applied sciences

MDPI

Article

# AT-CRF: A Chinese Reading Comprehension Algorithm Based on Attention Mechanism and Conditional Random Fields

**Nawei Shi [1], Huazhang Wang [1,*] and Yongqiang Cheng [2]**

[1] School of Electrical Engineering, Southwest Minzu University, Chengdu 610041,China
[2] Computer Science, University of Hull, Hull HU6 7RX, UK
* Correspondence: wanghuazhang@swun.edu.cn

**Abstract:** Machine reading comprehension (MRC) is an important research topic in the field of Natural Language Processing (NLP). However, traditional MRC models often face challenges of information loss, lack of capability to retain long-distance dependence, and inability to deal with unanswerable questions where answers are not available in given texts. In this paper, a Chinese reading comprehension algorithm, called the Attention and Conditional Random Filed (AT-CRF) Reader, is proposed to address the above challenges. Firstly, RoBERTa, a pre-trained language model, is introduced to obtain the embedding representations of input. Then, a depthwise separable convolution neural network and attention mechanisms are used to replace the recurrent neural network for encoding. Next, the attention flow and self-attention mechanisms are used to obtain the context–query internal relation. Finally, the conditional random field is used to handle unanswerable questions and predict the correct answer. Experiments were conducted on the two Chinese machine reading comprehension datasets, CMRC2018 and DuReader-checklist, and the results showed that, compared with the baseline model, the F1 scores achieved by our AT-CRF Reader model has improved by 2.65% and 2.68%, and the EM values increased by 4.45% and 3.88%.

**Keywords:** machine reading comprehension; pre-trained model; attention mechanism; conditional random field

check for updates

**Citation:** Shi, N.; Wang, H.; Cheng, Y. AT-CRF: A Chinese Reading Comprehension Algorithm Based on Attention Mechanism and Conditional Random Fields. *Appl. Sci.* **2022**, *12*, 10459. https://doi.org/10.3390/app122010459

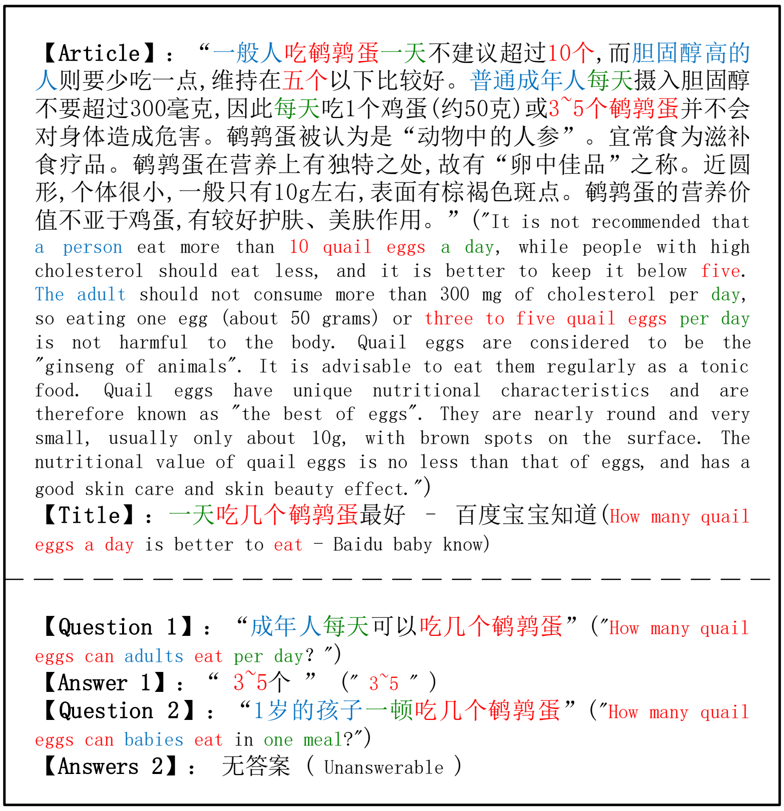
Academic Editors: Gui-Lin Qi, Tong Xu and Meng Wang

Received: 27 September 2022
Accepted: 15 October 2022
Published: 17 October 2022

## 1. Introduction

Reading comprehension plays an important role in people's daily life, and it is usually the core method for people to obtain information in many scenarios such as daily conversations, examinations, and information searches. With the development of technology and the requirements of intelligent applications, Machine Reading Comprehension (MRC) has become an important topic in the direction of Natural Language Processing (NLP), which is of great significance for guiding machines to understand human languages, improving the intelligence level of machines and enabling machines to acquire knowledge continuously [1].

MRC can help users find the most relevant information from a large number of complex texts by analyzing the given text and questions, guiding the machine to answer the questions around the text, and returning the correct answer [2]. At the same time, it is widely used because of its closeness to daily life. In recent years, there have been many improved models based on Bert [3], such as ALBERT [4], RoBERTa [5], etc. The use of pre-trained models has become a development trend of MRC. Because the pre-trained model can only learn the shallow semantic matching information of the context, most of the current models adopt the combination of the pre-trained language model and the attention mechanism [6], that is, they obtain the corresponding representation through the pre-trained model, and then use the attention mechanism for reasoning, so as to capture the deep semantic information of the context and predict more accurate answers. However, the original pre-trained models are designed for the English applications, and Chinese has different grammatical structures and characteristics from English [7], so it cannot effectively process Chinese text. For example, “Yu Mao Qiu Pai Mai Shi Yuan” has completely different semantics when the participles are “Yu Mao Qiu Pai/ Mai/ Shi Yuan” (The badminton racket is worth 10 RMB) and “Yu Mao Qiu/Pai/Mai/Shi Yuan” (The badminton auctioned for 10 RMB), so even the model with excellent performance will perform poorly when applied to Chinese text. In addition, in Chinese reading comprehension, the attention mechanism in the model is very important to obtain the correct answer. The existing attention mechanism usually only focuses on specific words, and it is difficult to mine the overall semantics of the context, which is poor for complex problem recognition. At the same time, the premise of MRC assumes that there are snippets in the article that match the question, but in reality, this is sometimes not the case, as shown in Figure 1. It has become an urgent task to understand the content of the passage correctly and to make a correct judgment on the answer ability of the questions.



