applied sciences

MDPI

Article

A Multiple-Choice Machine Reading Comprehension Model with Multi-Granularity Semantic Reasoning

Yu Dai [1,*], Yufan Fu [1] and Lei Yang [2]

[1] Software College, Northeastern University, Shenyang 110169, China; 1971110@stu.neu.edu.cn
[2] Computer Science and Engineering, Northeastern University, Shenyang 110169, China; yanglei@mail.neu.edu.cn
* Correspondence: daiy@swc.neu.edu.cn

Abstract: To address the problem of poor semantic reasoning of models in multiple-choice Chinese machine reading comprehension (MRC), this paper proposes an MRC model incorporating multi-granularity semantic reasoning. In this work, we firstly encode articles, questions and candidates to extract global reasoning information; secondly, we use multiple convolution kernels of different sizes to convolve and maximize pooling of the BERT-encoded articles, questions and candidates to extract local semantic reasoning information of different granularities; we then fuse the global information with the local multi-granularity information and use it to make an answer selection. The proposed model can combine the learned multi-granularity semantic information for reasoning, solving the problem of poor semantic reasoning ability of the model, and thus can improve the reasoning ability of machine reading comprehension. The experiments show that the proposed model achieves better performance on the C^3 dataset than the benchmark model in semantic reasoning, which verifies the effectiveness of the proposed model in semantic reasoning.

Keywords: natural language processing; machine reading comprehension; semantic reasoning; pre-training model

Citation: Dai, Y.; Fu, Y.; Yang, L. A Multiple-Choice Machine Reading Comprehension Model with Multi-Granularity Semantic Reasoning. *Appl. Sci.* **2021**, *11*, 7945. https://doi.org/10.3390/app11177945

Academic Editor: Andrea Prati

Received: 23 July 2021
Accepted: 25 August 2021
Published: 27 August 2021

1. Introduction

How to make computers understand human language is the main goal of the field of Natural Language Processing (NLP) and has been a long-standing challenge for artificial intelligence research. Machine Reading Comprehension (MRC) tasks are similar to human reading comprehension tests in which the computer needs to answer questions based on the content of a given text [1]. In contrast to traditional NLP, MRC requires techniques that involve multiple aspects of lexical, grammatical and syntactic meanings and also requires a combination of feature representations analysis of the text context and semantic reasoning techniques, making it a very challenging NLP task.

In MRC tasks, deep learning models are often used to help machines learn and understand contextual content so that they can answer the corresponding questions correctly. If machines can perform reading comprehension tasks similarly to humans, and have reading comprehension capabilities similar to or better than those of the human brain, then they can be of great value in replacing traditional human reading comprehension tasks.

Among the several types of tasks in MRC (cloze test, span extraction and multiple choice, etc.), this paper focuses on multiple-choice style tasks. A multiple-choice MRC task differs from a span extraction task in that it requires not only the text and questions but also a set of candidate answers from which the machine needs to find the correct answer, taking into account the semantic information of the text [2]. In contrast to the cloze test MRC task where the answers are fixed words and phrases, the answers to the multiple-choice MRC task are artificially generated sentences that are manually rewritten with complete logic based on the content of the text. Typical English datasets of this type include MCTest [3], RACE [4] and MCScript [5], and a representative Chinese dataset is

C^3 [6]. A sample of data selected from the RACE dataset for a multiple-choice MRC task is shown in Figure 1. The RACE dataset is a representative benchmark dataset for multiple-choice reading comprehension, which is constructed using an English test bank for junior and senior high schools. The candidates in candidate answer set A in Figure 1 never appear in the passage, and the machine needs to fully understand the semantic information of the text context. The machine needs to fully understand the semantic information of the context and select semantically similar candidates from the candidate set as the answer.

Article:

Homework can put you in a bad mood, and that might actually be a good thing. Researchers from the University of Plymouth in England doubted whether mood might affect the way kids learn. To find out the answer, they did two experiments with children.

The first experiments tested 30 kids. Some shapes were hidden inside a different, larger picture. The kids had to find the small shapes while sitting in a room with either cheerful or sad music playing in the background. To test their mood, the scientists asked the kids to point to one of the smiley faces while the others pointed to the unhappy ones. The researchers found that sad kids took at least a second less to find the small shapes. They also found an average of three or four more shapes.

In the second experiments, 61 children watched one of two scenes from a film. One scene was happy and the other was sad. Just like in the first experiment, kids who saw the sad scene acted better compared to the others.

