applied sciences

MDPI

Article

Chinese Named Entity Recognition in Football Based on ALBERT-BiLSTM Model

Qi An, Bingyu Pan *, Zhitong Liu, Shutong Du and Yixiong Cui

School of Sports Engineering, Beijing Sport University, Beijing 100084, China; 2021210404@bsu.edu.cn (Q.A.); 2022210171@bsu.edu.cn (Z.L.); 15929176687@163.com (S.D.); cuiyixiong@bsu.edu.cn (Y.C.)
* Correspondence: panbingyu@bsu.edu.cn

Abstract: Football is one of the most popular sports in the world, arousing a wide range of research topics related to its off- and on-the-pitch performance. The extraction of football entities from football news helps to construct sports frameworks, integrate sports resources, and timely capture the dynamics of the sports through visual text mining results, including the connections among football players, football clubs, and football competitions, and it is of great convenience to observe and analyze the developmental tendencies of football. Therefore, in this paper, we constructed a 1000,000-word Chinese corpus in the field of football and proposed a BiLSTM-based model for named entity recognition. The ALBERT-BiLSTM combination model of deep learning is used for entity extraction of football textual data. Based on the BiLSTM model, we introduced ALBERT as a pre-training model to extract character and enhance the generalization ability of word embedding vectors. We then compared the results of two different annotation schemes, BIO and BIOE, and two deep learning models, ALBERT-BiLSTM-CRF and ALBERT BiLSTM. It was verified that the BIOE tagging was superior than BIO, and the ALBERT-BiLSTM model was more suitable for football datasets. The precision, recall, and F-Score of the model were 85.4%, 83.47%, and 84.37%, correspondingly.

Keywords: named entity recognize; ALBERT; BiLSTM; deep learning; football

Citation: An, Q.; Pan, B.; Liu, Z.; Du, S.; Cui, Y. Chinese Named Entity Recognition in Football Based on ALBERT-BiLSTM Model. *Appl. Sci.* **2023**, *13*, 10814. https://doi.org/10.3390/app131910814

Academic Editors: Jae-Hoon Kim and Kichun Lee

Received: 15 August 2023
Revised: 24 September 2023
Accepted: 25 September 2023
Published: 28 September 2023

1. Introduction

Football is the most popular sport in the world, and there is a massive amount of related data, such as news articles, match statistics, social media posts, and transfer histories. The data can be used to prepare scouting reports, improve match performances, and provide insight into the game. However, it also presents significant challenges for processing and extracting meaningful information from fragmented text. Recent researchers have proposed to apply named entity recognition (NER) into automatically identifying and extracting important entities in the context of football. Specifically, named entities such as football players, clubs, leagues, terminologies, and other related entities could be extracted via such an approach. Finally, it helps football coaches, players, fans, and researchers to efficiently access essential information in a game, apply analyses to the trends, and summarize shortcomings in the competitions.

NER is also called "named entity recognition", "named entity extraction", "named entity segmentation", and so on [1], one of the significant topics in Natural Language Processing (NLP). NER is the basis of other NLP researches, such as relation extraction [2], knowledge graph construction [3], etc. Research on NER is spread globally across the fields of engineering [1], biotechnology-applied microbiology [2], medical informatics [3], biochemistry molecular biology [4], and other fields. However, little research has been performed in sports or football. In natural language processing for football-related tasks, there are significant challenges, due to a lack of standardized and easily accessible corpora for identifying named entities. In addition, Chinese NER in football faces some unique challenges due to the complexities and specificity of the Chinese language and the football

context. The Chinese language has a vast and complex character set without evident segments, which makes it more challenging to identify named entities than other languages. As a result, we proposed a novel approach for NER in football to recognize information, transforming the unstructured data into structured triples, which, in certain extents, enriches the sports knowledge base and provides a football knowledge foundation for football domain experts and football fans. Football NER has significant importance in various areas related to football, from knowledge network construction, information retrieval, and analysis to fan engagement and business applications.

Next, we will introduce several key techniques of NER and its research in the Chinese language field and in football.

1.1. Word Embedding

Word embedding is equivalent to a pre-training process for text, converting it into a vector that can be computed by a computer. In a bid to preserve the directional relationship of the triples, word embedding can capture contextual, semantic, and syntactic similarities, as well as the relationship of words to other words to effectively train computers to understand natural language. Static word vectors and dynamic word vectors are the two forms of word embedding models. According to a predefined vocabulary, termed as static word vectors, such as NNLM, Word2Vec, GPT, and Glove, each word solely has a different vector representation. The most prominent and popular embedding models currently in use are dynamic word vectors based on machine learning that primarily use an RNN or Transformer model. An attention mechanism module is added to the Transformer model, enabling it to capture a wider range of dependencies between words in the sentence. BERT, RoBERTa, ALBERT, and many more word embedding models based just on the Transformer model exist and are accessible [4,5]. A bidirectional Transformer encoder called BERT incorporates a Next Sentence Prediction model (NSP) and a Masked Language Model (MLM). With a bidirectional network and attention mechanism, by merging characters and context, BERT is used to pre-process the text, augmenting the word vectors' capacity to extrapolate, fully characterizing the character, word, and sentence levels, as well as the relationships between them, and is capable of conveying a word's polysemy. The concept of dynamic programming enables us to obtain globally optimal output sequence labels by taking full advantage of the relevance of the adjacent labels, which further improves the efficacy of model entity recognition [6]. However, the BERT model generates an extensive number of hyper-parameters during pre-training and requires greater amounts of time to train [7]. A lite Bidirectional Encoder Representation from NER is accomplished using the ALBERT-BiLSTM-CRF model hybrid approach [8]. The ALBERT layer's architecture is similar to that of the Bert layer, except that it splits the entire vocabulary embedding matrix into two separate matrices and shares parameters with the encoder and decoder to determine whether the next statement is indeed the following statement or not. It is demonstrated that the ALBERT layer is more functionally efficient than the Bert layer, as it emits fewer parameters.

