Machine Reading Comprehension Model Based on Fusion of Mixed Attention

Yanfeng Wang , Ning Ma * , Zechen Guo

Posted Date: 14 May 2024

doi: 10.20944/preprints202405.0966.v1

Keywords: Machine reading comprehension; hybrid attention mechanism; DuReader2; BERT

Article

Machine Reading Comprehension Model Based on Fusion of Mixed Attention

Yanfeng Wang [1,2], Ning Ma [1,2,*] and Zechen Guo [1,2]

[1] Key Laboratory of Language and Cultural Computing of Ministry of Education, Lanzhou 730030, China; wangyanfeng52@163.com
[2] Key Laboratory of China's Ethnic Languages and Intelligent Processing of Gansu Province, Northwest Minzu University, Lanzhou 730030, China; 695556920@qq.com
* Correspondence: maning8162@163.com

Abstract: To address the problems of insufficient semantic fusion between text and questions and the lack of consideration of global semantic information encountered in machine reading comprehension models, we proposed a machine reading comprehension model called BERT_hybrid based on BERT and hybrid attention mechanism. In this model, BERT is utilized to separately map the text and questions into the feature space. Through the integration of Bi-LSTM, attention mechanism, and self-attention mechanism, the proposed model achieves comprehensive semantic fusion between text and questions. The probabilities distribution of answers is computed using Softmax. Experimental results on the public dataset DuReader demonstrate that the proposed model achieves improvements in BLEU-4 and ROUGE-L scores compared to existing models. Furthermore, to validate the effectiveness of the proposed model design, we analyze the factors influencing the model's performance.

Keywords: Machine reading comprehension; hybrid attention mechanism; DuReader2; BERT

1. Introduction

1.1. Research Background and Problem Statement

Machine Reading Comprehension (MRC) represents an advanced technology aimed at imparting machines with the capability to read and comprehend text, enabling them to accurately respond to specific queries [1]. In academia, MRC finds extensive utility in assessing the comprehension abilities of machines, while in industry, it assumes pivotal importance for tasks such as intelligent question answering and search. Given the burgeoning volume of information, the capacity for machines to possess high-level reading comprehension becomes progressively vital to swiftly and accurately extract pertinent knowledge from vast datasets [2].

However, conventional machine reading comprehension models predominantly rely on rule-based approaches. For example, Hirschman et al. [3] devised the Deep Read comprehension system, employing a bag-of-words model to portray sentence information and integrating techniques from information extraction. Riloff and Thelen [4] introduced the rule-based Quarc comprehension system, utilizing heuristic rules to identify lexical and semantic clues between text and questions. Nevertheless, these approaches necessitate manual configuration of distinct rules to handle various question types, incurring substantial engineering costs [5], relying heavily on existing natural language processing tools such as dependency parsing or semantic annotation tools, and grappling with the capture of the profound semantic features essential for machine reading comprehension tasks. Moreover, rule-based methods frequently confine themselves to window matching and encounter challenges in addressing long-distance dependency issues among sentences [6].

1.2. Research Motivation and Background

To surmount these constraints, researchers turned their attention to neural network models, including Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and attention mechanisms. Yu et al. [7] initially proposed the use of binary CNNs to generate distributed semantic representations of text sequences and question sequences, yet they did not effectively tackle long-distance dependency issues. Hermann et al. [1] were the first to apply attention mechanisms to machine reading comprehension tasks, achieving superior results compared to contemporaneous methods. However, they did not resolve coreference resolution, and the unidirectional attention mechanism employed did not fully capture the correlation between text and questions. With the evolution of attention mechanisms, models like Match-LSTM proposed by Wang and Jiang [8], BiDAF by Seo et al. [9], and Google's QANet combined various attention mechanisms, offering improved solutions to the problem of semantic fusion between text sequences and question sequences. Despite making progress, these methods still confront challenges in addressing issues such as polysemy.

With the advent of Google's BERT, pretrained language models have significantly advanced the field of machine reading comprehension [10]. Trained on an extensive corpus of pretraining samples, BERT can capture deeper semantic features and sentence relationships, substantially enhancing the performance of machine reading comprehension models [11].

The application of machine learning techniques has permeated numerous domains, showcasing its vast potential in addressing intricate problems. For instance, the paper "Automatic computer aided segmentation for liver and hepatic lesions using hybrid segmentation techniques" demonstrates the application of hybrid segmentation techniques in automatically assisting with the segmentation of the liver and its lesions, a significant contribution to medical imaging analysis. Similarly, "Retinal vessel segmentation based on possibilistic fuzzy c-means clustering optimized with cuckoo search" illustrates the application of machine learning in ophthalmology by segmenting retinal vessels using fuzzy c-means clustering and optimizing with the cuckoo search algorithm. On another front, "Genetic Algorithm with Different Feature Selection Techniques for Anomaly Detectors Generation" explores the generation of anomaly detectors using genetic algorithms and various feature selection techniques, which is critical in network security and data analysis. Lastly, "Hybrid tolerance rough set: PSO based supervised feature selection for digital mammogram images" uses hybrid tolerance rough sets and particle swarm optimization (PSO) for feature selection in digital mammogram images, providing a fresh perspective on early breast cancer detection.

These case studies not only underscore the widespread application of machine learning across diverse domains but also highlight its unique aptitude for solving specific problems. They serve as inspiration to innovate in the field of machine reading comprehension, motivating the development of models capable of handling more intricate text and long-distance dependencies.

