applied sciences

MDPI

Article

A Study of Entity Relationship Extraction Algorithms Based on Symmetric Interaction between Data, Models, and Inference Algorithms

Ping Feng [1,2,3,4,5], Nannan Su [6], Jiamian Xing [6], Jing Bian [2] and Dantong Ouyang [1,*]

[1] College of Computer Science and Technology, Jilin University, Changchun 130012, China; fengping@ccu.edu.cn
[2] College of Computer Science and Technology, Changchun University, Changchun 130022, China; ccbianjing@126.com
[3] Ministry of Education Key Laboratory of Intelligent Rehabilitation and Barrier-Free Access for the Disabled, Changchun 130022, China
[4] Jilin Provincial Key Laboratory of Human Health State Identification and Function Enhancement, Changchun 130022, China
[5] Jilin Rehabilitation Equipment and Technology Engineering Research Center for the Disabled, Changchun 130022, China
[6] College of Cybersecurity, Changchun University, Changchun 130022, China; a1414301021@gmail.com (N.S.); 18332765510@163.com (J.X.)
[*] Correspondence: dantong1206@163.com; Tel.: +86-13504313045

Abstract: The purpose of this paper is to address the extraction of entities and relationships from unstructured Chinese text, with a particular emphasis on the challenges of Named Entity Recognition (NER) and Relation Extraction (RE). This will be achieved by integrating external lexical information and utilizing the abundant semantic information available in Chinese. We utilize a pipeline model that is applied separately to NER and RE by introducing an innovative NER model that integrates Chinese pinyin, characters, and words to enhance recognition capabilities. Simultaneously, we incorporate information such as entity distance, sentence length, and part-of-speech to improve the performance of relation extraction. We also delve into the interactions among data, models, and inference algorithms to improve learning efficiency in addressing this challenge. In comparison to existing methods, our model has achieved significant results.

Keywords: named entity recognition; relationship extraction; data; model; inference algorithm

check for updates

Citation: Feng, P.; Su, N.; Xing, J.; Bian, J.; Ouyang, D. A Study of Entity Relationship Extraction Algorithms Based on Symmetric Interaction between Data, Models, and Inference Algorithms. *Appl. Sci.* **2024**, *14*, 1058. https://doi.org/10.3390/app14031058

Academic Editors: Xianglong Liu, Shan An, Yue He and Yuqing Ma

Received: 30 December 2023
Revised: 19 January 2024
Accepted: 23 January 2024
Published: 26 January 2024

1. Introduction

Extracting entities and relations from unstructured text is a fundamental task in natural language processing, which is typically divided into two subtasks: named entity recognition (NER) and relation extraction. There are two main approaches in which a pipelined approach is employed to perform named entity recognition and relation extraction on text separately. The other is to model the two tasks jointly. In this study, we use the pipeline model to model a named entity recognition model and a relationship extraction model, respectively, and achieve good results on the relevant datasets, respectively. Information extraction is an important part of the natural language processing task, which is the process of converting unstructured textual information into valuable and organized structured information. Among them, Named Entity Recognition (NER), Entity Relationship Extraction (RE), and Event Extraction are the main tasks of information extraction. Named entity recognition is the basis of complex NLP tasks, which aim to identify entity elements in a text and categorize them into pre-defined entity classes [1]. In general, named entity recognition can be implemented in three different ways: rule-based methods, statistical machine learning-based methods, and deep learning-based methods [2]. Entity relationship extraction is currently used more often based on deep learning models. Deep learning is currently more successful for two reasons: on the one hand neural networks can memorize the trained model, and on the other hand deep learning can further discover deep features [3]. The essence is to migrate the weights of a multilayer network to other networks so that other networks can learn the features of this network. Currently, research in Chinese named entity recognition predominantly focuses on character-based approaches. However, when dealing with intricate entity relationships, the task of Chinese relation extraction confronts formidable challenges.

Currently, research on Chinese named entity recognition and relation extraction primarily focuses on character-based or word-based approaches. However, when dealing with intricate entity relationships, the Chinese entity relation extraction task faces significant challenges.

How to introduce vocabulary beyond entity characters:

The semantics contained in Chinese sentences are composed of words, meaning that Chinese words contain a vast amount of information [4]. If word embeddings are learned through large unlabeled text corpora, the impact of encountering unknown words in the text corpus will be significantly reduced. Introducing external vocabulary further helps to better address the issue of unclear lexical boundaries.

How to better utilize the rich semantic information in Chinese:

In the field of deep learning, models based on characters, lexical properties, sounds, etc., are extensively studied to improve the accuracy and generalization of relation extraction. Fully leveraging semantic information aids in a better understanding of relationships between entities.

Consistency of symmetry with data distribution:

In academic discussions, ensuring the consistency of symmetry within the triad of data, model, and inference algorithm with the symmetry of data distribution is crucial for achieving optimal learning efficiency. Addressing this issue requires in-depth exploration of the relationships between these symmetries and the development of more flexible models to enhance learning efficacy.

The main contributions of our study are outlined as follows:

1. Introduction of a novel NER model combining Chinese pinyin, characters, and words to enhance recognition capabilities;
2. The fusion of entity distances, sentence lengths, and part of speech improves the performance of the relation extraction model;
3. Exploration of the interaction between data, models, and inference algorithms to investigate synergies and symmetry in deep learning, resulting in enhanced learning efficiency.

As the introduction concludes, we provide an outline of the paper's structure for clarity. Specifically, Section 2 delves into related works, encompassing the current landscape of information extraction research, pre-trained models, decoding the classification layer, and previous work on the (D, M, I) triplet. In Section 3, we expound on the proposed models, detailing the named entity recognition model based on characters, words, and pinyin, and the relation extraction model based on multiple feature binding. Section 4 presents the experimental results and analysis, covering the performance evaluation of the Mengzi-BiLSTM-Crf and BERT-BiGRU-Attention models. Subsequently, Section 5 provides an in-depth analysis, focusing on the performance of both NER and relation extraction models. Finally, Section 6 synthesizes the study and discusses prospects for future research.

