EveMRC: A Two-stage Evidence Modeling For Multi-choice Machine Reading Comprehension

Anonymous ACL submission

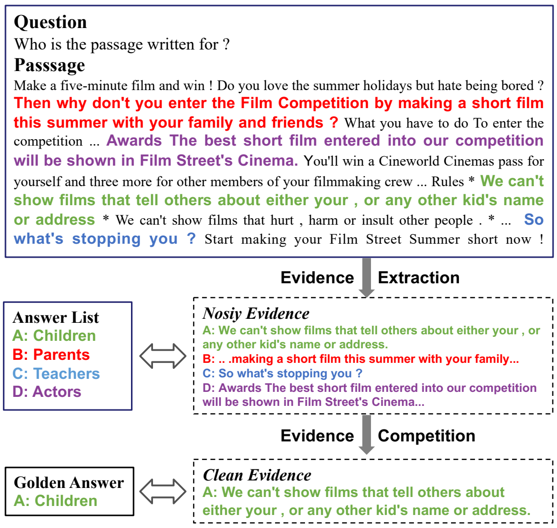
Abstract

Many impressive works have been proposed to improve the performance of Machine Reading Comprehension (MRC) systems in recent years. However, it is still difficult to interpret the predictions of existing MRC models, which makes the predictions unconvincing. In this work, we propose a two-stage explainable framework for multi-choice MRC to model not only the correlation between answers and evidence, but also the competition among evidence. In stage 1, we select evidence sentences for both the right answer and wrong answers using the semi-supervised evidence selector. In stage 2, we employ an evidence discriminator to compare among the competitive evidence set and make final judgements. Moreover, we propose an evidence-enabled data augmentation method. Experiments on four multi-choice MRC datasets show that: stage 1 provides strong explainability for MRC systems and stage 2 improves both the performance and robustness of MRC systems meanwhile.

1 Introduction

Machine Reading Comprehension (MRC), which aims to teach machines to read and comprehend the given passages and answer the questions. With the help of many effective architectures (Seo et al., 2016; Yu et al., 2018) and pre-trained language models (Devlin et al., 2018), reading comprehension systems are making rapid progress on many challenging datasets (Rajpurkar et al., 2016, 2018). However, though state-of-the-art systems could achieve better performances than humans, it's unclear to which extent these systems truly understand the language when simple adversarial examples can lead to a large performance drop (Jia and Liang, 2017; Gan and Ng, 2019).

As the need to build more convincing MRC systems, the research interests on explainability (Miller, 2019; Kratzwald et al., 2020) are rapidly growing. Models are required to expose the underlying mechanisms adopted to arrive at the final answers, whether by giving knowledge-based explanations from passages, or by giving operational explanation such as execution process of symbolic programs (Thayaparan et al., 2020). In this work, we focus on retrieving evidence sentences from passages as knowledge-based explanation.



Question

Who is the passage written for ?

Passsage

Make a five-minute film and win ! Do you love the summer holidays but hate being bored ? **Then why don't you enter the Film Competition by making a short film this summer with your family and friends ?** What you have to do To enter the competition ... **Awards The best short film entered into our competition will be shown in Film Street's Cinema.** You'll win a Cineworld Cinemas pass for yourself and three more for other members of your filmmaking crew ... Rules * **We can't show films that tell others about either your , or any other kid's name or address** * We can't show films that hurt , harm or insult other people . * ... **So what's stopping you ?** Start making your Film Street Summer short now !

Answer List

A: Children
B: Parents
C: Teachers
D: Actors

Nosiy Evidence

A: We can't show films that tell others about either your , or any other kid's name or address.
B: .. .making a short film this summer with your family...
C: So what's stopping you ?
D: Awards The best short film entered into our competition will be shown in Film Street's Cinema...

Golden Answer

A: Children

Clean Evidence

A: We can't show films that tell others about either your , or any other kid's name or address.

Figure 1: An example from the RACE dataset. We first extract evidence sentences for each answer, then acquire evidence for golden answer through competition.

However, most evidence selection methods only model the positive correlation between answers and passage sentences. On the one hand, most explainable MRC datasets only provide sentence-level supporting facts required for reasoning to the right answer (Yang et al., 2018). Sentences which may mislead the model to wrong answers were not given. On the other hand, the widely adopted pipeline methods always use evidence sentences as the substitution of full passages for more efficient reading over long articles (Min et al., 2018). The selected sentences are supposed to be most relevant to the right answer to give precise prediction which conversely influence the generation method of pseudo-evidence label (Wang et al., 2019).

By selecting evidence sentences for both right and wrong answers (Perez et al., 2019) in multi-choice MRC, some works modeled both positive and negative correlations. As shown in Figure 1, extracting evidence sentences not only for the right answer but also for wrong answers provides stronger explainability. We call the evidence set **Noisy Evidence** which consists of adversarial sentences for wrong answers and evidence sentences for the golden answer (**Clean Evidence**). However, they only modeled evidence-answer correlations and ignored the competitive relationship among selected evidence. In multi-choice MRC, evidence for each answer choice supports its own position and competes with each other. Modeling the competition relationship among evidence is crucial for rightly answering the question and building more robust and intelligent question answering systems.

To address the aforementioned problems, we propose EveMRC, a two-stage evidence modeling framework for multi-choice reading comprehension inspired by the Competition Model (see §2). We first model the correlation between answers and evidence sentences for both the right answer and wrong answers, then model the competitive correlation among evidence sentences of all answers. More specifically, in stage 1, we train an evidence selector to select evidence sentences for each answer under the supervision of pseudo-evidence label. In stage 2, we employ an evidence discriminator to compare among the competitive evidence set and make judgements about what the clean evidence and golden answer are.