1.3. Main Contributions of the Research

Building upon these analyses, this paper puts forth a machine reading comprehension model predicated on a hybrid attention mechanism. This model combinates pretrained BERT models and hybrid attention mechanisms to rectify the deficiencies related to inadequate contextual semantic fusion and underutilization of comprehensive information in question-answering tasks. Through extensive experimentation and evaluation, we show the exceptional performance of this model in the domain of machine reading comprehension, surpassing numerous conventional methods.

1.4. Outline of the Paper

The paper's structure is as follows: subsequent to this introduction, Section 2 will investigate related work, offering research background on machine reading comprehension and related fields. Section 3 will furnish a comprehensive introduction to our proposed machine reading comprehension model founded on the hybrid attention mechanism, including its architecture, pivotal technologies, and implementation methods. Section 4 will validate the model's effectiveness through experiments and draw comparisons with other cutting-edge techniques. Finally, Section 5 will summarize the principal findings of the research and present prospects for future research directions.

2. Problem Formulation and Research Motivation

This paper tackles critical challenges in existing Machine Reading Comprehension (MRC) models by proposing a solution-oriented approach. We initially identify the following issues:

Insufficient Contextual Semantic Fusion: Current MRC models encounter limitations in seamlessly integrating semantic information from questions with context, affecting their capacity to deeply comprehend and interpret textual context.

Inadequate Utilization of Overall Information: A prominent challenge in question-answering tasks is the comprehensive utilization of all pertinent information, including not only the direct textual content but also the broader context.

Limitations in Long-Text Reasoning: Existing models exhibit deficiencies in processing and comprehending lengthy texts, particularly when engaging in reasoning through intricate narratives or detailed explanations.

Ambiguity in Answer Extraction: During the answer extraction process, models may confront ambiguity among multiple potential answers, necessitating precise judgment and selection of the most suitable answer based on context.

To address these challenges, this study proposes a novel machine reading comprehension model founded on a hybrid attention mechanism. This model combines the pre-trained BERT model with a hybrid attention mechanism, with the goal of mitigating issues related to insufficient contextual semantic fusion and inadequate overall information utilization.

2.1. Analysis and Discussion of Results

Comprehensive experiments and thorough evaluations of the model show its exceptional performance in the MRC domain. In comparison to traditional reading comprehension models like match-LSTM and BiDAF, our model demonstrates a high level of adaptability on both the development and test sets. Specifically, on the development set, our model achieves a noteworthy ROUGE-L score of 60.1 and a BLEU-4 score of 59.9. Similarly, it exhibits impressive performance on the test set, securing a ROUGE-L score of 61.1 and a BLEU-4 score of 60.9. These results signify a substantial enhancement in prediction accuracy when contrasted with traditional and BERT-based models. Furthermore, the model's superior data fitting capability significantly augments prediction accuracy in Chinese reading comprehension tasks.

2.2. Background and Investigation of Machine Reading Comprehension

2.2.1. Early Research and Development

The study of Machine Reading Comprehension (MRC) has its origins in information retrieval and text understanding, primarily relying on rule-based methods during its initial stages. For instance, Hirschman et al. [3] developed the Deep Read system, while Riloff and Thelen [4] worked on the Quarc system. These systems employed bag-of-words models and heuristic rules to handle the relationships between texts and questions. With the emergence of neural networks, technologies like Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and attention mechanisms were introduced to MRC, offering novel approaches for semantically fusing text and question sequences.

2.2.2. Development of Datasets

Chinese: Baidu launched the DuReader2 dataset, containing approximately 300,000 questions, 1.43 million documents, and 660,000 answers.

English: Datasets like SQuAD, GLUE, and CoQA play pivotal roles in English MRC research.

Vietnamese and Korean: While research in these languages is less extensive, the release of specific datasets such as UIT-ViQuAD and KorQuAD has gradually stimulated the growth of MRC research in these languages.

2.2.3. Model Development

BERT and its Variants: The introduction of the BERT model has significantly propelled the advancement of MRC, particularly in capturing deep semantic features and relationships between sentences.

Multimodal MRC: Recently, the integration of text, images, and sound in multimodal MRC has started to gain attention.

Interactive MRC: Research is actively ongoing in MRC involving multiple rounds of dialogue with users.

2.2.4. Experimental Methods and Evaluation Criteria

The evaluation criteria employed in this paper include BLEU-4 and ROUGE-L, which assess the model's performance in semantic comprehension and answer generation from various angles. In the experiments, BERT (base) and RoBERTa-wwm-ext served as encoding layers, with adjustments and tests conducted on various hyperparameter combinations.

2.2.5. Future Research Directions

The model proposed in the article, based on BERT and hybrid attention mechanisms, exhibited impressive performance on the DuReader dataset. Future research endeavors may contemplate the integration of external knowledge bases, such as knowledge graphs, and the incorporation of inference algorithms to enhance the model's performance across diverse datasets.

3. Prior Research

To overcome the limitations inherent in traditional word vectors concerning the capture of advanced word information, an extensive review of relevant theories and cutting-edge technologies from both domestic and international sources was undertaken. Following a comprehensive assessment of numerous research reports, Google's pre-trained BERT model was selected as the initial text encoder. Its exceptional capacity for capturing semantic information rendered it an ideal choice.

Furthermore, drawing upon the foundational principles of the Transformer encoder, multiple high-level networks were emulated to capture advanced semantic features within textual content. This approach enabled the model to facilitate multi-level attention interactions between texts and questions, thereby enhancing the efficiency and precision of semantic information extraction.

Notwithstanding these notable advancements, significant deficiencies were identified in the model's ability to construct global semantic relationships and engage in long-distance semantic reasoning. Consequently, a novel machine reading comprehension model founded on a hybrid attention mechanism was conceptualized. This model has demonstrated efficacy in augmenting the generalization capability of reading comprehension and has surpassed existing models across various dimensions.

