FY2021Master's Thesis

HRCA+: An Advanced Multi-Choice Machine Reading Comprehension Method

A Thesis Submitted to the Department of Computer Science and Communications Engineering,
the Graduate School of Fundamental Science and Engineering of Waseda University in Partial Fulfillment of the Requirements for the Degree of Master of Engineering

Submission Date: Jan 24th, 2022

Yuxiang Zhang
(5120FG09-7)
Advisor: Prof. Hayato Yamana
Research guidance: Research on Bioinformatics

Abstract

Multiple-choice question answering (MCQA) for machine reading comprehension (MRC) is challenging. It requires a model to select a correct answer from several candidate options related to text passages or dialogue. To select the correct answer, such models must have the ability to understand natural languages, comprehend textual representations, and infer the relationship between candidate options, questions, and passages. Previous models calculated representations between passages and question-option pairs separately, thereby ignoring the effect of other relation-pairs. In this thesis, I propose a human reading comprehension attention (HRCA) model and a passage-question-option (PQO) matrix-guided HRCA model called HRCA+ to increase accuracy. The HRCA model updates the information learned from the previous relation-pair to the next relation-pair. HRCA+ utilizes the textual information and the interior relationship between every two parts in a passage, a question, and the corresponding candidate options. My proposed method outperforms other state-of-the-art methods without the need for extra training data. On the Semeval-2018 Task 11 dataset, my proposed method improved accuracy levels from 95.8% to 96.6%, and on the DREAM dataset, it improved accuracy levels from 90.4% to 90.6%.

Contents

List of Tables

List of Figures

1. Introduction

Machine reading comprehension (MRC) is a challenging task that involves training a model to comprehend the meaning of documents written in natural languages. MRC has attracted significant attention in the field of artificial intelligence, and it was developed to measure how deeply a machine understands context (Liu et al., 2019a). Note that MRC requires a model, especially a supervised learning model, to answer questions based on a specific context. Researchers are expected to train a model to orientate a passage and question pair towards the corresponding answer. MRC tasks are classified into four types (Chen, 2018): cloze test, multiple-choice, span extraction, and free answering.

In this thesis, I tackle the multiple-choice question answering task because this task does not require the model to do language generation and has a wider range of text types (such as articles, dialogues, etc.), making the model focused more on the level of natural language understanding and relational reasoning. Multiple-choice tasks, which are commonly used in language proficiency exams, require selecting one correct answer among multiple candidate options according to a passage. Because the ranges of questions and options are not limited in a passage, some questions require inference combined with commonsense, and this cannot be achieved only using an information retrieval system or through pattern matching. Therefore, pre-trained language models (PrLMs) for understanding a passage, together with matching networks for capturing the relationship between a passage, a question, and the candidate options, are helpful in and crucial to the tackling of multiple-choice tasks. MMM (Jin et al., 2020), DCMN+ (Zhang et al., 2020), and DUMA (Zhu et al., 2020) are three state-of-the-art methods that adopt PrLMs as the encoders of their models. Although all these methods combine questions and options as the entire textual input for their models, candidate options are not always guaranteed to make sense when combined with the question. For example, in some datasets, such as the CosmosQA dataset (Huang et al., 2019), a common candidate option might be ”None of

the above choices." Combining such an unrelated option with a question affects a model's performance. Additionally, questions are not always guaranteed to be related to a passage's parts or span. For example, some questions, such as those in the DREAM dataset (Sun et al., 2019a), involve commonsense knowledge, and such questions cannot be solved only according to the contents of a passage. In addition, previous methods, including the three methods mentioned above, consider the relationships between passages, questions, and the candidate options separately. However, the relationships between every two parts of a passage, a question, and the candidate options are not independent. For example, depending on the differences in the candidate options, the importance of each word in the related questions will differ. Therefore, handling the logical relationships between passages, questions, and the candidate options is indispensable.

To solve the aforementioned problems, in this thesis, I propose a novel method called human reading comprehension attention (HRCA), which is inspired by the ways in which humans achieve a high score in multiple-choice tasks. The HRCA approach simulates the reading strategy employed by humans in the following order: confirming the question, checking the candidate options, and combining the information learned from the question with the candidate options to read the entire passage. Unlike the currently existing methods, my proposed method adopts an updating strategy instead of conventional parallel approaches. Conventional parallel approaches calculate the relationships among passages, questions, and the candidate options individually, and as a result, they do not consider further interrelation. However, the relationships among passages, questions, and the candidate options are not parallel. For example, determining the relationship between a question and the corresponding candidate options helps in the improved inference of the relationship between a passage and a question, and this aspect also applies to other differently related pairs. After calculating the relationship between each related pair, my proposed method updates the relationship information to the calculation of the next related pair. Moreover, to tackle the problem of unrelated options, such as "None of the

above choices," and to address the problem of solving questions that require commonsense knowledge, my proposed method handles and updates the information of the passage, the question, and the candidate options separately, instead of combining them as question-option pairs, thereby ensuring the enhanced performance of my proposed method.

Finally, I extend the operations of my proposed HRCA method to extract nine relationships among every pair of elements in the passage, the question, and the candidate options, thereby enhancing the accuracy levels of my proposed approach. Subsequently, all the relationships are represented using my proposed 3 × 3 passage-question-option (PQO) matrix-guided framework called HRCA+.

The remainder of this thesis is organized as follows: In Chapter 2, I introduce the related studies on multiple-choice tasks. In Chapter 3, I introduce my proposed HRCA model and the PQO matrix-guided framework called HRCA+. In Chapter 4, I describe the datasets used in this thesis, and I present their corresponding hyperparameters. I also describe the experimental settings, and I provide the evaluation results compared to some baselines and various state-of-the-art methods. Finally, I present the discussions and conclusions in Chapter 5.

2. Related Work

Training machines to understand natural language documents remains a daunting challenge. Due to the lack of large-scale human-labeled datasets, conventional MRC methods are based on hand-designed syntax (Riloff and Thelen, 2000) or information extraction approaches (Poon et al., 2010). Between 2013 and 2015, researchers published many human-labeled datasets and promoted reading comprehension tasks to a supervised learning problem (Chen, 2018). After 2015, MRC approaches evolved from machine learning-based approaches to deep learning-based approaches.

In Section 2.1, I describe well-known machine learning-based approaches. In Section 2.2, I describe deep learning-based approaches designed to prevent the influence of noise in hand-engineered linguistic features. I summarize the entire development process in Section 2.3.