2. Related Work

2.1. Research Status of Information Extraction

After statistical machine learning algorithms, named entity recognition tasks have entered a new era and are now being integrated with neural networks. Researchers started to use recurrent neural networks (RNN) based on character embedding and word embedding to identify named entities in sentences, solving the feature engineering problem that exists in conventional statistical methods. With the rapid development of deep learning, using deep learning methods to solve named entity recognition problems has become a popular research topic. The advantage of this class of methods is that the neural network model can automatically learn sentence features without complex feature engineering [5]. In word vector-based processing, the text is first divided into words, and then the words are mapped into the form of vectors so that there is a feature vector for each word. The main problems faced by this approach are mainly word division errors leading to named entities being incorrectly sliced, large dimensionality of the lexicon, and the inability to effectively characterize the correlation between similar words [6]. However, the advantages are mostly the word boundary information and semantic features. Bidirectional Long Short-Term Memory (BiLSTM) in NER achieves comprehensive sequence modeling by considering the context of each word, thereby improving the accuracy of named entity recognition. The model possesses the capability to extract sequence features, handle long-distance dependencies, and generate dynamic sequence representations, providing flexibility to the model. By learning advanced language features, BiLSTM performs exceptionally well in NER tasks and has become an effective sequence labeling, similarly, in the early days, to named entity recognition. Later on, feature-based methods are used to classify relationships using machine learning methods with input entities and corresponding textual, syntactic, or language-related features. The effectiveness of the aforementioned methods is greatly influenced by the quality of the features derived from the available natural language processing tools [7]. The emergence of neural networks has introduced a new direction for feature extraction. Many neural networks, including recurrent neural networks, convolutional neural networks (CNN), graph neural networks, and long- and short-term memory networks, are applied in named entity recognition tasks and relationship extraction tasks. In natural language processing (NLP) tasks, text input is usually represented as a sequence of information, and RNN are suitable for processing sequence information and are widely used in NLP tasks. Graph neural networks are suitable for exploiting the relationship between points and point sets [8]. This enables them to effectively capture the structural information of text, and they have also received attention in natural language processing tasks in recent years. Daojian Zeng et al. [9] introduced the PCNN (segmental convolutional neural network) approach. In the case of relatively small training data, a multi-sample learning approach is employed to allow the presence of mislabeling in the tagged sentences or examples. Additionally, a segmented maximum pooling convolutional architecture is used to automatically learn the relevant features.

2.2. Pre-Trained Model

The pre-trained language model is based on the transformer architecture, such as BERT [10]. BERT has shown outstanding performance in Chinese sequence labeling tasks. In recent studies, researchers have adopted different strategies to exploit the advantages of BERT. Yang [11] simply added a softmax layer on top of BERT, achieving state-of-the-art performance in Chinese Word Segmentation (CWS) tasks. This indicates the significance of BERT's contextual representation for Chinese word segmentation tasks, and a simple classification head is sufficient to bring significant performance improvements. Additionally, research by Meng et al. [12] has shown that utilizing character features in BERT can significantly surpass methods based on static embeddings, particularly in Chinese Named Entity Recognition (NER) and Chinese Part-of-Speech (POS) tagging tasks. This suggests that BERT not only improves overall performance but also that its character-level information is valuable for structured tasks in Chinese text. Additionally, the utilization of external lexical features to augment character features has been demonstrated to be effective, as exemplified by SoftLexicon [13] and LEBERT [14]. In SoftLexicon, lexical features are merged into character representations, avoiding the necessity for complex architectures to integrate lexical features. LEBERT is based on this approach, integrating lexical features into the lower BERT layers. Lower BERT layers facilitate a more hierarchical interaction between lexical features and BERT. The main idea of the LEBERT is to integrate contextual representations derived from BERT and lexical features into a Chinese NER model.

This paper modifies the embedded model of LEBERT, replacing it with the lightweight and more powerful Mengzi model [15]. Compared to models of the same size or even larger dimensions, it shows significant performance improvements. Modifying the details of pre-trained models and fine-tuning strategies has been found to be effective in improving the baseline results. Without modifying the underlying model architecture, the Mengzi model is one of the advanced Chinese pre-trained language models currently.

2.3. Decode the Classification Layer

The current mainstream classifiers are mainly the Conditional Random Field (CRF) decoding classifier [16] and softmax decoding classifier [17]. CRF can obtain a globally optimal sequence of tags by learning the dependencies between tags and the constraints in sentences. CRF (Conditional Random Fields) is an undirected graph model for modeling the joint probability distribution of a sequence or grid data consisting of some random variables and is commonly applied in sequence annotation, natural language processing, and computer vision, etc. In the sequence labeling task, we need to label each position in a given input sequence. This task can be translated into modeling the annotation of each position as a sequence of random variables. CRF represents the interdependence between these variables by building a probabilistic graphical model between them; i.e., given the annotation of one position, the conditional probabilities of the annotations of other positions are modeled. The CRF model builds a global joint probability distribution by considering the annotations of neighboring positions. The training process of the CRF model is usually performed using the maximum likelihood estimation method. That is, the weights of each feature function are learned by maximizing the likelihood function of the training data. In summary, the CRF model describes the generation process of annotated sequences by establishing a global joint probability distribution, and it models the interaction between annotated and observed sequences through the feature functions.

In addition, the principle of softmax is to map the input to the probability of named entity classification labels. The maximum probability classification label corresponding to each state is obtained by a greedy algorithm, which is applicable to multi-classification problems. Its implication lies in the fact that the softmax function acts as a classifier, assigning a probability value to each output classification result, reflecting its likelihood of belonging to each category instead of just determining a certain maximum value. In this paper, the Conditional Random Field (CRF) is used as a decoding layer placed after the BiLSTM to improve the accuracy of the model prediction.

2.4. (D, M, I) Triplet

Lechao Xiao et al. [18] proposed an integrated system that considers machine learning as data, models, and inference algorithms. The relationship between the performance and consistency of the triplet is studied from a symmetry perspective. Learning is most effective when the algorithm is consistent with the inherent distribution of the data. This symmetry allows the algorithm to recognize relevant patterns and relationships in the data, thereby improving performance and generalization. There exist three main components:

1. Data are the basis of machine learning models, and the effectiveness of a model depends heavily on the quality and quantity of data used. Data augmentation is particularly important in natural language processing tasks because the complexity and diversity of natural language data prevent models from fully understanding the true meaning of language. Choosing high-quality word embeddings is advantageous for high-performance natural language processing models. High-quality word embeddings can improve the model's accurate grasp of words and increase the model's adaptability to novel data sources. Moreover, the number of word embeddings has a significant impact on the performance of the model, and more word embeddings usually allow the model to learn more features. Nevertheless, it is not enough to rely on more word embeddings, but the quality of the word embeddings is also important. If the embeddings contain errors, noise, or inconsistencies, these issues will be learned

by the model and will affect the performance of the model. Hence, to build a high-performance natural language processing model, it is necessary to select a suitable word embedding dataset and carefully clean and process the data to ensure that both its quality and quantity of the data can maximize the performance of the model;

2. Designing machine learning models that maximize data efficiency (M) is crucial in solving the problem of natural language processing tasks. In the modeling framework, model selection is important, and determining a suitable model for the task requires consideration of several aspects. In this paper, we aim to analyze the advantages and disadvantages of neural network models in the LSTM model and GRU model that are suitable for named entity recognition and relation extraction tasks. Compared with the LSTM model, the GRU model has fewer parameters and a faster training speed, but the LSTM performs better when dealing with longer sequences. Therefore, symmetry and synergy between data and models need to be considered when selecting a model that is more suitable for the given task;
3. The inference process (I) refers to the process in machine learning methods that can perform learning. A strong performance in machine learning performance may come from the symmetric role between (M, I), (D, I), or (D, M, I). Different inference algorithms need to be chosen for different tasks. In this study, we can see experimentally that various inference algorithms have different performance in the case of different models.

