applied sciences

MDPI

Article

A Chinese Named Entity Recognition Method Based on ERNIE-BiLSTM-CRF for Food Safety Domain

Taiping Yuan [1], Xizhong Qin [1,2,*] and Chunji Wei [1]

[1] College of Information Science and Engineering, Xinjiang University, Urumqi 830049, China
[2] Xinjiang Signal Detection and Processing Key Laboratory, Urumqi 830049, China
* Correspondence: qinxz@xju.edu.cn

Abstract: Food safety is closely related to human health. Therefore, named entity recognition technology is used to extract named entities related to food safety, and building a regulatory knowledge graph in the field of food safety can help relevant authorities to regulate food safety issues and mitigate the hazards caused by food safety problems. However, there is no publicly available named entity recognition dataset in the food safety domain. In contrast, the non-standardized Chinese short texts generated from user comments on the web contain rich implicit information that can help identify named entities in specific domains (e.g., food safety domain) where the corpus is scarce. Therefore, in this paper, named entities related to food safety are extracted from these unstandardized texts on the web. However, the existing Chinese named entity identification methods are mainly for standardized texts. Meanwhile, these unstandardized texts have the following problems: (1) their corpus size is small; (2) there are various new and wrong words and noise; (3) and they do not follow strict syntactic rules. These problems make the recognition of Chinese named entities for online texts more challenging. Therefore, this paper proposes the ERNIE-Adv-BiLSTM-Att-CRF model to improve the recognition of food safety domain entities in unstandardized texts. Specifically, adversarial training is added to the model training as a regularization method to alleviate the influence of noise on the model, while self-attention is added to the BiLSTM-CRF model to capture features that significant impact entity classification and improve the accuracy of entity classification. This paper conducts experiments on the public dataset Weibo NER and the self-built food domain dataset Food. The experimental results show that our model achieves a SOTA performance of 72.64% and a good performance of 69.68% for F1 values on the public and self-built datasets, respectively. The validity and reasonableness of our model are verified. In addition, the paper further analyses the impact of various components and settings on the model. The study has practical implications in the field of food safety.

Keywords: food safety supervision; named entity recognition; pre-trained language model; ERNIE; adversarial training; BiLSTM-CRF; self-attention

Citation: Yuan, T.; Qin, X.; Wei, C. A Chinese Named Entity Recognition Method Based on ERNIE-BiLSTM-CRF for Food Safety Domain. *Appl. Sci.* **2023**, *13*, 2849. https://doi.org/10.3390/app13052849

Academic Editor: Valentino Santucci

Received: 9 January 2023
Revised: 15 February 2023
Accepted: 20 February 2023
Published: 22 February 2023

1. Introduction

Food is the material basis for human survival, and food safety is closely related to human health. Especially in recent years, food safety problems (such as dyed steamed buns, expired meat, tainted bean sprouts, etc.) have occurred frequently. Some food safety incidents (e.g., the "melamine" milk powder, the "poisoned ginger", and the "smelly feet salt") were not brought to the attention of the public until they were exposed on the Internet and finally resolved. The issue of food safety has attracted a lot of public attention and has been a hot topic in the society. Food safety is not only related to people's health and safety but also has a significant impact on social stability and the credibility of the government. Although the state from the legislation to the supervision of all levels attach great importance to this topic, food safety issues such as excessive pesticide residues, genetically modified issues with melamine, Sudanese red pigment, "lean meat essence", "gutter oil",

plasticizers, and other prohibited issues have occurred repeatedly, stinging people's sensitive nerves. These are a reflection of the many food safety supervision problems that still exist.

With the rapid development of the Internet era, a huge amount of information is also being generated. People are used to expressing their various opinions through social media (e.g., Weibo) and e-commerce platforms (e.g., Jingdong). These short, unregulated texts generated from user comments imply a wealth of useful information that has not yet been fully explored. There are parts of this information that are critical to the management of food safety issues.

With the increasing standard of living of the people, people are eager to enhance their physical health and improve their medical care. However, in reality, intelligent healthcare in China is still in its infancy [1]. Artificial intelligence technology can be applied to health management to achieve intelligent health management [2–4]. By obtaining information about food safety from internet users' comments and analyzing them using AI technology, risks that may lead to diseases (e.g., food-borne diseases) can be identified, and referenceable risk prevention measures can be provided.

By building a knowledge graph for food safety, the following functions can be achieved:

1. To assist the relevant authorities in making rapid decisions and risk warnings;
2. To inform and alert the relevant authorities to take timely measures to reduce the probability of diseases caused by food safety problems;
3. To help the general public obtain food safety information from a large amount of cluttered and redundant data and to take suitable precautions in advance.