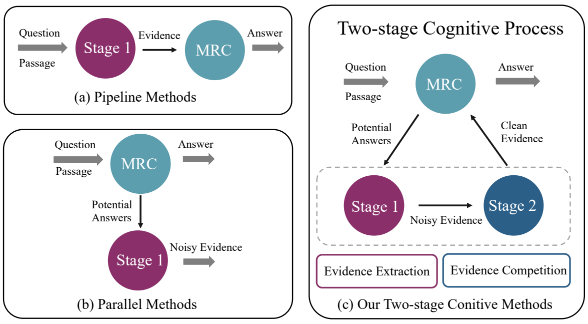
Due to the lack of ground truth evidence sentences in most multiple-choice MRC tasks, we propose a heuristic pseudo-evidence label generating algorithm based on model prediction. Our proposed algorithm is more effective compared to existing methods on multi-choice MRC which need complex handcrafted rules or extensive training iterations. Moreover, we propose a novel data augmentation method inspired by evidence. Concretely, we retain the evidence sentences for the golden answer of each question while replacing or reducing the adversarial evidence sentences for wrong answers or non-evidential sentences.

Our main contributions are as follows:

- We propose a two-stage evidence modeling framework for multi-choice MRC which not only models the positive and negative correlation between answers and evidence, but also models the competitive correlation among evidence.
- We propose an efficient pseudo-evidence label generating algorithm to train the evidence selector, and an evidence-enabled data augmentation method to improve the performance and robustness of MRC systems.
- We conduct thorough experiments on our framework and the experimental results show that our framework not only improves the performance of MRC models, but also shows strong explainability and robustness.

2 The Competition Model

The Competition Model is a psycholinguistic theory (Bates; MacWhinney, 1997) which mainly focus on the competition process of sentence processing and language acquisition. It argues that people understand a sentence by first searching various linguistic cues, such as word order, morphology, and semantic characteristics, for supporting each possible interpretation, eventually choosing the interpretation with the highest likeihood. Thus sentence processing can be viewed as a choice among different interpretations with different probabilities given by supporting cues.



(a) Pipeline Methods

(b) Parallel Methods

(c) Our Two-stage Conitive Methods

Figure 2: Comparison between our proposed framework and previous works.

Inspired by the Competition Model on sentence processing, we argue that the human cognitive process of reading comprehension can also be modeled as a two-stage process: (i) Evidence Extraction (ii) Evidence Competition. As shown in Figure 2(c), we propose an explainable framework for machine reading comprehension, which builds a closed loop between MRC system and two-stage evidence modeling. In stage 1, we collect supporting evidence for each possible answer. In stage 2, we conduct evidence competition among all evidence with their

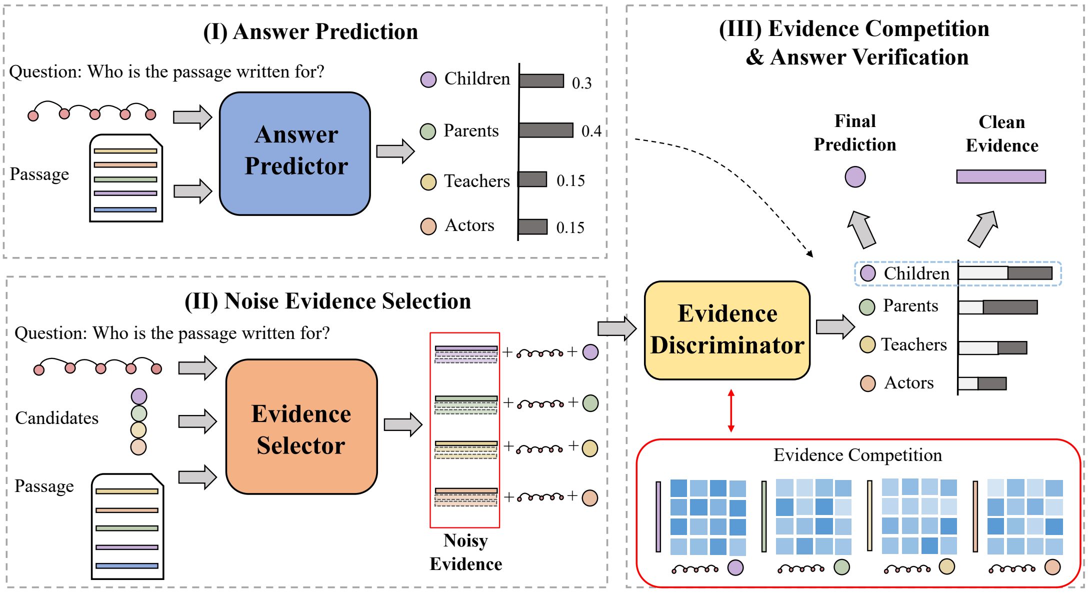


Figure 3: Overview of proposed EveMRC Framework. **Answer Predictor (AP)** gives the preliminary prediction among answers after reading through the whole passage; **Evidence Selector (ES)** then selects evidence sentences for each candidate answer; **Evidence Discriminator (ED)** discriminate among evidence or between evidence-answer pairs to choose final answer and corresponding evidence. Our EveMRC framework presents a two-stage selection-competition method for evidence modeling and also a two-stage answer-verify process for MRC.

corresponding answers, and we arrive at the final answer with its corresponding evidence. Stage 1 helps MRC model to retrieve relevant information for answering the question, both positively or negatively. Stage 2 makes the judgement among different answers with different probabilities given by supporting evidence which is very similar to the human cognitive process of Competition Model.

