applied sciences

MDPI

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 (Y.W.); 279021731@xbmu.edu.cn (Z.G.)
[2] Key Laboratory of China's Ethnic Languages and Intelligent Processing of Gansu Province, Northwest Minzu University, Lanzhou 730030, China
* Correspondence: maning8162@163.com

Abstract: To address the problems of the 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 the 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, an attention mechanism, and a self-attention mechanism, the proposed model achieves a comprehensive semantic fusion between text and questions. The probability distribution of answers is computed using Softmax. The 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

check for updates

Citation: Wang, Y.; Ma, N.; Guo, Z. Machine Reading Comprehension Model Based on Fusion of Mixed Attention. *Appl. Sci.* **2024**, *14*, 7794. https://doi.org/10.3390/app14177794

Academic Editors: Chilukuri K. Mohan and Antonio Pagliaro

Received: 29 March 2024
Revised: 23 August 2024
Accepted: 26 August 2024
Published: 3 September 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

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 a pivotal importance for tasks such as intelligent question-answering and searching. 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, Qian Z 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. Kakulapati and Reddy [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 the 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].

In fact, MRC can be seen as a basic task of essay format Q&A, which aims to enable computers to understand natural language texts like humans and be able to answer related

questions. In order to complete this task, the computer needs to find the correct answers from a given article, which can be related to the facts, reasons, explanations, or reasoning in the article. The formal definition of an MRC task is as follows: given context C and problem Q, the model needs to learn the function F to make $A = F(C, Q)$, in order to obtain the correct answer A to problem Q. Based on the research of Chen Yu et al., we divide MRC tasks into four types: cloze filling, multiple choice, fragment extraction, and free question-answering [7], and introduce the definitions and representative datasets of these four types of tasks.

As early as the 1970s, Lehnert et al. believed that MRC was an important method for testing computer programs to understand human language, and proposed the early representative work of MRC research, the QUALM system [7]. This study designed a theory of a question-answering system and emphasized the importance of context in answering questions, attempting to construct a framework for the human understanding of stories through scripts and strategy simulations. However, the scale of the system was very small and limited to specific fields, making it difficult to generalize to a wider range of application scenarios. Therefore, research on machine reading was basically stagnant at that time. Until 1999, Hirschman et al. collected primary school reading materials from grades 3 to 6 and constructed the first MRC dataset [8]. This dataset contains 60 stories and requires the machine reading system to be able to find the sentence corresponding to the answer based on the question. Meanwhile, the study also proposes a word bag model that represents sentences with questions and the context as phrases and takes words that appear simultaneously in both questions and the context as answers. However, early MRC systems did not perform well and could not be used in practical applications due to the lack of appropriate text representation methods and strong fitting capabilities [9]. During this period, the accuracy of MRC models based on rules and simple statistical models was only 30% to 40%, far from meeting human expectations.

Several related studies have proposed Chinese machine reading understanding models based on multiple high-level networks capturing advanced semantic feature information and integrating self-attention mechanisms and using a mixed attention mechanism to improve model performance. On the basis of pretraining language models, an improved Chinese machine reading understanding model was proposed by combining the mixed attention mechanism, and an experimental verification was conducted on multiple datasets, showing an improvement in model performance [9]. This can complement the research content of this paper.

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. Nguyen 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 the 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 [1], 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. In encyclopedic question-answering robots, a knowledge base is constructed by inputting a large number of encyclopedic articles. When users ask questions, the extraction-based MRC technology can quickly search for answer content from the knowledge base and quickly return it to the user [11]. In terms of intelligent customer service, by entering business documents to build a knowledge base, extraction-based MRC technology can analyze user questions and find the most suitable content to answer them from the knowledge base. This technology not only ensures the accuracy of intelligent customer service Q&A, but also greatly reduces operational costs. In the field of intelligent education, extractive MRC technology can assist human learning processes, assist students in writing through automatic essay grading models, understand semantics, correct grammar, and summarize easily erroneous knowledge points. In addition, extractive MRC technology can also be combined with online education to play a greater role. Therefore, the extraction-based MRC technology has extensive practical application value in fields such as information retrieval, question-answering systems, chatbots, and intelligent education, and has profound significance [12].

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 an 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 [12].