【Article】： “一般人吃鹌鹑蛋一天不建议超过10个，而胆固醇高的人则要少吃一点，维持在五个以下比较好。普通成年人每天摄入胆固醇不要超过300毫克，因此每天吃1个鸡蛋(约50克)或3~5个鹌鹑蛋并不会对身体造成危害。鹌鹑蛋被认为是“动物中的人参”。宜常食为滋补食疗品。鹌鹑蛋在营养上有独特之处，故有“卵中佳品”之称。近圆形，个体很小，一般只有10g左右，表面有棕褐色斑点。鹌鹑蛋的营养价值不亚于鸡蛋，有较好护肤、美肤作用。” (″It is not recommended that a person eat more than 10 quail eggs a day, while people with high cholesterol should eat less, and it is better to keep it below five. The adult should not consume more than 300 mg of cholesterol per day, so eating one egg (about 50 grams) or three to five quail eggs per day is not harmful to the body. Quail eggs are considered to be the ″ginseng of animals″. It is advisable to eat them regularly as a tonic food. Quail eggs have unique nutritional characteristics and are therefore known as ″the best of eggs″. They are nearly round and very small, usually only about 10g, with brown spots on the surface. The nutritional value of quail eggs is no less than that of eggs, and has a good skin care and skin beauty effect.″)

【Title】：一天吃几个鹌鹑蛋最好 － 百度宝宝知道(How many quail eggs a day is better to eat - Baidu baby know)

【Question 1】： “成年人每天可以吃几个鹌鹑蛋”(″How many quail eggs can adults eat per day? ″)

【Answer 1】： “ 3~5个 ” (″ 3~5 ″ )

【Question 2】： “1岁的孩子一顿吃几个鹌鹑蛋”(″How many quail eggs can babies eat in one meal?″)

【Answers 2】： 无答案 ( Unanswerable )

**Figure 1.** An example of the DuReader-checklist dataset. The same color in the article and the questions indicate that the corresponding content has relevant information; question 1 can find the corresponding answer fragment in the text, and question 2 has a similar correspondence in the text but is not the correct answer.

To address the above problems and to better accomplish the machine reading comprehension task, this paper proposes a Chinese machine reading comprehension algorithm, the AT-CRF Reader. This model introduces a pre-trained language model to fully acquire the underlying semantic features, combines CNN, attention mechanism, and position encoding to learn dependency information over long-distance. It also uses self-attention mechanism to obtain the internal representation of the context, and introduces Conditional Random Fields (CRF) [8] as the output layer to deal with unanswerable questions. Finally, the experimental results on two reading comprehension datasets, CMRC2018 [9] and DuReader-checklist [10], verify the effectiveness of the model in this paper.

The main contributions of this paper can be summarized as follows:

- We use a Chinese pre-trained language model in combination with multiple attention mechanisms to acquire input representations and fully learn the connection between the local and global contexts;
- We use the CRF to convert the output prediction into the sequence labeling question to predict the starting and ending position of the answer, and output the correct answer sequence when the question is answerable, while identifying the unanswerable questions and responding appropriately;
- We validate the performance of the model by comparing the AT-CRF Reader with state-of-the-art MRC models on CMRC2018 and DuReader-checklist datasets.

## 2. Related Works

### 2.1. Machine Reading Comprehension Research

Early MRC tasks were mainly based on hand-crafted rules, such as the QUALM system [11] using manually coded scripts, which is tedious and difficult for the system to generalize. Since then, some small corpora have appeared, such as the datasets released at the DEEP READ and ANLP2000 conferences [12]. These corpora are mostly based on the bag-of-words method and adopt pattern matching technology, which is difficult to construct effective features, and the effect is not satisfactory. In recent years, benefiting from the development of deep learning and the release of large-scale reading comprehension datasets (e.g., CNN/Daily Mail [13], SQuAD [14,15]), MRC models based on deep learning have shown obvious superiority and have gradually become the mainstream in the research field. The AOA Reader [16] applies the attention mechanism to the text level. Match-LSTM [17] combines a Long Short-Term Memory (LSTM) network with a one-way attention mechanism, and uses a pointer network as the result output for the first time. BiDAF [18] uses a bidirectional attention flow mechanism to extract textual interaction information. R-NET [19] introduces a self-attention mechanism based on a gated recurrent unit for text self-matching. Transformer [20] exclusively uses attention mechanisms to obtain information. Bert introduces dynamic coding on the basis of the two-layer Transformer to obtain deeper bidirectional semantic representation, and the performance was greatly improved. RoBERTa uses the whole word masking to further improve robustness. With the introduction of Chinese datasets such as PD&CFT [21], CMRC2018 and DuReader, the development of Chinese reading comprehension has been promoted. D-Reader [22] uses fastText to train word embedding to improve encoding for full-text prediction. RoBERTa-wwm-ext [23] considers Chinese characteristics for training. ERNIE [24] introduces external knowledge word-level embedding, phrase-level embedding, and entity-level embedding based on Bert. ELECTRA [25] uses a generator network to replace characters instead of masking the input to improve the prediction speed. MacBERT [26] adds N-grams for synonym replacement to improve model text modeling capabilities. FNN-MRC [27] uses GRU to aggregate frame semantic knowledge combined with neural network to facilitate question answering. ChineseBERT [28] integrates the phonetic and glyph into the pre-trained process to enhance the modeling ability of Chinese corpus. At present, pre-trained models have become the focus of research in the field of MRC, but a single pre-trained language model can only obtain shallow semantic information. Therefore, this paper combines the attention mechanism and pre-trained model to fully extract text interaction information and capture deeper feature relationships.

### 2.2. Conditional Random Field

The MRC task is essentially a supervised learning problem: learning a predictor $f$ from a given triple <context $C$, query $Q$, answer $A$>, which then outputs the corresponding answer $A$ through the input $Q$ and $C$, and the process is expressed as follows:

$$f : (C, Q) \rightarrow A \tag{1}$$

Most of the traditional MRC models use a Pointer Network to predict the final result, i.e., the network outputs the index of the starting and ending position of the contextual fragment containing the answers. However, this approach is only valid when the query has answer fragments in the text. If there are no answers that exist for the query, the training process cannot be performed, and hence is not able to handle the unanswerable situations. The CRF is a probabilistic graphical model of discriminant class proposed by Lafferty based on the Maximum Entropy Model and Hidden Markov Model [29]; it has been widely used in the field of nomenclature recognition and machine translation. The linear chain CRF is mainly for sequence tagging and segmentation, transforming the output into a BIO sequence tagging task and taking a certain sequence from it as the answer. Thus, it can help the machine make correct judgments when no answer exists, and the working principle is shown in Figure 2.