We also compare our two-stage framework with other explainable MRC methods. As shown in Figure 2(a), pipeline methods first extract evidence from passage, then substitute the passage with evidence sentences for more efficient reading. Due to the limitation of pipeline structure, the extracted evidence is more relevant to the correct answer and lacks diversity of explanation, which may also cause the performance degradation of MRC model. Parallel methods employ the evidence selector to select evidence for each answer which brings stronger explainability but ignores the exploitation of extracted evidence on model performance.

3 EveMRC

As multi-choice MRC need to discriminate among confusing candidate answers where the comparison between evidence is crucial for rightly answering the question, we choose it as the testbed for our framework. For multiple-choice MRC, machines are required to select the correct answer from the answer set $\mathrm{A} = \{a_1, a_2, ..., a_k\}$ (e.g. k=4) given with a passage $\mathrm{P} = \{s_1, s_2, ..., s_n\}$ with n sentences and a question Q.

In the following, we first present the overview of our framework as Figure 3, then describe the individual components in our framework: (i) Answer Predictor **(AP)**, (ii) Evidence Selector **(ES)** and (iii) Evidence Discriminator **(ED)**. Furthermore, we will introduce the unsupervised pseudo-evidence generating algorithm to initialize the evidence selector in §3.2, followed by the evidence-enabled data augmentation method in §3.3.

3.1 System Overview

As shown in Figure 3, our system pipeline is composed of three stages: 1) Answer Prediction. 2) Noisy Evidence Selection. 3) Evidence Competition & Answer Verification. The pipeline of our proposed framework can be formulated as follows:

$$\begin{aligned} p_A, \mathrm{A} &= \mathbf{AP}(\mathrm{P}, \mathrm{Q}) \\ \mathrm{E} &= \{E_i : \mathbf{ES}(\mathrm{P}, \mathrm{Q}, a_i) \ \text{ for } \ a_i \ \text{ in } \ \mathrm{A}\} \\ p_D &= \mathbf{ED}(\mathrm{E}, \mathrm{Q}, \mathrm{A}) \\ p &= \alpha * p_D + (1 - \alpha) * p_A \end{aligned} \tag{1}$$

where $\mathrm{AP}, \mathrm{ES}, \mathrm{ED}$ are Answer Predictor, Evidence Selector, and Evidence Discriminator, respectively, $\mathrm{P}, \mathrm{Q}, \mathrm{A}, \mathrm{E}$ are passage, question, answer set and evidence set, respectively, p_A, p_D, p are probability distribution over candidate answers for Answer Predictor, Evidence Discriminator, and final prediction, respectively, α is the weighting coefficient.

3.1.1 Answer Predictor

For multiple-choice MRC tasks, we construct the input sequence by concatenating `[CLS]`, P, `[SEP]`, Q, `[SEP]`, a_i, and `[SEP]`, where `[CLS]` and `[SEP]` are the classifier token and sentence separator in a pre-trained language model, respectively. Tokens before the first `[SEP]` (inclusive) is grouped as the first segment and the rest of the tokens are treated as the second segment. After feeding the input sequence into a pre-trained language model (e.g., BERT), we can arrive at the final hidden state for the first token in the input sequence as $h_i^{\mathrm{AP}} \in \mathbb{R}^{1\times N}$, where N is hidden size. A linear classification layer $\mathrm{W_{AP}} \in \mathbb{R}^{N\times 1}$ is applied to get the unnormalized score of each candidate answer a_i, and the final prediction is obtained by applying a softmax layer over the unnormalized scores of all candidate answers, i.e.

$$p_A = \mathrm{softmax}(\{h_1^{\mathrm{AP}}\mathrm{W_{AP}}, \cdots, h_k^{\mathrm{AP}}\mathrm{W_{AP}}\}) \quad (2)$$

3.1.2 Evidence Selector

Intuitively, evidence sentences of different candidate answers may not be the same. Thus, the evidence selector chooses evidences for each candidate answer independently. Given candidate answer a_i, sentence s_j and question Q, we construct the input sequence by concatenating `[CLS]`, s_j, `[SEP]`, Q, `[SEP]`, a_i, and `[SEP]`. Similar to the answer predictor, we denote the final hidden state for the first token in the input sequence as $h_{ij}^{\mathrm{ES}} \in \mathbb{R}^{1\times N}$. A linear classification layer $\mathrm{W_{ES}} \in \mathbb{R}^{N\times 1}$ is applied to achieve the evidence score $p_E(i,j)$, i.e.,

$$p_E(i,j) = \mathrm{sigmoid}(h_{ij}^{\mathrm{ES}}\mathrm{W_{ES}}) \quad (3)$$

After scoring each sentence for answer a_i, top K scored sentence are selected as the evidence sentences of a_i.

3.1.3 Evidence Discriminator

Given all associated evidence sentences E_i with candidate answer a_i, we verify it by feeding the concatenation of E_i, Q, and a_i into a model with the same structure as the Answer Predictor. Similarly, we denote the final hidden state for the first token in the input sequence as $h_i^{\mathrm{ED}} \in \mathbb{R}^{1\times N}$. A linear classification layer $\mathrm{W_{ED}} \in \mathbb{R}^{N\times 1}$ is applied to achieve the unnormalized score for a_i and the discriminator probabilities are computed by applying a softmax function over the unnormalized scores, i.e.,

$$p_D = \mathrm{softmax}(\{h_1^{\mathrm{ED}}\mathrm{W_{ED}}, \cdots, h_k^{\mathrm{ED}}\mathrm{W_{ED}}\}) \quad (4)$$

Then the final answer score will be:

$$p = \alpha * p_D + (1-\alpha) * p_A \quad (5)$$

where α is the weighting coefficient.

