Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6059–6068
Marseille, 20-25 June 2022

HRCA+: Advanced Multiple-choice Machine Reading Comprehension Method

Yuxiang Zhang, Hayato Yamana
Waseda University
3 Chome-4-1 Okubo, Shinjuku, Tokyo 169-8555, Japan
joel0495@asagi.waseda.jp, yamana@yama.info.waseda.ac.jp

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 study, we 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. Our proposed method outperforms other state-of-the-art methods. On the Semeval-2018 Task 11 dataset, our proposed method improved accuracy levels from 95.8% to 97.2%, and on the DREAM dataset, it improved accuracy levels from 90.4% to 91.6% without extra training data, from 91.8% to 92.6% with extra training data.

Keywords: natural language processing, machine reading comprehension, multiple-choice question answering

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 study, we tackle the multiple-choice question answering task. 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 study, we 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, our

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, our 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, our 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 our proposed method.

Finally, we extend the operations of our 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 our proposed approach. Subsequently, all the relationships are represented using our proposed 3×3 passage-question-option (PQO) matrix-guided framework called HRCA+.

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

2. Related Work

Conventional MRC methods are based on hand-designed syntax (Riloff and Thelen, 2000) or information extraction approaches (Poon et al., 2010). After 2013, MRC approaches evolved from machine learning-based approaches to deep learning-based approaches.

In Section 2.1, we describe well-known machine learning-based approaches. In Section 2.2, we describe deep learning-based approaches designed to prevent the influence of noise in hand-engineered linguistic features.

2.1. Machine Learning-based Approaches

(Richardson et al., 2013) first proposed a sliding window 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 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 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).

Such approaches require 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 multi-stage 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 87.8%, 88.9% and 90.4% in the DREAM dataset (Sun et al., 2019a), respectively.

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.

3. Proposed Method

The remaining problems associated with multiple-choice question answering (MCQA) for MRC include 1) the problem of incomplete correspondence and 2) the helping-relation problem. 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. 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.

Passage (dialogue form)
M: Excuse me. How can I get to the Prince Street?
W: Take Bus No. 13 and get off at Prince Street stop.
M: Can you tell me where I can buy such kind of shirt?
W: Oh, that's easy. There's a man's shop just around the corner.
M: Thank you.

Question1:
Which bus should the man take to get to Prince Street?
Candidate options:
A. Bus No. 12.
B. Bus No. 30.
C. Bus No. 13. $\surd$

Question2:
What does the man want to buy?
Candidate options:
A. A shirt. $\surd$
B. A bag.
C. A tie.

Figure 1: MCQA sample extracted from the DREAM (Sun et al., 2019a) dataset

To solve the problem of incomplete correspondence, we propose the HRCA method. 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, our proposed HRCA model considers the relationships between passages, questions, and the corresponding candidate options separately to solve the problem of incomplete correspondence. Our proposed 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, we extend the operations of our 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 our 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 Figure 1. MCQA tasks aim to build a model for calculating the probability of correctness for each candidate option.

$$\begin{aligned} F :&(P,\ Q,\ \{O_1, O_2, ..., O_n\}) \\ &\rightarrow \{Pr(O_1), Pr(O_2), ..., Pr(O_n)\} \end{aligned} \tag{1}$$