1.2. Main Contributions of the Research

Traditional machine learning models have shortcomings in handling text representation and have low model complexity. Therefore, in the complete AI problem of MRC, traditional models are often in an underfitting state [12]. With the successful application of deep learning in fields such as speech, image, and pattern recognition, deep learning methods have been introduced into various subtasks of natural language understanding. The MRC model based on deep learning effectively alleviates the problems of traditional models and promotes the further development of MRC.

Building upon these analyses, this paper puts forth a machine reading comprehension model predicated on a hybrid attention mechanism. This model combines pretrained BERT models and hybrid attention mechanisms to rectify the deficiencies related to inadequate contextual semantic fusion and the 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.

1.3. Outline of the Paper

The paper's structure is as follows: subsequent to this introduction, Section 2 will investigate related work, offering a 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 the context, affecting their capacity to deeply comprehend and interpret the 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 the precise judgment and selection of the most suitable answer based on the context.

To address these challenges, this study proposes a novel machine reading comprehension model founded on a hybrid attention mechanism. This model combines the pretrained 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 an 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 [13]. 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, Qian et al. [3] developed the Deep Read system, while Kakulapati and Reddy [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.

On the one hand, MRC based on deep learning can better capture contextual information and perform significantly better than traditional models; on the other hand, the public availability of large-scale datasets has made it possible to use deep neural network architectures to solve MRC tasks, such as Hermann et al.'s large-scale supervised fill in the blank MRC dataset CNN/DailyMail [2], Rajourka et al.'s famous Stanford question-answering dataset SQuAD [3] in the field of extractive MRC proposed in 2016, and Nguyen et al.'s real question-answering dataset MSMARCO [12] from the Bing search engine proposed in 2016, which provide a wide range of performance evaluation and testing platforms for MRC systems. The SQuAD dataset has a large scale and high quality of data, making it the mainstream testing benchmark for MRC models. Researchers have proposed many MRC models based on this dataset. With the continuous emergence of new models, the performance of deep neural network-based learning models has exceeded the human average by 91.2% [13].

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.

In fact, there is a problem of a weak model processing ability for complex texts in current MRC tasks. Most pretrained language models used in research are aimed at short texts and cannot handle complex texts well. At the same time, existing datasets also have shortcomings. Therefore, this article aims to improve the accuracy of MRC models in complex text understanding through methods such as context purification and problem type encoding. The aim is to build an effective extraction-based MRC model for complex text processing based on existing deep learning model architectures.

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

This article delves into the current situation of MRC both domestically and internationally. Firstly, it introduces the main task types, datasets, and corresponding evaluation indicators of MRC. At the same time, it explores the relevant technologies and implementation of extractive MRC. Subsequently, the difficulties in extracting MRC for complex texts were analyzed, and the question of how to improve these issues was studied. The research content of this article mainly includes the following three parts: (1) the MRC model based on a self-attention mechanism and convolutional neural network; (2) the MRC model based on a gated self-matching attention mechanism; (3) an extractive MRC model for complex texts.

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 pretrained 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.

4. Method and Models

Based on the analysis presented, this paper proposes a machine reading comprehension model that incorporates a hybrid attention mechanism. This model combines the pretrained BERT model with the hybrid attention mechanism, with the goal of addressing the challenges related to the insufficient semantic fusion of the context and the 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 the 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.

$$E^i_{input} = TE^i + SE^i + PE^i \tag{1}$$

In Equation (1), the input feature vector $E^i_{input} \in R^{(|Q|+|C|+3)\times d}$ 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 the 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 $\{W_1W_2W_3 \ldots W_d\}$ with the same word prefix into a new finite

subset $\{Prefix, \#\#Root_1, \#\#Root_2 \ldots \#\#Root_x\}$, 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 [13].

$$E^i_{petkn} = E^i_{token} + e^i, \tag{2}$$

$$PE(pos, 2i) = \sin\left(\frac{pos}{10,000^{\frac{2i}{d_{model}}}}\right), \tag{3}$$

$$PE(pos, 2i) = \sin\left(\frac{pos}{10,000^{\frac{2i}{d_{model}}}}\right). \tag{4}$$