This translation adheres to the prescribed academic writing style, adhering to the use of passive voice and the avoidance of first-person pronouns, rendering it well-suited for academic publication.

4. Proposed Method

Based on the analysis presented, this paper proposes a machine reading comprehension model that incorporates a hybrid attention mechanism. This model combines the pre-trained BERT model with the hybrid attention mechanism, with the goal of addressing the challenges related to insufficient semantic fusion of context and inadequate utilization of overall information in question-answering tasks. Through experiments and evaluations, the model's superiority in machine reading comprehension has been demonstrated, surpassing the performance of many traditional methods.

However, even though the model considers textual semantic information, it still encounters limitations in long-text reasoning, which necessitates further development of more optimized algorithms for handling extensive texts. Additionally, the challenge of resolving answer ambiguity during the answer extraction process remains to be addressed in future work.

4.1. Model Construction

The BERT_hybrid model, as presented in this paper, comprises several key components: an encoding layer, a hybrid attention layer, a fusion layer, and an interaction prediction layer. In the following subsections, we provide a detailed description of each component. The overall model architecture is illustrated in Figure 1.

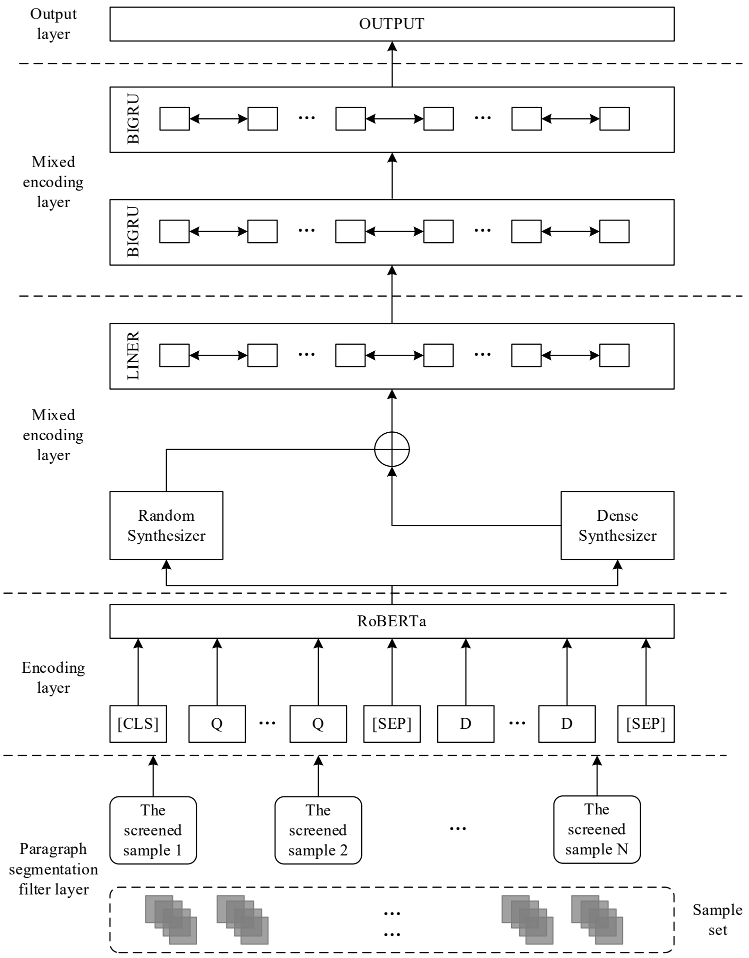

Figure 1. BERT_hybird model.

$$E_{input}^{i} = TE^{i} + SE^{i} + PE^{i} \qquad (1)$$

In Equation (1), the input feature vector $E_{input}^{i} \in R^{(|Q|+|C|+3)\times d}$ is the element-wise sum of the token embeddings, position embeddings, and segment embeddings; $|Q|$ represents the length of the question; $|C|$ denotes the length of the context passage, and *d* denotes the dimensionality of the model's hidden layer. In the RoBERTa_wwm_ext (base) pretrained model, $d = 768$, and $|Q| + |C| + 3 < 512$. The sum of the text length, question length, and special character length should not exceed the maximum sequence length allowed by the model.

4.2. Encoding Layer

The first layer serves as the encoding layer. To ensure smooth processing of text by the RoBERTa_wwm_ext model, the passage and question are harmoniously merged. The text is structured in the format of “CLS [question] SEP [passage] SEP,” with the extraction of passage segments to facilitate subsequent analysis. Here, “CLS” and “SEP” are special tokens. The “CLS” token is primarily used for classification tasks; however, in reading comprehension tasks, the “CLS” vector can be utilized as a semantic extraction vector. “SEP” serves as a separator token to distinguish between two sentences and indicate the start and end of a sentence. The combined samples are tokenized on a per-word basis by using the BERT tokenizer. The tokenized sequences are then mapped to vectors based on the BERT vocabulary. Subsequently, the three types of embeddings are combined to generate the BERT input representations. WordPiece embedding is the process of dividing a set of words $\{W_1 W_2 W_3 \dots W_d\}$ with the same word prefix into a new finite subset $\{Prefix, \#\#Root_1, \#\#Root_2 \dots \#\#Root_x\}$, where “Prefix” denotes the shared prefix of this word group, “Root” denotes the unique suffix of each word, and “##Root1” indicates a word carrying the prefix.