Furthermore, the evidence sentences corresponding to the final answer will be the predicted Clean Evidence.

3.2 Pseudo-evidence Generating Algorithm

Due to the lack of evidence annotation in most MRC datasets especially on multi-choice MRC datasets, we propose a heuristic pseudo-evidence label generating algorithm based on the prediction of Answer Predictor to train the Evidence Selector. The algorithm is inspired by how human regard a sentence as evidence that evidence contributes most to the process of human prediction.

Algorithm 1 (see appendix A) describes the procedure of generating pseudo-label. We first use Answer Predictor, i.e. a MRC model trained on multi-choice MRC, to get the original probability distribution over answers. To reveal the importance of each sentence s_i in the passage, we conduct sentence-level masking on input passages and acquired the masked prediction $\hat{p}_i$ after replacing the original input passage of **AP** with $\hat{P}_i$. We believe that if the masked sentence is critical for answering this question, the answer distribution will change a lot after masking. Thence, we use Kullback-Leibler (KL) Divergence between two distributions as the overall evidential score. To make sure the corresponding answer of each sentence, we also calculate the difference between two distributions and assign the sentence to the answer with the largest decrease in probability score. For each passage, we select top N sentences with their corresponding answer as positive examples. Moreover, we randomly sample non-evidential answer-sentence pairs as negative examples for training Evidence Selector.

Model	RACE		Model	DREAM	
	dev	test		dev	test
HAF (Zhu et al., 2018)	47.2	46.0	Sliding Window (Sun et al., 2019)	42.6	42.5
DFN (Xu et al., 2017)	-	47.4	DSW++ (Sun et al., 2019)	51.4	50.1
MRU (Tay et al., 2018)	-	50.4	GBDT++ (Sun et al., 2019)	53.3	52.8
GPT (Radford et al., 2018)	-	59.0	FTLM++ (Sun et al., 2019)	57.6	57.4
BERT-base†	-	65.0	BERT-base (Zhu et al., 2020)	61.2	61.5
ALBERT-base (Lan et al., 2019)	-	66.8	ALBERT-base (Zhu et al., 2020)	64.5	64.4
BERT-base (re-run)	66.5	65.6	BERT-base (re-run)	61.8	62.0
+DA	67.4	66.2	+DA	62.4	62.6
+Discriminator	67.5	67.3	+Discriminator	63.1	63.3
+EveMRC	68.2	67.4	+EveMRC	63.3	63.5
ALBERT-base (re-run)	71.0	69.9	ALBERT-base (re-run)	66.6	66.4
+DA	71.9	70.6	+DA	67.7	67.5
+Discriminator	72.5	71.2	+Discriminator	67.9	67.4
+EveMRC	72.8	71.4	+EveMRC	68.0	67.6

Table 1: Experimental results on RACE and DREAM datasets.†: Results are token from leaderboard; DA: using data augmentation method; Discriminator: using Evidence Discriminator module; EveMRC: both using data augmentation and Evidence Discriminator.

3.3 Evidence-enabled Data Augmentation

After we obtain Noisy Evidence and Clean Evidence from Evidence selector and Evidence Discriminator, we propose a data augmentation method based on evidence. We argue that evidence sentences not only provide strong explainability but also indicate the intrinsic structure of information. Concretely, we classify the sentences of passage into three categories: (i) non-evidential sentences (ii) evidence sentences corresponding to wrong answers (iii) evidence sentences corresponding to the golden answer. Non-evidential sentences contain background information that is not essential for answering the question. Evidence sentences for wrong answers will disturb the answer choice of MRC model. Consequently, we propose the following two data augmentation methods:

- **Clean Evidence Preservation.** For each question, we keep the Clean Evidence (evidence sentences of the golden answer) remained and substitute other sentences in the passage with retrieved passages.
- **Noisy Evidence Preservation.** Similar to Clean Evidence Preservation, we only retain Noisy Evidence (evidence sentences of both right and wrong answers) instead.

The algorithm details are described below. First, for a given example with passage P, question Q and answer list A, we retrieve examples with similar passage by embedding text into its corresponding TFIDF-weighted bag-of-words vector. We compute the cosine similarity S of the embeddings for two passages P_1 and P_2:

$$\mathrm{S}(\mathrm{P}_1, \mathrm{P}_2) = \cos(\mathrm{TFIDF}(\mathrm{P}_1), \mathrm{TFIDF}\ (\mathrm{P}_2)) \tag{6}$$

We conduct data augmentation only for passage pairs with a matching score S between α_1 and α_2. In this way, we can filter the irrelevant passages or over-similar passages for a given example.

4 Experiments

We evaluate our framework on four multi-choice datasets: RACE, DREAM, RACE$^+$ (with evidence annotation) and AdvRACE (with adversarial attacks). Our experiments are divided into three parts: (i) performance evaluation (ii) robustness evaluation and (iii) explainability evaluation, also with some analytical studies.

For the evaluation on RACE, DREAM, AdvRACE and answer prediction tasks in RACE$^+$, the standard accuracy is applied. As for the evaluation of evidence, F1 score that measures the weighted average of the precision and recall rate at a character level is used. Besides, the authors of RACE$^+$(Cui et al., 2021) also provide an overall F1 metric that reflects the correctness of both answers and its evidence.

4.1 Implementation Details

To evaluate our framework, we use two pre-trained language models: BERT-base-uncased(Lan et al., 2019) and Albert-base-v2(Lan et al., 2019) of which the implementation is based on the public Pytorch implementation from Transformers[1]. Due to the RACE$^+$ only provides development set and hidden test set, we use RACE train set for training. The sampling numbers N of evidence sentences and negative examples are both 2 for RACE and 1 for DREAM. The numbers of evidence sentences selected for evidence discriminator are 3, 3, 2, 1 for RACE, $RACE^+$, DREAM and AdvRACE, respectively. See more implementation details in appendix C.