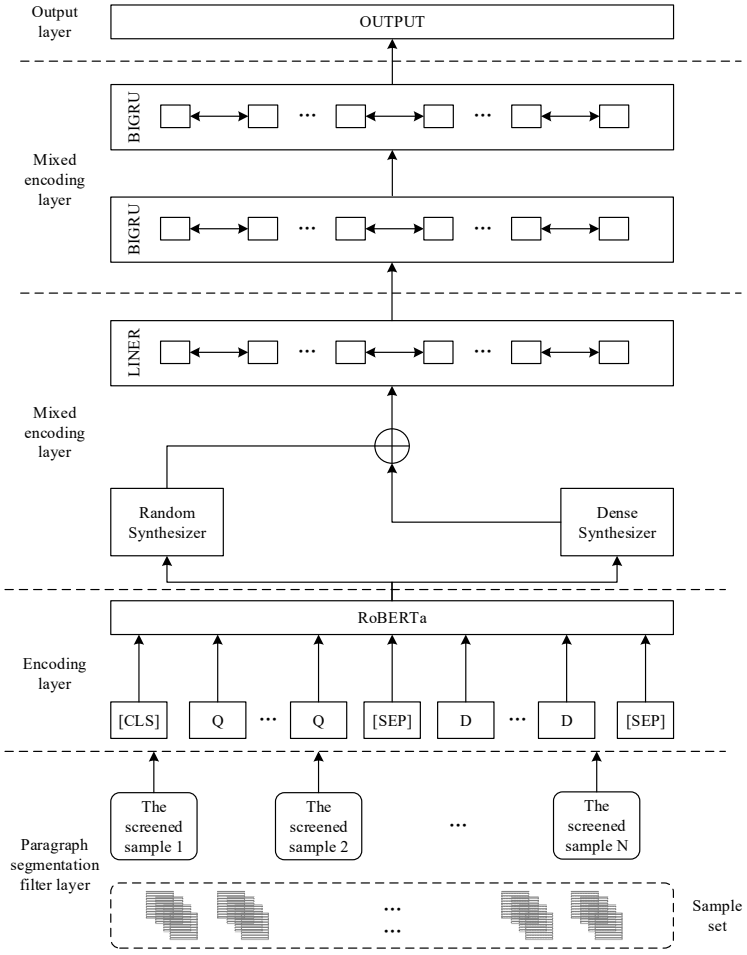


Figure 1. BERT_hybrid model.

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. This is inspired by the research of Tay et al. [12], which found that self-attention mechanisms with a 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.

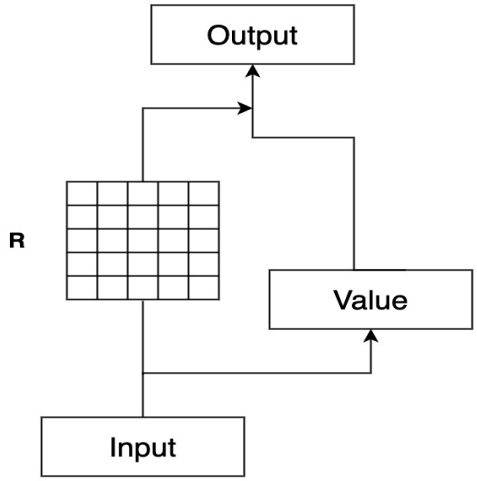


Figure 2. Structure of the self-attention mechanism.

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 mechanism is inherently unstable and sensitive to specific instances, compromising the model's ability to learn universal and generalizable features. To address this limitation, the Random Synthesizer attention mechanism was proposed to avoid an excessive focus on specific instance tokens and provide a certain degree of feature generalization capabilities.

In the Random Synthesizer, the random combination attention is initialized with random values and jointly trained with the model. It exhibits a 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. The Random Synthesizer primarily focuses on global attention; its structure is displayed in Figure 3.