2.1 Machine Learning-based Approaches

(Richardson et al., 2013) first proposed a sliding window approach and textual entailment (TE) approach to tackle problems associated with reading comprehension tasks. The sliding window approach is used to match a bag of words extracted from the question and the candidate options related to a specific passage. The TE approach combines the question and each candidate option into question-option pairs, then selects the question-option pair that has the highest similarity to the passage. The distance-based sliding window approach achieved an accuracy level of 61% on the

MCTest MC500 dataset (Richardson et al., 2013). The performance of current state-of-the-art methods is 95.3% in terms of accuracy (Jin et al., 2020). The TE approach later inspired a series of models until recently, e.g., multi-step attention network (MAN) (Jin et al., 2020), dual multi-head co-attention (DUMA) model (Zhu et al., 2020).

The models published later were mainly based on a max-margin framework. This framework posits a hidden relationship between passages, questions, and the corresponding candidate options. (Wang et al., 2015) augmented the initial baseline features based on syntactic dependencies, frame semantics, coreference resolutions, and word embeddings, and they combined all the hand-engineered linguistic features in a max-margin learning framework. As a result, the accuracy of their proposed machine learning-based approach improved from 61% to 70% on the MCTest MC500 dataset (Richardson et al., 2013).

Despite the improvement made by machine learning approaches, there exist several weaknesses. First, datasets are relatively small to support expressive statistical machine learning models. For example, AI2 ProcessBank dataset (Berant et al., 2014) includes only 585 examples, and MCTest dataset (Richardson et al., 2013) includes only 1,480 training examples. Besides, these approaches use hand-engineered linguistic features, some of which rely on existing linguistic tools, such as frame semantic parsing (Das et al., 2010). However, current linguistic tools are far from achieving a solution, and they are only trained in a few domains. Because multiple-choice MRC tasks focus on passages associated with various fields, using such linguistic tools will add noise and affect a model's performance.

2.2 Deep Learning-based Approaches

Deep learning-based approaches do not rely on linguistic features. However, the improvement of the performance of simple deep learning-based models in the completion of multiple-choice tasks is limited. After (Vaswani et al., 2017) proposed a transformer-based structure and demonstrated its enhanced performance in the field of natural language processing (NLP), different types of PrLMs trained using different approaches have been used to suppress and update state-of-the-art approaches for completing NLP tasks. The direct fine-tuning of PrLMs on the downstream task, defining new pre-training tasks, and adding task-specialization networks based on PrLMs are common approaches for ensuring the enhanced performance of such models in the completion of NLP tasks.

Models that are designed to complete multiple-choice tasks must have the ability to understand natural languages on a high level, and such models must be able to capture and infer the relationships among passages, questions, and the corresponding candidate options. Numerous methods for enhancing the performance of PrLMs in the completion of multiple-choice tasks have been proposed and applied. Following this direction, (Zhang et al., 2020) proposed a dual co-matching network, and they integrated two reading strategies into their proposed network. One strategy involved using the key sentence selection mechanism, which is used to determine the most salient supporting sentences for answering a specific question. The other strategy involved encoding the comparison information between candidate options. (Jin et al., 2020) proposed a multi-stage multi-task-based learning framework. The multistage multi-task learning approach relies on two out-of-domain (general) datasets and one large in-domain (same type) dataset to help the model achieve improved generalization using a limited amount of data. Additionally, a multi-step attention

network was proposed to dynamically calculate the attention scores between the passage and question or the passage and candidate options pairs step by step. (Zhu et al., 2020) proposed a dual multi-head co-attention approach for calculating the attention score between passage and question-option pairs, and their proposed approach considered the passage and question-option pairs, with a major focus on the standpoint of each pair. Such ideas promote the effective solving of multiple-choice tasks, with accuracy levels of 60% to 90% in the DREAM dataset (Sun et al., 2019a).

However, for complicated datasets, such as the C3 (Sun et al., 2020) and DREAM (Sun et al., 2019a) datasets, the question will not always be related to a part of or the span of a passage. In addition, some of the candidate options might not correspond to the question, and this means that if the combination of questions and the corresponding candidate options is considered part of the PrLM encoder's input text, it might affect the model's performance. Additionally, for every two parts, the relationships between passages, questions, and the corresponding candidate options are not independent of each other. For example, the relationship between a question and its corresponding candidate options helps in the enhanced inference of the relationship between a passage and a question, and this aspect applies to other differently related pairs.

2.3 Summary of Related Work

The entire development process of the multiple-choice machine reading comprehension task can be summarized as:

1) Early stage:

   Due to the lack of large-scale human-labeled datasets, the MRC methods in this stage rely on hand-designed syntax or information extraction approaches.

2) Machine learning stage:

   In this stage, researchers published many human-labeled datasets and promoted reading comprehension tasks to a supervised learning problem. This helped the development of machine learning methods. The two most famous and obvious improvements are sliding window approach and max-margin framework-based models.

3) Deep learning stage:

   Machine learning-based approach still fails to get rid of the dependence on hand-engineered linguistic features. With the development of PrLMs, a top-level network plus the PrLM has become the mainstream method to solve the MRC task. Among the models that follow this direction, MMM (Jin et al., 2020), DCMN (Zhang et al., 2020) and DUMA (Zhu et al., 2020) are the three latest and most representative models. Current state-of-the-art method is DUMA.

The proposals and limitations of those methods, together with the performance on MCTest 500 dataset (Richardson et al., 2013) and DREAM dataset (Sun et al., 2019a) are listed in Table 1.

	Proposal	Limitation	Performance [Accuracy]	
			MCTest 500	DREAM
Sliding Window (Richardson et al., 2013)	Match a bag of words extracted from the question and the candidate options related to a specific passage	Only consider the word overlap rate without any reasoning framework.	60.8♠	44.6†
Latent Structural SVM (Sachan et al., 2015)	1) Present a unified max-margin framework that learns to find the answer-entailing structure 2) Use a top-level question-type classification	Max-margin framework -based model is insufficient to find enough hidden information in natural language	69.9*	-
DCMN (Zhang et al., 2020)	1) Dual co-matching network 2) Integrate two reading strategies including key sentence selection and information comparison encoding of candidate options	Only consider three basic relationship pair among the passage, the question and options	93.4♡	87.8◊
MMM (Jin et al., 2020)	1) Use some out-of-domain (general) datasets and one in-domain (same type) dataset 2) Use a multi-step attention network to calculate the attention scores between the passage and (question, option) pair	Require huge amount of extra related datasets and resources	95.8♣	88.9♣
DUMA (Zhu et al., 2020)	Respectively considering each other's focus from the standpoint of passage and (question + candidate option)	Some of the candidate options might not corresponding to the question, information not updating	-	90.4◊

Table 1: MRC Approaches and Performance in accuracy (%) on MCTest 500 dataset and DREAM dataset.

(♠: reported by (Richardson et al., 2013),

*: reported by (Sachan et al., 2015),

†: reported by (Sun et al., 2019a),

♣: reported by (Jin et al., 2020),

◊: reported by (Zhu et al., 2020),

♡: reported by (Zhang et al., 2020).)

