MSS. w/ Ques. (B)	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 (B)	**63.6**	**45.7**	**31.7**	**60.1**	**43.5**	27.1	**70.9**	59.9	43.5	**69.0**	**57.5**	40.6
Most Similar Sent. (L)	**69.0**	37.6	29.9	68.1	36.8	28.9	73.1	59.4	49.9	72.0	52.7	43.9
MSS. w/ Ques. (L)	**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 (L)	**69.0**	45.9	32.6	**70.4**	41.3	30.8	**76.4**	**64.3**	50.7	**74.4**	**59.9**	**47.3**

For span-extraction MRC, traditional token similarity methods seem to be more effective as the answer is already a passage span, and its evidence often lies around its context. In contrast, the pseudo-data training approach is more effective in multi-choice MRC, where the options are not composed of the passage span, which is not capable of direct mapping, and it requires similarity calculation in semantics but not only in the token-level calculation.

Improving both answer and evidence prediction does NOT necessarily improve the overall score. For example, in the C^3 development set, pseudo-data training at a large-level baseline yields better performance on both answer and evidence prediction than the others. However, its overall score of 50.7 is lower than the best-performing baseline of 50.9. After checking the prediction file, we discovered that there are more samples that have either better evidence spans for the wrong answer prediction or worse evidence spans for correct answer prediction, which decreases the overall score.

Another interesting observation is that although pseudo-data training baselines do not yield better overall scores mostly, we see almost consistent improvements in the answer prediction accuracy, such as in C^3 using large-level PLM (e.g., dev +3.3, test+2.4). This suggests that using pseudo evidence helps improve answer prediction, and we expect there will be another improvement when we use a more effective method for extracting high-quality pseudo evidence.

6.3. Answer and evidence balance

To balance the ratio between the answer and evidence loss, we apply a lambda term to the evidence loss. To explore the effect of the lambda term, we select different $\lambda \in [0,1]$ and plot the 5-run average dev performance of each task using base-level PLMs. The results are shown in Fig. 3.

Overall, by increasing the lambda term, the evidence score and overall score decrease, suggesting that the pseudo-data training cannot be regarded as important as the original supervised task training (answer prediction), as the pseudo-data is not constructed by the ground truth evidence. However, in regard to the answer score, we observe that the span-extraction MRC tasks are less sensitive to the lambda term than the multi-choice MRC tasks. The optimal lambda value differs in span-extraction and multi-choice MRC tasks, where SQuAD and CMRC 2018 show a smaller value than RACE$^+$ and C^3. A possible guess is that two subtasks (answer extraction and evidence extraction) are the same in span-extraction MRC, and thus, the evidence extraction task benefits from the learning of answer extraction. However, as the evidence labels are not accurate enough, increasing the lambda term hurts the learning of evidence extraction.

Table 4. Upper bound performance of evidence F1 on the development sets.

	SQuAD	CMRC 2018	RACE$^+$	C^3
Most Similar Sent. w/ Ques.	81.9	76.3	48.0	63.2
Predicted Answer Sent.	85.4	77.7	-	-
Ground Truth Answer Sent.	88.2	82.1	49.9	66.8
Ground Truth Evidence Sent.	91.6	85.2	86.9	89.1
Human Performance	*92.1*	*94.6*	*92.4*	*95.7*

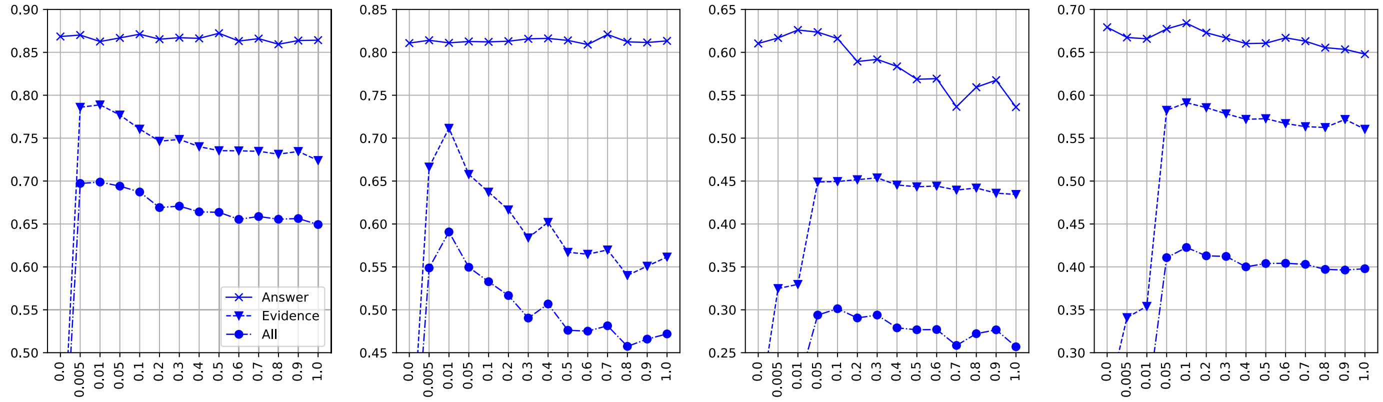
6.4. Upper bound for evidence extraction