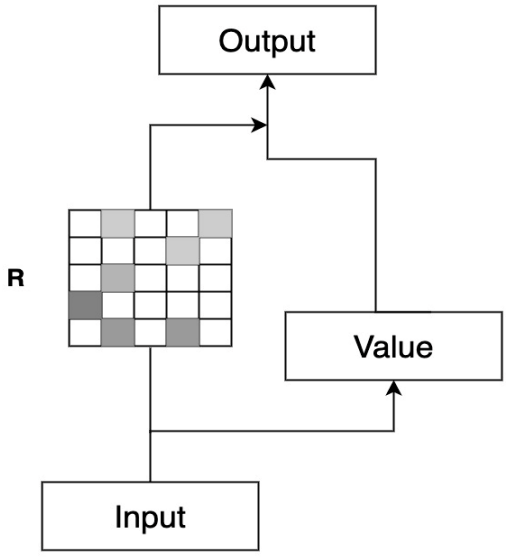


Figure 3. Structure of Random Synthesizer.

The 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):

$$Q_N = softmax(L_N(T_{i,N})) \tag{5}$$

The Dense Synthesizer 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 [13].

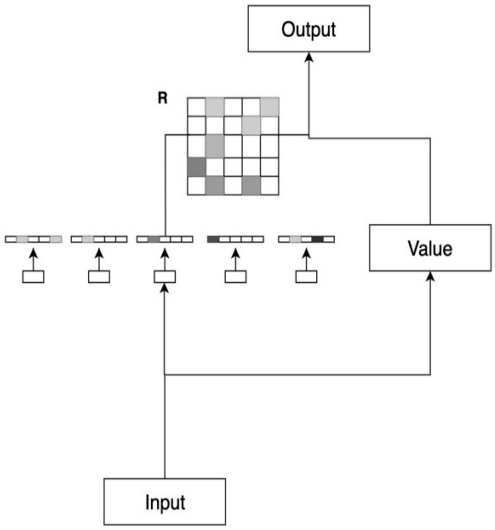


Figure 4. Structure of Dense Synthesizer.

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, the model's problem-solving efficiency is enhanced.

In the annotation process of the BLEU-4 dataset, annotators need to follow the following rules: (1) During the Q&A pair collection annotation stage, annotators need to carefully read each article and ask at least five questions. (2) The answer to each question must be found in the corresponding article. (3) The annotator cannot directly copy the sentence in the text as a question but needs to rephrase it in their own language to raise the question. (4) The annotator needs to ask more difficult questions as much as possible, and the answers can be descriptions or sentences in the article, which may increase the difficulty compared to entity-based questions. (5) The annotator should propose multiple types of questions as much as possible to avoid situations where the answer types are too singular. The types of questions that can be referred to include the following: person type (who), time type (when), place type (where), reason type (why), fact type (what), choice type (which), method type (how), quantity type (quantity), and other types (others). (6) It is suggested that the annotators ask more specific questions to prevent the occurrence of multiple possible answers [14].

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 a 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):

$$S_N^R = Q_N^R Liner(P_N^t), \tag{6}$$

$$S_N^D = Q_N^D Liner(P_N^t), \tag{7}$$

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 denoted as Q_N^Rr and Q_N^D, respectively, 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 Equations (8)–(11):

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

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

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

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

where γ_1 and γ_2 are trainable model parameters, and $\hat{S_r}$ and $\hat{S_d}$ 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 $\hat{S_r}$ and $\hat{S_d}$ after applying the BiGRU modeling. Furthermore, J^f represents the ultimate feature vector representation obtained by applying the Bi-LSTM modeling to $\hat{S}$.

4.5. Output Layer

The output layer utilizes the softmax function 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" (Equation (12)).

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

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 [15].

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 (see Table 1 for detailed statistics). 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 tradi-

tional machine reading comprehension models by including unanswered questions. The dataset comprises a series of quadruples, where each quadruple $T \leq Q, t, C, A>$ 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 [16].

Table 1. Details of the DuReader2 dataset.

Statistical Item	Article	Question	Answer
Number	1,431,429	301,574	665,723
The Average Length	1793 (char)	26 (char)	299 (char)

5.2. Evaluation Criteria

In this study, BLEU-4 [13] and ROUGE-L [17] metrics were simultaneously computed to evaluate the model performance:

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

In terms of length statistics, the average number of sentences in the BLEU-4 dataset is 15.86, with an average of 7.26 questions per article, an average question length of 13.02 words, and an average answer length of 9.86 words. In addition, the ROUGE-L dataset was also collected, with an average number of paragraph sentences of 9.93 and an average paragraph length of 435.8 words. The average number of questions corresponding to each article was 3.44, with an average question length of 21.07 words and an average answer length of 4.86 words. And the DuReader2 dataset constructed in this article has an average paragraph sentence count of 32.93, with an average number of questions corresponding to each article of 11.47. The average question length and answer length are consistent with the DRCD dataset. According to the statistical results, it can be found that the paragraph length in the BLEU-4 dataset and DuReader dataset is much longer than that in the benchmark dataset ROUGE-L.