The researchers guessed that feeling down makes people more likely to focus on a problem or difficult situation. Not all scientists agree with them, however. Other studies argued that maybe, that cheerful music in the first experiment distracted kids from finding shapes.

While scientists work on finding out the answers, it still might be wise to choose when to do your tasks according to your mood. After eating a delicious ice cream, for example, write an essay.

Question 1. Researchers did experiments on kids in order to find out __ .

A. how they really feel when they are learning
B. Whether mood affects their learning ability
C. what methods are easy for kids to learn
D. the relationship between sadness and happiness

Answer: B

Figure 1. Example of multiple-choice MRC task data.

While the answers to the cloze test and span extraction tasks must come from the context of the given passage, and the answers to the multiple-choice tasks are not necessarily sequences in the text. Those answers are manually rewritten and summarized based on the content of the text, and some answers even need to be inferred together with external knowledge.Questions and answer candidates for multiple-choice Chinese MRC are written by humans, which means the content is more flexible and difficult to find out the correct answer by simple matching.

Through our analysis of the multiple-choice Chinese MRC task, we found these three factors to make the task challenging: (1) few training data and lack of external knowledge severely limit the accuracy of the model; and (2) the answer selection for many questions requires deep semantic interaction to find out the corresponding answer. Repeated semantic interactions between articles, questions and candidates are crucial, but the learning of them is inadequate. (3) MRC requires a high level of semantic reasoning, and answer selection must not only take into account the local information of the related passages but also consider the global information of the article.

Based on the characteristics and challenges mentioned above, we propose a deep neural network (DNN) based model. The main contributions of this paper can be summarized as follows:

1. We design a multi-granularity semantic information extractor and apply it to our proposed MRC model to enhance the comprehension of local sematic meanings, which have been proved beneficial to the model performance in our experiments.
2. We investigate the semantic interactional reasoning aspect and leverage attention mechanism to extract semantic perceptual information between articles, questions and candidates. By fusing multiple semantic interaction information, we have further improved the performance of our multiple-choice MRC model.
3. We model the learning process of global semantic information and local semantic information, respectively, and jointly construct a deep global and local semantic multiple-choice MRC model to achieve better deep semantic learning and reasoning for articles, questions and answers in multiple-choice MRC tasks.

In the first section of this paper, we introduce the background of our research and the up to date researching progress in related fields. By analyzing existing works, we illustrate the significance of our proposed idea in Section 2. Then we define the task in formulaic language and describe the proposed model structure with multi-granularity semantic reasoning in Section 3. Section 4 describes the dataset, evaluation metrics and settings used in our experiments. Our experimental results are presented and comprehensively analyzed in detail in Section 5, and we summarize our work in this paper in the last section.

2. Related Work

For nearly half a century, research on MRC has gone through three stages of development: the early era of rule-based MRC, the era of machine-learning-based MRC and the era of neural networks that use deep learning to build MRC models.

2.1. Rule-Based MRC

When the MRC task was firstly proposed in the 1970s, most of the early approaches were limited by hand-coded scripts and rules, making them difficult to apply widely in real-world scenarios. In the late 20th century, Hirschman et al. [7] proposed an MRC dataset for development and testing that contained 120 reading materials for primary school students and a number of short question-answer pairs, such as who, where, when, why and what, consisting of questions and answer pairs. They did not require the model to give an exact answer, but only needed it to find the sentence where the answer is located in the article. They also proposed the DEEP READ model for this dataset (which primarily uses a rule-based bag-of-words model). Charniak [8] et al. fused a rule-based bag-of-words model with a lexical and semantic similarity-based approach, ultimately achieving an accuracy rate of 30% to 40% in the reading comprehension task of searching answer location.

2.2. Machine-Learning-Based MRC

In 2013, Richardson et al. [9] proposed the MCTest dataset on which the weighted distances between questions and answers were calculated to predict the correct answer. The presentation of this dataset has rapidly advanced the development of machine learning models [10–12]. In 2015, Wang et al. proposed a max-margin learning framework based on a heuristic sliding window approach, which improved the model accuracy from 63% to around 70% on the MCTest dataset by converting each question-answer pair into a textual implication system for the corresponding utterance. Similar to Wang's model, most of the models at that time were based on a simple max-margin learning framework with some rich linguistic features (such as syntactic dependencies, denotational disambiguation, semantics, word embeddings, etc.) to fit into passages, questions and answers. Compared to earlier rule-based MRC approaches, machine learning-based MRC models have shown good performance. However, we can find that the existing machine learning models still have significant limitations in terms of performance improvement, and there are two main reasons affecting the performance improvement: (1) the machine learning models mainly rely on existing language tools for feature extraction, such as dependency parsers and semantic role annotator, but these language tools are trained from data in a single domain,

and their generalization capability is relatively weak; therefore, for MCTest data, there is a lot of noise in the obtained features; (2) the size of the dataset is too small, thus it cannot support the adequate training of machine learning models.