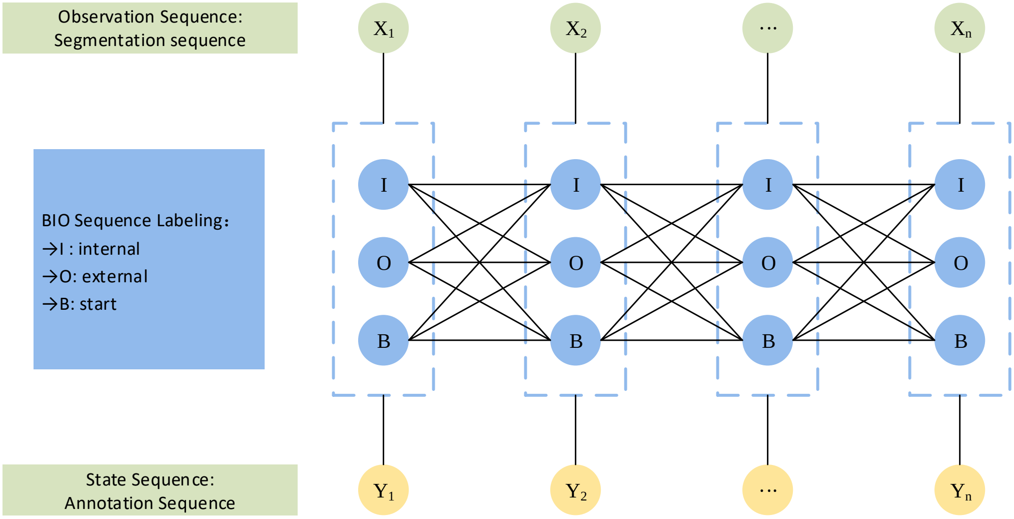


**Figure 2.** Working principle of the BIO sequence labeling task. The sequential tagging task can select the labeling result with the highest overall probability among all possible labeling results, and the probability of each item is the sum of the output probability of the upper layer and the labeling conversion probability. This method can not only obtain the overall optimal solution, but also avoid the situation where the answer is unreasonable.

The CRF model assumes that the input observation value is the random variable $X$, and the output sequence observation value is the random variable $Y$, and $P(Y \mid X)$ is the conditional probability of variable $Y$ given variable $X$. If the variable $P(Y \mid X)$ constitutes an undirected graph structure $G = (V, E)$ with the node $v \in V$ and the edge $e \in E$, $n(v)$ is the set of adjacent nodes to $v$, $v/\{\ v\}$ is all nodes excluding vertex $v$. Then, the conditional random field formed by the probability distribution of the random variable $X$ and the random variable $Y$ satisfies the following condition:

$$P(Y_v \mid X, Y_{v/\{\ v\}}) = P(Y_v \mid X, Y_{n(v)}) \tag{2}$$

## 3. Model Architecture

In order to effectively improve the accuracy of the Chinese reading comprehension model, deepen the model's ability to mine text information, and solve unanswerable problems, this paper proposes a Chinese machine reading comprehension model—AT-CRF Reader.

The model contains six layers, namely, Embedding Layer, Encoder Layer, Semantic Interaction Layer, Contextual Self-Attentive Layer, Modeling Layer, and Output Layer, as shown in Figure 3. First, the features of input are obtained through the Chinese pre-trained word vectors in the embedding layer, then the generated sequence vector is fed into the encoder layer, and further into the positional encoding, depthwise separable convolution neural network, and the attention mechanism to capture the internal structural and interaction information. Then, the attention flow mechanism is used to obtain the Query-aware Context representation and the Context-aware Query representation, and then the contextual self-attention representation and the output of the attention flow are combined through a fusion mechanism, so the interaction can be captured by the modeling layer. Finally, the sequence of answers is predicted by the conditional random field as the final result.