Named Entity Recognition (NER) technology can quickly and accurately acquire implicit information related to food safety in internet users' comments. It is used as data support for comprehensive research and judgment, intelligent decision-making, and dynamic early warning of food safety risk situations.

NER aims to identify predefined semantic types (e.g., names of people, places, organizations, and domain proper names) from text [5]. NER is one of the most fundamental and essential aspects of natural language processing (NLP), which has a wide range of applications in many scenarios, such as information extraction, question and answer systems, and machine translation. Named entities (NEs) are classified into generic classes (e.g., names of people, places) and domain-specific classes (e.g., proteins, drugs, diseases). The NER task is usually regarded as a sequence annotation task and is solved by statistical or neural network approaches [6].

With the rapid development of neural networks, deep learning-based methods [7] are widely used in NER tasks [8–11] and achieve state-of-the-art results. Deep learning-based methods do not rely on the manual construction of features and can automatically obtain models by training large amounts of data. Among them, based on the excellent sequence-modeling ability of one-way Long-Short Term Memory (LSTM) models, many methods use LSTM-conditional random field (CRF) as the main framework of NER tasks and the fusion of various relevant features on this basis. BiLSTM-CRF [8] is the most common method. In [12,13], state-of-the-art performance was achieved with this method as the main framework. In the low-resource domain, to improve entity recognition for the NER task, adversarial training is added to make the model more robust and improve the generalization ability during the training process [14,15].

Compared with English, there are no natural boundaries in Chinese. The blurred boundaries of unitary vocabulary, complex entity structures, and diverse expressions make Chinese NER more difficult. In the field of Chinese NER, a distinction between words and characters exists in Chinese. A character-based tagging strategy is generally used to tag named entities [16,17]. Compared with word-based and character-word union-based methods, the character-based method can avoid the problem of word separation error transmission and generally have superior performance [18,19]. However, the character-based Chinese NER methods have a general problem: they cannot characterize the polysemy of words. It means that the same word expresses different meanings in different

scenarios, but the representation of the word vector is the same, which is not consistent with the objective fact.

In recent years, pre-trained language models have become a critical fundamental technology in NLP, especially the Bidirectional Encoder Representations from Transformers (BERT) [20] model proposed by Google. BERT utilizes the multilayer self-attention bidirectional modeling capability of Transformer [21] to achieve state-of-the-art (SOTA) results in several NLP tasks by predicting masked characters. Pre-trained Language Models can be used for character representation. This is because better character representations not only contain rich syntactic and semantic information but also allow the modeling of polysemous words. However, the modeling objects of BERT are mainly focused on original language signals and less on the use of semantic knowledge units for modeling. This problem is particularly evident in Chinese. For example, BERT is modeled by predicting Chinese characters, and it is difficult for the model to learn the intact semantic representation from the larger semantic units. Therefore, Baidu improved the Chinese direction of BERT by proposing the Enhanced Representation from Knowledge Integration model (ERNIE), a semantic representation model for knowledge enhancement [22]. It empowers the model to learn the hidden knowledge implied in a large amount of text. As shown in Figure 1, the ERNIE enables the model to learn the semantic representation of complete concepts by masking semantic units such as words and entities. Compared to BERT, ERNIE directly models priori semantic knowledge units to enhance the semantic representation of the model.

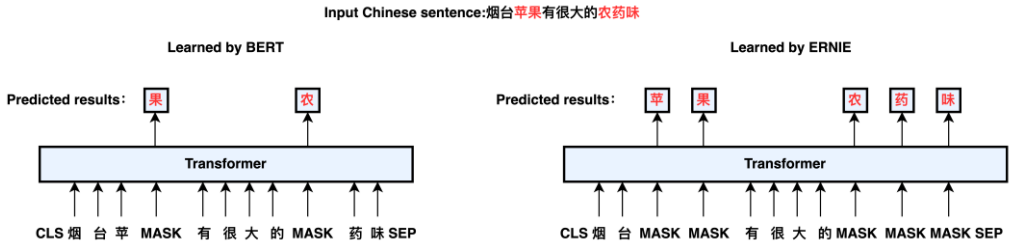

Figure 1. Comparison of learning methods between BERT and ERNIE model. ERNIE learns the semantic representation of complete concepts by masking Chinese characters and entities (“ 苹果” and “ 农药味”), while BERT learns the semantic representation by masking individual Chinese characters (“ 果” and “ 农”).

The NER mission has penetrated various areas, such as healthcare, finance, and law. However, there is no publicly available dataset in the food safety domain. Therefore, in domains where corpus resources are scarce, few-shot learning has emerged as an effective entity extraction method. [23] constructed a food recognition classifier based on few-shot learning by limiting the training samples. However, compared to standard news-like canonical texts, named entity recognition based on the content of Internet users' comments has the following main challenges:

- The corpus is small and contains many types of entities;
- Users write comments as they wish, with frequent new words and mistakes, containing noise such as Internet phrases and emoticons;
- The text does not follow strict syntactic rules [24].