2.3. Deep-Learning-Based MRC

In 2015, Hermann et al. [13] presented a large-scale fill-in-the-blank MRC dataset CNN/Daily Mail for the first time (about 1.26 million training data), and also proposed the ATTENTIVE READER neural network model for this dataset, which is based on the attention mechanism model and compared to the traditional ATTENTIVE READER neural network model, which is based on an attention mechanism and achieves a 12.9% improvement in accuracy compared to traditional NLP models. This marks the beginning of the DNN era for MRC. In 2016, Rajpurkar et al. [14] proposed SQuAD, an English dataset for extractive answer-based MRC tasks, which is the first canonical dataset containing large-scale natural language question-answer pairs in the MRC research community. Relying on the SQuAD dataset, Wang et al. [15] proposed the Match-LSTM and Answer Pointer models, which use a bidirectional LSTM model to encode questions and articles, and a one-way attention mechanism to perform semantic matching between article and question. Yu et al. [16] proposed QAnet, which uses multi-layer convolution and a self-attentiveness in the encoding module mechanism to integrate local and global interactions of articles and questions to improve the performance of the model. Basafa el al. [17] use Longformer [18], a long document transformer, to learn the abstract meaning of the context. It has been proved that deep learning-based MRC models have stronger text semantic representation ability and answer reasoning ability in English compared with traditional machine learning models.

A set of large-scale datasets for different Chinese MRC tasks and datasets have also been proposed, such as ReCO [19] for Chinese reading comprehension, ChID [20] for the cloze-style task on Chinese idioms, CMRC2018 [21] for the extractive task, and C^3 [6] for the multiple-choice task. The proposal of a large number of high-quality MRC datasets has driven the development of deep neural MRC models. Knowledge-enhanced pretrained models, such as ERNIE 3.0 [22] and Kepler [23], are able to integrate factual knowledge into PLMs to achieve better performance. Instead of striving for better objective evaluation, Cui et al. [24] try to improve the explainability for MRC tasks on multiple-choice datasets. To the best of our knowledge, few studies have focused on semantic reasoning on different levels of granularity to address the Chinese MRC challenge. In this paper, we build on previous work to construct models that have greater comprehension and generalization capabilities for natural language in the field of MRC.

3. Method

3.1. Task Definition

The multiple-choice MRC task requires the model to select the appropriate answer from the candidate answers based on the given context, and the answers are not only limited to words or entities present in the context, which makes the answer format more flexible. By giving the machine a Document (denoted as D) and a Question (denoted as Q), which corresponds to a set of options (denoted as O), the goal of the model is to be able to infer the correct answer from the set of candidate answers.

Then we can define the task as follows: the relationship between document $D = d_1, d_2, \ldots, d_m$ (m denotes the number of words in the article), questions $Q = q_1, q_2, \ldots, q_n$ (n denotes the number of words in the question) and answers $O_i = o_1, o_2, \ldots, o_k$ (k denotes the number of words in the i-th candidate answer) is shown in Equation (1).

$$F(D, Q, O) = O_i \tag{1}$$

In this task, our model should find the best answer from all k candidates by learning from the aforementioned relationship.

3.2. Model

In this section, we construct a multiple-choice MRC model incorporating multi-granularity semantic reasoning. Firstly, the articles, questions and candidates are input to the BERT model for learning, and the semantic information in the articles, questions and candidates is learned through the multi-layer transformer structure in the BERT model; secondly, the feature vectors output from the final hidden layer of BERT become the input to a convolutional neural network (CNN) with multiple windows of different sizes, and the convolutional kernels with different sizes of windows are used to learn different lengths of semantic paths. Then, the output of the multi-granularity semantic reasoning information is spliced with the global feature information output from the CLS position in BERT to obtain the information of global reasoning and local multi-granularity semantic reasoning; finally, the spliced information is used to complete the answer selection by using the fully connected layer and softmax function, and the model structure is shown in Figure 2.