4.2 Accuracy Evaluation

4.2.1 Datasets

RACE (Lai et al., 2017): RACE is a dataset collected from the English exams for middle and high school Chinese students. RACE are generated by human experts, and covers a variety of topics that are carefully designed for evaluating the students' ability in understanding and reasoning.

DREAM (Sun et al., 2019): DREAM is the first dialogue-based multiple-choice reading comprehension dataset, which is collected from English as a Foreign Language examinations designed by human experts to evaluate the ability of reading comprehension of Chinese English learners.

4.2.2 Experiment Results

Table 1 shows our results on RACE and DREAM with BERT-base and ALBERT-base as baselines. EveMRC achieves comparable and consistent improvement over RACE + BERT-base with +1.7%, +1.8%, RACE + ALBERT-base with +1.8%, +1.5%, DREAM + BERT-base with +1.5%, 1.5% and DREAM + ALBERT-base with +1.4%, 1.2% for development set and test set, respectively. Also, only using data augmentation or only using the evidence discriminator can achieve comparable improvements. Notably, we observe that evidence discriminator achieves better results than data augmentation method and contributes most to our EveMRC Framework.

4.3 Explainability Evaluation

4.3.1 Datasets

RACE$^+$ (Cui et al., 2021): RACE$^+$ is a subset of ExpMRC and similar to RACE, which is designed for evaluating the explainability of the MRC systems. The evidence of each case in RACE$^+$ is a minimal passage span that can support the answer.

Model	RACE+ (dev)			RACE+ (test)		
	Ans.	Evi.	All	Ans.	Evi.	All
BERT-base Baselines♣						
Most Similar Sent.	62.4	36.6	28.2	59.8	34.4	26.3
Most Similar Sent. w/Ques.	62.4	44.5	31.5	59.8	41.8	27.3
Pseudo-data training	63.6	45.7	31.7	60.1	43.5	27.1
BERT-large Baselines♣						
Most Similar Sent.	69.0	37.6	29.9	68.1	36.8	28.9
Most Similar Sent. w/Ques.	69.0	48.0	36.8	68.1	42.5	31.3
Pseudo-data training	**69.0**	45.9	32.6	**70.4**	41.3	30.8
Bert-base + Search agents◇	63.6	35.9	29.6	-	-	-
Bert-base + Learned agents◇	63.6	44.7	37.0	-	-	-
Our Method						
Bert-base + EveMRC	66.7	**58.5**	**47.2**	66.7	**52.5**	**40.7**

Table 2: Experimental results on RACE$^+$ dataset. **Ans.:** answer accuracy. **Evi.**: F1 score between golden evidence label and selected evidence sentences. **All** reflects the correctness of both answer and its evidence. ♣: Results are taken from Cui et al. (2021). ◇: Our implementation of Perez et al. (2019); we only report the results on dev set due to the submission interval of ExpMRC leaderboard.

4.3.2 Baselines

Most Similar Sent. and *Most similar Sent. w/Ques* (Cui et al., 2021) select the sentence with the highest F1 score calculated with predicted answer text or concatenation of predicted answer and question. *Pseudo-data training* is a stronger baseline that employs golden answers and questions to retrieve the most similar sentences as pseudo-data. Besides, we implement two algorithms proposed by (Perez et al., 2019). *Search agents* search the sentences that most convince the model while learned agents employ another model to learn from the prediction of search agents.

4.3.3 Experiment Results

Table 2 compare our framework on RACE$^+$ with several baselines which are proposed in (Cui et al., 2021). Experimental results on RACE$^+$ show our framework achieves significant improvement over the same base-level baselines both on evidence selection (+12.8%, +9% for development set and test set respectively) and answer prediction (+3.1%, +6.6% for development set and test set respectively). Although we did not submit compared to large-level baselines, the improvement on evidence selection is substantial where F1 score of evidence selection is increased by 10.5 points and 10 points for development set and test set respectively. On the other hand, we observe that search agents and learned agents have relatively poor performance considering that they only selected evidence for each answer independently and did not model the

[1] https://github.com/huggingface/transformers

	Original	AS	CS	DE	DG	Ave.
BERT-base	65.6	19.8	49.3	30.7	53.1	38.2
+ DA	66.2	21.8	48.8	30.4	52.5	38.4
+ ED	67.3	25.0	**50.9**	**33.0**	55.2	41.0
+ EveMRC	**67.4**	**26.7**	50.6	33.0	**55.4**	**41.4**

Table 3: Experimental Results on AdvRACE dataset. DA: use data augmentation. ED: use evidence discriminator. EveMRC: both with data augmentation and evidence discriminator. AS: AddSent for adversarial augmentation. CS: CharSwap. DE: Distractor Extraction. DG: Distractor Generation.

competition process among evidence.

4.4 Robustness Evaluation

4.4.1 Datasets

AdvRACE (Si et al., 2021): AdvRACE is a multi-choice style benchmark for evaluating the robustness of MRC models under four different types of adversarial attacks, i.e., Distractor Extraction, Distractor Generation, AddSent and CharSwap.

4.4.2 Experiment Results

Table 3 shows the experimental results on AdvRACE. The performance of BERT-base dramatically drops during all types of attack methods. The most violent attack method is AddSent where BERT-base exhibits a nearly 70% reduction in accuracy. However, both data augmentation and evidence discriminator improve the model robustness on AddSent by +2.0 point and + 5.2 point, respectively. Although sentence-level data augmentation does not defend against other types of attack, evidence competition by discriminator can significantly improve the model robustness comprehensively for all attacks (agerage +3.2 point).