3. Overview of the Model

3.1. Named Entity Recognition Model Based on the Combination of Characters, Words, and Pinyin

The model employed in this paper utilizes character-level input, which means that some features are influenced by ambiguous boundaries. To address this issue, the paper uses pre-trained word embeddings to augment the accuracy of named entity recognition. By introducing word embeddings, the model can better capture semantic relationships between words and comprehend words in context. The Chinese named entity recognition model is based on the LEBERT model. It integrates semantic information of over 12 million Chinese words pre-trained on a large-scale, high-quality dataset [19]. Additionally, it utilizes unique pinyin vectors of Chinese characters. It uses a BiLSTM to make predictions on character sequence labels and finally optimizes the predicted labels using CRF. The model demonstrates better performance compared to the original model on multiple Chinese named entity recognition datasets, showing significant improvements in recognition rates for diverse categories and heightened precision for each label. The model is illustrated in Figure 1.

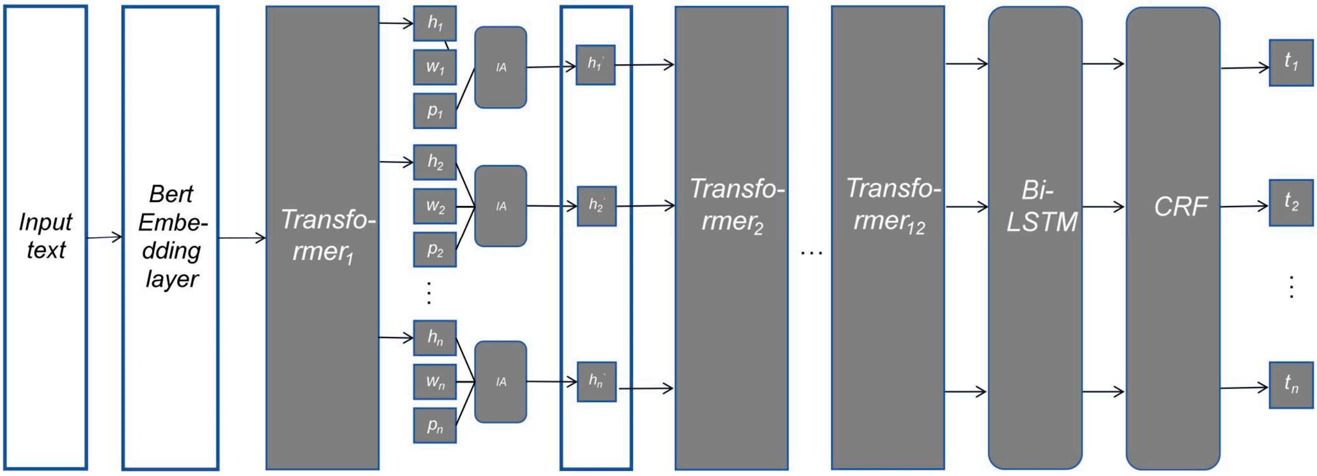

Figure 1. Named entity recognition model based on the combination of characters, words, and pinyin.

In the embedding of pinyin representation, this paper adopts a similar approach to the generation of pinyin embeddings as ChineseBERT. Firstly, the open-source pypinyin package is used to obtain the pinyin sequence for each character. Then, for each character's pinyin, completion is applied. A CNN model with a width of 2 is utilized to process the sequence, and the resulting sequence embedding is derived through max-pooling. This guarantees that the output dimension is not affected by the input sequence length. The input sequence length is fixed at 8. When the length of the sequence is less than 8, the remaining slots are filled with the special character "-". This is illustrated in Figure 2.

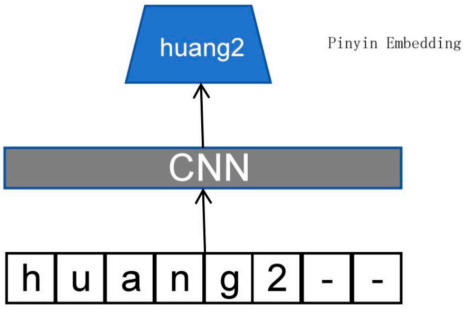

Figure 2. Generation of pinyin embedding.

For the given input text, it undergoes the initial processing step where it is fed into the embedding layer of BERT. This layer generates token embeddings, position embeddings, and segment embeddings for the sentence. Subsequently, the sentence is fed into the initial transformer layer of BERT, which yields the partial word features $h_1, h_2, \ldots h_n$. Next, the word set embedding vector $w_i = \{w_{i1}, w_{i2}, \ldots w_{im}\}$ is obtained, where w_i represents the word embedding vector set for the i-th character. The words embedding w_{i1}, w_{i2}, and so on are derived from the pre-trained word embedding lookup table of Tencent AI Lab Embedding Corpus. Additionally, there is a discrepancy between the dimensions of the word embeddings in the lookup table and those of the model. To ensure the alignment between the word vectors w_{ij} and the character vectors h_i, a non-linear transformation of the word vectors is required:

$$x_{ij} = W_2\big(tanh\big(W_1 w_{ij} + b_1\big)\big) + b_2$$

where W_1 is a d_c by d_w matrix, W_2 is a d_c by d_c matrix, and b_1 and b_2 are scalar bias. $\mathrm{d_c}$ represents the hidden size of BERT, and d_w represents the dimension of word embedding. The above parameters, together with the tanh function, implement the transformation of word embedding. In order to combine the lexical features and pinyin features with character information, inspired by LEBERT, we designed a novel information adapter (as shown in Figure 3).