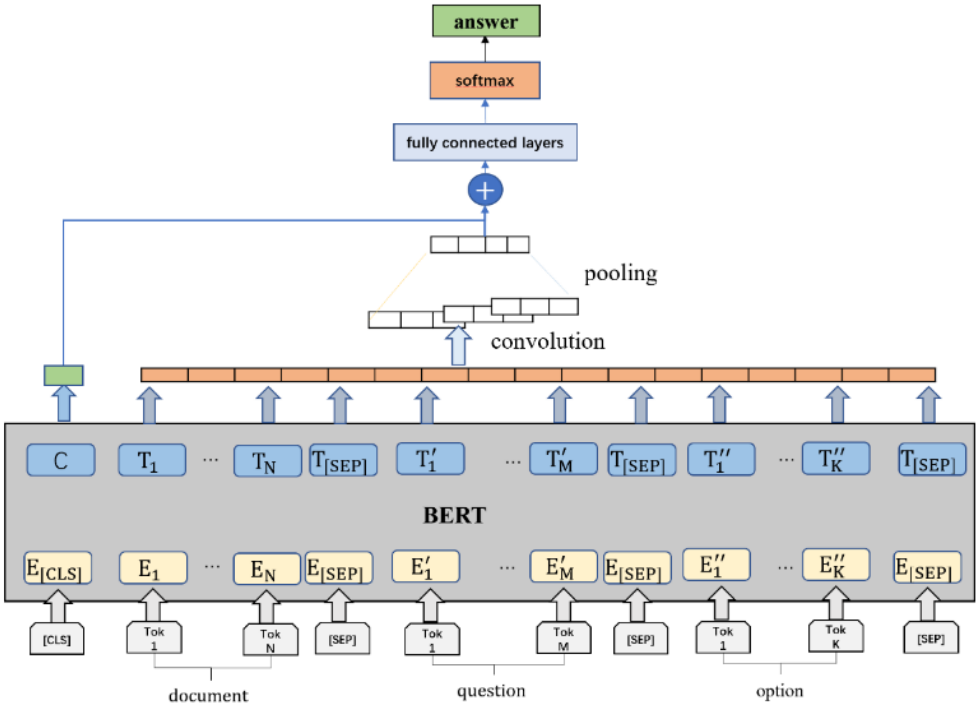

Figure 2. Model diagram of multiple-choice MRC with multi-granularity semantic reasoning.

The model structure is divided into six main layers from bottom to top: input layer, embedding layer, encoding layer, multi-granularity semantic reasoning layer, information fusion layer and answer prediction layer.

3.2.1. Input Layer

This layer mainly represents the inputs of the documents, questions and candidates. According to the input characteristics of the BERT model, the sequence of inputs of documents, questions and candidates is represented as shown in Equation (2).

$$S = [CLS]D[SEP]Q[SEP]O_i[SEP] \tag{2}$$

where D denotes the set of token sequences of documents, Q denotes the set of token sequences of questions, and O_i denotes the set of token sequences of the i-th candidate.

3.2.2. Embedding Layer

The embeddings we use in our model are divided into three layers: token Embeddings, Segment Embeddings and position embeddings, Token Embeddings is the conversion of S in the input layer into a vector of fixed dimensions; Segment Embeddings is used to

distinguish the front and back parts of the sentence pairs; Position Embeddings is to encode the position information in the input layer S. The formulae are shown in Equations (3)–(6).

$$w_S = TokenEmbeddings(S) \tag{3}$$

$$seg_S = SegmentEmbeddings(S) \tag{4}$$

$$p_S = PositionEmbeddings(S) \tag{5}$$

$$Input_S = w_S + seg_S + p_S \tag{6}$$

where $Input_S$ represents the overall output after the BERT embedding layer.

3.2.3. Encoding Layer

This layer is used to encode the input embedded layer sequence through the multi-layer transformer in BERT. There is a dependency relationship between the multi-layer transformer and the output of the previous layer transformer is the input of the current layer transformer, which is calculated as shown in Equations (7) and (8).

$$h_1 = Transformer(Input_S) \tag{7}$$

$$h_i = Transformer(h_{i-1}), i \in [1, N] \tag{8}$$

where h_i denotes the output of the transformer's i-th layer and N is the number of layers of the transformer in BERT.