3. Human Reading Comprehension Attention Models

The remaining problems associated with multiple-choice question answering (MCQA) for MRC include:

Problem 1 The problem of incomplete correspondence

The problem of incomplete correspondence involves the mismatch between a passage and a question and the mismatch between a question and its corresponding candidate options. In other words, a question is not guaranteed to be related to a part or the span of a passage. Additionally, some of the candidate options might not correspond to the question.

Problem 2 The helping-relation problem

The helping-relation problem involves the interlinkages problem, which is associated with the relationship between every two parts of a passage, a question, and the corresponding candidate options. For example, determining the relationship between a question and its corresponding candidate options helps in the enhanced inference of the relationship between a passage and a question, and this aspect applies to other differently related pairs.

To solve the problem of incomplete correspondence, I propose the HRCA method. HRCA method tackle the corresponding two problems through the following two solutions:

Solution 1:

Instead of only considering the passage and question or the question and candidate options pairs, to prevent the influence of inconsistent pairs of candidate options and questions, my proposed HRCA model considers the relationships between passages, questions, and the corresponding candidate options separately to solve the problem of incomplete correspondence.

Solution 2:

HRCA method also addresses the helping-relation problem by updating the learned information obtained from the previous relation-pair to the next relation-pair when executing the attention mechanism. However, previous models update such information in parallel. Moreover, I extend the operations of my proposed HRCA method to fully utilize the information extracted using the PrLM.

Section 3.1 shows the definition of a multiple-choice question answering task. Next, the overall architecture of my proposed model is presented in Section 3.2, each of which is explained in Sections 3.3–3.6.

3.1 Task Definition

MCQA tasks comprise passages (P) of text, questions (Q) related to P, and n candidate answer options (O) for each Q. An example is presented in Table 2. MCQA tasks aim to build a model for calculating the probability of correctness for each candidate option.

$$F: (P, Q, \{O1, O2, \dots, On\}) \rightarrow \{\Pr(O1), \Pr(O2), \dots, \Pr(On)\} \quad (1)$$

Table 2. MCQA sample extracted from the DREAM (Sun et al., 2019a) dataset

Passage (dialogue form)

W: Good morning, Mr. Jacob. Is everything all tight?

M: No, it's not. Someone's stolen some of my valuables two rings and a gold necklace.

W: I'm very sorry to hear that, sir. Where were they?

M: In my room. And the door was locked. It can only be one of your staff. I want my things back. And fast.

W: Well, I can certainly understand that you're upset about losing them and we'll do all we can to help. If they really are missing, it's a matter for the police.

M: What do you mean, if they are missing? I told you they were.

W: Yes, Sir. But first I'll have one of the housekeeping staff look through your room in case they're still there. But I must say that we can't be held responsible. You should have deposited the valuables with Reception. It says so on the Key Card.

M: That's not good enough. I want to see the manager immediately.

W: I'll be glad to call the duty manager for you, sir. But he'll certainly say the same. We have clear instructions about valuables and we must follow them."

Question1: Where does this conversation most probably take place?

Candidate options:

A. In a shop.

B. At a hotel. √

C. At a restaurant.

Question2: Why does the man want to see the manager?

Candidate options:

A. He isn't satisfied with the woman's answer. √

B. He is angry.

C. He is sad.

3.2 Model Architecture

Figure 1 illustrates the overview of my proposed model's architecture, which simulates the strategies employed by humans in their attempt to achieve high scores in reading comprehension exams. Based on the PrLM encoder, I first generate the word embedding of the combination of passages, questions, and each candidate option. Next, all the word embeddings are divided into three parts corresponding to the passage, question, and each candidate option separately. Further, K HRCA layers repeat the following three steps by K times, thereby simulating the way in which humans attempt to achieve high scores in reading comprehension exams.

Step 1:

I perform multi-head self-attention on the question (I regard the question as a query, key, and value).

This step is aimed to allow the model confirm the question again.

Step 2:

I perform multi-head attention on the option and the updated question presented in Step 1 (I regard the option as a query, and I regard the updated question as the key and value).

This step is executed to simulate the understanding of the candidate options after confirming the question.

Step 3:

I perform multi-head attention on the passage and the updated options presented in Step 2 (I regard the passage as a query, and I regard the updated option as the key and value).

This step allows for the model to comprehend the passage with the question and the corresponding candidate options after understanding the candidate options.

Through this approach, I obtain the attention score for each text, which is then transformed into a probability distribution for each candidate option using a multi-layer perceptron. Finally, I choose the option with the highest probability as the answer.

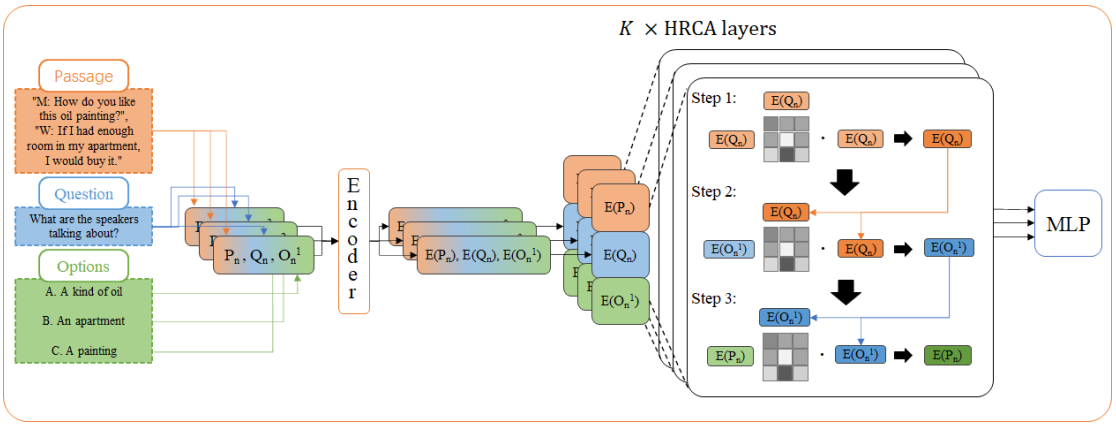

Figure 1. Architecture of HRCA model.

3.3 Contextualized Encoding

In my proposed model, I adopt a PrLM to generate a global contextualized representation. Let a passage be $P = [p_1, p_2, \dots, p_i, \dots, p_k]$, a question be $Q = [q_1, q_2, \dots, q_i, \dots, q_l]$, and a candidate option be $O = [o_1, o_2, \dots, o_i, \dots, o_m]$, where p_i, q_i, and o_i represent tokens processed using the PrLM as a word in the text of the passage, the question, and the corresponding candidate option, respectively. I concatenate each candidate option O with its corresponding question Q and its corresponding passage P into one sequence. After feeding the concatenated sequence into the PrLM's encoder function $Encode(\cdot)$, the output $\mathrm{E} = \mathrm{Encode}(\mathrm{P} \oplus Q \oplus O)$ can be obtained. The PrLM's encoder output E has the following form: $[e_1, e_2, \dots, e_{k+l+m}]$. Note that if we have one passage, one question, and four candidate options, we obtain four different concatenated sequences in total.