1.2. Named Entity Recognition Techniques

NER is a crucial component of information extraction, the foundation for constructing knowledge graphs, and the primary method for discovering a domain's knowledge ontology, as well as the fundamental task of building a knowledge graph through the extraction of specific categories and critical information from a variety of unstructured texts.

Computer technology has contributed to the categorization of NER techniques into three distinct types: rule-based, statistical machine learning, and deep learning approaches. Firstly, the rule-based method of NER [9] requires specialists to manually spend a lot of time creating the rule templates [10] used to select entities that match or not in the early stages. Currently, the rule-based methods were applied to NER of various languages, including Korean [11], Arabic [12,13], Persian [5], Chinese [14], Turkish [15], and others. It is noteworthy that the rule-based model achieved F-Score of 89.5% in Arabic NER [13] and

85.93% in Persian NER [5]. This approach is suitable for languages in which the rules are not particularly evident. The rule-based method has a high accuracy but a weak generalization ability. It fits to better deduce from the context of the entity made up by proper names and grammatical terms. Moreover, this method will have ambiguities because word polysemy causes completely distinct formulations to result from different contexts. Therefore, this approach is not suitable for named entity recognition for complex data.

Secondly, statistical machine learning techniques like the HMM [16], CRF [17], SVM [18–20], and others regard NER as a classification question. This method performs better on small datasets but poorly on complex and huge datasets, and it calls for a feature matrix that has been predefined by experts. This method treats the entity recognition problem as a character label classification problem, and it uses labeled datasets and predetermined features to guide computers how to recognize entities in strings. As an example, HMM has been used in NER of Korean [21,22], Chinese, and Urdu [23]. Malik et al. [23] classified Urdu into person names, location names, organization names, and so on based on the HMM, precision, recall, and f-measure values of 55.98%, 83.11%, and 66.90%. Imam et al. [24] did an experiment on software entities such as system, use case, and actor on based on an SVM model that achieved 72.1% for the F-Score. Statistical machine learning-based approaches have a fairly narrow range of applications, as they are only relevant to small datasets and have poor performances on datasets with complex entity types and vast amounts of data.

Thirdly, deep learning models consisting of many stacked neural networks are widely used in NER, such as RNN [25,26], Long Short-Time Memory (LSTM) [27], Bidirectional LSTM [28,29], BiLSTM-CRF [30,31], BERT-BiLSTM-CRF [6,32], ALBERT-BiLSTM-CRF [8], and attention mechanisms [33], which have replaced statistical machine learning techniques. This method can handle high-dimensional data with excellent accuracy. Deep learning is utilized the most frequently in Korean [34,35], Polish [36], Chinese [37], Arabic [38], and so on. The F1 value of the NER algorithm, the BI-LSTM-CRF model, on weaponry equipment names in Chinese reached 93.88% [37]. Based on the review of NER methods, the approach of stacking deep learning models was reliable in terms of accuracy and capability for processing complicated data.

1.3. Named Entity Recognition for Chinese Language

Named entity recognition in the Chinese language is used in agriculture [39], natural hazards [40], the military [41], engineering [42], chemicals [43], and mainly in medicine, covering electronic health records [43,44] and clinical texts [45,46]. Although named entity recognition has many applications in Chinese, it still has a variety of challenges. Due to the intrinsic characteristics of Chinese, research on Chinese named entity recognition is not as mature as that on English named entity recognition. Firstly, English uses the derivation method to form words, whereas Chinese uses the compound method. From the point of view of word formation elements (root, prefix, and suffix), English roots, prefixes, and suffixes have strong abstraction; derivation words are good at describing abstract relations, actions, things, etc. Chinese, which often uses specific morphemes to create words in a compound way, is not as good as English at describing the internal laws of abstract things. B. Ji [47] mentioned that the performance of a model can be affected by differences in the Chinese descriptions of the same thing when named entity recognition is used in Chinese electronic medical records, but the labels remain the same in the data annotations. In addition, Wei Liu [48] referred to Chinese sequence labeling being more challenging because of the lack of explicit word boundaries in Chinese sentences.

1.4. Studies on Named Entity Recognition in Sports

Although the research on named entity recognition in sports has been conducted over a long period of time, it remains limited in quantity and lacks cohesion among different approaches. In the earlier study, scholars adopted a rule-based method due to limited computer technology development. T. Yao [49], Q.-M. Nguyen [50], L. Chiticariu [51] iden-

tified players and organizations in football and sports events in football transfer news by formulating rules and constructing dictionaries to identify named entities, so as to continuously enrich the knowledge base of football. Recently, machine learning and deep learning merged together in the research of NER in sports. Xieraili Seti [52] took 10,000 entity samples from text data of major sports news websites in China as data sets, and a character-based graph convolution neural network model based on self-attention mechanism was proposed to identify 10 entities in the field of sports, such as name of competition, team name, place name, name of athlete and duties, competition level, competition level, and so on. In order to test the learning ability of authors' model, they performed verification with a customized sports data set, comprising the Bakeoff-3, OntoNotes, and ResumeNER data sets, respectively, and confirmed that the method proposed by the author was higher than the statistical machine learning method in accuracy, recall rate, and F-Score, which were 90.73%, 94.36%, and 92.51%, respectively. Moreover, P. Liu [53] combined the dictionary in the LTP word segmentation tool and a dictionary containing a large number of proper nouns and terms for winter sports, receiving a 93.56% F-Score based on RoBERTa's pre-training model and the BiLSTM-CRF model on the sports news of the Beijing Winter Olympics.

Above all, at present, named entity recognition based on deep learning has been widely used in various fields, but there are few studies on entity recognition in football texts. To solve the problem of named entity recognition in football, we employed a deep learning model combined with word embedding. It solved the problem wherein traditional methods relied heavily on dictionaries and rules and avoided artificial feature extraction.