These challenges make it difficult for few-shot learning to achieve good entity recognition results. Therefore, to obtain as many practical features as possible from the noisy mixed small-scale corpus to improve the recognition performance of food-safe named entities in unstandardized Chinese texts, this paper proposes the ERNIE-Adv-BiLSTM-Att-CRF model.

Our work has the following contributions:

1. We mine food safety-related information in unstandardized texts by building a food safety domain dataset to train a food safety domain NER model. Mining food safety-

related information in unstandardized texts helps build a regulatory knowledge graph in the food safety domain;
2. We use ERNIE, a model pre-trained by a large-scale corpus, to generate context-based word vectors so that the model learns the semantic representation of complete concepts and enhances the semantic representation of the model. We are adding self-attention to the BiLSTM-CRF sequence annotation model to capture the most critical semantic information in sentences and improve the entity recognition performance of unstandardized texts;
3. We added adversarial training as a regularization method to the model training to make the model more robust and improve the model's generalization ability.

In Section 4.2, the ERNIE-Adv-BiLSTM-Att-CRF model will improve NER tasks' performance with the unstandardized Chinese. As can be seen from the experimental results, our model outperforms other SOTA models on the public dataset Weibo NER with an F1 value of 72.64% and outperforms other baseline models on a self-built food domain dataset with an F1 value of 69.68%.

2. Related Work

A deep learning-based approach is used for unstandardized Chinese NER in our work. Collobert et al. [8] proposed a CNN-CRF model based on a convolutional neural network (CNN), which obtained competitive performance compared to various best statistical models. Huang et al. [9] proposed a BiLSTM-CRF model to solve the text sequence labeling problem as a benchmark model. Ma et al. [25] and Chiu et al. [11] infused character features extracted by CNN to enhance the word-level representation based on the BiLSTM-CRF framework. In the low-resource domain, Zhou et al. [16] enhanced the robustness of NER models by adding adversarial perturbations to the original samples. In the Chinese NER task, Dong et al. [26] composed a sequence of word roots for each character and used LSTM networks to obtain the root information of Chinese characters. Zhang et al. [27] proposed the Lattice-LSTM method by replacing the traditional LSTM cells with a lattice LSTM. It cleverly encodes the Chinese characters and all the potential words matched with the lexicon. Based on the Lattice-LSTM, Wei et al. [28] proposed the word-character LSTM (WC-LSTM) model to alleviate the impact of word separation errors by adding word information to the start and end characters of a word.

Chinese characters, as pictographs, contain potential glyph information. Ref. [29] proposed the fused glyph network FGN to extract the interaction information between distributed representations of characters and glyph representations through a fusion mechanism. Ref. [30] proposed the FLAT model by converting the Lattice structure into a planar structure composed of spans. This model makes full use of Lattice information and has good parallelization capability based on the power of the Transformer and well-designed positional encoding. The SLK-NER model proposed in [31] uses global semantic information to fuse lexical knowledge through an attention mechanism to integrate more informative words into a character-based model and alleviate lexical boundary conflicts. In NER domain research, introducing extra knowledge is a common way to improve model performance. Ref. [32] proposed an AESINER model that efficiently uses attention integration to encode and fuse different types of syntactic information (e.g., lexical annotation, constituent syntactic information, and dependent syntactic information) to help the model identify named entities. In social media such as Weibo and Twitter, many short texts generated by users contain various types of entities. Some entities are not written following standard syntactic conventions (e.g., abbreviated by users at will), resulting in a small probability of such entities showing sparsity, making recognizing such entities more difficult. In response to this question, previous studies have used domain information (e.g., gazetteers and embeddings trained on prominent social media texts) and external features (e.g., lexical tags) to help improve the performance of NER on social media [33,34]. However, these methods require extra work to obtain this information, and there is noise in the results. For example, training embeddings in the social media domain can bring

many unique expressions to the vocabulary. Therefore, [35] proposed the SA-NER model to enhance the recognition of named entities with semantic expansion. However, the F1 value of this model only reached 69.8% on the Weibo dataset.

However, most of the studies in the literature above are based on canonical texts and large-scale corpora. In contrast, the food safety domain corpus is scarce, and the implicit information in non-canonical texts on the web is not fully utilized. In addition, the above papers rarely enhance the entity recognition performance based on the features of the irregular Chinese text.