$\mathrm{BLEU_n}$ is the general calculating formula for N-gram scoring and is the evaluation index adopted in this chapter when the value is four, as shown in Formula (13). The scoring principle of the $\mathrm{BLEU_n}$ calculation emphasizes 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).

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

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

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

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.

$$\text{ROUGE–L} = \frac{(1+\gamma^2)\text{Recall}_{\text{LCS}}\text{Precision}_{\text{LCS}}}{\text{Recall}_{\text{LCS}} + \gamma^2\text{Precision}_{\text{LCS}}}. \quad (16)$$

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):

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

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

Typically, a larger value is chosen for γ. However, to achieve a balanced 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. (Hefei, China). 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

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

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 the prediction accuracy in Chinese reading comprehension tasks.

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 underwent 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

Based on the complex Chinese MRC dataset, the previous MRC model R-Net was further extended, and two methods were proposed: text purification based on key sentence extraction and the encoding of problem type information. Text purification technology is the process of extracting key sentences from complex text MRC datasets before predicting answers, in order to achieve the goal of purification. The specific implementation is as follows: firstly, select the appropriate key sentence extraction algorithm; secondly, conduct information statistics on the dataset used in the experiment, including the number of questions, the number of questions included in each article, the number of sentences in the text segment, the length of questions, and the length of answers; once again, filter out the optimal number of extracted sentences for each dataset; finally, conduct experiments on a complex text MRC dataset. Problem classification is the addition of the encoding of problem type information in the MRC model, predicting the category of problems through problems and articles, and providing the model with more problem information. The specific implementation is as follows: firstly, classify the problems in the dataset, analyze the proportion of each problem type in the dataset, and conduct experiments based on different problem types to obtain the advantageous problem types; secondly, integrate the problem classification module into the model; finally, compare the performance of the baseline model and the model integrated with the problem classification module. Both methods have shown some improvement in the original accuracy.

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 the 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

The machine reading comprehension model based on mixed attention fusion has shown significant performance improvements in understanding complex texts and capturing contextual associations, but its research and application still have multiple limitations. On the one hand, model training relies on a large amount of high-quality annotated data, and the acquisition and annotation of such data are often time-consuming and labor-intensive, and it is difficult to ensure the coverage of all domains and contexts, which limits the generalization ability of the model. Although the hybrid attention mechanism can enhance the model's capture of key information, it is challenging to achieve the efficiency and accuracy of attention allocation when dealing with long texts or cross document understanding, which may lead to information omission or misunderstanding. On the other hand, the current model has shortcomings in terms of interpretability, and its decision-making process is difficult to intuitively understand, which reduces the credibility and practicality of the model. The accuracy of reading comprehension in cross-linguistic and cross-cultural contexts still needs to be improved, and the performance of the model in processing non-native texts or culturally specific content is still unstable, limiting its application breadth in multilingual environments.

6.7. Prospect Research Developments

Given the aforementioned limitations, future research will focus on exploring more efficient, generalized, and interpretable hybrid attention models to overcome existing bottlenecks. Researchers can fully develop adaptive data augmentation and semi-supervised learning techniques to reduce reliance on large amounts of annotated data and improve the performance of models in small or zero sample learning scenarios. Secondly, future research can efficiently optimize attention mechanisms, such as introducing hierarchical attention and dynamic attention allocation strategies, to improve the attention efficiency

and accuracy of the model when processing long texts and complex contexts. Furthermore, the interpretability of the model can be enhanced by visualizing attention distribution, constructing interpretable inference paths, and other methods to improve the transparency of model decision-making and enhance its practical value in real-world scenarios. In addition, a universal framework for cross-linguistic and cross-cultural reading comprehension will be studied, utilizing techniques such as multilingual pretraining and cultural sensitivity learning to enhance the model's cross domain adaptability. The deep integration of hybrid attention models with knowledge graphs, common sense reasoning, and other technologies will be explored to construct a reading comprehension system with deep understanding and reasoning abilities, in order to address the more complex and diverse challenges of natural language processing. Through exploration and practice in these directions, hybrid attention models are expected to achieve more extensive and in-depth applications in the future, promoting continuous progress in the field of natural language processing.