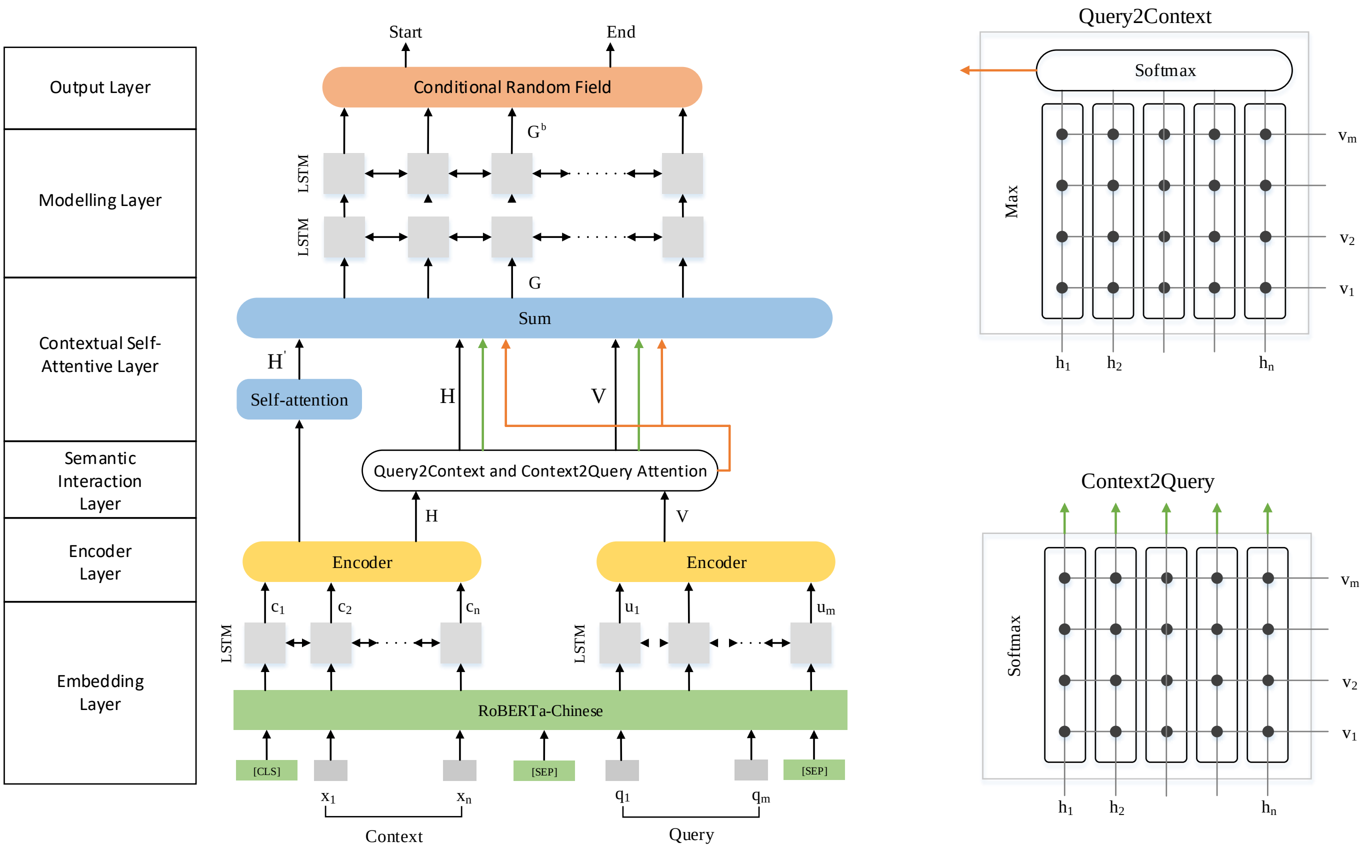


**Figure 3.** The overall architecture of the AT-CRF Reader reading comprehension model, with the bidirectional attention flow mechanism on the right, where $x$ and $q$ are the input context and query, $c$ and $u$ are the initial vector representations of the context and the query obtained after the embedding layer processing, $H$ and $V$ are the contextual vector representations of the context and the query respectively, $\tilde{H}$ contains the most important word information of the query in the context, $\tilde{V}$ contains the correlation representation of the query word and each context word, and $H^{'}$ is the self-attention representation of the whole context, $G$ is the splicing and fusion of the output of the semantic interaction layer and the contextual self-attentive layer. $G_b$ embeds information about the entire context and query through the Bi-LSTM.

### *3.1. Embedding Layer*

In this layer, in order to obtain richer semantic information and word embedding representations of different granularity for each input, we use the Chinese pre-trained model, RoBERTa. RoBERTa has the same model structure compared to the pre-trained model—Bert—but has the optimizations below:

- The model adopts the Whole Word Mask (WWM) mechanism, which improves the learning ability by constantly changing the masking position during training.
- The model uses the full-sentences and the doc-sentences as input to dynamically increase the batch size to replace the Next Sentence Prediction objective (NSP) in Bert.
- The model is trained using a larger batch size with more training data and longer training time.

- The model uses a larger Byte to construct the vocabulary table.

In practical applications, the RoBERTa outperforms the Bert on various downstream tasks. The model used in this paper contains a 12-layer Transformer encoder, and the vector dimension is 768 dimensions.First, splicing the input context $X = \{ x_1, x_2, \cdots, x_n\}$ with query $Q = \{ q_1, q_2, \cdots, q_m\}$, by adding a $[CLS]$ token at the start position, and an $[SEP]$ token at the end and split position. Then, take the corresponding positional encoding, word vector, and contextual vector as input, and then processing by RoBERTa, use the output of the last layer of the encoder as the vector representation $O$ of the context and query. Finally, the initial vector representations of the context $C = \{ c_1, c_2, \cdots, c_n\}$ and the query $U = \{ u_1, u_2, \cdots, u_n\}$ are output, respectively. The calculation process is as follows:

$$input = [CLS] + X + [SEP] + Q + [SEP] \tag{3}$$

$$O = RoBERTa\_wwm\_ext(input) \tag{4}$$

### *3.2. Encoder Layer*

The encoder layer is used to extract local and global information between words at different positions. It is composed of positional encoding, depthwise separable convolution, a self-attention mechanism, and a feed-forward network. Layer normalization is applied for each layer, and residual connections are used between layers. First, a position function calculates the encoded positional information, which is further weighted and summed with the word embedding of the previous layer to form the input of this layer—in the convolutional layer, a depthwise separable convolution network with fewer parameters and stronger generalization ability to obtain more contextual information, specifically, with a kernel size of 7, 128 filters, and 4 convolutional layers. The self-attention layer adopts a multi-head attention mechanism with eight heads to calculate the connection between each word and the input texts to obtain the global contextual information. Finally, the feed-forward network is used to generate the contextual vector representation of the context $H \in R^{d\times n}$ and the query $V \in R^{d\times m}$ as the output, where $d$ is the matrix dimension.

### *3.3. Semantic Interaction Layer*

Although in general the context in reading comprehension tasks can be very long, the answer often exists in a short segment related to the query. Hence, the semantic interaction layer can be used to correlate information of the words in the context and the query to capture the interactive information. This layer uses the bi-directional attention mechanism to calculate the attention of both context-to-query (Context2Query) and query-to-context (Query2Context), respectively. The input is the contextual vector representation of the context $H \in R^{d\times n}$ and the query $V \in R^{d\times m}$ both from the previous layer. The shared similarity matrix $S \in R^{n\times m}$ between the contextual embedding of $H$ and $V$ can be calculated as follows:

$$\delta(h, v) = w^{n}_{(S)}[h; v; h \odot v], \tag{5}$$

$$S_{nm} = \delta(H_{:n}, V_{:m}) \in R^{n\times m}, \tag{6}$$

where $S_{nm}$ is computed by the similarity of the n-th context word and the m-th query word; $\delta$ is a numerical function that encodes the similarity of the two input functions; $w_{(S)}$ is a trainable weight vector; $[;]$ represents the concatenation of matrix vectors on rows; $H_{:n}$ is the n-th column vector of $H$; and $V_{:m}$ is the m-th column vector of $V$. Then, the attention value in both directions can be calculated by the similarity matrix $S$.

Context-to-query Attention. Context2Query attention indicates which query word is the most relevant to each context word. First, the attention weight $\alpha_n$ is calculated by row normalization of the matrix $S$ using softmax. Then, the context-to-query vector $\tilde{v}$ is obtained by calculating the weighted summation of the query vectors, and the calculation process is as follows:

$$\alpha_n = softmax(S_{n:}) \in R^{m}, \tag{7}$$

$$\tilde{v} = \sum\nolimits_{m} \alpha_{nm} V_{:m} \in R^{d}, \tag{8}$$

where the $\alpha_n$ is the attention weight of the n-th context word to the query words; $\sum \alpha_{nm} = 1$ for all $n$; the query representation of the whole context is $\tilde{V} \in R^{d\times n}$.