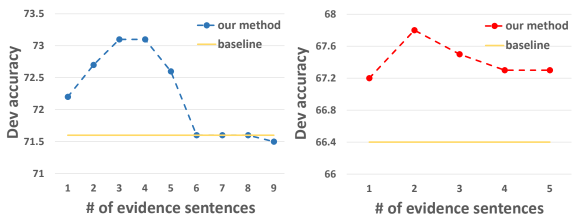
4.5 Analytical Studies

we design another three types of evidence discriminator for analytical studies:

Exclusive-Evidence Discriminator. Exclusive evidence means that one sentence can only be the evidence of one answer. Once multiple answers select the same sentence as evidence, we use the sentence-answer pair with highest evidential score.

Unified-Evidence Discriminator. In unified-evidence discriminator, all the answers are using the same evidence sentences. We select the sentence with the highest max-score among all the answers as the unified evidence.

Question-only Discriminator. To figure out the efficiency of evidence-question interaction and



(a) RACE (b) DRAEM

Figure 4: Different lengths of evidence for evidence discriminator. Baseline: Albert-base. Our method: Albert-base with the evidence discriminator.

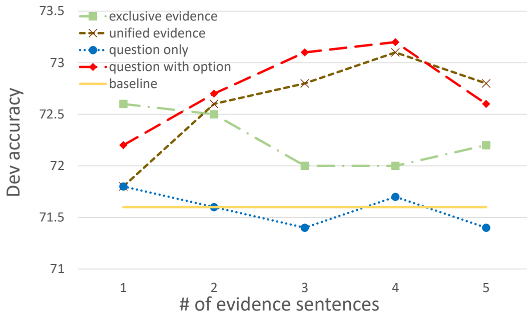


Figure 5: Comparison among different discriminator settings. Baseline: Albert-base. Question with option: Albert-base with standard evidence discriminator. The others are described in section 4.3.2.

evidence-answer interaction, we implement evidence discriminator only using question and evidence sentences for all answers.

4.5.1 Evidence Sentences of Different Lengths

When we extract evidence sentences for each answer candidate to do answer verification, a question that what length of evidence sentences is most suitable for answer verification comes naturally. It will introduce a lot of noise or ignore essential evidence in the case that extracted evidence is too long or too short. Figure 4(a) and Figure 4(b) shows the discriminator results of answer selection accuracy on RACE and DREAM dev-set with the growing number of evidence sentences. We see that our model achieves the best accuracy with 3 or 4 evidence sentences on RACE and 2 evidence sentences on DREAM which is reasonable considering that RACE has a much longer average number of sentences than DREAM (17.6 vs. 8.5).

4.5.2 Different discriminator Settings

In Section 4.3.2, we propose several verification settings with exclusive-evidence, unified-evidence, and question-only. The results for comparison on the RACE dev set are shown in Figure 5. From

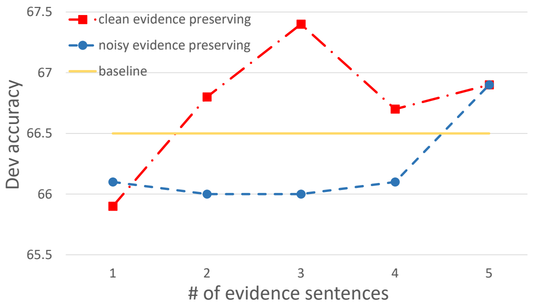


Figure 6: Different types of data augmentation methods. Baseline: BERT-base.

the overall results of four types of discriminator settings, we first observe that our standard evidence discriminator which employs answer-wise evidence with both question and option as verification achieves the best performance. Unified-evidence discriminator performs slightly worse than our standard discriminator but also improves a lot over baseline. One important difference between unified evidence and answer-wise evidence is that answer-wise evidence can provide more comprehensive evidential information for verification while unified evidence shares limited information. Similarly, the exclusive evidence ensures the comprehensiveness of evidential information in the case of short evidence but introduces more noise in the case of long evidence which can also be drawn from the accuracy curve over the number of evidence sentences of the exclusive-evidence discriminator. Furthermore, the question-only discriminator performs worst which reveals the indispensability of evidence-answer verification in our framework.

4.5.3 Clean Evidence vs. Noisy Evidence

Figure 6shows the results of two types of data augmentation methods on RACE dev sets. Obviously, data augmentation method with the preserving of clean evidence performs better than the method with noisy evidence preserving. Surprisingly, we observe an accuracy increase of noisy evidence preserving also with an accuracy decrease of clean evidence preserving with the growth of the number of evidence sentences. We argue that clean evidence with short length can already contain the most important evidential information and it will introduce more noise as the consequence of increasing evidence length. On the contrary, noisy evidence requires a longer length to include the essential information for answering the question.

5 Related Work

Building MRC systems with stronger explainability are more urgent due to the lack of robustness (Jia and Liang, 2017; Mudrakarta et al., 2018). On the one hand, researchers build benchmarks with labeled data for training or evaluation. HotpotQA(Yang et al., 2018) provides sentence-level supporting facts and introduces a leaderboard for evaluating the explanations. CoQA(Reddy et al., 2019) contains free-form answers and each answer has a span-based rationale for each answer. CoS-E(Rajani et al., 2019) collect human explanations for commonsense reasoning. ExpMRC(Cui et al., 2021) annotated several datasets for explainability evaluation. On the other hand, attention mechanisms have been frequently used for revealing the prediction process with attended sentences(Seo et al., 2016). Moreover, Niu et al. (2020) train a self-supervised evidence extractor with auto-generated labels in an iterative process for multi-hop reasoning MRC. Zhang et al. (2020); Min et al. (2018) use two-stage pipeline methods which extract evidence for build more efficient and robust MRC systems. Wang et al. (2019) employ linguistic knowledge to extract evidence sentences for multiple-choice MRC. Perez et al. (2019) select evidence sentences for all answers by learning to convince Q&A models. Although Perez et al. (2019) proves that the evidence for all answers can be generalizable for training MRC models, they do not model the competition process among evidence.