3.2.4. Multi Granularity Semantic Reasoning Layer

This layer is mainly used to perform multi-granularity semantic reasoning on the vectors encoded by BERT using CNN. This process simulates the process of human reading comprehension by repeatedly focusing on the important semantic information before and after reasoning, and finally completing the answer selection. The CNN mainly contains a convolutional layer and a pooling layer, and the convolutional kernel windows used in our mode are 2, 3 and 4, and the pooling method uses the maximum pooling method. The calculation method is shown in Equations (9)–(11).

$$T_2 = con\_and\_maxpooling(h_i)_2 \tag{9}$$

$$T_3 = con\_and\_maxpooling(h_i)_3 \tag{10}$$

$$T_4 = con\_and\_maxpooling(h_i)_4 \tag{11}$$

where T_2, T_3 and T_4 represent the results of convolution kernels for 2, 3 and 4 convolutions with maximum pooling, respectively.

3.2.5. Information Fusion Layer

This layer fuses the output of multi-granularity layer with the feature vector acquired by the CLS embeddings from BERT. The CLS embedding vectors represent the global feature information obtained through the BERT model, while the output results of the multi-granularity layer represent the feature information obtained by reasoning at multiple local granularities. By fusing the two pieces of information, the model is able to learn more comprehensive information, which is more conducive to the subsequent answer prediction. The calculation method is shown in Equation (12).

$$x = C \bigoplus T_2 \bigoplus T_3 \bigoplus T_4 \tag{12}$$

where C denotes the feature vector output from the CLS position, and x is the result of information fusion operation.

3.2.6. Answer Prediction Layer

This layer focuses on the prediction of answers for multiple-choice MRC, and after the fully connected layer, the answer prediction is performed by the softmax function. The final output is calculated as shown in Equation (13).

$$\hat{y} = softmax(W_x x + b) \tag{13}$$

where W_x denotes the weight and b denotes the bias.

3.3. Optimization

Our model uses the cross-entropy loss function as the loss function, which is calculated by Equation (14) as below:

$$\mathcal{L} = -\frac{1}{N}\sum_{c=1}^{M} y_{ic} \log(\hat{y}_{ic}) \tag{14}$$

where N is the total number of inputs, M is the number of categories. y_ic is the expected output, which is 1 when the categories are the same with the actual output and 0 when they are different, and $\hat{y}_ic$ is the probability of predicting sample i to category c.

4. Experiment

4.1. Dataset

C_3: We use the multiple-choice Chinese MRC dataset C_3 and perform a statistical analysis about it. In 2019, researchers at Tencent AI Lab presented the first free-form multiple-choice Chinese MRC dataset, which contains 13,369 documents (containing both formal and informal forms) collected from questions in the general domain of the Chinese Proficiency Test and 19,577 multiple-choice MRC questions associated with these documents.

In order to evaluate the generalization ability of different domain models, the dataset contains two document types, conversational form documents and non-dialogical documents with mixed topics (e.g., stories, news reports, monologues, or advertisements). MRC tasks can be classified into two categories based on the different document types: C^3-Dialogue(C^3_D) and C^3-Mixed(C^3_M), and within these two task types, each document corresponds to a diversity of question types, such as complete fill-in-the-blank questions formed by removing spans or sentences from the text, closed-form questions that can be answered with minimal answers (e.g., yes or no), or free-form questions that reason from multiple sentences of the text. With 86.8% of the questions in this dataset requiring a combination of internal and external knowledge of the document (general world knowledge) to better understand the given text, we can say that most questions in this dataset require rich external knowledge to assist the machine in answering the question.

There is a significant difference between C^3_D and C^3_M in that most of the documents in C^3_M are formal written texts, while there is a lot of spoken language in the dialogue documents in C^3_D, so there is a larger vocabulary in C^3_M compared to the dialogue documents. The average document length in C^3_M is 180.2, and the vocabulary size is 4120, while the average document length in C^3_D is 76.3, and the vocabulary size is 2922. Due to the longer document length in C^3_M, it may be better for assessing MRC for verbose texts.

4.2. Metrics

Multiple-choice MRC tasks generally use accuracy to measure the performance of the model. The accuracy rate indicates the number of samples that made the correct choice as a percentage of the total number of samples. A higher value of accuracy means that the model answered more questions correctly. Accuracy is calculated by Equation (15).

$$accuracy = -\frac{1}{N}\sum_{c=1}^{M} I(y'_i = y_i) \tag{15}$$

where I is a function to determine whether the predicted value y_i' and the actual value y_i are equal. The output is 1 or 0 for equal or not equal, respectively.

4.3. Experimental Settings

In this paper, our model is built using the pytorch deep learning framework. After many experiments, we adjust the parameters in our model to what is shown in Table 1.