Query-to-context Attention. Query2Context attention indicates which context word has the closest similarity to the word in the query. First, we take the maximum value of each column of the similarity matrix $S$; then, the attention weight $b_n$ is obtained through softmax normalization, and, finally, the query-to-context vector $\tilde{h}$ is obtained after the weighted summation of the context vectors. The context representation of the query is $\tilde{H} \in R^{d\times n}$, and the calculation process is as follows:

$$b_n = softmax(max_{col}(S)) \in R^{n} \tag{9}$$

$$\tilde{h} = \sum\nolimits_{n} b_{nm} H_{:n} \in R^{d} \tag{10}$$

### 3.4. Contextual Self-Attention Layer

In addition to the query-context interaction in reading comprehension tasks, there are also correlations between the related contexts. Therefore, the contextual self-attention layer is used to capture the internal structure of the context, learned the feature relationship between its phrases, and obtained its global attention. First, we calculate the similarity matrix $S'_{ij}$ between each context word and words in other segments of the context, and normalize the matrix by softmax to obtain the attention weight $\eta_i$ between context words. Then, through the weighted summation of each context vector, the contextual self-attention representation $h'$ is obtained, and the entire contextual self-attention representation is $H' \in R^{d\times n}$. The process is shown in the following formula:

$$S'_{ij} = h \odot h \in R^{n\times n} \tag{11}$$

$$\eta_i = softmax(S'_i) \in R^{n} \tag{12}$$

$$h' = \sum\nolimits_{n} n_i H_n \tag{13}$$

Finally, the new query representation $\tilde{V}$ and the new context representation $\tilde{H}$ in the two directions of the attention mechanism obtained by the semantic interaction layer are combined with the contextual vector representation $H$. Then, the input is fused through a splicing function $\beta$ to obtain the output matrix representation of this layer $G$. The calculation process is as follows:

$$\beta(h', \tilde{v}, \tilde{h}) = [h'; \tilde{v}; h' \odot \tilde{h}] \in R^{3d\times n} \tag{14}$$

$$G_n = \beta(H'_n, \tilde{V}_n, \tilde{H}_n) \in R^{3d} \tag{15}$$

### 3.5. Modeling Layer

In this layer, we use two Bi-directional Long Short-Term Memory networks (Bi-LSTM) [30] in both directions, the input is $G$, which was output by the previous layer, and the output is the matrix $M \in R^{2d\times n}$, which represents the contextual information for each word about the sequence containing the entire context-query interaction. The calculation process is as follows:

$$G^{f} = BiLSTM(G) \tag{16}$$

$$G^{b} = BiLSTM(G^{f}) \tag{17}$$

$$M = \left[\, G^{f}, G^{b} \right] \tag{18}$$

### 3.6. Output Layer

In this layer, in order to avoid the situation where the model cannot answer unanswerable questions, the output layer uses a CRF model for prediction, it determines the relationship of adjacent labels by the transition score, and obtains the global optimal sequence as the final answer. First, we take the $M$ obtained by the modeling layer as the input of this layer, set the sentence sequence with the predicted label as $y = (y_1, y_2, \cdots, y_n)$, and define its probability as $p(y \mid m)$, where $Z(m)$ is the normalization (partition function), $\lambda_k$ and $\mu_l$ are the weights of different features, $t_k$ is the transition probability matrix, representing the transition score from label $y_{i-1}$ to label $y_i$, and $s_l$ is the status transition matrix, representing the score at which character $m_i$ is predicted to be label $y_i$. The calculation process is as follows:

$$p(y \mid m) = \frac{1}{Z(m)} exp\Big[ \sum\nolimits_{i,k} \lambda_k t_k(y_{i-1}, y_i, m, i) + \sum\nolimits_{i,l} \mu_l s_l(y_i, m, i) \Big] \tag{19}$$

$$Z(m) = \sum\nolimits_{y} exp\Big[ \sum\nolimits_{i,k} \lambda_k t_k(y_{i-1}, y_i, m, i) + \sum\nolimits_{i,l} \mu_l s_l(y_i, m, i) \Big] \tag{20}$$

During the training process, we use the maximum likelihood estimation to obtain the conditional probability, and use the Viterbi algorithm to predict the global highest scoring label sequence $y^*$. For simplicity, $w$ is used to represent the unified weight determined by $\lambda_k$ and $\mu_l$, and $F(y, m)$ is used to represent the global eigenvector determined by $t_k$ and $s_l$. The calculation formula is as follows:

$$y^* = argmax(w, F(y, m)) \tag{21}$$

## 4. Experimental Results and Discussion

### *4.1. Experimental Dataset*

In this paper, we choose the CMRC2018 dataset and the DuReader-checklist dataset to test our model. Among them, the CMRC2018 is a span-extraction machine reading comprehension dataset based on ground truth proposed by iFLYTEK and Harbin Institute of Technology. The data come from Wikipedia, and the data format is a triple <context, query, answer>. The data scale statistics are shown in Table 1.

**Table 1.** Scale statistics of the CMRC2018 dataset.

| | **Training Set** | **Validation Set** | **Test Set** |
|---|---|---|---|
| Question # | 10321 | 3351 | 4895 |
| Answers per Q | 1 | 3 | 3 |
| Max P tokens | 962 | 961 | 980 |
| Max Q tokens | 89 | 56 | 50 |
| Max A tokens | 100 | 85 | 92 |
| Avg P tokens | 452 | 469 | 472 |
| Avg Q tokens | 15 | 15 | 15 |
| Avg A tokens | 17 | 9 | 9 |

In order to test the prediction of the model for unanswerable questions, we conducted experiments on the DuReader-checklist dataset proposed by Baidu. The dataset is designed to use the evaluation system to evaluate the model ability, and contains a large number of unanswerable questions. The data format is a four-tuple <query, context, title, answer>, and the data scale statistics are shown in Table 2.

**Table 2.** Scale statistics of the DuReader-checklist dataset.

| | **Training Set** | **Test Set** |
|---|---|---|
| Question # | 3000 | 1130 |
| Answers per Q | 1 | 1 |
| Q with A | 1404 | 539 |
| Unanswerable Q | 1596 | 591 |

## *4.2. Evaluation Indicators*

Two commonly used evaluation indicators in reading comprehension tasks have been used: Exact Match value (EM) and Fuzzy Match score (F1-score), which are used to evaluate the exact match and the degree of match between the predictive value and the true value, respectively.

## *4.3. Experimental Parameter Settings*

In our experiments, AT-CRF Reader is implemented in the PyTorch framework. Four GeForce GTX 1080Ti 11GB GPUs are used for model training, and the Adam optimizer is used. The parameter settings used in the experiment are shown in Table 3.