In this section, we analyze the possible steps to achieve better evidence extraction performance. In addition to the 'Most Similar Sentence with Question' and 'Predicted Answer Sentence' (PA Sent.), we also provide two additional baselines for large-level PLMs. We extract the sentence that contains the ground truth answer (GA Sent.) and evidence (GE Sent.) to measure the upper bounds for those systems that only extract sentence-level evidence. The results are shown in Table 4.

As can be seen, the PA-GA and GA-GE gaps in span-extraction MRC are very small (approximately 3%∼5%), suggesting that the current system is about to reach the ceiling performance when only using sentence-level evidence extraction. In contrast, in multi-choice MRC, we see a large gap between GA and GE, indicating that only using the answer sentence is not enough to achieve strong evidence extraction performance. The gap between GE and human performance indicates the gains from expanding sentence-level evidence to a free-form evidence span. In addition to the SQuAD task, the others yield a 5.5%∼9.4% gap, which demonstrates that finding the exact evidence span in these tasks can still achieve a decent improvement.

7. Conclusion

In this paper, we propose a comprehensive benchmark for evaluating the explainability of MRC systems. The proposed ExpMRC benchmark contains four datasets, including SQuAD, CMRC 2018, RACE$^+$, C^3, covering span-extraction MRC and multiple-choice MRC in both En-



(a) SQuAD (b) CMRC 2018 (c) RACE$^+$ (d) C^3

Fig. 3. Effect of the lambda term in the evidence loss. X-axis: lambda, Y-axis: average F1.

glish and Chinese. ExpMRC aims to evaluate the MRC system to give not only correct predictions on the final answer but also extract correct evidence for the answer. We set up several baseline systems to thoroughly evaluate the difficulties of ExpMRC. The experimental results show that both traditional and state-of-the-art pre-trained language models still underperform human performance by a large margin on most of the subsets, indicating that more efforts should be made on designing an effective approach for evidence extraction. We hope the release of the dataset will further accelerate the research on the explainability and interpretability of MRC systems, especially for the unsupervised approaches.

Declarations

Author contribution statement

Yiming Cui: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Ting Liu, Shijin Wang: Analyzed and interpreted the data.

Wanxiang Che: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Zhigang Chen: Contributed reagents, materials, analysis tools or data.

Funding statement

This work was supported by National Key Research and Development Program of China (2018YFB1005100).

Data availability statement

Data associated with this study (ExpMRC datasets and baselines) have been deposited at https://github.com/ymcui/expmrc.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

Acknowledgements

We would like to thank anonymous reviewers and section editor for their thorough review and providing constructive comments to improve our paper. Yiming Cui would like to thank Google TPU Research Cloud (TRC) program for Cloud TPU access.

References

[1] K.M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, P. Blunsom, Teaching machines to read and comprehend, in: Advances in Neural Information Processing Systems, 2015, pp. 1684–1692.

[2] F. Hill, A. Bordes, S. Chopra, J. Weston, The Goldilocks principle: reading children's books with explicit memory representations, in: International Conference on Learning Representations, 2016, pp. 1–17, https://arxiv.org/abs/1511.02301.

[3] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100,000+ questions for machine comprehension of text, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2016, pp. 2383–2392.

[4] G. Lai, Q. Xie, H. Liu, Y. Yang, E. Hovy, Race: large-scale reading comprehension dataset from examinations, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2017, pp. 796–805.

[5] Y. Cui, T. Liu, W. Che, L. Xiao, Z. Chen, W. Ma, S. Wang, G. Hu, A span-extraction dataset for Chinese machine reading comprehension, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5886–5891.