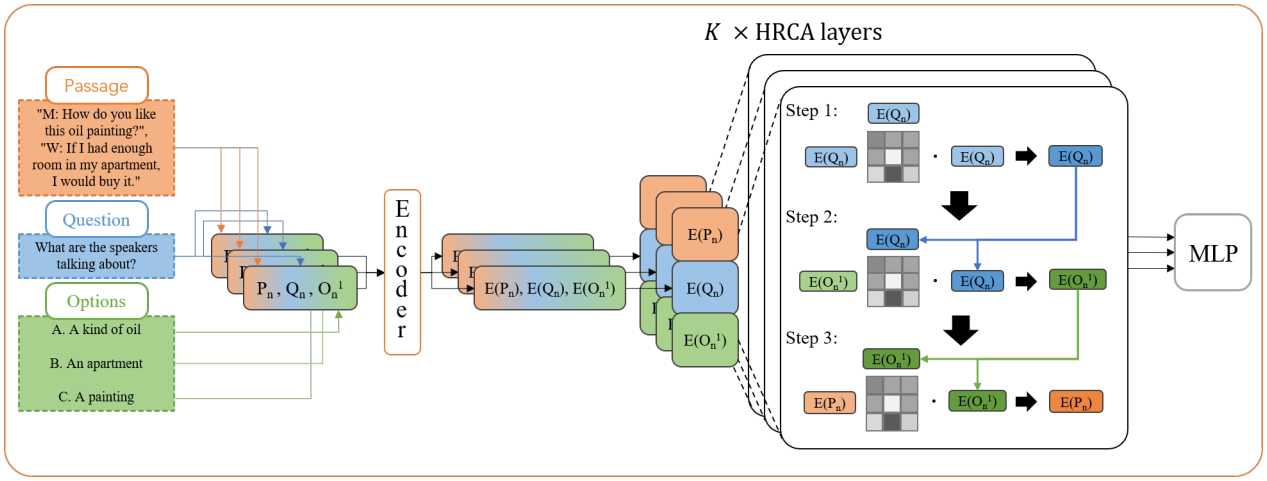

Figure 2: Architecture of HRCA model.

3.2. Model Architecture

Figure 2 illustrates the overview of our proposed model's architecture. Our model simulates the strategies employed by humans in their attempt to achieve high scores in reading comprehension exams. That is, first confirming the question, then checking the candidate options, and finally combining the information learned from the question with the candidate options to read the entire passage. Based on the PrLM encoder, we 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:
We perform multi-head self-attention on the question (We regard the question as a query, key, and value). This step is aimed to allow the model confirm the question again.
Step 2:
We perform multi-head attention on the option and the updated question presented in Step 1 (We regard the option as a query, and we 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:
We perform multi-head attention on the passage and the updated options presented in Step 2 (We regard the passage as a query, and we 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, we 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, we choose the option with the highest probability as the answer.

3.3. Contextualized Encoding

In our proposed model, we adopt a PrLM to generate a global contextualized representation. Let a passage be $P = [p_1, p_2, \ldots, p_i, \ldots, p_k]$, a question be $Q = [q_1, q_2, \ldots, q_i, \ldots, q_\ell]$, and a candidate option be $O = [o_1, o_2, \ldots, o_i, \ldots, o_m]$, 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. We 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 $Encode(\cdot)$, we can obtain the output $E = Encode(P \bigoplus Q \bigoplus O)$. The PrLM's encoder output E has the following form: $[e_1, e_2, ..., e_{k+\ell+m}]$. Note that if we have one passage, one question, and four candidate options, we obtain four different concatenated sequences in total.

3.4. Human Reading Comprehension Attention

As shown in Figure 2, based on the multi-head attention module (Vaswani et al., 2017), we 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. We first use a question as a query, key, and value, to reconfirm the question. Afterwards, we use

the answer as a query and the updated question as the key and value to understand the candidate option. Finally, we 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. We 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: E^P, E^Q, E^O
function MHSA(Q, K, V)
▷ The query, key, and value.
$Attention(Q, K, V) \leftarrow Softmax(\frac{Q(K)^T}{\sqrt{d_k}}) \cdot V$
▷ d_k *represents the dimension of* K.
$head_i \leftarrow Attention(QW_i^Q, KW_i^K, VW_i^V)$
▷ W^x *represents the weight matrix of* x.
return $Concat(head_1, head_2, ..., head_l)W^O$
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)
$E^{\mathbf{Q^U}} \leftarrow$ MHSA(E^Q, E^Q, E^Q)
$E^{\mathbf{O^U}} \leftarrow$ MHSA$(E^O, E^{\mathbf{Q^U}}, E^{\mathbf{Q^U}})$
$E^{\mathbf{P^U}} \leftarrow$ MHSA$(E^P, E^{\mathbf{O^U}}, E^{\mathbf{O^U}})$
return R$(E^{\mathbf{P^U}}, E^{\mathbf{Q^U}}, E^{\mathbf{O^U}})$
end function

3.5. PQO Matrix

As shown in Figure 3, 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 we 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 3, "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.