2. Materials and Methods

2.1. Datasets

Currently, there is a lack of research on football-related named entity recognition, and no relevant datasets are available. Therefore, we constructed a football data set of 1 million characters of football-related information in Chinese, crawled from various websites. Our data set comprised detailed information on 111 clubs from the Premier League, La Liga, Serie A, Bundesliga, Ligue 1, the Chinese Football Association Super League, 47 national teams, 44 leagues such as the Asian Cup (AFC) and the FIFA World Cup, as well as more than 3300 football athletes. We divided the total amount of data into training sets, validation sets, and testing sets; we used an 8:1:1 ratio.

The football entities were divided into five categories by our research: football organizations (ORG), place name (LOC), names of football players and person served in football competitions (PER), football terminology (TERM), and football competitions (SCPT). Table 1 shows some examples of the various types of entities.

Table 1. The entity categories of football datasets.

Parameters	Entity Instance	Entity Symbol
Football organization	皇家马德里 (Real Madrid Baloncesto)	ORG
Football location	英国、西班牙 (Britain, Spain)	LOC
Person and sports job name	内马尔·达席尔瓦·桑托斯 (Neymar da Silva Santos Júnior)	PER
Football rule and ranking	冠军、亚军 (Champion, second place)	TERM
Football competitions	意大利甲级联赛 (Serie A)	SCPT

2.2. Annotation Scheme

Different annotation schemes also affect the effect of the model. Currently, two popular annotation schemes on NER have been used in different works, and those schemes were as follows:

BIO: Three tags were assigned for text to estimate whether the word was the beginning (B) of a named entity, inside (I) of a named entity, or outside (O) of all named entities.

BIOE: This scheme worked almost identically to BIO annotation; four tags were assigned for text to estimate whether the word was the beginning (B) of a named entity, inside (I) of a named entity, end (E) of a named entity or outside (O) of all named entities.

The inter-annotator agreements were as follows: Non-overlapping, namely, the same string could not be annotated with two different labels; Unnested, i.e., one entity could not be inside another; and Factual, i.e., we only annotated actual facts. For instance, B-SCPT, I-SCPT, and E-SCPT denoted the beginning, middle, and end of the league entity, respectively. Partial examples are shown in Table 2. The sentence in Table 2, "此外，即将前往国足的申花体能教练欧文威尔克也接受了东体的采访", means "Besides, Oriental Sports News interviewed Owen Wilke, the fitness coach of Shanghai Shenhua football club, who will be charge of the national football team of People's Republic of China". "国足" refers to the national football team of the People's Republic of China, whereas "申花" refers to the Shanghai Shenhua Football Club of the China Premier Football League. Additionally, "欧文威尔克" serves as the name of Shenhua's fitness coach, and "东体" is the Oriental Sports Daily. Table 3 demonstrates the distribution of the data labels divided for our research.

Table 2. Entity annotation sample of football.

Word	Tagging	Word	Tagging
此	O	练	O
外	O	欧	B-PER
，	O	文	I-PER
即	O	威	I-PER
将	O	尔	I-PER
前	O	克	E-PER
往	O	也	O
国	B-ORG	接	O
足	E-ORG	受	O
的	O	了	O
申	B-ORG	东	B-ORG
花	E-ORG	体	E-ORG
体	O	的	O
能	O	采	O
教	O	访	O

Table 3. The number of entities in train set, test set, and datasets.

Entity Symbol	Train Set	Test Set	Datasets
LOC	7065	698	7763
PER	21,739	3214	24,953
TERM	6903	2721	9624
SCPT	21,778	624	22,402
ORG	14,922	1271	14,922

2.3. ALBERT-BiLSTM Model

The entire model structure is shown in Figure 1. The first layer of the model is the ALBERT layer. Based on BERT, ALBERT achieves parameter reduction by slicing the word embedding matrix and parameter sharing. The BERT layer is a pre-trained multi-layer bidirectional Transformer encoder that converts input characters into a matrix for computer processing. BERT incorporates an encoder, masked language modeling (MLM), and next sentence prediction (NSP), which ALBERT converts to sentence order prediction (SOP). The encoder includes the segment embedding, token embedding, and position embedding of each character. For segment embedding, we utilized a specific embedding token to represent the symbols "[CLS]" and "[SEQ]" for the start and end of a sentence, respectively. Each word in position embedding has a unique position feature based on its location in

the sentence. Additionally, we used one-hot word embedding to create token embedding. The word embedding of each word was the sum of these three values. Based on word embedding, we transformed each Chinese character into a vector that could be processed by computers. Figure 2 illustrates the structure of word embedding.

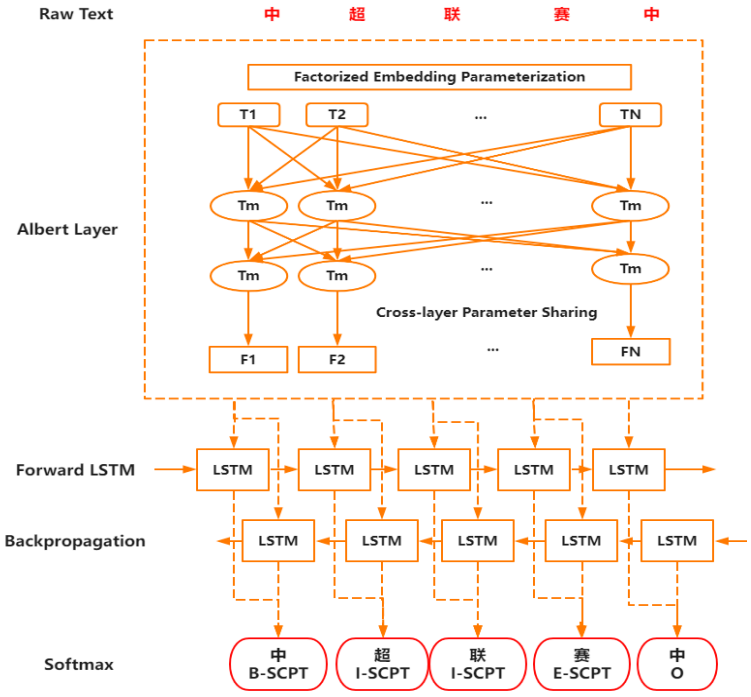

Figure 1. The architecture of the ALBERT-BiLSTM model.