The notations defined in this section are listed in Table 3.

Table 3: Notations of variable.

Variable	Interpretation
P, Q, O	The **P**assage, the **Q**uestion and candidate **O**ptions
p_i, q_i, o_i	The token in the text of the **P**assage, the **Q**uestion, and the corresponding candidate **O**ption processed by the PrLM
$Encode(\cdot)$	The PrLM's encoder
E	The embedding output of PrLM encoder
e_i	The token of the PrLM encoder's output

3.4 Human Reading Comprehension Attention

As shown in Figure 1, based on the multi-head attention module (Vaswani et al., 2017), I propose the HRCA method to enlighten the model and enable it to learn the relationships between every two parts in a passage, a question, and the corresponding candidate options and to update the learned information to the next learning step. The output E of the PrLM's encoder is separated into E^P, E^Q, and E^O, each of which represents the embedding of the passage, the question related to the passage, and the candidate option to the question, respectively. I first use a question as a query, key, and value, to reconfirm the question. Afterwards, I use the answer as a query and the updated question as the key and value to understand the candidate option. Finally, I regard the passage as a query and the updated answer as the key and value to comprehend the passages with the question and the corresponding candidate options. I then update the information obtained using the HRCA layer in the order of question, option, and passage, and the pseudo-code is presented in Algorithm 1.

Algorithm 1 HRCA calculation process

Require The PrLM's encoder output of the passage E^P, the question E^Q and the options E^O

1: **function** MHSA(Q, K, V) ▷The query, key and value.

2: $\text{Attention}(\text{Q}, \text{K}, \text{V}) \leftarrow \text{Softmax}\left(\frac{Q(K)^T}{\sqrt{d_k}}\right) \cdot V$

▷d_k represents the dimension of K

3: $head_i \leftarrow Attention(QW_i^Q,\ KW_i^K,\ VW_i^V)$

▷W^x represents the weight matrix of x

4: **return** $Concat(head_1, head_2, \ldots, head_i)W^O$

end function

function R(E^P, E^Q, E^O)

    return $Concat(E^P, E^Q, E^O)$

end function

function HRCA(E^P, E^Q, E^O)

    $E^{\boldsymbol{Q^U}} \leftarrow$ MHSA(E^Q, E^Q, Q)

    $E^{\boldsymbol{O^U}} \leftarrow$ MHSA($E^O, E^{\boldsymbol{Q^U}}, E^{\boldsymbol{Q^U}}$)

    $E^{\boldsymbol{P^U}} \leftarrow$ MHSA($E^P, E^{\boldsymbol{O^U}}, E^{\boldsymbol{O^U}}$)

    return R($E^{\boldsymbol{P^U}}, E^{\boldsymbol{Q^U}}, E^{\boldsymbol{O^U}}$)

3.5 PQO Matrix

As shown in Figure 2, the PQO matrix is a 3 × 3 matrix that includes all the possible combinations of the relationships between passages, questions, and their corresponding candidate options. Although HRCA method already considered the three types of relations, i.e., the question and the question itself, the candidate options and the question, and the passage and candidate options, in the HRCA layer, we still have six unused relationship pairs among these relations. Therefore, the PQO matrix is used to list the nine possible relation-pairs for confirming the parts that remain used or unused through the HRCA method.

In Figure 2, “Self” represents self-attention, and “AtoB” shows the way in which B is used to calculate the attention score of A, where A and B represent one passage,

question, and the corresponding candidate option. Additionally, the dark–red-colored cells represent the attention process used in the HRCA layer. The light–red-colored cells (cells in the diagonal axis) represent the self-attention of the corresponding element, which is normally calculated using the PrLM's self-attention module. The grey-colored cells represent the attention processes that are not used in the HRCA layer.

	Passage	Question	Option
Passage	Self	PtoQ	PtoO
Question	QtoP	Self	QtoO
Option	OtoP	OtoQ	Self

Figure 2. PQO Matrix for calculating the attention

3.5.1 HRCA+: PQO Matrix-guided HRCA

I extend the operations of my proposed HRCA method to adopt all the passage, question, and corresponding candidate option relationships. Because such relationships are not limited to the three relationships adopted during the implementation of the HRCA method, it is expected that the performance of the

proposed HRCA approach increases during the use of entire relationships. Therefore, I propose an advanced multi-choice MRC method called HRCA+ to adopt the unused relationships in the proposed HRCA, i.e., the light-red-colored cells and the grey-colored cells presented in Figure 2.

In HRCA+, I update the adjacent cells in the order presented in Figure 3. Because updating the attention scores of the target relation pairs from their previous relation pairs relies on the adjacent connection between both relation pairs, I update the adjacent cells in a manner that allows the updating of only one sequence. For example, suppose that the previous relation pair is a question-to-option pair. In such a case, the next relation pair must satisfy the appearance of at least one element in the previous relation pair, i.e., the question or option, to ensure the update's validity, where A-to-B represents using B to calculate the attention score of A.

	Passage	Question	Option
Passage	7	8	9
Question	6	1	2
Option	5	4	3

Figure 3. Updating order in PQO Matrix

3.6 R Function and Multi-Layer Perceptron

HRCA+ updates the PrLM's outputs of the passage, the question, and the corresponding candidate options. Next, a reduce function (R function) is required to combine those three outputs. The common reduce function includes concatenation, element-wise summation, and element-wise production. (Zhang et al., 2020) used concatenation to combine the final representation outputs. I investigate and compare the reduction functions mentioned above, and the results are presented in Section 4.3.

Additionally, I must consider using the combined outputs to generate the probability distribution for each option. Because PrLM outputs have a much larger dimension than that of the candidate options, reducing the dimension is required. Note that 512 is a common embedding dimension for most PrLMs, and for most multiple-choice tasks, the number of candidate options ranges from 2–4.

To generate one feature map for each corresponding PrLM token, (Zhang et al., 2020) used row-wise max pooling, and (Zhu et al., 2020) used mean pooling. My proposed model adopts global average pooling to retain additional information from the previous output. Meanwhile, as a multi-class classification task, multiple-choice requires the model to predict the label with the highest confidence score, which works well using a Softmax function. This is how my proposed multi-layer perceptron (MLP) was formulated.

4. Experiments

In this chapter, I evaluate the performance of my proposed method on multiple-choice reading comprehension examination datasets.

4.1 Datasets

I used the following two datasets: DREAM (Sun et al., 2019a) and SemEval-2018 Task 11 (Ostermann et al., 2018).