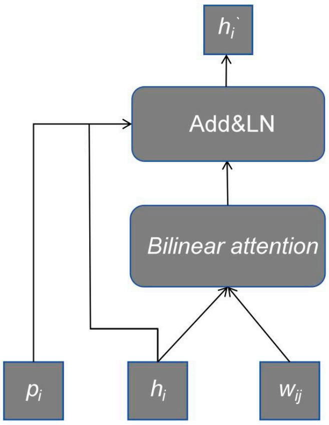

Figure 3. Information adapter.

The information adapter uses attention mechanism to integrate character features and lexical features into BERT, and then adds pinyin information to the adapter. We input the pinyin embedding vector p_i, the character embedding vector h_i, and the word embedding vector of the word into the information adapter (IA) layer. To calculate the relevance weight of each matching word, we introduce LEBERT's character-to-word attention mechanism as a reference. Specifically, we represent the i-th character vector as h_i and the set of word embedding vectors corresponding to the i-th character as x_i, where x_{ij} represents the embedding representation of the j-th word corresponding to the i-th character (aligned). Assign all x_i to h_i , as $x_i = \{x_{i1}, x_{i2}...x_{im}\}$, with a size of m by d_c. The method for calculating the relevance of each word is as follows:

$$A_i = softmax\left(h_i W_{attn} x_{ij}^T\right)$$

where W_{attn} is the weight matrix of bilinear attention, and $A_i = \{\alpha_{i1}, \alpha_{i2}, \ldots \alpha_{in}\}$ is a vector of weights for each word. Then, we can obtain the weighted sum of the vocabulary features as follows:

$$h_i' = \sum_{j=1}^{m} \alpha_{ij} x_{ij}$$

Weighted word embeddings and pinyin features are injected into the character vectors through the following methods:

$$h_i' = h_i + h_i' + p_i$$

The integration of vocabulary, character, and pinyin features is accomplished by utilizing dropout layers and layer normalization. The information obtained is processed through a series of 11 transformer blocks in order to derive the final output of BERT. The aforementioned output is subsequently fed into a BiLSTM to acquire the sequence features of the input vectors. Finally, the ultimate prediction results are obtained through CRF.

3.2. Relational Extraction Model Based on Multiple Feature Binding

For one of the relationship extraction tasks, this paper obtains more semantic feature information by BERT encoder in order to make full use of the information contained in Chinese text based on the BERT model to improve the accuracy of relationship extraction. This paper classifies the textual relations based on a BERT-based Chinese pre-trained model. After analyzing the data in the dataset, the relationship of most sentences is contained between two entities. If two entities are close to each other and there is no relationship in between, then the relationship may be hidden among the second entity behind. For example, in a news article that mentions a relationship between two people, the names of these people may be closely related, or there may be some words spaced apart. By adding the distance information between two entities, the location of the entity relationship information can be better determined. Therefore, the distance between entities (expressed in number of words or characters, for example) can be used as a feature to help the model determine the relationship between them. Second, the relationship between two entities is usually related to the length of the sentence. Length features can help identify differences between texts of different lengths. For example, in shorter texts, relationships may be easier to identify because there is less other information in the text, while in longer texts, relationships may be more difficult to identify because there is more interfering information. Therefore, in many cases, length features can provide useful clues to help the model perform relationship extraction more accurately. The lexical properties of Chinese often contain rich information features. In this paper, we divide the sentences into words and extract the lexical properties by THULAC [20]. (THULAC is an efficient Chinese lexical analysis toolkit.) These properties are then compared with other lexical properties, such as adjectives and prepositions. Nouns contain relationships with a greater possibility. In general, nouns or verbs are more likely to contain relations compared to other word types. For example, Liu Bei, Guan Yu, and Zhang Fei swore brotherhood, becoming brothers.

Using THULAC for part-of-speech tagging results in the following (Liu Bei\noun, \w Guan Yu\noun, and\c Zhang Fei\noun swore brotherhood\verb, \w becoming\verb brothers\noun). Here, "swore brotherhood" is a verb, and "brothers" is a noun, indicating a relationship between each pair of entities. Finally, multiple features are stitched together and fed into the model. The experimental results show that the F1 values of the method are improved on several Chinese relationship extraction datasets, and the recognition rates of relationships for different categories are greatly improved, respectively. In addition, the model obtains more features by introducing word vectors that have been externally passed, where the word vectors are pre-trained. The word vector of the *i*-th word is V_i^w, which is obtained by looking up the table using the pre-trained word vector, and the other feature vectors are V_i^{wj}, where V_i^{wj} represents is the *j*-th feature. The final word representations are stitched together as word vectors and feature vectors as follows:

$$X_i = \left[V_i^w; V_i^{w1} \ldots; V_i^{wm} \right]$$

Among them, m means that in addition to the word vector, there are m kinds of features. In this paper, $m = 3$. For sentence-level representation, directly input the word representation into BiLSTM for encoding, use F (forward) and B (backward) to represent the two directions. h_i (hidden) and c_i (cell) represent hidden information and global information; then the output at the *i*-th moment is:

$$F_i = [\, F_{hi}; F_{ci}; B_{hi}; B_{ci}]$$

Sentence-level features and vocabulary-level features are extracted and concatenated to form the final extracted feature vector. Lexical-level features mainly focus on Entity 1 (N1) and Entity 2 (N2). In this paper, the vectors obtained from feature embedding and BiGRU are concatenated to represent the two entities as $[X_{n1}, F_{n1}, X_{n2}, F_{n2}]$. Sentence-level features focus on contextual information, which is constructed from the output of BiGRU layers, as shown in Figure 4.

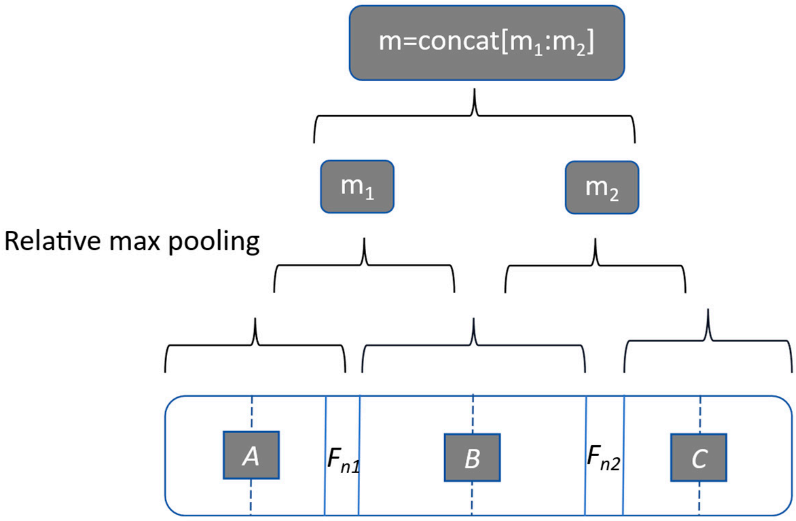

Figure 4. Construction of sentence-level feature vectors.