6 Conclusion

With the emerging research interest in explainable MRC systems, this paper proposes an explainable MRC framework for evidence extraction and answer verification. We tackle the problem of lacking labeled evidence data by proposed a heuristic method to generate the pseudo-evidence label and propose two impressive applications of evidence sentences: answer verification and data augmentation. The experimental results show the effectiveness and strong explainability of our framework. In the future, we will explore more unsupervised methods to utilize and enhance the explainability of MRC systems.

References

Elizabeth Bates. Functionalist approaches to grammar.

Yiming Cui, Ting Liu, Wanxiang Che, Zhigang Chen, and Shijin Wang. 2021. Expmrc: Explainability evaluation for machine reading comprehension. *arXiv preprint arXiv:2105.04126*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Wee Chung Gan and Hwee Tou Ng. 2019. Improving the robustness of question answering systems to question paraphrasing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6065–6075.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Bernhard Kratzwald, Stefan Feuerriegel, and Huan Sun. 2020. Learning a cost-effective annotation policy for question answering. *arXiv preprint arXiv:2010.03476*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Brian MacWhinney. Functionalism and the competition model.

Brian MacWhinney. 1997. Second language acquisition and the competition model. *Tutorials in bilingualism: Psycholinguistic perspectives*, pages 113–142.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Sewon Min, Victor Zhong, Richard Socher, and Caiming Xiong. 2018. Efficient and robust question answering from minimal context over documents. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1725–1735, Melbourne, Australia. Association for Computational Linguistics.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the model understand the question? *arXiv preprint arXiv:1805.05492*.

Yilin Niu, Fangkai Jiao, Mantong Zhou, Ting Yao, Jingfang Xu, and Minlie Huang. 2020. A self-training method for machine reading comprehension with soft evidence extraction. *arXiv preprint arXiv:2005.05189*.

Ethan Perez, Siddharth Karamcheti, Rob Fergus, Jason Weston, Douwe Kiela, and Kyunghyun Cho. 2019. Finding generalizable evidence by learning to convince q&a models. *arXiv preprint arXiv:1909.05863*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

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.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Chenglei Si, Ziqing Yang, Yiming Cui, Wentao Ma, Ting Liu, and Shijin Wang. 2021. Benchmarking robustness of machine reading comprehension models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 634–644, Online. Association for Computational Linguistics.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2018. Multi-range reasoning for machine comprehension. *arXiv preprint arXiv:1803.09074*.

Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2020. A survey on explainability in machine reading comprehension. *arXiv preprint arXiv:2010.00389*.

Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, Dong Yu, David McAllester, and Dan Roth. 2019. Evidence sentence extraction for machine reading comprehension. *arXiv preprint arXiv:1902.08852*.

Yichong Xu, Jingjing Liu, Jianfeng Gao, Yelong Shen, and Xiaodong Liu. 2017. Dynamic fusion networks for machine reading comprehension. *arXiv preprint arXiv:1711.04964*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.

Shuailiang Zhang, Hai Zhao, Yuwei Wu, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. 2020. Dcmn+: Dual co-matching network for multi-choice reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9563–9570.

Haichao Zhu, Furu Wei, Bing Qin, and Ting Liu. 2018. Hierarchical attention flow for multiple-choice reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Pengfei Zhu, Hai Zhao, and Xiaoguang Li. 2020. Duma: Reading comprehension with transposition thinking. *arXiv preprint arXiv:2001.09415*.

A Pseudo-evidence label Generating Algorithm

Algorithm 1 Pseudo-evidence generating algorithm

```
Require: Passage $P = \{s_1, s_2, \dots s_n\}$, Question
    $Q$ , answer list $A = \{a_1, a_2, ..., a_k\}$, Answer
    Predictor AP
Ensure: Pseudo-evidence label set $E$
    {each element in $E$ is a sentence-answer pair}
Initialize three empty list $S_{KL}$ ,$S_\Delta$ and $E$
$p_A = \mathbf{AR}(P, Q, A)$
{$p_A$ is the probability distribution on candidate answers}

for $s_i$ in $P$ do
    $\hat{P_i} = \{s_1, \dots, s_{i-1}, [\text{MASK}], s_{i+1}, \dots, s_n\}$
    $\hat{p_i} = \mathbf{AR}(\hat{P_i}, Q, A)$
    $s = KL(p_A, \hat{p_i})$
    Add $s$ to $S_{KL}$
    Add $\hat{p_i} - p_A$ to $S_\Delta$
end for
for $i$ in $\arg\max(S_{KL}, n)$ do
    $m_i = \arg\min S_{\Delta i}$
    Add $P_i : A_{m_i}$ to $E$
end for
return Pseudo-evidence label $E$
```

B Dataset Statistics

		RACE	DREAM	RACE+	AdvRACE
# of documents	Train	25137	3869	-	-
	Dev	1389	1288	167	-
	Test	1407	1287	168	1407
# of questions	Train	87866	6116	-	-
	Dev	4887	2040	561	-
	Test	4934	2041	564	19736
Average # of sentences per document		17.6	8.5	23 (dev)	19
Average # of tokens per evidence		-	-	23	-

Table 4: Datasets Statistics.