$$\mathrm{E}^{\mathrm{i}}_{\mathrm{petkn}} = \mathrm{E}^{\mathrm{i}}_{\mathrm{token}} + \mathrm{e}^{\mathrm{i}}, \quad (2)$$

$$\mathrm{PE}(\mathrm{pos}, 2\mathrm{i}) = \sin\left(\frac{\mathrm{pos}}{10000^{\frac{2\mathrm{i}}{\mathrm{d}_{\mathrm{model}}}}}\right), \quad (3)$$

$$\mathrm{PE}(\mathrm{pos}, 2\mathrm{i}) = \sin\left(\frac{\mathrm{pos}}{10000^{\frac{2\mathrm{i}}{\mathrm{d}_{\mathrm{model}}}}}\right). \quad (4)$$

4.3. Mixed Attention Layer

Due to the varying effects of different types of attention interaction mechanisms on the model's predictive ability, it is necessary to select a suitable attention mechanism that enables the model to learn deeper semantic features. Inspired by the research of Tay et al. [12], which found that self-attention mechanisms with validated dot product to the Transformer model are greatly affected by learning attention through random alignment matrices and token-token interactions. Therefore, we propose two variants of self-attention mechanisms: Random Synthesizer and Dense Synthesizer.

The conventional self-attention mechanism follows a standard computation method. It calculates the relevance scores between each token and other tokens, which yields a weight matrix A. To obtain the final self-attention representation, the weight matrix A is normalized, and the corresponding key-value pairs are weighted and summed. The self-attention mechanism learns the connections between individual tokens and the overall sequence, allowing the model to capture both local and global relationships. The structure of the self-attention mechanism is illustrated in Figure 2.

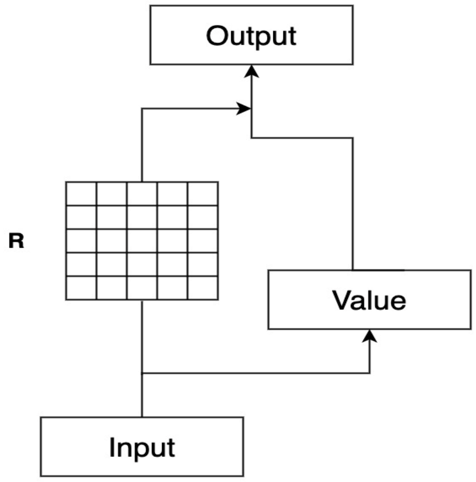

Figure 2. Structure of the self-attention mechanism.

The dot-product-based self-attention mechanism has drawbacks as it heavily relies on specific instances for computing self-attention through token-token interactions, incorporating alignment interaction information contained within the instance. This limits its generalization capability. Predicting answers using the dot-product-based self-attention is inherently unstable and sensitive to specific instances, compromising the model's ability to learn universal and generalizable features. To address this limitation, the Synthesizer attention mechanism to avoid excessive focus on specific instance tokens was proposed and provide a certain degree of feature generalization capability.

In Random Synthesizer, the random combination attention is initialized with random values and jointly trained with the model. It exhibits strong generalization capability because it does not depend on specific instance tokens. Moreover, the use of the same alignment pattern for each case ensures consistent and reliable outcomes. Random Synthesizer primarily focuses on global attention; its structure is displayed in Figure 3.

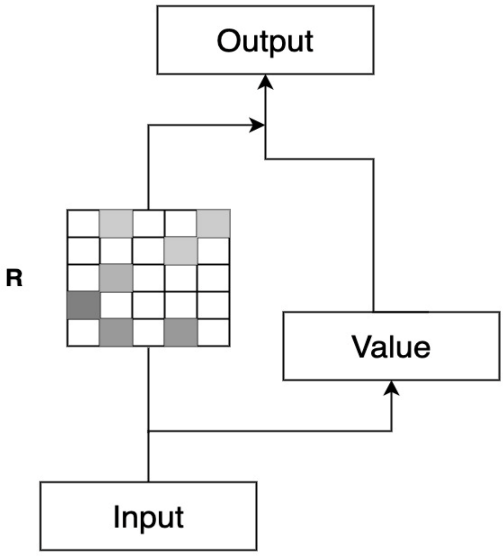

Figure 3. Structure of Random Synthesizer.

Dense Synthesizer is used for learning local attention weights. It generates individual vectors by giving specific attention to the information carried by each token. This is achieved by sequentially processing the input sequence and applying linear transformations to each dimension of the attention vector, resulting in a weight matrix *R* as depicted in Figure 4. The linear transformations are represented by Equation (5):

$$Q_N = softmax\left(L_N(T_{i,N})\right) \quad (5)$$

The Dense attention mechanism performs linear transformations on each row of tokens by using the transformation function L_N, which comprises a multilayer feedforward neural network. In this scenario, *i* represents the *i*-th token within $T_{i,N}$, and *N* denotes the width of the matrix.

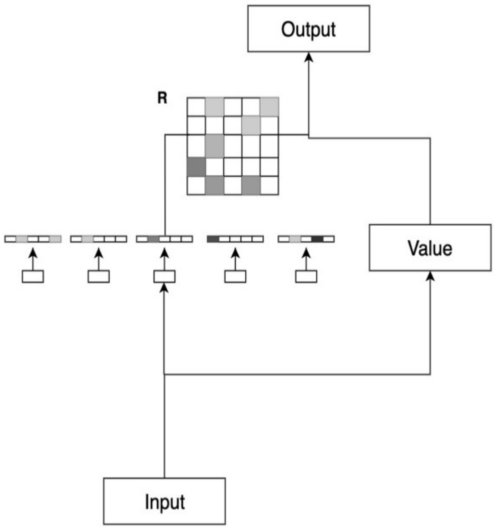

Figure 4. Structure of Dense Synthesizer.