The framework diagram of the model is shown in Figure 5. The matrix obtained from BiGRU can be divided into three parts, A, B, and C, consisting of n1 and n2. The vectors and B and C parts are extracted from the maximum pooling operation matrix. The vectors m1 and m2 are concatenated to form a representation of the output, which later enters the attention layer for weighted summation, filtering out the redundant information from the large amount of information through the attention mechanism, making the model more focused on the needed information so that it can improve the model's ability to focus on more important features [21].

In this paper, we use BERT-based Chinese trained on a Chinese dataset. BERT is pre-trained with a large amount of unlabeled data, which allows for a better understanding of the linguistic context in the text. This helps to identify specific relationships between entities in the relation extraction task, such as subject–predicate relations, object relations, etc. The model learns rich linguistic representations during pre-training and can represent entities and relations as high-quality vector representations. These vector representations can be used for a variety of relationship extraction tasks, such as relationship classification, relationship extraction, etc. As the pre-trained model is trained on large-scale data, rich linguistic representations and linguistic knowledge are learned. This allows the model to learn new types of relations without large amounts of labeled data.

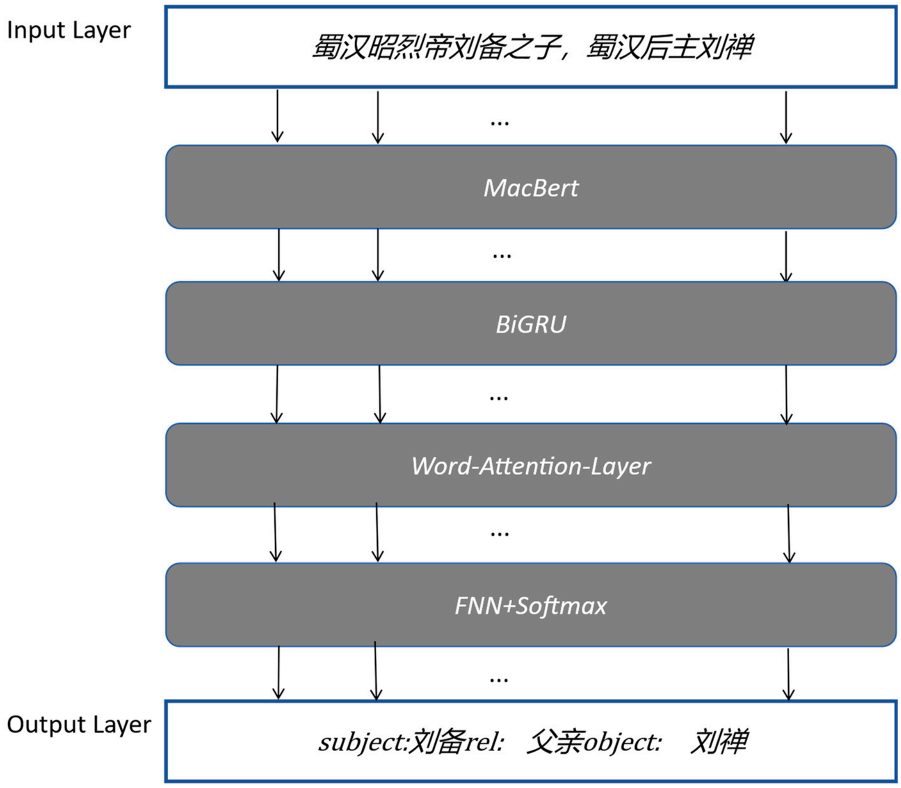

Figure 5. Relationship extraction model framework diagram. The Input Layer in the figure contains a Chinese sentence, which translates to: "The later lord Liu Shan of Shu Han is the son of Emperor Zhao Lie, Liu Bei". In the Output Layer, the meaning of the triplet is as follows: subject: Liu Bei (personal name), rel: father relationship, object: Liu Shan (personal name).

4. Results and Analysis of the Experimental Results

4.1. Mengzi-BiLSTM-Crf

For the named entity recognition task, three Chinese named entity recognition datasets are used in this paper to verify the effectiveness of the model. Each of these datasets is sequentially classified as training set, validation set, and test set. The performance of the model in the named entity recognition task is evaluated using the F1 score (F1) as an evaluation metric and is compared with other mainstream Chinese named entity recognition models.

4.1.1. Datasets

This article uses three commonly used Chinese named entity recognition datasets: Weibo [22], Resume [23], and MSRA [24]. The Weibo dataset's corpus comes from social media and includes four entity types: individuals, locations, organizations, and geopolitical entities. The MSRA-NER dataset is released by Microsoft Research Asia and mainly includes personal names, place names, and organization names. Resume is a dataset of resumes of senior executives listed on the Chinese stock market, including personal names, place names, and positions.

4.1.2. Experimental Environment

The experiments were built based on the PyTorch, and the environment was configured according to the configuration in Table 1.

Table 1. Named entity recognition experimental environment.

Operating System	Version
CPU	AMD Ryzen 5 5600X 6-Core Processor 3.70 GHz
GPU	3060 Ti
Python	3.6
Pytorch	1.7.1

4.1.3. Experimental Parameter Configuration

In this paper, the parameters of the named entity recognition model are tuned by referring to several related studies, and the parameters are finally set to the values shown in Table 2.

Table 2. Named entity recognition experiment parameters.

Parameters	Max Length	Crf_lr	Adapter_lr	Epoch	Batch Size	Loss_Type	Learning Rate
values	150	1×10^{-4}	1×10^{-4}	20	12	Ce	1×10^{-4}

In order to more effectively prevent the instability of the results and cause overfitting, the output of the fully connected layer is set to dropout = 0.5 in this work.

4.1.4. Experimental Result

The experimental results in Table 3 show that the entity model proposed in the paper performs substantially better than the comparison models in the three datasets of Resume, Weibo, and MSRA in terms of comprehensive performance. The smaller the dataset is, the more obvious the effect improvement is.

Table 3. Experimental results' F1 value in multiple Chinese named entity recognition datasets.