DREAM DREAM is a dialogue level multiple-choice reading comprehension dataset collected from English-as-a-foreign-language Examinations. DREAM is a challenging dataset since 85% of questions require reasoning beyond a single sentence, 34% of questions involve commonsense knowledge. The statistical data of DREAM dataset is shown in Table 4.

SemEval-2018 Task 11 The SemEval-2018 Task 11 dataset assesses the way in which the inclusion of commonsense knowledge, i.e., script knowledge, benefits MRC systems. Script knowledge is defined as the knowledge regarding daily activities, such as baking a cake or taking a bus. In addition to what is mentioned in the text, many questions require inference using script knowledge regarding different scenarios, such as answering questions that require additional knowledge beyond the facts mentioned in the text. The statistical data of SemEval-2018 Task 11 dataset is shown in Table 5.

Table 4: Statistical data of DREAM dataset. “Extractive” means the answers are spans of the passage, and “Abstractive” means the answers are not spans.

	DREAM
Number of source documents	6,444
Number of total questions	10,197
Train/Dev/Test split	3:1:1
Extractive (%)	16.3
Abstractive (%)	83.7
Average answer length	5.3
Number of answers per question	3
Avg./Max. number of turns per dialogue	4.7 / 48
Avg. passage length	85.9
Vocabulary size	13,037

Table 5: Statistical data of SemEval-2018 Task 11 dataset.

	SemEval-2018 Task 11
Number of source documents	2,199.0
Number of total questions	13,939.0
Average passage length	196.0
Average question length	7.8
Average answer length	3.6
Number of answers per question	2.0
Number of questions per passage	6.7

4.2 Experimental Settings

My proposed method is an improvement of PrLMs. I use a PrLM as the encoder for generating the hidden states of concatenated text. For the layers of HRCA and HRCA+, I use $K = 4$ for both the DREAM and the SemEval-2018 Task 11 datasets.

Baseline ALBERTbase, and ALBERTxxlarge (Lan et al., 2019) using multiple-choice models are selected as the baselines. The learning rate is 8e-6 for both the DREAM and Semeval-2018 Task 11 datasets. The batch size is set to two for the DREAM dataset, and it is set to four for the Semeval-2018 Task 11 dataset. The proposed model is trained for three epochs on the DREAM dataset and for two epochs on the Semeval- 2018 Task 11 dataset. Note that previous methods (Zhang et al., 2020; Zhu et al., 2020) used a batch size of eight for the DREAM dataset. Therefore, I also implement DUMA (Zhu et al., 2020), and I train both DUMA and vanilla ALBERT-xxlarge using a batch size of two to achieve highly intuitive performance comparison. The hyperparameters remain the same for all other compared methods, including PrLMs and PrLM-based models. The other PrLMs used for comparison in this study include BERT (Devlin et al., 2019), GPT (Radford et al., 2018), XLNet (Yang et al., 2019), and RoBERTa (Liu et al., 2019b). Other PrLM-based models used for comparison in this study include WAE (Kim and Fung, 2020), DCMN (Zhang et al., 2020), MMM (Jin et al., 2020), and DUMA (Zhu et al., 2020). The hyperparameters used are listed in Table 6.

Table 6: Hyperparameters used on DREAM dataset and SemEval-2018 Task 11 dataset.

Hyperparameter	DREAM	SemEval-2018 Task 11
Batch size	2	4
Learning rate	8e-6	8e-6
Epoch	3	2
Layers of HRCA+	4	4
# of attention head in HRCA+	12	8
Attention hidden size in HRCA+	64	64
HRCA+ dropout rate	0.1	0.2
Dropout rate in MLP	0.1	0.2

Data pre-processing For the DREAM dataset, I apply data pre-processing to maintain the consistency of gender representations, i.e., man and woman, among passages, questions, and the corresponding candidate options. The symbols, W and M, represent "woman" and "man" as the speaker attributes in a passage. However, the symbols, man and woman, are used in questions and their corresponding candidate options. Therefore, the symbols W and M are replaced with "woman" and "man" during data pre-processing.

4.3 Results

In this thesis I adopted accuracy as the evaluation metric. The experiments' results are listed in Tables 7–10.

I first evaluated the performance of my proposed model using the DREAM dataset. Table 7 shows the difference in accuracy, as it pertains to the development and test sets when using one to five MHSA layers on the HRCA and HRCA+ approaches. In the HRCA method, the performance first increases, after which it decreases as the number of layers increases. Contrarily, in HRCA+, the performance increases in proportion to the number of layers. For comparison, the performance of the related model, DUMA, begins to decrease when the number of layers increases to two. The HRCA method uses the three most efficient relation pairs, whereas the HRCA+ uses all possible relation pairs. This phenomenon reflects that HRCA+ learns additional information compared to the HRCA method. Even if the attention calculation mechanism is repeated multiple times, the proposed model can still learn useful information to improve accuracy.

Table 8 shows the accuracy difference in the test set when using element-wise production, element-wise summation, and concatenation as the reduce function for combining the final representation outputs with those of HRCA+. According to the results, concatenation demonstrates improved performance compared to that of the other two functions because it retains the previously learned information to the maximum extent possible.

Table 9 shows the publication results on the DREAM dataset. The state-of-the-art model that combines the ALBERT-xxlarge and DUMA approaches applies multi-task learning (Jin et al., 2020), thereby adopting extra training data. On the official leaderboard of the performance on the DREAM dataset, my proposed model achieves the highest accuracy of 90.6% among all the other models that do not require extra training data. My proposed model uses different batch sizes obtained from the previously mentioned methods. Owing to challenges associated with computational

resources, I only use a batch size of two, whereas the previous methods use a batch size of eight. To achieve enhanced intuitive performance for comparison purposes, I implement DUMA, which is the current state-of-the-art method, and I use similar parameters to train my implementation of the DUMA and vanilla ALBERTxxlarge-based methods. For each result, I train my proposed model five times, and I calculate the average value. Note that there is a gap of 1.3% in the performance of my implementation and the initial DUMA-based method. To verify the accuracy of my proposed method, I test the performance of my proposed implementation through the ALBERT-base using parameters that are similar to those of the initial DUMA-based method.

I also test the performance of my proposed model on the SemEval-2018 Task 11 dataset presented in Table 10. The best accuracy level was 84.1% (Ostermann et al., 2018) during the SemEval-2018 Task 11 competition. As shown in the second grid of Table 10, I adopted the PrLM results showing improved accuracy from 84.1%, i.e., from the results obtained using the GPT and RoBERTa-Large-based methods. By applying strategies, such as back and forth reading, highlighting, and self-assessment (Sun et al., 2019b), and by applying model ensembles using extra training data, the accuracy of my proposed model was improved to 95.8%. My proposed model achieves the highest accuracy level of 96.6% among all the previous models without the need for extra training data. Even in comparison to the performance of the best model relying on extra training data, i.e., the RoBERTa-Large+MMM (Jin et al., 2020) model, my proposed model achieves the highest accuracy levels.