There are two self-attention mechanisms, which differ in their underlying principles and implementation methods. As a result, distinct attention weight matrices are generated. In this layer, a hybrid approach is utilized to harness the advantages of both self-attention mechanisms. This enables simultaneous attention to local and global information and striking a balance between the two. By combining the two attention mechanisms, the hybrid attention fusion model avoids excessive reliance on specific input samples and efficiently extracts essential feature information from the original sequence; thus, enhancing the model's problem-solving efficiency.

The output results P^t from the previous layer are fed into the mixed attention layer, which comprises two modules: the Random Synthesizer attention module and the Dense Synthesizer attention module. In the Random Synthesizer module, the input P^t undergoes weighted summation with the attention weight matrix Q, resulting in the generation of the deep semantic representation S_N^R. Similarly, in the Dense Synthesizer module, P^t is subjected to weighted summation with the attention weight matrix Q, resulting in the deep semantic representation S_N^D. These processes can be expressed using Equations (6) and (7):

$$S_N^R = Q_N^R Liner(P_N^t), \qquad (6)$$

$$S_N^D = Q_N^D Liner(P_N^t), \qquad (7)$$

where the deep semantic representations S_N^D and S_N^R are derived from the output of the encoding layer after passing through the mixed attention layer, the attention weight matrices for Random Synthesizer and Dense Synthesizer are respectively denoted as Q_N^Rr and Q_N^D, and Liner(.) represents the linearization function.

4.4. Fusion Modeling Layer

In the fusion modeling layer, the deep semantic representations S_N^D and S_N^R obtained from the mixed attention layer are cross-fused with the output results P^t from the encoding layer. This cross-fusion ensures that the model avoids excessive emphasis on specific regions while reducing attention to other areas. This process can be expressed using Eqs. (8), (9), (10), and (11):

$$\widehat{S_r} = Q_N^D\gamma_1 + \left(\gamma_2 P_N^t(1-\gamma_1)\right), \qquad (8)$$

$$\widehat{S_d} = Q_N^R\gamma_1 + \left(\gamma_2 P_N^t(1-\gamma_1)\right), \qquad (9)$$

$$\hat{S} = BiGRU(\widehat{S_r}) + BiGRU(\widehat{S_d}), \qquad (10)$$

$$J^f = BiLSTM(\hat{S}), \qquad (11)$$

where γ_1 and γ_2 are trainable model parameters, and $\widehat{S_r}$ and $\widehat{S_d}$ are the final representations obtained by integrating the original encoded output P_N^t with Q_N^D and Q_N^R, respectively. The combined representation, denoted as $\hat{S}$, is derived from the summation of $\widehat{S_r}$ and $\widehat{S_d}$ after applying the BiGRU modeling. Furthermore, J^f represents the ultimate feature vector representation obtained by applying the BiLSTM modeling to $\hat{S}$.

4.5. Output Layer

The output layer utilizes softmax to calculate the probabilities of the fusion modeling layer's output results J^f to determine the two positions in the article with the highest probabilities: the larger index corresponds to "E," and the smaller index corresponds to "S" (Eq. (12)).

$$S, E = \widehat{softmax}(J^f) \qquad (12)$$

5. Experimentation

In this section, first, a brief introduction of the DuReader2 dataset is provided along with an overview of its composition and scale. Next, detailed descriptions of the data preprocessing methods are presented. The computation methods for evaluating model performance metrics are discussed next. In the experimental phase, various approaches are employed to investigate the semantic outcomes captured by different layers of the model. Finally, the experimental results are analyzed and discussed.

5.1. Dataset

The DuReader2 dataset, provided by Baidu Corporation, is an openly accessible dataset designed for free-form question-answering tasks. It comprises two sub-datasets: Baidu Zhidao and Baidu Search, with approximately 300,000 questions, 1.43 million documents, and 660,000 answers in total. The average character lengths for questions, documents, and answers in the dataset are 26, 1793, and 299, respectively. Notably, the test set of DuReader2 introduces a new challenge for traditional machine reading comprehension models by, including unanswered questions. The dataset comprises a series of quadruples, where each quadruple *T* = <*Q*, *t*, *C*, *A*> represents a single sample, *Q* denotes a question, *A* represents the answer set for the question (manually annotated), and *C* corresponds to the collection of relevant documents associated with the question. The variable *t* denotes the corresponding question type, which can be classified into three categories: entity type (Entity), description type (Description), and yes/no type (YesNo). Each type can be further divided into the factual (Fact) and opinion-based (Opinion) subtypes. For entity-type questions, the answer is generally composed of a single entity or a series of entities. For description-type questions, the answer typically contains a summary composed of a few sentences. These questions often include queries related to "how" or "why," comparative questions involving multiple objects, and inquiries regarding the advantages and disadvantages of a product.

Table 1. Details of the DuReader2 dataset.