Model	Resume	Weibo	MSRA
		F1	
BERT	95.33	67.27	94.71
ERNIE	94.82	67.96	95.08
ChineseBERT	/	70.80	/
MFE-NER	95.73	67.74	89.96
LeBERT	96.08	70.75	95.70
LeBert-Bilstm	**96.61**	**73.72**	**95.86**

The table shows the experimental results of multiple models on different Chinese named entity recognition datasets. As shown in the table, the F1 scores improve in all Three datasets, which indicates that combining word vectors and pinyin vectors to improve the accuracy of named entity recognition is effective. The first five rows list the variant models based on the BERT model, of which the first row is the BERT baseline model. The second row is Baidu's optimization model ERNIE based on the BERT model in April 2019 [25]. Next is ChineseBERT [26], which integrates the glyph information and pinyin information of Chinese characters. It integrates the connection between Chinese characters, glyphs, pronunciation, and context. The glyph vector is composed of several different structures, and the pinyin vector is obtained from the corresponding sequence of romanized pinyin characters. The two are fused together with the word vector to obtain the final fusion vector as the input to the pre-trained model. The fourth row follows a similar approach of merging glyph and pinyin, but MFE-NER employs a distinctive method for handling fonts. It captures glyph features and utilizes the "Wubi" encoding scheme to represent the structural patterns of Chinese characters, enhancing the proximity of characters with similar glyph structures in the embedding space. Lastly, the original LEBERT model incorporates lexical information into the model, enhancing named entity recognition capabilities. Building upon LEBERT, this study introduces modifications and compares the performance with several models, including BERT, ChineseBERT, and the enhanced LEBERT. The results demonstrate the efficacy of combining Chinese language models and dictionary features for Chinese named entity recognition. In this paper, we improve the LEBERT model based on it, compare it with several models for verification, and achieve better performance. These include BERT, ChineseBERT, and the improved LEBERT, which proves the effectiveness of combining Chinese speech models and lexical features for Chinese named entity recognition. Compared with the LEBERT model, this paper has a large improvement in entity recognition for different categories. The F1 values for entity recognition in the Weibo dataset for different categories are shown in Figure 6.

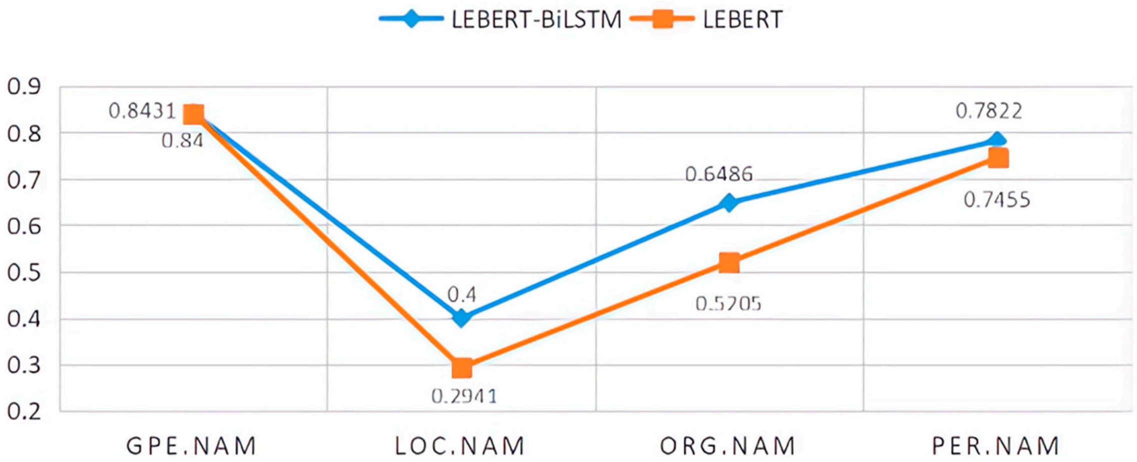

Figure 6. Named entity F1 value of different tags under Weibo dataset.

As shown in Figure 6, the recognition effect of the proposed model in this paper is significantly improved in several entity classifications, especially for specific location names and specific organization names.

4.1.5. Ablation Experiment

Through Table 4, it has been observed that in the context of BiLSTM-CRF, the BiLSTM sequentially models the vectors encoded by the pre-trained model for the input sequence, effectively capturing continuous sequential information. This type of continuous sequential information proves to be beneficial for entity recognition, resulting in favorable outcomes in the dataset.

Table 4. Experimental results of different models under the Weibo dataset.

Models	F1
LEBERT-CRF	70.75
LEBERT-LSTM-CRF	71.95
LEBERT-BiGRU-CRF	71.17
LEBERT-BiLSTM-Softmax	72.61
LEBERT-BiLSTM-CRF	**73.72**

4.2. BERT-BiGRU-Attention

In this approach, we begin by encoding the input text utilizing the BERT model, resulting in a vector sequence. Subsequently, this sequence serves as the input for the Bi-GRU model, generating the output of the Bi-GRU model. The output of the Bi-GRU model is then weighted using an attention layer to produce a weighted sum as the final output vector. Finally, the output vector is fed into a linear layer and mapped to generate the probabilities associated with each relationship.

To verify the effectiveness of the model, experiments are conducted on three Chinese relationship extraction datasets in this paper. The paper is divided into training set, testing set, and validation set according to their original proportions, and the experiments are used to evaluate the results of the model in named entity recognition and compare it with other Chinese relationship extraction models using precision, recall, and F1 score (F1) as evaluation metrics.

4.2.1. Datasets

In this paper, two Chinese relationship extraction datasets are used. They are the Duie Chinese Relation Extraction Dataset [27] and the Discourse-Level Relation Extraction Dataset for Chinese Literature Text via [28] modification of the obtained dataset. The main contents of one of the datasets are shown in Table 5.

Table 5. Number of Duie datasets and Chinese-Literature-RE-Dataset datasets.

Dataset Content	Duie	Chinese-Literature-RE-Dataset
Number of training sets	362,516	19,447
Number of test sets	50,000	2220
Number of validation sets	45,429	2220
Number of relationship types	49	10

For the Duie dataset, the entities a, relations, entities b, and texts containing entities in the dataset are first extracted, and these data are formatted and normalized into JSON format. By manually removing 2164 sentences with incorrect JSON file formatting due to the text containing multiple double quotes, a total of 362,516 Duie training datasets, 50,000 test set data, and 45,429 validation set data were obtained after cleaning the data. For the Chinese-Literature-RE-Dataset dataset, since the text of the dataset is composed of one literary article, the dataset is first processed by dividing the text of the article into sentences by periods. The entities and relationship annotations in it are extracted from the corresponding ANN files and normalized into new JSON files. In total, 19,447 articles of the collated data Duie training dataset, 2220 articles of the test set data, and 2220 articles of the validation set data were obtained.

4.2.2. Experimental Environment

The hardware configuration, system environment, and development environment of the experimental environment in this paper are shown in Table 6.