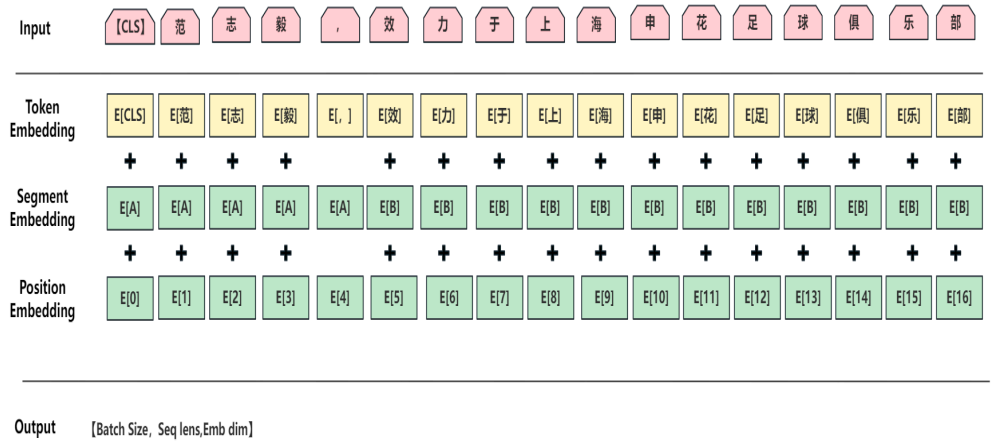

Figure 2. The structure of word embedding.

MLM is a context-based bidirectional prediction model that acts as the equivalent of the attention mechanism. The core idea of the masked language model is to randomly select a certain percentage of tokens in the input sequence and replace them with special MASK tokens. During model training, MLM predicts the actual content of the masked tokens based on the previous tokens. This approach allows the model to learn the global representation of the sentence during training, which, in turn, improves the performance of the MLM. By selecting a random 15% of the characters to be masked, 80% of which will be genuinely masked, 10% will continue unmarked, and 10% will be any other character. Figure 3 depicts the MLM model's architecture. NSP is used to determine whether the current sentence is behind the previous sentence. If the result is no, it is not sure about the

position of the current sentence. Therefore, the NSP task is difficult to learn the coherence between sentences, whereas SOP is of greater concern in coherence between sentences.

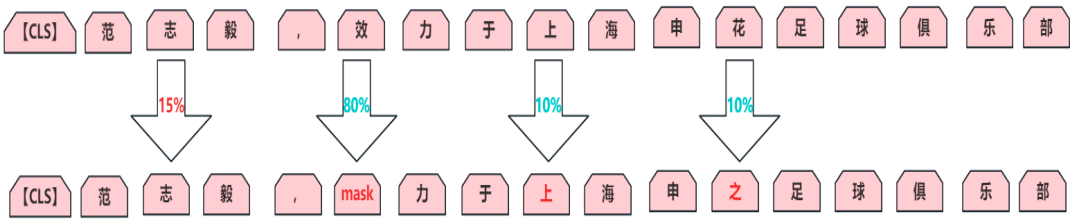

Figure 3. The architecture of MLM.

Despite BERT's remarkable learning ability, the large number of characters in our research led to an explosion of hyperparameters. In order to lower the number of parameters and the training time because of the substantial amounts of parameters generated during pre-training, ALBERT developed the following modifications based on BERT: factorized embedding parameterization and cross-layer parameter sharing. The factorized embedding parameterization separated the overall embedding matrix into two tiny matrices via a transition hidden layer with fewer layers since the word vocabulary was too huge to compute. Between the feed forward layer and the attention layer, ALBERT shared all parameters.

The second layer of the model was the BiLSTM layer. BiLSTM was made up of a forward and a backward LSTM model, so BiLSTM could not only infer from front to back based on the information in front of it but also infer from back to front based on the information behind it. The forward hidden layer vector and the backward hidden layer vector collaboration generated the BiLSTM output. The input gate, forget gate, and output gate of the LSTM regulated to input, forget, and output the input embedding, respectively, which contributed to the capacity for long-term memory. The below were the formulas:

$$i_t = \sigma\left(x_t \bullet w_{xh}^{j} + h_{t-1} \bullet w_{hh}^{j} + b_h^{i}\right), \tag{1}$$

$$f_t = \sigma\left(x_t \bullet w_{xh}^{f} + h_{t-1} \bullet w_{hh}^{f} + b_h^{f}\right), \tag{2}$$

$$o_t = \sigma(x_t \bullet w_{xh}^{o} + h_{t-1} \bullet w_{hh}^{o} + b_h^{o}), \tag{3}$$

$$\overrightarrow{c_t} = tanh(x_t \bullet w_{xh}^{c} + h_{t-1} \bullet w_{hh}^{c} + b_h^{c}\Big), \tag{4}$$

$$c_t = i_t \otimes \overrightarrow{c_t} + f_t \otimes c_{t-1}, \tag{5}$$

$$h_t = o_t \otimes tanh(c_t), \tag{6}$$

where σ is the sigmoid activation function; $\otimes$ is a dot product; *tanh* is the activation function; i_t,f_t, and o_t represent the memory gate, forget gate, and output gate at time t, respectively; c_t denotes the cell state at time t; and h_t denotes the hidden state at time t.

The BiLSTM formulation is as follows:

$$\overrightarrow{h_i} = LSTM\left(\overrightarrow{h_{i-1}}, x_i\right), \tag{7}$$

$$\overleftarrow{h_i} = LSTM\left(\overleftarrow{h_{i-1}}, x_i\right), \tag{8}$$

$$h_t = \overrightarrow{h_i} \oplus \overleftarrow{h_i}, \tag{9}$$

In which, $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$ represent the forward hidden layer vector and the backward layer vector, respectively, and $\oplus$ denotes the joint operation.