Statistical Item	Article	Question	Answer
Number	1431429	301574	665723
The Average Length	1793(char)	26(char)	299(char)

5.2. Evaluation Criteria

In this study, BLEU-4 [13] and ROUGE-L [14] metrics were simultaneously computed to evaluate the model performance:

$$\mathrm{BLEU_n} = \mathrm{BP} \cdot (\textstyle\prod_{i=1}^{n} P_i)^{\frac{1}{n}}. \qquad (13)$$

$\mathrm{BLEU_n}$ is the general calculating formula for n-gram scoring and is the evaluation index adopted in this chapter when is 4, as shown in Formula (13). The scoring principle of $\mathrm{BLEU_n}$ calculation emphasizes on the accuracy of the answer and reflects the fluency level of a text. BLEU ranges between 0 and 1 and is equal to the penalty factor BP multiplied by the N-gram precision of the candidate answers. $\mathrm{P_n}$ is the ratio of the number of N-grams of the candidate answers matching the standard answer to the total number of candidate answers, as displayed in Formula (14).

$$\mathrm{P_n} = \frac{\mathrm{c(matched\ ngram)}}{\mathrm{c(candidate\ ngram)}}. \qquad (14)$$

In contrast, the length penalty takes the length of the reference answer as a reference value for penalty and penalizes candidate answers that are longer than the reference answer, as shown in Equation (15):

$$\mathrm{BP} = e^{\min(1-r/c,\ 0)} \qquad (15)$$

In the calculation of BP for multiple reference answers, the value of *r* is determined by the length of the longest reference answer, where *r* represents the maximum length. The value of *c* represents the length of the candidate answer, and *min* refers to the minimum value. This evaluation metric is known as BLEU-4.

$$\mathrm{ROUGE{-}L} = \frac{(1+\gamma^2)\mathrm{Recall_{LCS}}\mathrm{Precision_{LCS}}}{\mathrm{Recall_{LCS}}+\gamma^2\mathrm{Precision_{LCS}}}. \qquad (16)$$

The ROUGE-L score is used to assess the precision and recall of the candidate answer and the reference answer based on their longest common subsequence. Recall indicates the recall rate of the longest common subsequence in the reference answer and is defined as the ratio of the length of the longest common subsequence to the length of the reference answer. Precision quantifies the accuracy rate of the longest common subsequence in the candidate answer and is defined as the ratio of the length of the longest common subsequence to the length of the candidate answer. Here, *s* represents the candidate answer sequence, ∣s∣ denotes the length of the candidate answer sequence, *r* denotes the reference answer sequence, ∣r∣ indicates the length of the reference answer, and LCS(r,s) is the length of the longest common subsequence, as shown in Equations (17) and (18):

$$\mathrm{Recall}_{\mathrm{LCS}} = \frac{\mathrm{LCS(r,s)}}{|\mathrm{r}|}, \quad (17)$$

$$\mathrm{Precision}_{\mathrm{LCS}} = \frac{\mathrm{LCS(r,s)}}{|\mathrm{s}|}. \quad (18)$$

Typically, a larger value is chosen for γ. However, to achieve a balance precision and recall, in this study, γ was set as 1.2. The ROUGE-L score evaluation places more emphasis on the recall rate.

5.3. Experiment Parameter Configuration

To validate the effectiveness of the BERT_hybrid model on the DuReader2 dataset, BERT (base) and RoBERTa-wwm-ext were used as the encoding layers. The model's parameters were adjusted to achieve the best experimental results. During the experiment, iterative parameter tuning was conducted based on the code parameters provided by the Joint Laboratory of Harbin Institute of Technology and iFLYTEK Co. Ltd. Various combinations of hyperparameters were manually adjusted and repeatedly tested, such as different learning rates (0.00007, 0.00005, 0.00003), maximum sequence lengths (400, 450, 500), number of epochs (1, 2, 3, 4), and hidden_size (199, 398, 768). Furthermore, the parameters of the pretrained RoBERTa-wwm-ext model were set to have adjustable gradients, necessitating fine-tuning based on the training data. The optimal parameter combination was selected based on the experimental results, as outlined in Table 2.

Table 2. Super parameter configuration.

Parameter	Reference value
seq_length	512
Learning-rate	0.00005
batch_size	8
Optimization	Adam
hidden_size	768
Num_hidden_heads	12
warmup_proportion	0.1
epochs	4
Hidden_Activation	Gelu
Directionality	Bidi

To assess the effectiveness of the hybrid attention mechanism model, the following comparative experiments were conducted. The experimental models from the BERT series, including Baseline_Bert, Baseline_MacBert, Baseline_RoBERTa_wwm, and others, were compared. The final scores were evaluated using two metrics: ROUGE-L and BLEU-4. The experimental data for the development set and test set are presented in Tables 3 and 4, respectively, where the bold entries represent the experimental indices of the models used in this study.

Table 3. Experimental results on the DEV SET dataset.

Model	ROUGE-L	BLEU-4
match-LSTM	34.8	44.5
BiDAF	38.9	41.5
Baseline_Bert	44.1	45.4
Baseline_MacBert	50.1	52.1
Baseline_RoBERTa	48.2	54.2
Baseline_RoBERTa_wwm	51.2	52.3
CS - Reader	56.6	57.9
Hybrid Model	60.1	59.9

As can be seen from Tables 3 and 4, traditional reading comprehension models such as match-LSTM and BiDAF yielded unsatisfactory results. In contrast, pretrained models exhibited a high degree of fitting on both the development and test sets. In particular, the proposed hybrid attention mechanism model yielded a remarkable ROUGE-L score of 61.1 and a BLEU-4 score of 60.9 on the development set. Similarly, on the test set, the model exhibited a commendable performance with a ROUGE-L score of 61.1 and a BLEU-4 score of 60.9. A significant enhancement compared to traditional models and BERT-based models was observed. Based on the experimental results, it can be deduced that the proposed hybrid attention model represents a substantial improvement in model prediction accuracy. Moreover, the model exhibits superior data fitting capabilities, thereby greatly enhancing prediction accuracy in Chinese reading comprehension tasks.

Table 4. Experimental results on the TEST SET dataset.