Table 7: Performance in accuracy (%) with various MHSA layers on DREAM dataset based on ALBERTbase. (Showing the accuracy as development dataset / test dataset)

Baseline	Layers	+ HRCA	+ HRCA+
	1	66.2/67.4	67.3/67.9
ALBERT	2	67.0/67.9	67.5/68.9
-base	3	67.6/68.8	68.0/69.0
64.5/64.4	4	68.2/67.7	68.5/69.7
	5	67.9/68.0	**69.6/69.8**

Table 8: Performance in accuracy (%) with different reduce functions on DREAM dataset.

Model	Reduce function	Test
ALBERT-base	-	64.4
+ HRCA+	element-wise production	66.9
+ HRCA+	element-wise summation	68.5
+ HRCA+	concatenation	68.9

Model	Dev	Test
Random	32.8†	33.4†
GBDT++ (Sun et al., 2019a)	53.3†	52.8†
FTLM++ (Radford et al., 2018)	57.6†	68.5†
Ensemble of 3 FTLM++	58.1†	68.9†
Ensemble of 1 GBDT++ and 3 FTLM++	59.6†	59.5†
BERT-base	63.2♣	63.2♣
ALBERT-base	64.5◊	64.4◊
BERT-large	66.2♣	66.9♣
XLNet-large	-	72.0◊
RoBERTa-large	85.4♣	85.0♣
ALBERT-xxlarge	89.2◊	88.5◊
BERT-large + WAE (Kim and Fung, 2020)	-	69.0◊
ALBERT-xxlarge + DCMN (Zhang et al., 2020)	-	87.8◊
RoBERTa-large + MMM (Jin et al., 2020)	88.0♣*	88.9♣*
ALBERT-xxlarge + DUMA (Zhu et al., 2020)	89.9◊	90.4◊
ALBERT-xxlarge + DUMA + Multi-Task Learning (Wan, 2020)	91.9♡*	91.8♡*
ALBERT-base + DUMA	-	67.6◊
ALBERT-base + DUMA (my implementation)	-	67.5
ALBERT-xxlarge (my implementation)	88.2	88.0
ALBERT-xxlarge + DUMA (my implementation)	89.5	89.1
ALBERT-xxlarge + HRCA+ (my model)	**90.1**	**90.6**
Human Performance (Sun et al., 2019a)	93.9†	95.5†
Ceiling Performance (Sun et al., 2019a)	98.7†	98.6†

Table 9: Performance in accuracy (%) on DREAM dataset.

(†: reported by (Sun et al., 2019a),

♣: reported by (Jin et al., 2020),

◊: reported by (Zhu et al., 2020),

♡: reported by (Wan, 2020),

*: using extra training data when training.)

Model	Test
Sliding Window (Richardson et al., 2013)	55.0†
Attentive Reader (Chen et al., 2016)	72.0†
Best score in competition (Ostermann et al., 2018)	84.1†
GPT	88.0♣
BERT-base	88.1◊
GPT(2x)	88.6♣
BERT-large	88.7◊
RoBERTa-large	94.0◊
GPT+Strategies (Sun et al., 2019b)	88.8♣
GPT+Strategies (2x) (Sun et al., 2019b)	89.5♣
RoBERTa-large + MMM (Jin et al., 2020)	95.8◊*
ALBERT-xxlarge + HRCA+ (my model)	**96.6**
Human Performance (Ostermann et al., 2018)	98.0†

Table 10: Performance in accuracy (%) on SemEval- 2018 Task 11 dataset.

(†: reported by (Ostermann et al., 2018),

♣: reported by (Sun et al., 2019b),

◊: reported by (Jin et al., 2020),

*: using extra training data when training.)

5. Conclusions

In this thesis, I propose a method called human reading comprehension attention (HRCA) for simulating the reading strategies employed by humans. Compared to other state-of-the-art methods, my proposed approach achieves a higher score when tackling multiple-choice comprehension tasks. I further propose a passage-question-option matrix-guided HRCA approach called HRCA+ to fully utilize the information between passages, questions, and the corresponding candidate options extracted using PrLMs. The experiments' results on the DREAM and Semeval-2018 Task 11 datasets show that my proposed method achieves the highest accuracy among other existing state-of-the-art methods without the need for extra training data.

In my future studies, I shall focus on the application of multi-task learning (Jin et al., 2020) and ensemble methods to improve the performance of my proposed HRCA method. Additionally, I shall integrate the applications of my proposed method to other tasks, such as the extraction of relationships between passages, questions, and their corresponding candidate options.

Reference

Chen, D., Bolton, J., and Manning, C. D. (2016). A thorough examination of the CNN/Daily Mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany, August. Association for Computational Linguistics.

Chen, D. (2018). *Neural reading comprehension and beyond*. Stanford University.

Das, D., Schneider, N., Chen, D., and Smith, N. A. (2010). Probabilistic frame-semantic parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 948–956, Los Angeles, California, June. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). 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)*, pages 4171– 4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Huang, L., Le Bras, R., Bhagavatula, C., and Choi, Y. (2019). Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. 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)*, pages 2391– 2401, Hong Kong, China, November. Association for

Computational Linguistics.

Jin, D., Gao, S., Kao, J.-Y., Chung, T., and Hakkanitur, D. (2020). Mmm: Multi-stage multi-task learning for multi-choice reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8010–8017.

Kim, H. and Fung, P. (2020). Learning to classify the wrong answers for multiple choice question answering (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13843–13844.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Liu, S., Zhang, X., Zhang, S.,Wang, H., and Zhang,W. (2019a). Neural machine reading comprehension: Methods and trends. *Applied Sciences*, 9(18):3698.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019b). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ostermann, S., Roth, M., Modi, A., Thater, S., and Pinkal, M. (2018). Semeval-2018 task 11: Machine comprehension using commonsense knowledge. In *Proceedings of the 12th International Workshop on semantic evaluation*, pages 747–757.

Poon, H., Christensen, J., Domingos, P., Etzioni, O., Hoffmann, R., Kiddon, C., Lin, T., Ling, X., Ritter, A., Schoenmackers, S., et al. (2010). Machine reading at the university of washington. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, pages 87–95.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.

Richardson, M., Burges, C. J., and Renshaw, E. (2013). Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 193–203.

Riloff, E. and Thelen, M. (2000). A rule-based question answering system for reading comprehension tests. In *ANLP-NAACL 2000 Workshop: Reading Comprehension Tests as Evaluation for Computer- Based Language Understanding Systems*.