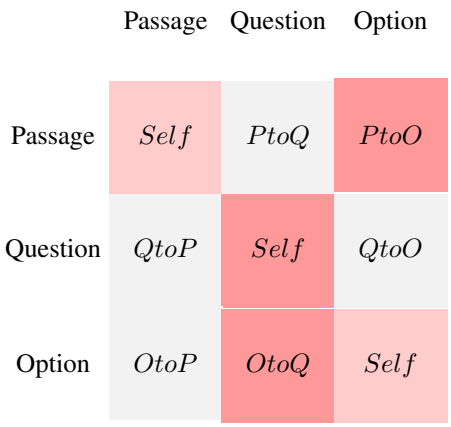

Figure 3: PQO Matrix for calculating the attention

3.5.1. HRCA+: PQO Matrix-guided HRCA

We extend the operations of our 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, we 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 3.

In HRCA+, we update the adjacent cells in the order presented in Figure 4. Because updating the attention scores of the target relation pairs from their previous relation pairs relies on the adjacent connection between both relation pairs, we 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.

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. We investigate and compare the reduction functions mentioned above, and the results are presented in Section 4.4.

Additionally, we must consider using the combined outputs to generate the probability distribution for each

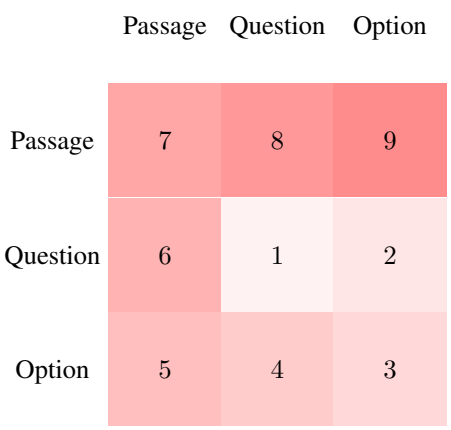

Figure 4: Updating order in PQO Matrix

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. Our 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 our proposed multi-layer perceptron (MLP) was formulated.

4. Experiments

In this section, we evaluate the performance of our proposed method on multiple-choice reading comprehension examination datasets.

4.1. Datasets

We 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 because 85% of the questions require reasoning beyond a single sentence, and 34% of the questions involve commonsense knowledge.

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.

4.2. Experimental Settings

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

Baselines 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, we also implement DUMA (Zhu et al., 2020), and we 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).

Data pre-processing For the DREAM dataset, we apply data pre-processing to maintain the consistency of gender representations, i.e., man and woman, between passages and questions. 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. Therefore, the symbols W and M are replaced with "woman" and "man" during data pre-processing.

4.3. Multi-task learning

Existing large-scale PrLMs are more oriented towards acquiring the corresponding syntactic and semantic features of the text in the pre-training task. Although our proposed model can maximize the inference information and interrelationships on the downstream task, the performance may still be limited by the size of the downstream task. Therefore, we designed multi-task training, aiming to increase the inference information and inference capability of the PrLM.

We do not want to increase the cost of training on downstream tasks, and at the same time want the PrLM to have a more general reasoning ability. Therefore, we chose four large-scale general natural language inference datasets including SNLI(Bowman et al., 2015), Multi-NLI(Williams et al., 2018), NLI-version of FEVER(Nie et al., 2019) and ANLI(Nie et al., 2020) for multi-task training. We first fine-tuned the PrLM on these four datasets and then used the fine-tuned PrLM for subsequent downstream tasks.

4.4. Results

In this study we adopted accuracy as the evaluation metric. The experimental results are listed in Tables 1–5.

We first evaluated the performance of our proposed model using the DREAM dataset. Table 1 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 2 shows the accuracy difference 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 3 shows the results of ablation experiments for HRCA model on DREAM dataset. To verify the validity of all the three steps in our HRCA model, we tested several combinations. The results show that all the three steps are necessary.