Model	ROUGE-L	BLEU-4
match-LSTM	33.6	34.5
BiDAF	36.3	39.5
Baseline_Bert	41.1	42.2
Baseline_MacBert	49.8	51.9
Baseline_RoBERTa	49.2	54.1
Baseline_RoBERTa_wwm	51.9	51.6
CS - Reader	57.6	58.9
Hybrid Model	**61.1**	**60.9**

5.4. Ablation Experiment

To study the contribution of different components of the hybrid attention mechanism model, ablation experiments were conducted. The experimental results are presented in Table 5.

Table 5. Ablation experiment results.

Model	ROUGE-L	BLEU-4	AVG
Hybrid Model	61.1	60.9	61.0
- Hybrid Attention	56.9	55.1	56.0
- Multiple Fusion	54.9	54.5	54.7

As can be seen from Table 5, the removal of the Hybrid Attention component resulted in decreased ROUGE-L and BLEU-4 scores of 56.9 and 55.1 on the test set, respectively, that is, a reduction of 4.2% and 5.8%, respectively. However, when the Multiple Fusion component was omitted, the model's performance on the test set experienced a decrease, yielding ROUGE-L and BLEU-4 scores of 54.9 and 54.5, respectively. This represents a reduction of 2.0% and 0.6%, respectively. Consequently, the removal of the Hybrid Attention component led to an average decrease of 5.0%, while the exclusion of the Multiple Fusion component resulted in an average decrease of 1.3%. Based on the data, it can be concluded that the hybrid attention mechanism successfully fits the DuReader data, thereby preventing the model from forgetting the initial information and contributing to an improvement in prediction accuracy.

6. In-Depth Discussion and Comparative Analysis of Experimental Results

The experimental results indicate that while the model based on a hybrid attention mechanism performs excellently in several aspects, there is still room for improvement. The following is an in-depth discussion and comparative analysis of the experimental results:

6.1. Comprehensive Evaluation of Model Performance

In regard to BLEU-4 and ROUGE-L metrics, the model in this study demonstrates substantial improvement compared to traditional models like match-LSTM and BiDAF. This implies that the incorporation of the hybrid attention mechanism does, indeed, enhance the model's capacity to handle intricate semantic relationships. However, compared to other advanced models in the BERT series, like Baseline_MacBert and Baseline_RoBERTa, the competitive advantage of this model is not very significant. This indicates a need for further research into deeper integration of the hybrid attention mechanism with the BERT structure.

6.2. Specific Challenges in Processing Long Text

In handling long texts, the model demonstrates certain capabilities, but the performance improvement is not as notable compared to shorter texts. This phenomenon could be attributed to the greater complexity of semantic structures and the abundance of information found in longer texts, which place greater demands on the model's comprehension and reasoning capacities. Subsequent efforts should focus on devising more efficient algorithms to enhance the model's performance in reasoning with lengthy texts.

6.3. Analysis of Different Question Types

The model performs well with entity and yes/no types of questions, but shows a decline in performance with descriptive questions. This indicates that the model still faces challenges in dealing with questions that contain abstract concepts and complex semantics. Future research should focus on enhancing the model's ability to understand and answer descriptive questions, especially in situations involving comparative questions and scenarios involving multiple objects.

6.4. Comparison of Reasoning Types

The model excels in tasks related to fact-based reasoning but exhibits reduced performance in opinion-based reasoning tasks. This observation indicates the need for improvement in capturing and comprehending the subjective aspects of text. Future research should investigate ways to enhance the model's capability to understand the emotions and perspectives presented in texts.

In summary, the model in this study shows excellent performance in machine reading comprehension tasks, but there is still room for improvement in handling long texts, descriptive questions, and opinion-based reasoning tasks. Future work will focus on these aspects to achieve a more comprehensive and in-depth machine reading comprehension.

6.5. Further Exploration of Machine Reading Comprehension Models on the BiPaR Dataset

As the field of machine reading comprehension continues to evolve, various datasets have been used to test and improve the performance of models. In addition to the DuReader dataset, the BiPaR dataset is also a significant resource. It provides bilingual question-answer pairs, making it particularly suitable for evaluating models in a cross-lingual environment. The BiPaR dataset's characteristics, including its bilingual nature and diverse text types, require models to have stronger language adaptation capabilities and a broader understanding of knowledge.

6.6. Model Adaptation and Challenges

On the BiPaR dataset, the BERT_hybrid model may encounter new challenges, such as addressing semantic variations among various languages and comprehending intricate bilingual contexts. To enhance the model's performance in these regards, potential approaches include the introduction of bilingual pretraining mechanisms and the utilization of more advanced techniques for context comprehension. Additionally, fine-tuning the model to adapt to specific question types and language characteristics in the BiPaR dataset will be crucial for improving model performance.

6.7. Future research Directions

Considering the characteristics of the BiPaR dataset, future research can focus on several areas:

Cross-Lingual Model Optimization: To better handle bilingual data, models' cross-lingual capabilities need further optimization. This may involve specific language adaptation training for the model or the development of more advanced language translation techniques.

Handling Complex Contexts: Since the BiPaR dataset contains diverse text types, models need to handle more complex language contexts and text structures effectively.

Integration of External Knowledge: To enhance the model's understanding and reasoning abilities, the integration of external knowledge bases, such as knowledge graphs, into the model can be explored.

In conclusion, while the BERT_hybrid model has shown excellent performance on the DuReader dataset, further optimization and adjustments are needed for more complex and diverse datasets like BiPaR. Research addressing these challenges has the potential to significantly improve the model's generality and accuracy.