Table 6. Environment configuration for relational extraction model.

Name	Configuration Information
CPU	AMD Ryzen 5 5600X 6-Core Processor 3.70 GHz
RAM	32 G
Memory	1 T
Operating System	Windows 11 22000.1574
Development Languages	Python 3.6
Development Framework	PyTorch 1.7.1

4.2.3. Experimental Parameter Setting

The same model performs differently under different hyperparameter settings. Therefore, it is necessary to conduct multiple experiments with multiple different parameter settings and to select model parameters with better performance as the final parameters based on the results of the experiments. The optimal parameters set in this paper are shown in Table 7.

Table 7. Experimental parameters of the relational extraction model.

Parameter Name	Parameter Information
Torch.Size	[x, 1, 128]
Self Attention Hidden layer dimension	768
Dropout	0.5
Learning Rate	0.001
Epoch	10
Batch Size	8

4.2.4. Experimental Results

In order to verify the effect of the relationship extraction of the BERT-BiGRU-Attention model proposed in this paper, this paper uses different models to conduct experiments on the same dataset. The comprehensive comparison results of each model are shown in the table below. From the results in the table below, it can be seen that the model proposed in this paper has improved the precision, recall, and F1 values in the Duie and Chinese literature text discourse-level relationship extraction datasets. For datasets with a small number of relations, the effect of lifting is higher. The experimental results under the Duie dataset are shown in Table 8.

Table 8. Experimental results under Duie dataset.

Experiment Number	Model Name	Metrics		
		Precision	Recall	F1
1	BiLstm-Attention	86.15	85.20	85.67
2	PCNN-Attention	87.32	87.16	87.24
3	BERT	94.20	92.71	93.45
4	**BERT-BiGRU-Attention**	**94.60**	**94.61**	**94.62**

The experimental results under the Chinese-Literature-RE-Dataset dataset are shown in Table 9.

Table 9. Experimental results under the Chinese-Literature-RE-Dataset dataset.

Experiment Number	Model Name	Metrics		
		Precision	Recall	F1
1	BiLstm-Attention	88.90	88.23	88.56
2	PCNN-Attention	95.73	67.74	89.96
3	BERT	91.78	90.63	91.20
4	**BERT-BiGRU-Attention**	**93.35**	**91.69**	**92.51**

It can be seen from the experimental results that the model performs well in the Chinese relation extraction task, and the performance of the model in the Chinese relation extraction task does not mean that it performs equally well in other datasets. Therefore, a comprehensive evaluation of the model is required in the next work to determine its actual effect on other datasets. In addition, continuous efforts are needed to improve the performance of this model to meet the ever-changing demands of Chinese relation extraction tasks.

5. Analysis

The paradigm of the entity model and relation extraction model in this paper is simple, but the experiments show that the model has achieved good results in related datasets. In this section, we will focus on what causes the impact of model performance.

5.1. NER Model

Experimental comparison is effective in improving the effect of the named entity recognition model by adding the pinyin vector, grapheme vector, and word vector to the character vector. Table 10 shows the fusion results of different vectors under the same configuration environment.

Table 10. Experimental results after fusion of different vectors.

Models	F1
Pinyin + Lettering	69.61
Pinyin + Lettering + Word vectors	70.86
Lettering + Word vectors	68.33
Pinyin + Word vectors (12 million word vectors)	72.13
Pinyin + Word vectors (2 million word vectors)	**73.72**

The glyph features utilized in this study involve breaking down Chinese characters into radicals and components based on a dictionary, thereby leveraging Chinese glyph characteristics. Through comparative experiments, it is evident that the integration of additional features does not necessarily yield superior results. Jordan Hoffmann et al. [29] mentioned in their work that the model size and training data scale should align. Comparative experiments reveal that the effectiveness of using large word vectors with 12 million words is not necessarily better than using small word vectors with 2 million words. When the dataset is relatively small, the use of small word vectors can prove to be more effective than employing larger ones. This is because large word vectors often encompass a considerable amount of redundant information, which may not be essential for addressing smaller-scale tasks and could potentially interfere with the model's learning. This also underscores the importance of having balanced data in triplet systems, as well-curated data contribute significantly to enhancing the efficiency of natural language processing tasks.

In natural language processing tasks, the vector dimension usually contains the characteristics of the characters in the corpus. High-dimensional vectors typically contain more accurate information, which is a further help to natural language processing tasks. In Table 11, it can be observed that embedding 200-dimensional vectors is beneficial for enhancing the performance of named entity recognition.

Table 11. F1 values of word vector embeddings with different dimensions for the Weibo dataset.

Models	F1
Pinyin + Word vectors (12 million word vectors) (100 dimensions)	67.47
Pinyin + Word vectors (2 million word vectors) (100 dimensions)	68.13
Pinyin + Word vectors (12 million word vectors) (200 dimensions)	72.13
Pinyin + Word vectors (2 million word vectors) (200 dimensions)	**73.72**

5.2. RE Model

With respect to the relation extraction task, this paper tests the embedding of word embedding dimension, as shown in Table 12, and demonstrates that for natural language processing tasks, the higher the dimension selected, the better the processing results. Choosing a data type that is symmetrical to the model and the inference algorithm is more helpful to improve the effect of the model.

Table 12. Effect of word embedding with different dimensions in Chinese-Literature-RE-Dataset dataset.

Models	F1
BERT-BiGRU-Attention (200 dimensions)	87.56
BERT-BiGRU-Attention (100 dimensions)	**92.51**

In conclusion, various attempts have proven to be beneficial in enhancing the efficacy of natural language processing tasks. This paper focuses on the dimensions of word vectors without delving into discussions on whether substituting other layers with different dimensions could yield superior results. In future work, we will actively explore alternative approaches to enhance the performance of the model.

6. Conclusions

The present study examines the effects of various feature fusion methods on Chinese Named Entity Recognition (NER) and Relation Extraction (RE). A method is proposed to improve NER and RE performance by integrating character vectors with domain-specific word vectors. The proposed method's effectiveness is validated through experiments on multiple Chinese entity and relation extraction datasets. Furthermore, making symmetrical adjustments to the trio of data, model, and inference algorithm leads to improved experimental outcomes. However, it is noted that the model may have certain limitations, as the experiments are currently confined to the data subset of the trio. Experimental comparisons have been conducted only on multiple Chinese entity relation datasets, lacking validation in domain-specific datasets. Future research will focus on addressing these issues by exploring the integration of Chinese font features and vocabulary-based enhancement methods to further improve model accuracy and inference speed. The investigation of symmetry and synergy in data, models, and inference algorithms aims to improve learning efficiency. On the application front, the proposed method can be further validated using specific domain datasets, such as medical and financial domains. Additionally, the method can be applied to sentiment analysis [30] in particular domains. It is anticipated that future research will lead to the development of more advanced models.