Table 4 shows the publication results on the DREAM dataset. Our proposed model achieves the highest accuracy of 92.6% among all models that use extra training data, while it also achieves the highest accuracy of 91.6% among all models that do not use extra training data. Owing to challenges associated with computational resources, we only use a batch size of two, whereas several previous methods use a batch size of eight. To achieve enhanced intuitive performance for comparison purposes, we implement DUMA, which is the current state-of-the-art method, and we use similar parameters to train our implementation of the DUMA and vanilla ALBERT-xxlarge methods. For each result, we train our proposed model five times, and we calculate the average value. Note that there is a gap of 1.3% in the performance of our implementation and the initial DUMA method. To verify the accuracy of our implementation, we test the performance of our implemented DUMA through the ALBERT-base using parameters that are similar to the initial DUMA method.

We also test the performance of our proposed model on the SemEval-2018 Task 11 dataset presented in Table 5. 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 5, we adopted the PrLM results showing improved accuracy from 84.1%, i.e., from the results obtained using the vanilla GPT and RoBERTa-Large methods. By applying strategies, such as back and forth reading, highlighting, and self-assessment (Sun et al., 2019b), and by applying model ensembles, multi-task learning, the accuracy level was improved to 95.8%. Even in comparison to the performance of the best model relying on extra training data, i.e., the MMM (Jin et al., 2020) model, our proposed model is able to achieves the highest accuracy levels without extra training data.

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

Table 1: Performance in accuracy (%) with various MHSA layers on DREAM dataset based on ALBERT-base. (Showing the accuracy as development dataset / test 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

Table 2: Performance in accuracy (%) with different reduce functions on DREAM dataset.

Model	Steps	Test
ALBERT-base	-	64.4
+ HRCA	step1	66.9
+ HRCA	step1 & step2	67.8
+ HRCA	step1 & step3	67.2
+ HRCA	step2 & step3	67.1
+ HRCA	step1 & step2 & step3	68.8

Table 3: Ablation experiments for HRCA on DREAM dataset.

Model	Dev	Test
Random	32.8†	33.4†
GBDT++ (Sun et al., 2019a)	53.3†	52.8†
FTLM++ (Radford et al., 2018)	57.6†	57.4†
Ensemble of 3 FTLM++	58.1†	58.2†
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 (our implementation)	-	67.5
ALBERT-base + DUMA	-	67.6♢
ALBERT-base + HRCA	-	68.8
ALBERT-base + HRCA+	-	69.8
ALBERT-xxlarge	88.2	88.0
ALBERT-xxlarge + DUMA (our implementation)	89.5	89.1
ALBERT-xxlarge + HRCA+	**90.8**	**91.6**
ALBERT-xxlarge + HRCA+ + Multi-Task Learning	**92.1**	**92.6**
Human Performance (Sun et al., 2019a)	93.9†	95.5†
Ceiling Performance (Sun et al., 2019a)	98.7†	98.6†

Table 4: 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 (2×)	88.6♣
BERT-large	88.7♢
RoBERTa-large	94.0♢
ALBERT-xxlarge	95.4
GPT+Strategies (Sun et al., 2019b)	88.8♣
GPT+Strategies (2×) (Sun et al., 2019b)	89.5♣
RoBERTa-large + MMM (Jin et al., 2020)	95.8♢*
ALBERT-xxlarge + HRCA+	**96.6**
ALBERT-xxlarge + HRCA+ + Multi-Task Learning	**97.2**
Human Performance (Ostermann et al., 2018)	98.0†

Table 5: 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 study, we 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, our proposed approach achieves a higher score when tackling multiple-choice comprehension tasks. We 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 our proposed method achieves the highest accuracy among other existing state-of-the-art methods. In our future studies, we shall integrate the applications of our proposed method to tasks in other fields, such as the extraction of relationships between passages and given argument pairs.

6. References

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September. Association for Computational Linguistics.

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.

Nie, Y., Chen, H., and Bansal, M. (2019). Combining fact extraction and verification with neural semantic matching networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., and Kiela, D. (2020). Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online, July. Association for Computational Linguistics.

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*.

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 *Proceedings 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.

Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. 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.