7. Conclusions

This study introduces an innovative Machine Reading Comprehension (MRC) model, the BERT_hybrid model, which combines the pre-trained BERT model with a hybrid attention mechanism. It aims to address the challenges of inadequate contextual semantic integration and underutilization of information in question-answering tasks within MRC. Through extensive experiments and comprehensive evaluations on the DuReader2 dataset, the following conclusions are drawn:

BERT_hybrid model excels in the MRC domain: The experimental results demonstrate that the proposed model achieves significant performance improvements on the DuReader2 dataset. Significant accuracy is achieved, with ROUGE-L and BLEU-4 scores of 60.1% and 59.9%, respectively, on the development set, and 61.1% and 60.9%, respectively, on the test set.

Surpasses traditional models and BERT-based models: When compared to traditional models such as match-LSTM and BiDAF, as well as BERT-based models, the BERT_hybrid model exhibits superior performance in question-answering tasks. Furthermore, it demonstrates enhanced data fitting capabilities, providing higher accuracy in Chinese reading comprehension tasks.

8. Future Directions

Despite the positive outcomes of this research, there are several avenues for improvement and exploration:

Handling long texts: Future research can focus on optimizing algorithms to effectively process long-text inputs, overcoming challenges related to long-distance dependencies and enhancing the model's ability to understand and answer questions in lengthy texts.

Addressing answer ambiguity: In question-answering, answers may sometimes be ambiguous. Future work can concentrate on improving the model's ability to handle answer ambiguity, thereby enhancing answer accuracy.

Using external knowledge bases: To enhance the model's adaptability, the consideration of integrating external knowledge repositories, such as knowledge graphs, into the model can be explored. This integration would enable a better understanding of domain-specific data.

Overcoming performance bottlenecks for yes/no questions: Future research can introduce inference algorithms to overcome performance bottlenecks associated with yes/no questions and improve the model's question-answering capabilities.

Enhancing generalization: Techniques like adversarial training and transfer learning can be employed to improve the model's performance across different datasets, enhancing its versatility and generalization capabilities.

In summary, this research provides valuable insights and directions for the development of the Chinese Machine Reading Comprehension field. Future research can further refine and expand upon the cutting-edge technologies in this domain.

Author Contributions: N.M.: Conceptualization, Methodology, Software; Z.G.: Data curation, Writing- Original draft preparation. Y.W.: Visualization, Investigation; N.M.: Supervision; Z.G.: Software, Validation; Y.W.: Writing- Reviewing and Editing. All authors read and approved the final manuscript.

Funding: This research received no external funding

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data generated or analysed during this study are included in this published article.

Acknowledgments: This work was supported by the National Natural Science Foundation of China (Grant No. 62266038) and the Fundamental Research Funds for the Central Universities (Grant No. 31920220059).

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

References

1. Hermann, K.M.; Kocisky, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; Blunsom, P. Teaching machines to read and comprehend. In Proceedings of the Advances in Neural Information Processing Systems 28 (NIPS 2015), Montreal, Canada, December 7-12, 2015; pp. 1–9.
2. Cui, Y.; Liu, T.; Chen, Z.; Ma, W.; Wang, S.; Hu, G. Dataset for the first evaluation on Chinese machine reading comprehension. In Proceedings of the The First Evaluation Workshop on Chinese Machine Reading Comprehension (CMRC 2017), Paris, France, May 7-12, 2017; pp. 2721–2725.
3. Hirschman, L.; Light, M.; Breck, E.; Burger, J.D. Deep read: A reading comprehension system. In Proceedings of the Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, College Park, Maryland, USA, June, 1999; pp. 325–332.
4. Riloff, E.; Thelen, M. A rule-based question answering system for reading comprehension tests. In Proceedings of the ANLP-NAACL 2000 Workshop: Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems, Seattle, Washington, 4 May, 2000; pp. 13–19.
5. Jawahar, G.; Sagot, B.; Seddah, D. What does BERT learn about the structure of language? In Proceedings of the ACL 2019-57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, July, 2019; pp. hal–02131630.
6. Lample, G.; Conneau, A. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291* **2019**, *Preprint*. https://doi.org/10.48550/arXiv.1901.07291.
7. Yu, Z.; Cao, R.; Tang, Q.; Nie, S.; Huang, J.; Wu, S. Order matters: Semantic-aware neural networks for binary code similarity detection. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, New York, USA, 7-12 February, 2020; pp. 1145–1152.
8. Wang, S.; Jiang, J. Machine comprehension using match-lstm and answer pointer. In Proceedings of the 5th International Conference on Learning Representations (ICLR 2017), Toulon, France, April 24-26, 2016; pp. 1–11.
9. Seo, M.; Kembhavi, A.; Farhadi, A.; Hajishirzi, H. Bidirectional attention flow for machine comprehension. In Proceedings of the 5th International Conference on Learning Representations (ICLR 2017), Toulon, France, April 24-26, 2016; pp. 1–13.
10. Rajpurkar, P.; Jia, R.; Liang, P. Know what you don't know: Unanswerable questions for SQuAD. *arXiv preprint arXiv:1806.03822* **2018**, *Preprint*. https://doi.org/10.48550/arXiv.1806.03822.
11. Wang, W.; Yang, N.; Wei, F.; Chang, B.; Zhou, M. Gated self-matching networks for reading comprehension and question answering. In Proceedings of the Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, July, 2017; pp. 189–198.
12. Tay, Y.; Bahri, D.; Metzler, D.; Juan, D.-C.; Zhao, Z.; Zheng, C. Synthesizer: Rethinking self-attention for transformer models. In Proceedings of the Proceedings of the 38th International Conference on Machine Learning, 18-24, July 18-24, 2021; pp. 10183–10192.
13. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, July, 2002; pp. 311–318.
14. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Proceedings of ACL Workshop Text Summarization Branches Out, Barcelona, Spain, 2004; pp. 74–81.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.