2.4. Metrics

Three criteria were used to measure the degree to which the ALBERT-BiLSTM model performed while extracting entities. F-Score could be regarded as a harmonic average of "Precision" and "Recall," in which "Precision" denoted the ratio of the number of correct entities recognized by the model to the number of all entities recognized by the model, and "Recall" denoted the ratio of the number of correct entities recognized by the model to the number of all entities that the model should recognize.

The three metrics' formula were as follows:

$$Precision = \frac{TP}{TP + TF} \times 100\%, \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \times 100\%, \tag{11}$$

$$F - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{12}$$

Here, *TP* indicates the number of entities in the data that the model correctly identified, *FP* indicates the amounts of incorrect entities it correctly identified, and *FN* represents the number of entities that the model should have correctly identified but did not. When the beginning and end of a recognized entity match those of the matching entity in the testing data set, the *TP* has been achieved. The better the *F-Score*, the more accurate the learning outcomes and the more capable the model.

3. Results

To evaluate the performance of the model, experiments were conducted on another data set with the same features. Here, we selected the data set the PFR People's Daily Corpus, whose scale was the same as that of the football data set, including one million characters. The entities in the public data set were classified into three categories: people's names (PER), places' names (LOC), and organizations' names (ORG). However, in our data set, entities were divided into five categories. The weighted test results with BIO annotation schemes are shown in Table 4. It can be seen from this table that the model achieved lower performance than those on "The PFR People's Daily corpus".

Table 4. The test results on "The PFR People's Daily corpus" using BIO annotation schemes based on ALBERT-BILSTM.

Dataset	P (%)	R (%)	F-Score (%)
Public Dataset	93.46	89.23	91.26
Football Dataset	85.40	83.47	84.37

We also compared the results of two different annotation methods BIO and BIOE, based on the ALBERT-BiLSTM model, as shown in Table 5. From the results, it can be seen that the BIOE annotation method had significantly higher accuracy, recall, and F-Score for all five categories of data. The BIOE tagging scheme was more suitable for the datasets and had better performance in identifying the boundaries of entities. Therefore, the BIOE annotation method was chosen for the subsequent analysis. We then compared the performance of the two models ALBERT-BiLSTM and ALBERT-BiLSTM-CRF in name entity recognition, and the results are shown in Figure 4. It is evident that the ALBERT-BiLSTM

model consistently exhibited higher precision, recall, and F-Score values across all entity categories when compared to ALBERT-BiLSTM-CRF. The F-Score of the ALBERT-BiLSTM model was 1.27% higher than that of the ALBERT-BiLSTM model.

Table 5. The test results with different annotation schemes, based on ALBERT-BILSTM.

Annotation Scheme	Entity Category	P (%)	R (%)	F-Score (%)
BIO	LOC	80.96	72.45	76.47
	ORG	78.63	79.27	78.95
	PER	81.92	80.34	81.12
	SCPT	82.70	80.93	81.80
	TERM	63.83	52.29	57.49
BIOE	LOC	87.10	81.69	84.31
	ORG	86.27	86.74	86.50
	PER	86.02	84.08	85.04
	SCPT	85.59	85.74	85.66
	TERM	74.50	59.84	66.37

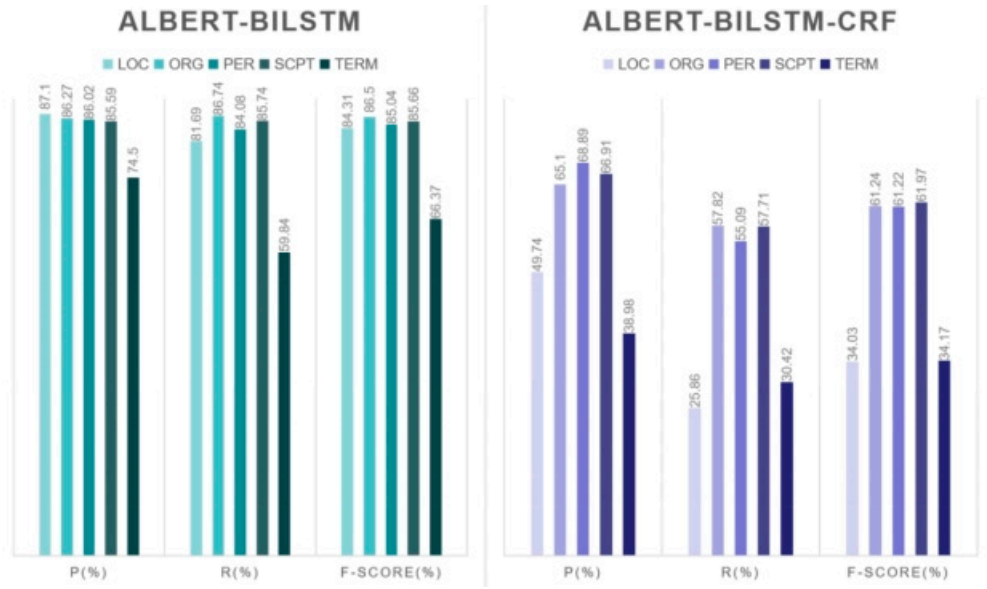

Figure 4. Experimental results of five entity categories in ALBERT-BiLSTM and ALBERT-BiLSTM-CRF model.

Additionally, we compared the results of previous studies on Chinese named recognition in football, which is shown in Table 6. We discovered that the authors used the most primitive named entity recognition method in the previous study, namely, rule based, which resulted in a low recall rate in the experimental results. The authors identified only three entity categories. From the comparison results, the model we adopted in football was more suitable for the downstream applications, relation extraction, information retrieval, event tracking, and intelligent question answering system.

Table 6. Comparison of the experiments between the current study and the one by Nguyen.