**Table 3.** Training parameters.

| Parameter | Setting Value |
|---|---|
| Epoch | 3 |
| Batch size | 32 |
| Learning-rate | $3 \times 10^{-5}$ |
| Dropout | 0.01 |
| Max A tokens | 50 |
| Max-seq-length | 512 |

## *4.4. Experimental Results and Analysis*

### 4.4.1. Comparison Experiment

To verify the effectiveness of the proposed AT-CRF Reader, we conduct comparative experiments on the CMRC2018 and DuReader-checklist datasets to evaluate the overall performance of the model and the ability to deal with unanswerable questions, respectively. We selected and reproduced six state-of-the-art machine reading comprehension models: Bert-base(Chinese), ERNIE, RoBERTa-wwm-ext, ELECTRA, MacBERT-Chinese, and ChineseBERT, and compared them with the AT-CRF Reader to obtain the final results.

The experimental results on the overall performance of the models on the CMRC2018 dataset are shown in Table 4. It can be seen from the results that the AT-CRF Reader in this paper has an EM value of 65.52%, and an F1 score of 87.51%. Compared with the Bert-base(Chinese), ERNIE, RoBERTa-wwm-ext, ELECTRA, and MacBERT-Chinese, the EM values are increased by 4.45%, 2.61%, 1.65%, 1.09%, and 0.18%, and the F1 scores are increased by 2.65%, 0.99%, 1.15%, 0.78%, and 0.67%, respectively. In addition, the AT-CRF Reader proposed in this paper has almost the same performance as the ChineseBERT proposed in the 2021 ACL Conference, which confirms that the model proposed in this paper has certain performance advantages.

The experimental results of the performance of unanswerable questions on the DuReader-checklist dataset are shown in Table 5. The results show that the AT-CRF Reader achieves an EM value of 74.45%, and an F1 score of 87.39% on this dataset. Compared with Bert-base(Chinese), ERNIE, RoBERTa-wwm-ext, ELECTRA, and MacBERT-Chinese, the EM values are increased by 3.88%, 2.97%, 1.34% 0.63%, and 0.49%, and the F1 scores are increased by 2.68%, 1.62%, 0.81%, 0.56%, and 0.25%, respectively. In addition, the performance of the AT-CRF Reader proposed in this paper is basically the same as that of the ChineseBERT proposed in the 2021 ACL Conference, which shows that this model can not only predict correct answers, but also has a better response ability for unanswerable questions.

Table 4. The performance on the CMRC2018 dataset.

| Model | EM/% | F1/% |
|---|---|---|
| Bert-base(Chinese) | 61.07 | 84.86 |
| Ernie | 62.91 | 86.52 |
| RoBERTa-wwm-ext | 63.87 | 86.36 |
| ELECTRA | 64.43 | 86.73 |
| MacBERT-Chinese | 65.34 | 86.84 |
| ChineseBERT | 65.83 | 88.09 |
| AT-CRF Reader | 65.52 | 87.51 |

Table 5. The performance of the DuReader-checklist dataset.

| Model | EM/% | F1/% |
|---|---|---|
| Bert-base(Chinese) | 70.57 | 84.71 |
| Ernie | 71.48 | 85.77 |
| RoBERTa-wwm-ext | 73.11 | 86.58 |
| ELECTRA | 73.82 | 86.83 |
| MacBERT-Chinese | 73.96 | 87.14 |
| ChineseBERT | 74.62 | 87.76 |
| AT-CRF Reader | 74.45 | 87.39 |

### 4.4.2. Ablation Experiment

To demonstrate that, using the CRF layer as the output module can better handle unanswerable questions, we conducted a set of ablation experiments on the DuReader-checklist dataset. The results shown in Table 6 are achieved by the baseline model Bert-base(Chinese), the AT-CRF Reader without using CRF as the output layer (AT Reader), and the full AT-CRF Reader. It can be seen from the results that the AT Reader has improved the performance compared to the baseline model, with EM value and F1 score being improved by 2.75% and 2.4%, respectively. The performance has been further improved by 1.13% and 0.28% by the full model AT-CRF Reader, on the EM value and F1 score, respectively. This result shows that, using CRF instead of pointer networks as the output layer, can effectively improve the understanding ability of the model, and it is better to deal with unanswerable questions.

Table 6. The effect of CRF on the AT-CRF Reader.

| Model | EM/% | F1/% |
|---|---|---|
| Bert-base(Chinese) | 70.57 | 84.71 |
| AT Reader | 73.32 | 87.11 |
| AT-CRF Reader | 74.45 | 87.39 |

### 4.4.3. Other Experiments

To further analyze the model performance, we also conducted supplementary experiments on CMRC2018 and DuReader-checklist datasets. Figure 4 shows the performance plot of F1 score and EM value in the two datasets as the epoch change. It can be seen from the figure that, when the epoch is 5, F1 scores in the two datasets reach 87.51% and 87.39%, and the EM values reach 65.52% and 74.45%, respectively. After five iterations, the model performance gradually tends to be stable, Therefore, through the performance curve, it can be concluded that the model in this paper can converge to the optimal result with only a few iterations.