Sachan, M., Dubey, K., Xing, E., and Richardson, M. (2015). Learning answer-entailing structures for ma-chine comprehension. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 239–249, Beijing, China, July. Association for Computational Linguistics

Sun, K., Yu, D., Chen, J., Yu, D., Choi, Y., and Cardie, C. (2019a). Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Sun, K., Yu, D., Yu, D., and Cardie, C. (2019b). Improving machine reading comprehension with general reading strategies. 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)*, pages 2633–2643, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Sun, K., Yu, D., Yu, D., and Cardie, C. (2020). Investigating prior knowledge for challenging Chinese machine reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:141– 155.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural*

information processing systems, pages 5998– 6008.

Wan, H. (2020). Multi-task learning with multi-head attention for multi-choice reading comprehension. *arXiv*:2003.04992.

Wang, H., Bansal, M., Gimpel, K., and McAllester, D. (2015). Machine comprehension with syntax, frames, and semantics. In P*roceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 700–706, Beijing, China, July. Association for Computational Linguistics.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32, pages 5754–5764. Curran Associates, Inc.

Zhang, S., Zhao, H., Wu, Y., Zhang, Z., Zhou, X., and Zhou, X. (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.

Zhu, P., Zhao, H., and Li, X. (2020). Dual multi-head co-attention for multi-choice reading comprehension. *ArXiv*, abs/2001.09415.

Berant, J., Srikumar, V., Chen, P.-C., Vander Linden,A., Harding, B., Huang, B., Clark, P., and Man-ning, C. D. (2014). Modeling Biological Processes for Reading Comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* pages 1499–1510, Doha, Qatar, October. Association for Computational Linguistics.

Appendix I

Case Study

In Table 9, my HRCA+ model improves the previous state-of-the-art method DUMA (Zhu et al., 2020) by 1.5% in accuracy on the DREAM dataset. To further demonstrate the superiority and performance of my model, four cases (including three good cases and one bad case) are shown in Table A1–A4.

Note that both my HRCA+ model and my implementation of the DUMA model are based on the ALBERTbase model. The same hyperparameters are used for two models, including: The batch size of 8, the learning rate of 1e-5 and the training epochs of 3. All cases are taken from the DREAM (Sun et al., 2019a) dataset, which is a more complicated dataset than SemEval-2018 Task 11 (Ostermann et al., 2018) dataset and requires a higher level of infer ability.

In cases 1–3, my HRCA+ model successfully selected all the correct answers. DUMA model chose the wrong answers. However, those wrong answers are highly related to the passage and the question. This phenomenon reflects that both my HRCA+ model and DUMA model successfully extracted the useful information from the passage, and my HRCA+ model has a much better infer ability when dealing with the complicated relationship between the passage, the question and the candidate options.

In case 4, both HRCA+ and DUMA models chose the wrong answer. Although my model is in the correct direction according to the passage, it is still hard for the model to give a compromise answer when a clearer answer has a higher degree of confidence.

Case 1

Table A1. Good case 1 from DREAM dataset

Passage (dialogue form)
M: Did you watch the football match on TV yesterday evening?
W: No I didn't. I had dinner with a friend and didn't go back home until eight o'clock.
Question: What did the woman do yesterday evening?
Candidate options:
A. She ate out. √
B. She watched TV.
C. She watched a match.
DUMA (my implementation): She watched a match. ×
HRCA+ (my model): She ate out. √

Case 2

Table A2. Good case 2 from DREAM dataset

Passage (dialogue form)
M: What day is it today?
W: It is Thursday, December the 5th.
Question: What day will it be tomorrow?
Candidate options:
A. Thursday, December the 5th.
B. Thursday, November the 5th.
C. Friday, December the 6th. √
DUMA (my implementation): Thursday, December the 5th. ×
HRCA+ (my model): Friday, December the 6th. √

Case 3

Table A3. Good case 3 from DREAM dataset

Passage (dialogue form)
W: Henry, why were you late this morning?
M: My neighbour had a sudden heart attack and I had to take him to the hospital.
Question: What's wrong with Henry?
Candidate options:
A. He was in hospital.
B. He had a heart attack.
C. He was late for work. √
DUMA (my implementation): He had a heart attack. ×
HRCA+ (my model): He was late for work. √

Case 4

Table A4. Bad case 4 from DREAM dataset

Passage (dialogue form)
M: Did you see that movie last night? I thought it was fantastic.
W: Really? I didn't think there was anything special about it.
Question: What did the woman think of the movie?
Candidate options:
A. Terrible.
B. Average. √
C. Fantastic.
DUMA (my implementation): Fantastic. ×
HRCA+ (my model): Terrible. ×

Appendix II

Attention weight heatmap

The case study shows that my model has a better inferential capability than the previous state-of-the-art method DUMA. To confirm the effectiveness of my model's inferential capability, the attention weight heatmap is shown.

We manually predict one example from the DREAM (Sun et al., 2019a) dataset using my HRCA+ model. In my HRCA+ model, the word embeddings generated by PrLM encoder are divided into three parts corresponding to the passage, question, and each candidate option separately to apply further attention calculation. We generate the heatmap using the calculated output of my PQO-matrix guided HRCA before the reduce function (concatenation). Table A5 shows the example taken from the DREAM dataset, Figure A1 shows the attention weight heatmaps of the example predicted by HRCA+. The column of each heatmap represents the attention weight of the corresponding token in passage, question and each candidate option. For example, option 3 is "A painting". In the heatmap of option 3, the attention weight of the token "A" and the token "painting" are shown in the corresponding column 0 and column 1. The row of each heatmap represents the magnitude of attention weight.

The heatmap of the wrong answer (option 1 and option 2) in Figure A1 does not have much varied attention weight in passage and question. However, the heatmap of the correct answer (option 3) has different attention weights for different tokens. This phenomenon reflects that my HRCA+ model is able to learn effective information from the passage and question.

Table A5. An example from DREAM dataset

Passage (dialogue form)
M: How do you like this oil painting?
W: If I had enough room in my apartment, I would buy it.
Question: What are the speakers talking about?
Candidate options:
A. A kind of oil
B. An apartment
C. A painting √