Methodology	Dataset	Entity Category	P (%)	R (%)	F1-Score (%)
Rule-based [46]	A total of 237 football transfer news from Sky Sports website	3	81.2	65.5	72.51
ALBERT-BiLSTM	Our data set	5	85.40	83.47	84.37

4. Discussion

The following are some factors that contributed to the lower performance than the public corpus:

In our corpus, there were more entity categories than in the public data set, and classification was somewhat more difficult. There were translation errors in the football datasets, especially in the long transliterations of English names. The data set contained a large number of football player names and football club names transliterated from different languages, which were subject to errors when translated into Chinese. We take "Lionel Messi" as an example. "Lionel Messi" can be translated into Chinese as "利昂内尔·梅西" or "莱奥·梅西"; although they are indeed the same entity, the Chinese expressions are different.

As each category of entity comprises numerous instances of that entity, and the entity names are too long to identify the whole entity, it is difficult to reliably differentiate between long organizational names, long addresses, and long names, such as "贝尔格莱德红星足球俱乐部", "费耶诺德足球俱乐部", "埃因霍温足球俱乐部", and so on.

The annotation datasets were carefully labeled, although, sometimes, it might be challenging to tell apart nested elements, such as the team "中国队" and the location "中国". Named entity recognition's effects are diminished by a number of elements working together.

The three categories of term entities, ranking (champions, Asian third-place teams, etc.), football skills (hat-tricks, penalties, etc.), and awards (Asian Footballer of the Year, Golden Boot Award, etc.), were disproportionately concentrated in the winning, second-place, and third-place, creating an imbalance in the data, which accounted for the three measures of TERM being lower than those of other entities.

Additionally, as the named entity extraction method model is a sentence-level model, each sentence in the file is treated as a separate sentence, and multiple instances of the same tag in different sentences in the file are treated as completely independent tagging entities, which will result in inconsistent marking issues; the same instance will be marked with different labels.

By comparing the results of the two models ALBERT-BiLSTM and ALBERT-BiLSTM-CRF (Figure 4), we found that the model was not as good after adding the CRF layer. This may have been due to the fact that the use of CRF alone differed significantly from the use of the CRF in the combined model. In the traditional CRF approach, the data itself is used to identify features, such as the features of the current word with the previous words, the next word, punctuation, English case, and numerical values. Additionally, the inputs and outputs are linked. In the ALBERT-BILSTM-CRF model, the CRF layer adopts the concept of transfer matrix from standard CRF. Probabilistic error already exists in BILSTM. After processing by CRF, errors may propagate. The football corpus contains extremely long words, such as club names, league names, and people's names, so adding the CRF layer contributes to lower accuracy.

Drawing from the aforementioned comparisons and error analysis, it becomes evident that future efforts should focus on building a more extensive and standardized football corpus, thereby mitigating the challenge of data imbalance. At the same time, the entity types should be further enriched in order to build a more comprehensive knowledge graph of football. Furthermore, to enhance the accuracy of name entity recognition model, advanced combined neural networks should be developed and incorporate refined loss functions and optimization techniques into the models.

5. Conclusions

With the continuous development of entity name recognition technology and deep learning related models, the recognition accuracy of Chinese named entity recognition has been greatly improved. However, the databases of these studies are in non-public states, and it is hoped that people can share their datasets to facilitate a better understanding of the current state of development of named entity recognition in specific domains. There

are few studies on named entity recognition in sports, and even fewer in football. In this paper, we have performed entity extraction based on the ALBERT model for the Chinese football domain. We constructed a named entity corpus for Chinese football and compared different annotation methods and deep learning models. The results demonstrated the superiority of the BIOE tagging scheme based on the ALBERT-BiLSTM model, with weighted average precision, recall, and F-Score percentages of 85%, 83%, and 84%, respectively. By accurately identifying entities such as player names, team names, and match events, we can facilitate information retrieval, event tracking, and sentiment analysis in the football domain.

NER is a key technology for knowledge graph construction. The results of our research allow us to build databases containing soccer matches, players, etc., which can then be analyzed statistically and trend predictions. We hope that we can assist in developing more effective game strategies by analyzing the historical data of players and teams. In the general domain, this research can be applied in the field of optimizing search engines and accurate content recommendations. NER in football helps to simplify and improve the speed and relevance of search results.

In the future, to achieve enhanced performance, we plan to incorporate optimization algorithms into our proposed method to finely adjust key hyperparameters. Furthermore, we will apply our scheme in various domains and improve the performance of NER in other sports fields. Additionally, we are going to extend the proposed NER approach to support multiple languages simultaneously and develop a comprehensive knowledge graph related to football, which we hope will enable player performance prediction and data visualization capabilities.

Author Contributions: Conceptualization, Q.A. and B.P.; methodology, Q.A. and S.D.; formal analysis, Q.A. and B.P.; investigation, Q.A.; data curation, S.D., Z.L. and Q.A.; writing—original draft preparation, Q.A.; writing—review and editing, Y.C., B.P., S.D., Z.L. and Q.A.; project administration, B.P.; funding acquisition, Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Key Research and Development Program of China, under grant 2020AAA0103404, and the National Natural Science Foundation of China, under grant 72101032, and the Fundamental Research Funds for the Central Universities of China, under grants 2021TD008, 2022YB005, and 2023RC001.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to their containing information that could compromise the privacy of research participants.

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

References

1. Cheng, J.R.; Liu, J.X.; Xu, X.B.; Xia, D.W.; Liu, L.; Sheng, V.S. A review of Chinese named entity recognition. *KSII Trans. Internet Inf. Syst.* **2021**, *15*, 2012–2030. (In English) [CrossRef]
2. Zhou, D.Y.; Zhong, D.Y.; He, Y.L. Biomedical Relation Extraction: From Binary to Complex. *Comput. Math. Method Med.* **2014**, *2014*, 298473. (In English) [CrossRef] [PubMed]
3. Qu, J. A Review on the Application of Knowledge Graph Technology in the Medical Field. *Sci. Program.* **2022**, *2022*, 3212370. (In English) [CrossRef]
4. Ruiz-Dolz, R.; Alemany, J.; Barbera, S.M.H.; Garcia-Fornes, A. Transformer-Based Models for Automatic Identification of Argument Relations: A Cross-Domain Evaluation. *IEEE Intell. Syst.* **2021**, *36*, 62–70. [CrossRef]
5. Moradi, H.; Ahmadi, F.; Feizi-Derakhshi, M.R. A Hybrid Approach for Persian Named Entity Recognition. *Iran. J. Sci. Technol. Trans. A-Sci.* **2017**, *41*, 215–222. (In Iranian) [CrossRef]
6. Ceovic, H.; Kurdija, A.S.; Delac, G.; Silic, M. Named Entity Recognition for Addresses: An Empirical Study. *IEEE Access* **2022**, *10*, 42108–42120. [CrossRef]