Table 1. Model parameter settings.

Parm Type	Parm Value
Batch size	32
Learning rate	2×10^{-5}
Epoch	10
Max Length	512
Dropout	0.1
Gradient Accumulation Steps	4
Optimization functions	Adam

In order to prevent model overfitting and excessive training time, the validation set is tested every round during the training phase of the model. If there is no further improvement in accuracy in two consecutive rounds on the validation set, the training process of the model is stopped, and the model with the highest accuracy round is used as the final model.

5. Results and Analysis

5.1. Experimental Results

Most of the answers to the questions in the C^3 dataset used in this paper require a combination of semantic reasoning, some of which need reasoning on a single sentence, and others require multiple sentences to be considered together to find the appropriate answer. Therefore, this dataset can better verify the effectiveness of multi-granularity reasoning.

To illustrate the effectiveness of the proposed model, the test results are compared with the test results of several models, and the detailed experimental comparison results are shown in Table 2.

Table 2. Comparison table of experimental results [a].

Model	C_M^3-*Test*	C_D^3-*Test*	C^3-*Test*
Random	27.8	26.6	27.2
Distance-Based Sliding Window	45.8	40.4	43.1
Co-Matching	48.2	51.4	49.8
ERNIE	63.7	64.6	64.1
BERT	64.6	64.4	64.5
Our model	65.234	65.238	65.236

[a] Results are measured by accuracy, using percentages (%).

From the table above, we can see that our proposed model achieve improvements of 0.634%, 0.838%, and 0.736% over the benchmark BERT model on C_M^3-*test*, C_D^3-*test*, and C^3-*test*, respectively, which indicate that the introduction of the multi-granularity module has a significant improvement over the benchmark BERT model. The results of our experiment also suggest that the fused multi-granularity semantic reasoning method we propose can improve the reasoning ability of the model. By convolution and maximum pooling on the convolution window size of 2, 3 and 4, our model can extract the local multi-granularity feature information and then combine it with the global granularity feature information, which can achieve the effect of reasoning from multi-granularity.

We have also counted the testing results of each round on the validation set to verify the relationship between the number of training rounds and model performance. The results are shown in Figures 3 and 4.

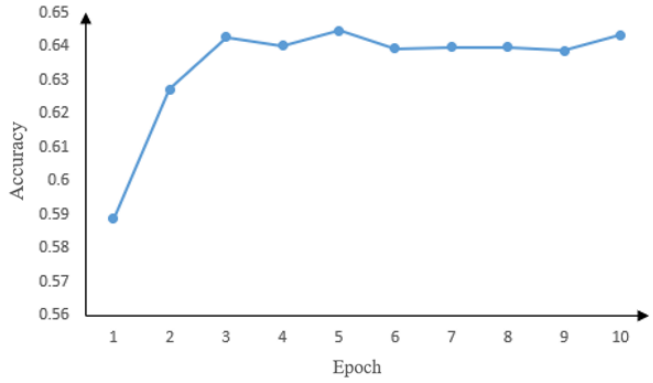

Figure 3. Accuracy changes graph with epoch.

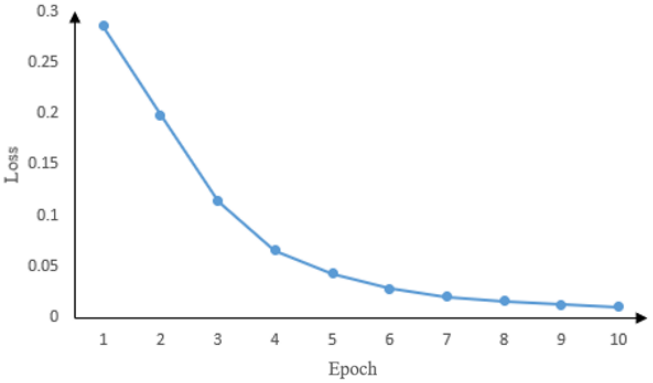

Figure 4. Loss changes graph with epoch.

From Figures 3 and 4, we can see that the model converges faster in the first three rounds of training, and the model convergence becomes steady when the training process reaches the third round. The model achieves the best performance at the fifth round, which indicates that the proposed model can achieve a better convergence effect and can complete the training of the model with fewer training rounds.

5.2. Ablation Studies

In order to verify the effectiveness of global granularity reasoning and local multi-granularity reasoning in our model, we design a local multi-granularity reasoning model without fusing the global granularity feature information of the CLS position and using only the BERT model with convolutional pooling. The experimental results are shown in Table 3.