7. Conclusions

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

The 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.

The BERT_hybrid model 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.

Author Contributions: N.M.: conceptualization, methodology, and software. Z.G.: data curation and writing—original draft preparation. Y.W.: visualization and investigation. N.M.: supervision. Z.G.: software and validation. Y.W.: writing—reviewing and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the National Natural Science Foundation of China (Grant No. 62466052) and the Fundamental Research Funds for the Central Universities (Grant No. 31920220059).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

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

References

1. Nguyen, A.D.; Li, M.; Lambert, G.; Kowalczyk, R.; McDonald, R.; Vo, Q.B. Efficient Machine Reading Comprehension for Health Care Applications: Algorithm Development and Validation of a Context Extraction Approach. *JMIR Form. Res.* **2024**, *8*, e52482. [CrossRef] [PubMed]
2. Duan, J.; Liu, S.; Liao, X.; Gong, F.; Yue, H.; Wang, J. Chinese EMR Named Entity Recognition Using Fused Label Relations Based on Machine Reading Comprehension Framework. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2024**. ahead of print. [CrossRef]
3. Qian, Z.; Zou, T.; Zhang, Z.; Li, P.; Zhu, Q.; Zhou, G. Speculation and negation identification via unified Machine Reading Comprehension frameworks with lexical and syntactic data augmentation. *Eng. Appl. Artif. Intell.* **2024**, *131*, 107806. [CrossRef]

4. Kakulapati, V.; Reddy, J.G.; Reddy, K.K.K.; Reddy, P.A.; Reddy, D. Analysis of Machine Reading Comprehension Problem Using Machine Learning Techniques. *J. Sci. Res. Rep.* **2023**, *29*, 20–31. [CrossRef]
5. Guan, B.; Zhu, X.; Yuan, S. A T5-based interpretable reading comprehension model with more accurate evidence training. *Inf. Process. Manag.* **2024**, *61*, 103584. [CrossRef]
6. Lample, G.; Conneau, A. Cross-lingual language model pretraining. *arXiv* **2019**, arXiv:1901.07291. [CrossRef]
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 AAAI Conference on Artificial Intelligence, New York, NY, 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, 24–26 April 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, 24–26 April 2016; pp. 1–13.
10. Rajpurkar, P.; Jia, R.; Liang, P. Know what you don't know: Unanswerable questions for SQuAD. *arXiv* **2018**, arXiv:1806.03822. [CrossRef]
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 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 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 38th International Conference on Machine Learning, Hong Kong, China, 18–24 July 2021; pp. 10183–10192.
13. Huang, Z.; He, L.; Yang, Y.; Li, A.; Zhang, Z.; Wu, S.; Wang, Y.; He, Y.; Liu, X. Application of machine reading comprehension techniques for named entity recognition in materials science. *J. Cheminform.* **2024**, *16*, 76. [CrossRef] [PubMed]
14. Liu, L.; Liu, M.; Liu, S.; Ding, K. Event extraction as machine reading comprehension with question-context bridging. *Knowl.-Based Syst.* **2024**, *299*, 112041. [CrossRef]
15. Liu, S.; Zhang, S.; Ding, K.; Liu, L. JEEMRC: Joint Event Detection and Extraction via an End-to-End Machine Reading Comprehension Model. *Electronics* **2024**, *13*, 1807. [CrossRef]
16. Ahmed, Y.; Telmer, C.A.; Zhou, G.; Miskov-Zivanov, N. Context-aware knowledge selection and reliable model recommendation with ACCORDION. *Front. Syst. Biol.* **2024**, *4*, 1308292. [CrossRef]
17. Haiyang, Z.; Chao, J. Verb-driven machine reading comprehension with dual-graph neural network. *Pattern Recognit. Lett.* **2023**, *176*, 223–229.

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.