Author Contributions: Software, writing—review and editing, P.F.; writing—original draft preparation, N.S. and J.X.; validation, N.S.; visualization, J.B.; supervision, D.O. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Science and Technology Development Plan Project of Jilin Provincial Science and Technology Department (Key Technology Research on Risk Prediction and Assessment of Old Chronic Diseases Based on Medical Knowledge Graph (2023JB405L07)).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in [22–24,27,28].

Acknowledgments: We thank the anonymous reviewers for their valuable feedback.

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

References

1. Kainan, J.; Xin, L.; Rongchen, Z. Overview of Chinese Domain Named Entity Recognition. *Comput. Eng. Appl.* **2021**, *57*, 1–15. [CrossRef]
2. Liu, L.; Wang, D. A Review on Named Entity Recognition. *J. China Soc. Sci. Tech. Inf.* **2018**, *37*, 329–340.
3. Kang, Y.L.; Sun, L.B.; Zhu, R.B.; Li, M.Y. Survey on Chinese named entity recognition with deep learning. *J. Huazhong Univ. Sci. Technol. Nat. Sci. Ed.* **2022**, *50*, 44–53. [CrossRef]
4. Zhong, S.S.; Chen, X.; Zhao, M.H.; Zhang, Y.J. Incorporating word-set attention into Chinese named entity recognition Method. *J. Jili Univ. Eng. Technol. Ed.* **2022**, *52*, 1098–1105. [CrossRef]
5. He, Y.J.; Du, F.; Shi, Y.J.; Song, L.J. Survey of Named Entity Recognition Based on Deep Learning. *Comput. Eng. Appl.* **2021**, *57*, 21–36.
6. Xie, W.R. *Research and Implementation of Named Entity Recognition Based on Character Multi-Semantic Features*; Jiangnan University: Wuxi, China, 2022.
7. Cui, M.J.; Li, L.; Wang, Z.H.; You, M.Y. A Survey on Relation Extraction. In *Knowledge Graph and Semantic Computing: Language, Knowledge, and Intelligence*; Springer: Singapore, 2017. Available online: https://link.springer.com/chapter/10.1007/978-981-10-7359-5_6 (accessed on 22 January 2024).
8. Xiong, S.; Li, B.; Zhu, S. DCGNN: A single-stage 3D object detection network based on density clustering and graph neural network. *Complex Intell. Syst.* **2022**, *9*, 3399–3408. [CrossRef]
9. Zeng, D.; Liu, K.; Chen, Y.; Zhao, J. Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1753–1762.
10. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K.J.A. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
11. Yang, H. BERT Meets Chinese Word Segmentation. *arXiv* **2019**, arXiv:1909.09292.
12. Meng, Y.; Wu, W.; Wang, F.; Li, X.; Nie, P.; Yin, F.; Li, M.; Han, Q.; Sun, X.; Li, J. Glyce: Glyph-vectors for Chinese Character Representations. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019.
13. Peng, M.; Ma, R.; Zhang, Q.; Huang, X. Simplify the Usage of Lexicon in Chinese NER. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019.
14. Liu, W.; Fu, X.; Zhang, Y.; Xiao, W. Lexicon Enhanced Chinese Sequence Labeling Using BERT Adapter. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; pp. 5847–5858.
15. Zhang, Z.; Zhang, H.; Chen, K.; Guo, Y.; Hua, J.; Wang, Y.; Zhou, M. Mengzi: Towards Lightweight yet Ingenious Pre-trained Models for Chinese. *arXiv* **2021**, arXiv:2110.06696.
16. Lafferty, J.D.; McCallum, A.; Pereira, F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the International Conference on Machine Learning, Williamstown, MA, USA, 28 June–1 July 2001.
17. Peng, J.Y. *The Research of the Chinese Named Entity Recognition Method with Glyph Feature*; Shanghai Jiao Tong University: Shanghai, China, 2019.
18. Xiao, L.; Pennington, J. Synergy and Symmetry in Deep Learning: Interactions between the Data, Model, and Inference Algorithm. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; Proceedings of Machine Learning Research. pp. 24347–24369.
19. Song, Y.; Shi, S.; Li, J.; Zhang, H. Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; pp. 175–180.
20. Sun, M.; Chen, X.; Zhang, K.; Guo, Z.; Liu, Z. THULAC: An Efficient Lexical Analyzer for Chinese. 2016. Available online: https://github.com/thunlp/THULAC (accessed on 22 January 2024).
21. Li, B.; Lu, Y.; Pang, W.; Xu, H. Image Colorization using CycleGAN with semantic and spatial rationality. *Multimed. Tools Appl.* **2023**, *82*, 1–15. [CrossRef]
22. Peng, N.; Dredze, M. Named Entity Recognition for Chinese Social Media with Jointly Trained Embeddings. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 548–554.
23. Zhang, Y.; Yang, J. Chinese NER Using Lattice LSTM. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 1554–1564.

24. Levow, G.-A. The Third International Chinese Language Processing Bakeoff: Word Segmentation and Named Entity Recognition. In Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, Sydney, Australia, 22–23 July 2006; pp. 108–117.
25. Zhang, Z.; Han, X.; Liu, Z.; Jiang, X.; Sun, M.; Liu, Q. ERNIE: Enhanced Language Representation with Informative Entities. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 1441–1451.
26. 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. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; pp. 2065–2075.
27. Li, S.; He, W.; Shi, Y.; Jiang, W.; Liang, H.; Jiang, Y.; Zhang, Y.; Lyu, Y.; Zhu, Y. DuIE: A Large-Scale Chinese Dataset for Information Extraction. In Proceedings of the Natural Language Processing and Chinese Computing, Dunhuang, China, 9–14 October 2019; pp. 791–800.
28. Xu, J.; Wen, J.; Sun, X.; Su, Q. A Discourse-Level Named Entity Recognition and Relation Extraction Dataset for Chinese Literature Text. *arXiv* **2017**, arXiv:1711.07010.
29. Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; Casas, D.D.L.; Hendricks, L.A.; Welbl, J.; Clark, A.; et al. Training Compute-Optimal Large Language Models. *arXiv* **2022**, arXiv:2203.15556.
30. Cauteruccio, F.; Kou, Y. Investigating the emotional experiences in eSports spectatorship: The case of League of Legends. *Inf. Process. Manag.* **2023**, *60*, 103516. [CrossRef]

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.