C Implementation Details

		RACE	DREAM
learning rate	AR	3e-5	2e-5
	ES / AV	3e-5	3e-5
batch size	all	32	32
epoch	AR / AV	3	30
	ES	2	20

Table 5: Training hyperparamters of different components.

We use AdamW(Loshchilov and Hutter, 2017) optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and without weight decay and warmup. The max input sequence length is set to 512. The training hyperparameters are shown in 5. We search for the best weighting coefficient of probability combination on dev set which ranges from 0.1 to 0.5 with 0.1 as the interval. The α_1 and α_2 for filtering TFIDF scores are 0.2, 0.8 respectively. We average the main results by running three random seeds and report the average scores. We use 4 NVIDIA 2080Ti for all the experiments.

D Case Study

Passage

The discovery of an ancient giant panda skull has confirmed its bamboo diet dates back more than 2 million years and may have played a key part in its survival.[: A Chinese-US research team reports its results today following studies on a fossil skull found in south China's Cuangxi Zhuang Autonomous Region in 2001. The six fossils unearthed in Jinyin Cave are dated between 2.4 and 2 million years ago, according to the report in Proceedings of the National Academy of Sciences, an influential US journal. **Jin Changzhu, of the chinese Academy of Sciences (CAS) and lead author of the paper, said the smaller fossil skull indicates the giant pandas were about a third smaller than today's pandas.** **Researchers knew the panda reached its maximum size about 500,000 years ago, when it peaked, and then gradually became smaller.** Jin, a paleontologist at the Institute of Vertebrate paleontology and Paleoanthropology attached to the CAS, said the size _ was a basic rule of evolution. "A species tends to grow bigger when it reaches the peak of its population , but becomes smaller when numbers decline," he said. ...

Q: According to the research of the CAS , there were most pandas in the world _

A: 2 million years ago

B: between 2.4 and 2 million years ago

C: 500,000 years ago (verifier prediction) ✓

D: Nowadays (original prediction)

Table 6: An example from RACE with one sentence as evidence for each answer candidate. We only mark the original prediction and verifier prediction for clarity(in color). Baseline model wrongly predict option D as the answer and our verifier successfully predict the golden option C. We can see that "CAS" occurred in both the evidence sentence of option D and question, which may be the reason of misjudgement. With the extracted evidence sentences, it's more effective at distinguishing between answers with subtle semantic differences.

Passage

I'm Mary.I have a piece of good news to tell you.My parents bought a new flat in the centre of the city.The rooms are not big, but they are all comfortable.There are more rooms than our old flat. **I am excited because I have my own bedroom.**In the old flat, I share the bedroom with my sister. **My favourite room in the new flat is my bedroom.** I can be alone in it.It is my own small world.I can listen to music, read comics and chat with my friends on the phone.I can also look for things on the internet and send e-mails to my e-friends. **Kitchen is my favourite room, too.**I like helping my mother with the cooking.She is not only a good teacher but also a good cook.She often teaches me how to make some different dishes.She lived in Sichuan when she was a child.So she likes hot food and she can cook very delicious hot food.

Q: Which room does Marry like best?

A. The kitchen

B. Her bedroom (original prediction)

C. The sitting room

D. Both A and B (verifier prediction) ✓

Table 7: In this example which is also taken from RACE, we select two evidence sentences for each candidate answer. The sentence in red is both the evidence sentence for option B and option D. Baseline model wrongly predict the option B while our verifier predict the right option D. The two evidence sentences for answer B are strongly related to question but are one-sided for answering the question. On the contrary, two evidence sentences for answer D are necessary and sufficient. It's easily misled to choose the wrong answer B when MRC models or humans only see the sentence in red. However, we can perform accurate reasoning with the help of comprehensive evidence sentences.

Passage
People around the world have their own ways of celebrating weddings. **Now let's compare Eastern and Western weddings.** Chinese and Indian brides normally wear red dresses and most of the wedding decorations are of the same color. This is because the color red is said to bring good fortune. In many Eastern weddings, especially Chinese weddings, the bride will change into a different dress after the ceremony. **White dresses only arrived in modern times because of the influence of Western wedding dress designs.** Western wedding dresses are different. Brides usually wear a wedding dress that is white in color and wear it throughout the whole wedding. **Wearing white wedding dresses is said to have started in the 1840s, beginning with Queen Victoria, Queen of Great Britain.** There are occasions when brides dress into more comfortable clothing so that they will be able to move more freely during the wedding reception. **Traditional practices are strictly followed for most Western weddings.** The couple follow up with any plans they have agreed and decided on to make their special wedding day unique and memorable. **Wedding receptions and other celebrations differ among the East and the West, but the concept is still the same.** They are held to show gratitude towards family members, friends and guests for being a part of the wedding ceremony. Another common custom that weddings of the East and the West have is the wearing of the wedding veil . According to superstitious beliefs, the bride wears a veil to protect her from being seen by evil spirits and the bridesmaids are decoys . **This is still followed even today.**
Q: The passage is mainly developed _ . **A: by time (original prediction)** B: by space C: by process **D: by comparison (verifier prediction)** ✓

Table 8: In this example, we select three evidence sentences for each candidate answer. Both the evidence sentences for option A and for option D are strongly related to their corresponding answer. It's difficult for typical MRC model to arrive at the final answer with these confusing sentences. On the other hand, reasoning over long distance is still strong challenge for MRC models. With the extracted evidence sentences and answer verification, we can perform long-distance reasoning by extracting evidence sentences from the whole passage and aggregating them together.