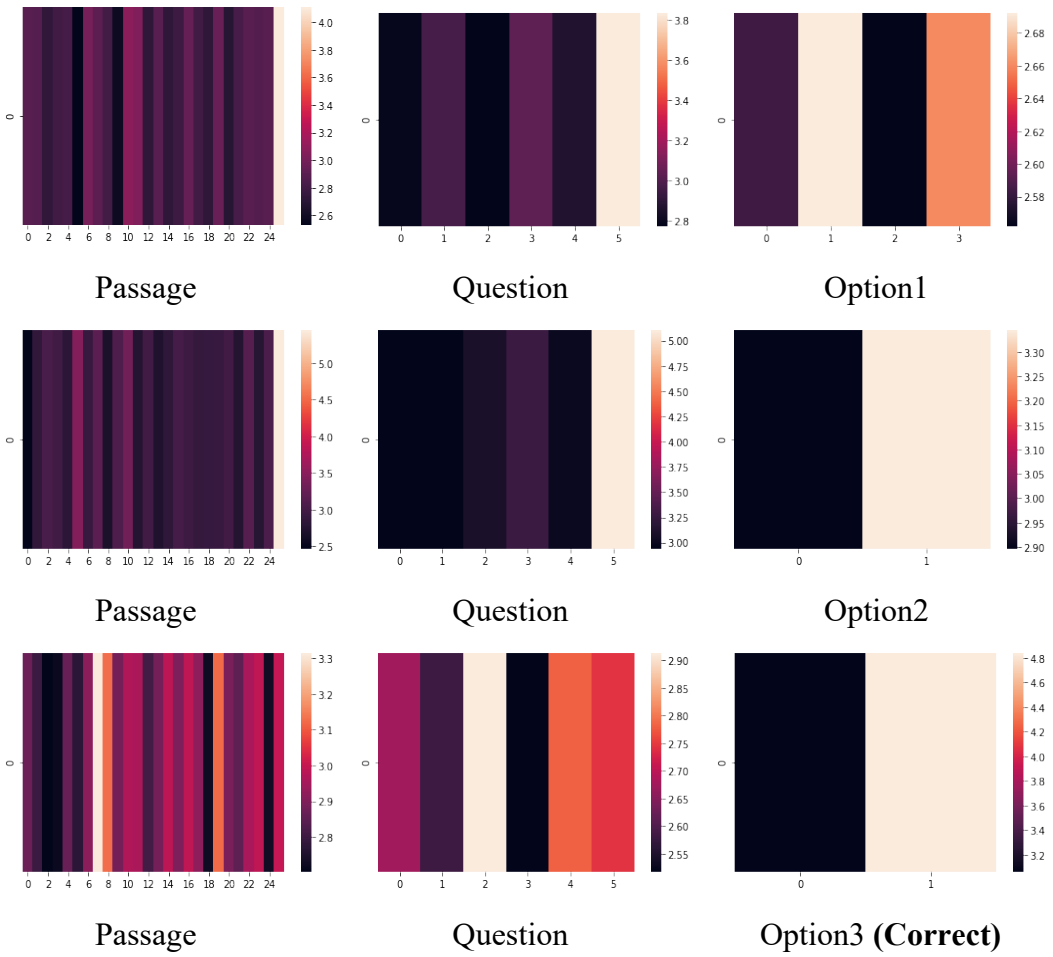

Figure A1. Attention weight heatmaps of one example predicted by HRCA+

Appendix III

Failure / limitation of the HRCA+ model

My HRCA+ model has already surpassed the current state-of-the-art model in accuracy by 1.5%. However, it still has a 4.9% of the gap in accuracy between my HRCA+ model and the human performance on the DREAM dataset.

To check how my model failed and consider the future direction of improvement, I summarize four typical failures of my model in Table A6–A9.

Failure type 1 in Table A6 represents the question type that does not have a definite answer and requires the model to exclude all wrong answers before the model can select the correct one which has the meaning of "We don't know." , "None of the above choices", etc.

Failure type 2 in Table A7 represents the question type that requires the model to make further inferences when there is already an almost correct answer. For this type of question, there are usually words like "suggest", "suppose to", "what can we learn from", etc., appearing in the question

Failure type 3 in Table A8 represents the question type that none of the choices are completely correct according to the passage and requires the model to select a relatively correct one. For example, in Table A8, the correct answer should be "sometime tomorrow afternoon after 3", and the provided option A "Sometime tomorrow afternoon." is not completely correct, however, the other two options are even more wrong thus only option A can be selected.

Failure type 4 in Table A9 represents the question type that requires both sufficient

reasoning ability and certain mathematical calculation ability. The PrLMs are trained well with their reasoning ability. However, the mathematical ability is limited.

To further improve my HRCA+ model, the listed four failure types should be considered. For failure types 1 and 3, a single-choice model inside a multiple-choice model structure can be considered to simulate the human strategy of exclusion.

For type 4, we have two solutions including

1) Pre-training the model with a large scale of the mathematical dataset
2) Identify those problems that require mathematical calculation skills and train an individual framework to deal with mathematical problems

Under the natural language understanding level of nowadays mainstream PrLMs, failure type 2 is complicated and hard to tackle from the direction of the model. A compromise method is to build a rule-based approach to identify this type of questions.

Failure type 1

Table A6. Failure question type 1 from DREAM dataset

Passage (dialogue form)
M: Don't go away. You haven't finished your homework.
F: I'll come back later to continue.
Question: What is the woman going to do?
Candidate options:
A. Do her homework.
B. Play with her classmates.
C. We don't know. √
HRCA+ (my model): Do her homework. ×

Failure type 2

Table A7. Failure question type 2 from DREAM dataset

Passage (dialogue form) M: This TV set is getting worse and worse. Now it doesn't work at all. W: Here's an advertisement about a big TV sale. There might be some good bargains in it.
Question: What does the woman suggest? **Candidate options**: A. They go and buy a big TV set. √ B. They have a look at the advertisement. C. They sell their TV set.
HRCA+ (my model): They have a look at the advertisement. ×

Failure type 3

Table A8. Failure question type 3 from DREAM dataset

Passage (dialogue form) W: Excuse me, Professor Davis. Could I talk to you about my paper now? M: I have a class in a few minutes. Why don't you come to my office after 3 tomorrow afternoon?
Question: When will the woman see the professor? **Candidate options**: A. Sometime tomorrow afternoon. √ B. After 3 o'clock. C. After class tomorrow.
HRCA+ (my model): After 3 o'clock. ×

Failure type 4

Table A9. Failure question type 4 from DREAM dataset

Passage (dialogue form) M: Tickets for the art museum are three dollars for adults and the children's tickets are half price. W: I see. I'd like two adults' and three children's tickets, please.
Question: How much will the woman pay for the museum? **Candidate options**: A. $6.00. B. $10.50. √ C. $15.00.
HRCA+ (my model): $15.00. ×

Acknowledgement

First and foremost, I would like to thank my supervisor, Prof. Yamana. I am not his best student, but he is my most respected professor. During my graduate study, I received all kinds of help from my professor. Professor's academic spirit, rigorous style, serious attitude, and kindness have deeply inspired me.

I would also like to thank everyone in the Yamana lab. I learnt various knowledge in many fields, and I am grateful for the variety of discussions I have had with them.

Publications

[1] Yuxiang Zhang and Hayato Yamana. "HRCA+: An Advanced Multi-Choice Machine Reading Comprehension Method", submitted to *the 13th International Conference on Language Resources and Evaluation (LREC 2022)*, European Language Resources Association (ELRA), Paris, France.

[2] Junjie Wang, Yuxiang Zhang, Tetsuya Sakai and Hayato Yamana. "SKYMN at the NTCIR-15 DialEval-1 Task", In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies*, December 8-11, 2020 Tokyo Japan.