7. Su, P.; Vijay-Shanker, K. Investigation of improving the pre-training and fine-tuning of BERT model for biomedical relation extraction. *BMC Bioinform.* **2022**, *23*, 120. [CrossRef]
8. Zhao, P.; Wang, W.; Liu, H.; Han, M. Recognition of the Agricultural Named Entities with Multifeature Fusion Based on ALBERT. *IEEE Access* **2022**, *10*, 98936–98943. [CrossRef]
9. Bao, P.; Zhu, S.L. System design for location name recognition in ancient local chronicles. *Libr. Hi Tech* **2014**, *32*, 276–284. [CrossRef]
10. Korkontzelos, I.; Piliouras, D.; Dowsey, A.W.; Ananiadou, S. Boosting drug named entity recognition using an aggregate classifier. *Artif. Intell. Med.* **2015**, *65*, 145–153. [CrossRef]
11. Kim, J.H. Rule-based named entity (NE) recognition from speech. *Malsori* **2006**, *1*, 45–66. (In Korean)
12. Oudah, M.; Shaalan, K. NERA 2.0: Improving coverage and performance of rule-based named entity recognition for Arabic. *Nat. Lang. Eng.* **2017**, *23*, 441–472. (In English) [CrossRef]
13. Salah, R.; Mukred, M.; Zakaria, L.Q.B.; Ahmed, R.; Sari, H. A New Rule-Based Approach for Classical Arabic in Natural Language Processing. *J. Math.* **2022**, *2022*, 7164254. (In English) [CrossRef]
14. Ye, X.N.N. Study on Text Preprocessing and Automatic Rule Learning Technology for Information Extraction. Ph.D. Thesis, 2004.
15. Küçük, D.; Yazıcı, A. A hybrid named entity recognizer for Turkish. *Expert Syst. Appl.* **2012**, *39*, 2733–2742. (In English) [CrossRef]
16. Zhao, S. Named entity recognition in biomedical texts using an HMM model. In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications (NLPBA/BioNLP), Geneva, Switzerland, 28–29 August 2004; pp. 87–90.
17. Mozharova, V.A.; Loukachevitch, N.V. Combining knowledge and CRF-based approach to named entity recognition in Russian. In *Analysis of Images, Social Networks and Texts: 5th International Conference, AIST 2016, Yekaterinburg, Russia, April 7–9, 2016, Revised Selected Papers 5*; Springer: Berlin, Germany, 2017; pp. 185–195. [CrossRef]
18. Patra, R.; Saha, S.K. A Kernel-Based Approach for Biomedical Named Entity Recognition. *Sci. World J.* **2013**, *2013*, 950796. [CrossRef]
19. Devi, G.R.; Kumar, M.A.; Soman, K.P. Co-occurrence based word representation for extracting named entities in Tamil tweets. *J. Intell. Fuzzy Syst.* **2018**, *34*, 1435–1442. [CrossRef]
20. Ju, Z.; Wang, J.; Zhu, F. Named entity recognition from biomedical text using SVM. In Proceedings of the 2011 5th International Conference on Bioinformatics and Biomedical Engineering, Wuhan, China, 10–12 May 2011; IEEE: Piscataway, NJ, USA; pp. 1–4.
21. Hwang, Y.-G.; Yun, B.-H. HMM-based Korean Named Entity Recognition. *KIPS Trans. Softw. Data Eng.* **2003**, *10*, 229–236. (In Korean)
22. Seok, L.H. Named Entity Boundary Recognition Using Hidden Markov Model and Hierarchical Information. *J. Korea Acad.-Ind. Coop. Soc.* **2006**, *7*, 182–187. (In Korean)
23. Malik, M.K. Urdu Named Entity Recognition and Classification System Using Artificial Neural Network. *ACM Trans. Asian Low-Resource Lang. Inf. Process* **2017**, *17*, 13. [CrossRef]
24. Imam, A.T.; Alhroob, A.; Alzyadat, W.J. SVM Machine Learning Classifier to Automate the Extraction of SRS Elements. *Int. J. Adv. Comput. Sci. Appl.* **2021**, *12*, 174–185. (In English)
25. Eligüzel, N.; Çetinkaya, C.; Dereli, T. Application of named entity recognition on tweets during earthquake disaster: A deep learning-based approach. *Soft Comput.* **2022**, *26*, 395–421. [CrossRef]
26. Goyal, A.; Gupta, V.; Kumar, M. Recurrent neural network-based model for named entity recognition with improved word embeddings. *IETE J. Res.* **2021**, 1–7. [CrossRef]
27. Guo, S.L.; Yang, W.T.; Han, L.N.; Song, X.W.; Wang, G. A multi-layer soft lattice based model for Chinese clinical named entity recognition. *BMC Med. Inform. Decis. Mak.* **2022**, *22*, 201. [CrossRef]
28. Tsinganos, N.; Mavridis, I. Building and Evaluating an Annotated Corpus for Automated Recognition of Chat-Based Social Engineering Attacks. *Appl. Sci.* **2021**, *11*, 10871. [CrossRef]
29. Shah, S.A.A.; Masood, M.A.; Yasin, A. Dark Web: E-Commerce Information Extraction Based on Name Entity Recognition Using Bidirectional-LSTM. *IEEE Access* **2022**, *10*, 99633–99645. [CrossRef]
30. Gridach, M. Character-level neural network for biomedical named entity recognition. *J. Biomed. Inform.* **2017**, *70*, 85–91. [CrossRef]
31. Wang, M.; Zhou, T.; Wang, H.H.; Zhai, Y.C.; Dong, X.B. Chinese power dispatching text entity recognition based on a double-layer BiLSTM and multi-feature fusion. *Energy Rep.* **2022**, *8*, 980–987. [CrossRef]
32. Zhou, S.; Liu, J.; Zhong, X.; Zhao, W. Named entity recognition using bert with whole world masking in cybersecurity domain. In Proceedings of the 2021 IEEE 6th International Conference on Big Data Analytics (ICBDA), Xiamen, China, 5–8 March 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 316–320. [CrossRef]
33. He, S.; Sun, D.; Wang, Z. Named entity recognition for Chinese marine text with knowledge-based self-attention. *Multimed. Tools Appl.* **2021**, *81*, 19135–19149. [CrossRef]
34. Min, K.G.; Seok, L.H. Constructing for Korean Traditional culture Corpus and Development of Named Entity Recognition Model using Bi-LSTM-CNN-CRFs. *J. Korea Converg. Soc.* **2018**, *9*, 47–52. (In Korean) [CrossRef]
35. Yeon-Soo, Y.; Park, H.-R. Syllable-based Korean named entity recognition using convolutional neural network. *J. Korean Soc. Mar. Eng.* **2020**, *44*, 68–74. (In English) [CrossRef]