Table 3. Results of the ablation studies [a].

Model	C^3_M-*Test*	C^3_D-*Test*	C^3-*Test*
BERT	64.6	64.4	64.5
Local Multi-Granularity Model	63.08	63.439	63.25
Local + Global-Granularity Model	65.234	65.238	65.236

[a] Results are measured by accuracy, using percentages(%).

From Table 3, it is clear that the model considering only local multi-granularity reasoning degrades in performance of by 1.52%, 0.961%, and 1.25% over the baseline BERT model

on C_M^3-*test*, C_D^3-*test*, and C^3-*test*, respectively. This result indicates that only considering local granularity reasoning without the global scope information will cause a certain decrease in the performance of the model. This is consistent with the fact that if humans reason only in terms of the partial information when answering reading comprehension questions, it will lead to inaccurate outcomes; thus, further demonstrating the correctness of our multiple-choice MRC model that incorporates multi-granularity semantic reasoning not only considering local multi-granularity reasoning but also global granularity reasoning.

5.3. Analysis

To further demonstrate the performance improvement of the proposed model, we randomly selected some questions from the Chinese multiple-choice MRC dataset that require different types of reasoning to give answers. Then we conducted experiments on the benchmark BERT model and our proposed model, and the results are shown below in Table 4.

Table 4. Comparison table of experimental results [a].

Model	**BERT**		**Our Model**	
Dataset	C_D^3-*Test*	C_M^3-*Test*	C_D^3-*Test*	C_M^3-*Test*
Semantic Reasoning	81.5	81.8	88.9	90.91
Implicative Reasoning	62.5	0	75.0	0
Causal Reasoning	55.6	57.1	66.7	57.1

[a] Results are measured by accuracy, using percentages (%).

From Table 4, we can see that our fused multi-granularity MRC model has considerable improvement in semantic reasoning, implicative reasoning, and causal reasoning, respectively, compared with the BERT model. All the results indicate that our proposed method has higher rationality and feasibility.

6. Conclusions and Future Work

This paper focuses on a multiple-choice Chinese MRC model that incorporates multi-granularity semantic reasoning. By designing the fusion method of global features and local semantic reasoning outputs, we effectively improve the performance of the model, which proves the effectiveness of the proposed method. This research has proved that studying the patterns of human reading, thinking and learning is an essential way to conduct research in the field of deep learning. A well-designed local-global semantic information interaction scheme can provide remarkable enhancement in model perception capabilities. Our study calls for the research community to go deeper into the utility of semantic meanings and explores further how to find out a better way to build up a stronger MRC model.

In the future, we plan to focus on two main aspects. Due to the deficiency discovered in other types of reasoning experiments, the first aspect is to improve the model's ability to handle the referential problems by adapting solutions used in Coreference Resolution tasks. Secondly, as the proposed model is less capable of processing excessively long context with various external knowledge, we will further leverage the latest promising knowledge enhanced approaches to overcome this shortcoming and extend the proposed model to deal with more challenging settings.

Author Contributions: Conceptualization, Y.D. and L.Y.; investigation, F.Y.; methodology Y.D. and Y.F.; supervision, Y.D. and Y.L.; visualization, Y.F. and L.Y.; writing—original draft preparation, Y.D., Y.F. and L.Y.; writing—review and editing, Y.D., Y.F. and L.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by the Chinese national key research plan: 2019YFB1405402.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The model is trained on the multiple-choice Chinese MRC dataset C^3 [25], which contains 13,369 documents (containing both formal and informal forms) collected from questions in the general domain of the Chinese Proficiency Test and 19,577 multiple-choice MRC questions associated with these documents. This dataset gives paragraphs of varying length and a number of questions, along with corresponding English translations. The average document length in C^3_M is 180.2, and the vocabulary size is 4120, while the average document length in C^3_D is 76.3, and the vocabulary size is 2922. Both types are used to assess the ability of MRC for verbose texts.