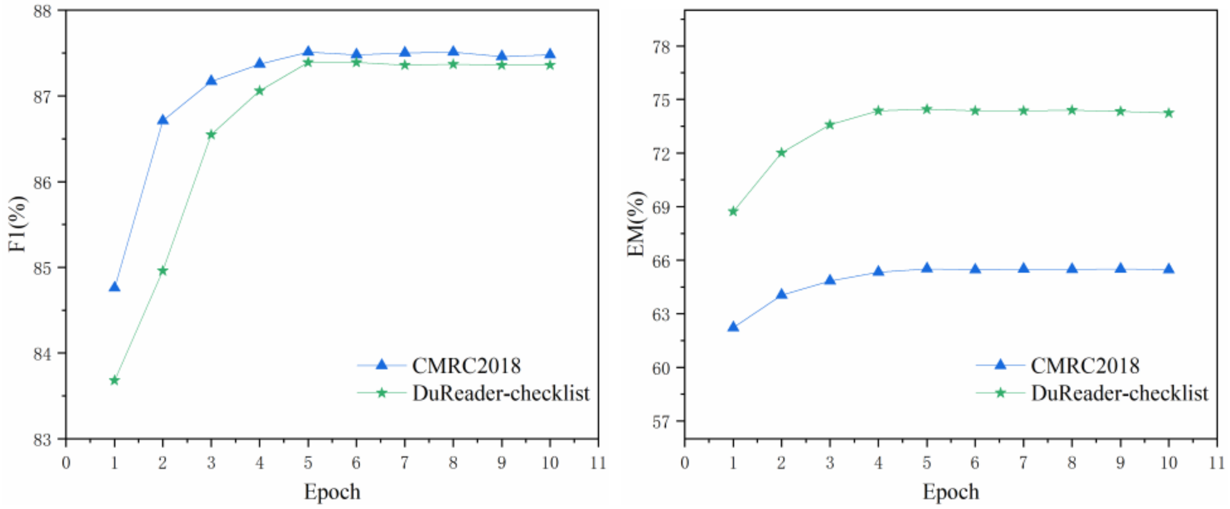


**Figure 4.** F1 scores and EM values on the two datasets under different training epochs.

In addition, we also designed a set of experiments to test the impact of passage lengths on the model. We chose an interval of (50, 550) with a step of 50 for the experiment, and plotted the curve of F1 score and EM value. The final results are shown in Figures 5 and 6. From Figure 5, it can be concluded that the model performs best on the CMRC2018 dataset with the highest F1 scores and EM values when the length of the passage is within the interval of 300–350. The same conclusion can be drawn from Figure 6 that the overall performance is optimal on the DuReader-checklist dataset when the model input passage length is within the 300–350 interval, with the highest F1 scores and EM values. In addition, the relative gap between the F1 scores and EM values of the two models, AT Reader and AT-CRF Reader, on the DuReader-checklist dataset is larger compared to that on the CMRC2018 dataset, which indicates that the model proposed in this paper is more applicable to the dataset with unanswerable questions.

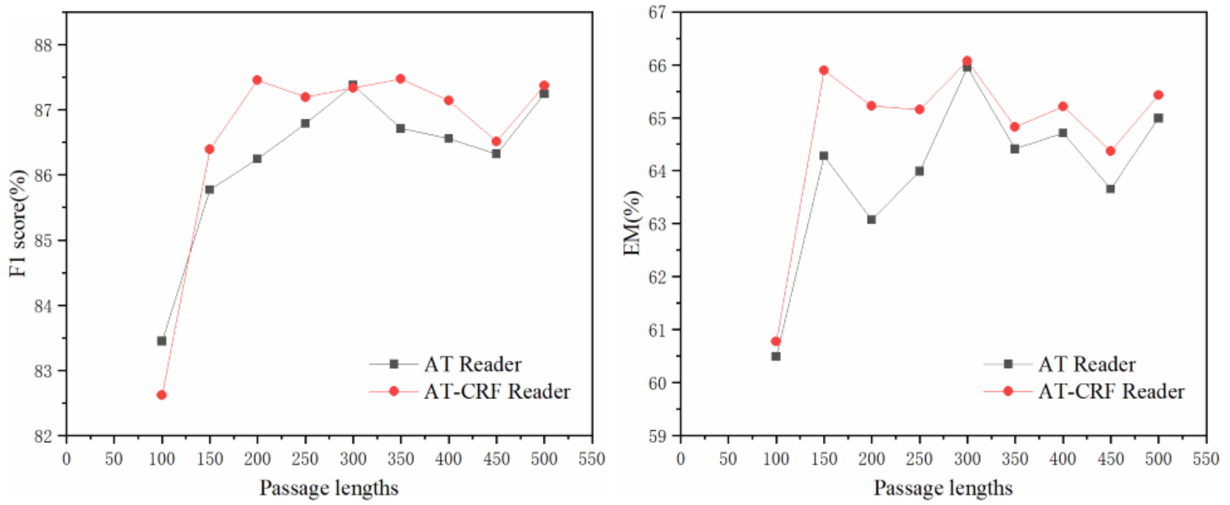


**Figure 5.** The effect of input length on model performance in the CMRC2018 dataset. In this experiment, only the input length parameter is changed, and other experimental parameters are kept constant.

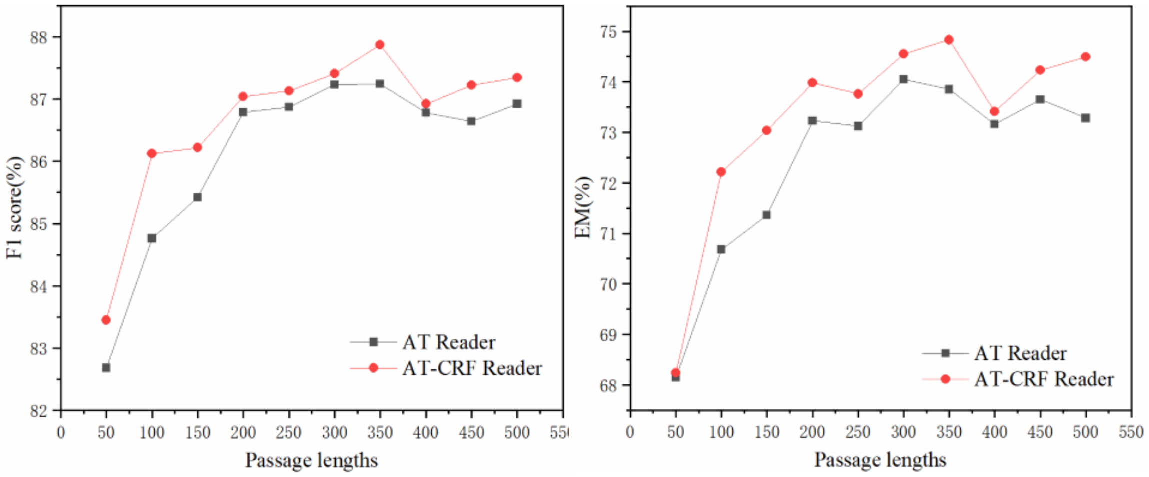


**Figure 6.** The effect of input length on model performance on the DuReader-checklist dataset. In this experiment, only the input length parameter is changed and other experimental parameters are kept constant.

## 5. Conclusions

Aiming at the problems of the traditional MRC model in Chinese machine reading comprehension tasks, such as information loss, lack of long-distance dependence ability and inability to deal with unanswerable questions, this paper proposes the AT-CRF Reader model. This model uses Chinese RoBERTa as an embedding process to fully obtain the underlying semantic features of the context, combines depthwise separable convolution neural networks and attention mechanisms to obtain long-distance dependency ability, and uses the transition score in the conditional random field to judge the relationship between adjacent labels, and then obtains the global optimal sequence as the answer output. To verify the effect of this model, we conduct experiments on two Chinese reading comprehension datasets, CMRC2018 and DuReader-checklist, and the results demonstrate the effectiveness of the AT-CRF Reader. In future research, the problem of excessive parameters caused by the introduction of pre-trained models can be solved by Knowledge Distillation, which is trained through a lightweight network. This method can effectively reduce the number of model parameters and improve the training speed.