36. Marcińczuk, M.; Wawer, A. Named entity recognition for Polish. *Pozn. Stud. Contemp. Linguist.* **2019**, *55*, 239–269. (In English) [CrossRef]
37. Liu, C.G.; Yu, Y.L.; Li, X.X.; Wang, P. Named Entity Recognition in Equipment Support Field Using Tri-Training Algorithm and Text Information Extraction Technology. *IEEE Access* **2021**, *9*, 126728–126734. (In English) [CrossRef]
38. Ali, M.N.A.; Tan, G.; Hussain, A. Bidirectional Recurrent Neural Network Approach for Arabic Named Entity Recognition. *Future Internet* **2018**, *10*, 123. (In English) [CrossRef]
39. Wang, C.; Gao, J.; Rao, H.; Chen, A.; He, J.; Jiao, J.; Zou, N.; Gu, L. Named entity recognition (NER) for Chinese agricultural diseases and pests based on discourse topic and attention mechanism. *Evol. Intell.* **2022**, 1–10. [CrossRef]
40. Sun, J.; Liu, Y.; Cui, J.; He, H. Deep learning-based methods for natural hazard named entity recognition. *Sci. Rep.* **2022**, *12*, 4598. [CrossRef] [PubMed]
41. Zhou, J.H.; Li, X.Q.; Wang, S.P.; Song, X. NER-based military simulation scenario development process. *J. Déf. Model. Simul. Appl. Methodol. Technol.* **2022**, 15485129221094842. [CrossRef]
42. Dai, H.; Zhu, M.; Yuan, G.; Niu, Y.; Shi, H.; Chen, B. Entity Recognition for Chinese Hazardous Chemical Accident Data Based on Rules and a Pre-Trained Model. *Appl. Sci.* **2023**, *13*, 375. [CrossRef]
43. Zhang, Y.; Xu, J.; Chen, H.; Wang, J.; Wu, Y.; Prakasam, M.; Xu, H. Chemical named entity recognition in patents by domain knowledge and unsupervised feature learning. *Database J. Biol. Databases Curation* **2016**, *2016*, baw049. [CrossRef]
44. Liu, Z.J.; Yang, M.; Wang, X.; Chen, Q.; Tang, B.; Wang, Z.; Xu, H. Entity recognition from clinical texts via recurrent neural network. *BMC Med. Inform. Decis. Mak.* **2017**, *17*, 67. [CrossRef]
45. Wu, Y.; Jiang, M.; Lei, J.; Xu, H. Named Entity Recognition in Chinese Clinical Text Using Deep Neural Network. *Stud. Health Technol. Inform.* **2015**, *216*, 624–628.
46. Wei, Q.; Ji, Z.; Li, Z.; Du, J.; Wang, J.; Xu, J.; Xiang, Y.; Tiryaki, F.; Wu, S.; Zhang, Y.; et al. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 13–21. [CrossRef] [PubMed]
47. Ji, B.; Liu, R.; Li, S.; Yu, J.; Wu, Q.; Tan, Y.; Wu, J. A hybrid approach for named entity recognition in Chinese electronic medical record. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 64. [CrossRef] [PubMed]
48. Liu, W.; Fu, X.; Zhang, Y.; Xiao, W. *Lexicon Enhanced Chinese Sequence Labeling Using BERT Adapter*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 5847–5858. [CrossRef]
49. Yao, T.; Ding, W.; Erbach, G. CHINERS: A Chinese named entity recognition system for the sports domain. In Proceedings of the Second Sighan Workshop on Chinese Language Processing, Sapporo, Japan, 11–12 July 2003; pp. 55–62.
50. Nguyen, Q.-M.; Cao, T.-D. A novel approach for automatic extraction of semantic data about football transfer in sport news. *Int. J. Pervasive Comput. Commun.* **2015**, *11*, 233–252. [CrossRef]
51. Chiticariu, L.; Krishnamurthy, R.; Li, Y.; Reiss, F.; Vaithyanathan, S. Domain adaptation of rule-based annotators for named-entity recognition tasks. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, MA, USA, 9–11 October 2010; pp. 1002–1012.
52. Seti, X.; Wumaier, A.; Yibulayin, T.; Paerhati, D.; Wang, L.; Saimaiti, A. Named-entity recognition in sports field based on a character-level graph convolutional network. *Information* **2020**, *11*, 30. [CrossRef]
53. Liu, P.; Cao, Y. A Named Entity Recognition Method for Chinese Winter Sports News Based on RoBERTa-WWM. In Proceedings of the 2022 3rd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), online, 15–17 July 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 785–790.