Acknowledgments: This work is supported by Chinese national key research plan: 2019YFB1405402.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gu, Y.; Gui, X.; Li, D.; Shen, Y.; Liao, D. A Review of Neural Network-based Machine Reading Comprehension. *J. Softw.* **2020**, *20*, 2095–2126.
2. Sheng, Y.; Lan, M. A Multiple-Choice Machine Reading Comprehension Model using External Knowledge Assistance and Multi-Step Reasoning. *Comput. Syst. Appl.* **2020**, *29*, 5–13.
3. Li, C. A Study of Machine Reading Comprehension Based on Semantic Reasoning and Representation. Ph.D. Thesis, East China Normal University, Shanghai, China, 2018.
4. Lai, G.; Xie, Q.; Liu, H.; Yang, Y.; Hovy, E. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017.
5. Ostermann, S.; Modi, A.; Roth, M.; Thater, S.; Pinkal, M. MCScript: A Novel Dataset for Assessing Machine Comprehension Using Script Knowledge. *arXiv* **2018**, arXiv:1803.05223.
6. Sun, K.; Yu, D.; Yu, D.; Cardie, C. Investigating Prior Knowledge for Challenging Chinese Machine Reading Comprehension. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 141–155, doi:10.1162/tacl_a_00305.
7. Hirschman, L.; Light, M.B.E. Deep read: A reading comprehension system. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, College Park, MD, USA, 20–26 June 1999; pp. 325–332.
8. Charniak, E.; Altun, Y.B.R.S. Reading Comprehension Programs in a Statistical-Language-Processing Class. In Proceedings of the 2000 ANLP/NAACL Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Sytems, Stroudsburg, PA, USA, 4 May 2000; Volume 6, pp. 1–5.
9. Richardson, M.; Burges, C.; Renshaw, E. 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, Seattle, WA, USA, 18–21 October 2013.
10. Narasimhan, K.; Barzilay, R. Machine Comprehension with Discourse Relations. 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), Beijing, China, 26–31 July 2015.
11. Sachan, M.; Dubey, K.; Xing, E.; Richardson, M. Learning Answer-Entailing Structures for Machine Comprehension. In Proceedings of the Meeting of the Association for Computational Linguistics & the International Joint Conference on Natural Language Processing, Beijing, China, July 26–31 2015.
12. Hai, W.; Bansal, M.; Gimpel, K.; Mcallester, D. 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), Beijing, China, 26–31 July 2015.
13. Hermann, K.M.; Kovcisk, T.; Grefenstette, E.; Espeholt, L.; Blunsom, P. Teaching Machines to Read and Comprehend. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1693–1701.
14. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016.
15. Wang, S.; Jiang, J. Machine Comprehension Using Match-LSTM and Answer Pointer. *arXiv* **2016**, arXiv:1608.07905.
16. Yu, A.W.; Dohan, D.; Luong, M.T.; Zhao, R.; Chen, K.; Norouzi, M.; Le, Q.V. QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. *arXiv* **2018**, arXiv:1804.09541.
17. Basafa, H.; Movahedi, S.; Ebrahimi, A.; Shakery, A.; Faili, H. NLP-IIS@UT at SemEval-2021 Task 4: Machine Reading Comprehension using the Long Document Transformer. *arXiv* **2021**, arXiv:2105.03775.
18. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The Long-Document Transformer. *arXiv* **2020**, arXiv:2004.05150.
19. Wang, B.; Yao, T.; Zhang, Q.; Xu, J.; Wang, X. ReCO: A Large Scale Chinese Reading Comprehension Dataset on Opinion. *arXiv* **2020**, arXiv:2006.12146,

20. Zheng, C.; Huang, M.; Sun, A. ChID: A Large-scale Chinese IDiom Dataset for Cloze Test. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 778–787, doi:10.18653/v1/P19-1075.
21. Cui, Y.; Liu, T.; Che, W.; Xiao, L.; Chen, Z.; Ma, W.; Wang, S.; Hu, G. A Span-Extraction Dataset for Chinese Machine Reading Comprehension. *arXiv* **2018**, arXiv:1810.07366.
22. Sun, Y.; Wang, S.; Feng, S.; Ding, S.; Pang, C.; Shang, J.; Liu, J.; Chen, X.; Zhao, Y.; Lu, Y.; et al. ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation. *arXiv* **2021**, arXiv:2107.02137.
23. Wang, X.; Gao, T.; Zhu, Z.; Liu, Z.; Li, J.; Tang, J. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. *arXiv* **2019**, arXiv:1911.06136.
24. Cui, Y.; Liu, T.; Wang, S.; Hu, G. Unsupervised Explanation Generation for Machine Reading Comprehension. *arXiv* **2020**, arXiv:2011.06737.
25. Sun, K.; Yu, D.; Yu, D.; Cardie, C. Probing Prior Knowledge Needed in Challenging Chinese Machine Reading Comprehension. *arXiv* **2019**, arXiv:1904.09679.