[6] K. Sun, D. Yu, D. Yu, C. Cardie, Investigating prior knowledge for challenging Chinese machine reading comprehension, Trans. Assoc. Comput. Linguist. 8 (2020) 141–155.

[7] R. Kadlec, M. Schmid, O. Bajgar, J. Kleindienst, Text understanding with the attention sum reader network, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2016, pp. 908–918.

[8] Y. Cui, Z. Chen, S. Wei, S. Wang, T. Liu, G. Hu, Attention-over-attention neural networks for reading comprehension, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2017, pp. 593–602.

[9] M. Seo, A. Kembhavi, A. Farhadi, H. Hajishirzi, Bidirectional attention flow for machine comprehension, in: ICLR 2017, 2017, https://openreview.net/pdf?id=HJ0UKP9ge.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: a robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692.

[12] K. Clark, M.-T. Luong, Q.V. Le, C.D. Manning, ELECTRA: pre-training text encoders as discriminators rather than generators, in: ICLR, 2020, https://openreview.net/pdf?id=r1xMH1BtvB.

[13] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.-Z. Yang, Xai-explainable artificial intelligence, Sci. Robot. 4 (37) (2019) eaay7120, https://www.science.org/doi/abs/10.1126/scirobotics.aay7120.

[14] S. Serrano, N.A. Smith, Is attention interpretable?, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2931–2951, https://www.aclweb.org/anthology/P19-1282.

[15] S. Jain, B.C. Wallace, Attention is not explanation, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Associ-

ation for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 3543–3556, https://www.aclweb.org/anthology/N19-1357.

[16] S. Wiegreffe, Y. Pinter, Attention is not not explanation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 11–20, https://www.aclweb.org/anthology/D19-1002.

[17] J. Bastings, K. Filippova, The elephant in the interpretability room: why use attention as explanation when we have saliency methods?, in: Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, 2020, pp. 149–155, Online, https://www.aclweb.org/anthology/2020.blackboxnlp-1.14.

[18] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible ai, Inf. Fusion 58 (2020) 82–115, http://www.sciencedirect.com/science/article/pii/S1566253519308103.

[19] Y. Cui, T. Liu, W. Che, Z. Chen, S. Wang, Teaching machines to read, answer and explain, in: IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022.

[20] B. Dhingra, H. Liu, Z. Yang, W. Cohen, R. Salakhutdinov, Gated-attention readers for text comprehension, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2017, pp. 1832–1846.

[21] P. Rajpurkar, R. Jia, P. Liang, Know what you don't know: unanswerable questions for SQuAD, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 784–789.

[22] O. Kovaleva, A. Romanov, A. Rogers, A. Rumshisky, Revealing the dark secrets of BERT, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 4365–4374, https://www.aclweb.org/anthology/D19-1445.

[23] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, C.D. Manning, HotpotQA: a dataset for diverse, explainable multi-hop question answering, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2369–2380, https://www.aclweb.org/anthology/D18-1259.

[24] L. Qiu, Y. Xiao, Y. Qu, H. Zhou, L. Li, W. Zhang, Y. Yu, Dynamically fused graph network for multi-hop reasoning, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistic, Association for Computational Linguistics, Florence, Italy, 2019, pp. 6140–6150, https://www.aclweb.org/anthology/P19-1617.

[25] N. Shao, Y. Cui, T. Liu, S. Wang, G. Hu, Is graph structure necessary for multi-hop question answering?, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020, pp. 7187–7192, Online, https://www.aclweb.org/anthology/2020.emnlp-main.583.

[26] H. Wang, D. Yu, K. Sun, J. Chen, D. Yu, D. McAllester, D. Roth, Evidence sentence extraction for machine reading comprehension, in: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 696–707, https://aclanthology.org/K19-1065.

[27] Y. Cui, W.-N. Zhang, W. Che, T. Liu, Z. Chen, S. Wang, Multilingual multi-aspect explainability analyses on machine reading comprehension models, iScience 25 (4) (2022).

[28] W. Wu, D. Arendt, S. Volkova, Evaluating neural model robustness for machine comprehension, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, 2021, pp. 2470–2481, Online, https://aclanthology.org/2021.eacl-main.

[29] Y. Cui, W. Che, T. Liu, B. Qin, Z. Yang, Pre-Training with Whole Word Masking for Chinese Bert, 2021.

[30] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, in: International Conference on Learning Representations, 2015, pp. 1–15.