**Author Contributions:** Conceptualization, N.S. and H.W.; methodology, N.S.; validation, N.S.; formal analysis, N.S.; investigation, N.S.; data curation, N.S.; writing—original draft preparation, N.S.; writing—review and editing, H.W. and Y.C.; visualization, N.S.; supervision, H.W. and Y.C.; funding acquisition, N.S. and H.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Southwest Minzu University Graduate Innovative Research Project Grant No. (CX2021SZ31, Master Program), the Sichuan Science and Technology Applied Foundation Project Grant No. (No. 20220016), the National Natural Science Foundation of China (Grant No. 72174172, 71774134), the Fundamental Research Funds for the Central Universities, Southwest Minzu University (Grant No. 2021101).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

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

## References

1. Liu, S.; Zhang, X.; Zhang, S.; Wang, H.; Zhang, W. Neural machine reading comprehension: Methods and trends. *Appl. Sci.* **2019**, *9*, 3698. [CrossRef]
2. Li, K.; Li, Y.; Lin, M. Review of Conversational Machine Reading Comprehension. *J. Front. Comput. Sci. Technol.* **2021**, *15*, 1607.
3. Kenton, J.D.M.W.C.; Toutanova, L.K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
4. Chi, P.; Chung, P.; Wu, T.; Hsieh, C.; Chen, Y.; Li, S.; Lee, H. Audio albert: A lite Bert for self-supervised learning of audio representation. In Proceedings of the IEEE Spoken Language Technology Workshop (SLT), Virtual, 19–22 January 2021; pp. 344–350.
5. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
6. Chen, D.; Ma, J.; Ma, Z.; Zhou, J. Review of Pre-training Techniques for Natural Language Processing. *Front. Comput. Sci. Technol.* **2021**, *15*, 31.
7. Yin, F.; Wang, Y.; Liu, J. Modeling multi-prototype Chinese word representation learning for word similarity. *Complex Intell. Syst.* **2021**, *7*, 2977–2990. [CrossRef]
8. Liao, W.; Huang, M.; Ma, P.; Wang, Y. Extracting Knowledge Entities from Sci-Tech Intelligence Resources Based on BiLSTM and Conditional Random Field. *IEICE Trans. Inf. Syst.* **2021**, *104*, 1214–1221. [CrossRef]
9. Cui, Y.; Liu, T.; Che, W.; Xiao, L.; Chen, Z.; Ma, W.; Wang, S.; Hu, G. A Span-Extraction Dataset for Chinese Machine Reading Comprehension. In Proceedings of the EMNLP/IJCNLP, Hong Kong, China, 3–7 November 2019.
10. He, W.; Liu, K.; Liu, J.; Lyu, Y.; Zhao, S.; Xiao, X.; Liu, Y.; Wang, Y.; Wu, H.; She, Q.; et al. DuReader: A Chinese Machine Reading Comprehension Dataset from Real-world Applications. In Proceedings of the QA@ ACL, Melbourne, Australia, 19 July 2018.
11. 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, WA, USA, 4 May 2000.
12. Hirschman, L.; Light, M.; Breck, E.; Burger, J.D. Deep read: A reading comprehension system. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics, College Park, MD, USA, 20–26 June 1999; pp. 325–332.
13. Chen, D.; Bolton, J.; Manning, C.D. A thorough examination of the cnn/daily mail reading comprehension task. *arXiv* **2016**, arXiv:1606.02858.
14. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. Squad: 100,000+ questions for machine comprehension of text. *arXiv* **2016**, arXiv:1606.05250.
15. Rajpurkar, P.; Jia, R.; Liang, P. Know what you don't know: Unanswerable questions for SQuAD. *arXiv* **2018**, arXiv:1806.03822.
16. Cui, Y.; Chen, Z.; Wei, S.; Wang, S.; Liu, T.; Hu, G. Attention-over-attention neural networks for reading comprehension. *arXiv* **2016**, arXiv:1607.04423.
17. Wang, S.; Jiang, J. Machine comprehension using match-lstm and answer pointer. *arXiv* **2016**, arXiv:1608.07905.
18. Seo, M.; Kembhavi, A.; Farhadi, A.; Hajishirzi, H. Bidirectional attention flow for machine comprehension. *arXiv* **2016**, arXiv:1611.01603.
19. 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 (Volume 1: Long Papers), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 189–198.
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
21. Cui, Y.; Liu, T.; Chen, Z.; Wang, S.; Hu, G. Consensus attention-based neural networks for Chinese reading comprehension. *arXiv* **2016**, arXiv:1607.02250.
22. Lai, Y.; Tseng, Y.; Lin, P.; Hsiao, V.; Shao, C. D-Reader: A Reading Comprehension Model by Full-text Prediction. *J. Chin. Inf. Process.* **2018**, *32*, 135–142.
23. Cui, Y.; Che, W.; Liu, T.; Qin, B.; Yang, Z. Pre-training with whole word masking for chinese bert. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3504–3514. [CrossRef]
24. Zhang, Z.; Han, X.; Liu, Z.; Jiang, X.; Sun, M.; Liu, Q. ERNIE: Enhanced Language Representation with Informative Entities. *arXiv* **2019**, arXiv:1905.07129.
25. Clark, K.; Luong, M.T.; Le, Q.V.; Manning, C.D. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv* **2020**, arXiv:2003.10555.
26. Cui, Y.; Che, W.; Liu, T.; Qin, B.; Wang, S.; Hu, G. Revisiting pre-trained models for Chinese natural language processing. *arXiv* **2020**, arXiv:2004.13922.
27. Guo, S.; Guan, Y.; Tan, H.; Li, R.; Li, X. Frame-based neural network for machine reading comprehension. *Knowl.-Based Syst.* **2021**, *219*, 106889. [CrossRef]

28. Sun, Z.; Li, X.; Sun, X.; Meng, Y.; Ao, X.; He, Q.; Wu, F.; Li, J. Chinesebert: Chinese pretraining enhanced by glyph and pinyin information. *arXiv* **2021**, arXiv:2106.16038.
29. Chopra, D.; Morwal, S.; Purohit, G. Hidden markov model based named entity recognition tool. *Int. J. Found. Comput. Sci. Technol. (IJFCST)* **2013**, *3*, 67–73.
30. Xie, C.; Chen, D.; Shi, H.; Fan, M. Attention-Based Bidirectional Long Short Term Memory Networks Combine with Phrase Convolution Layer for Relation Extraction. In Proceedings of the 2021 5th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI), Kelaniya, Sri Lanka, 6–7 December 2021; pp. 1–6.