Therefore, to better represent syntactic semantic information in different contexts, we use ERNIE, a pre-training model more suitable for the non-standardized Chinese text NER task. ERNIE is based on character representation word vectors. Inspired by the attention-based BiLSTM model proposed in [36] for the relationship classification task, self-attention is added to the sequence annotation model BiLSTM-CRF for the NER task mechanism to obtain the most critical features for entity classification. The difference is that we use character-level feature vectors for entity classification instead of sentence-level features. Meanwhile, to improve the robustness and generalization of the model, we train the model adversarially by adding appropriate adversarial perturbations to the original samples.

3. Methods

In this section, we will introduce the ERNIE-Adv-BiLSTM-Att-CRF model in detail, as shown in Figure 2, which consists of six main parts:

- Input layer: Input samples in terms of sentences into the model;
- Embedding layer: The input sentences are represented using the ERNIE pre-training model to obtain a context-based word vector that contains rich implicit information;
- Adversarial training: Creating adversarial samples by adding an adversarial perturbation to the word embedding layer to train the model adversarially;
- BiLSTM layer: Using the BiLSTM network to learn the dependencies on the observed sequences and selectively picking higher-order features to integrate;
- Attention layer: Generates a weight vector and multiplies the weight vector with the state of the hidden layer at each moment of BiLSTM to obtain the feature vector after self-attention;
- CRF layer: Output the best-predicted label sequence.

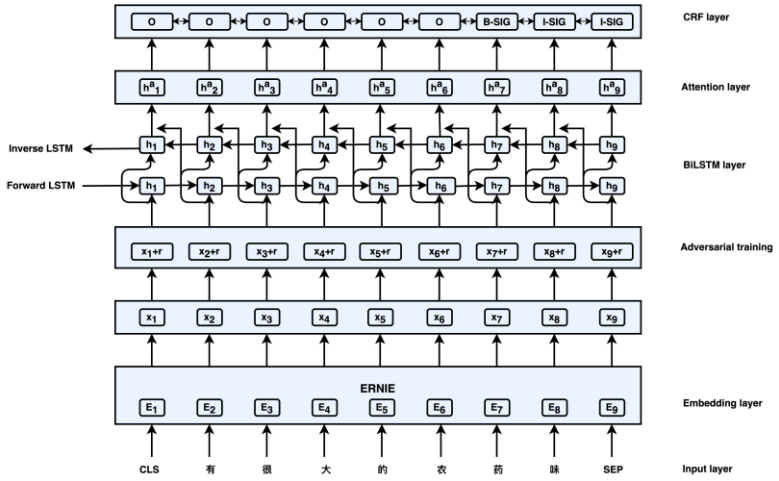

Figure 2. ERNIE-Adv-BiLSTM-Att-CRF model framework for named entity recognition. The model consists mainly of an input layer, an embedding layer, adversarial training, BiLSTM-CRF and self-attention.

3.1. Word Embedding

As with BERT, ERNIE uses a multi-layer Transformer as the primary encoder. A sentence $s = \{w_1, w_2, \cdots, w_t\}$ containing t tokens (characters) is connected with unique token CLS and SEP, where CLS denotes the beginning of the sentence and SEP denotes the end of the sentence. As shown in Figure 3, for each token in the sentence, ERNIE represents it as Embeddings constructed by summing Token Embeddings, Segment Embeddings, and Position Embeddings through the embedding layer, i.e., $E_{w_i} = E_{token} + E_{seg} + E_{pos}$. The sentence is vectorized into $emb = \{E_{w_1}, E_{w_2}, \cdots, E_{w_t}\}$ and input to the bidirectional Transformer for feature extraction. The contextual information of each token in the sentence is captured using the Self-Attention Mechanism of the Transformer to generate a sequence vector $x = \{x_1, x_2, \cdots, x_t\}$ containing rich semantic features. In other words, the rich text features are extracted using the pre-trained model ERNIE to obtain a $batch\_size * max\_seq\_len * emb\_size$ output vector, used as the classification task's sequence representation. Where $batch\_size$ is the batch size of the processed data, max_seq_len is the maximum length of the input sentences, and emb_size is the embedding dimension of each character.

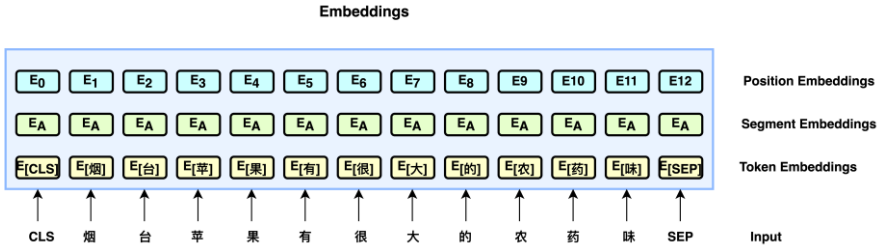

Figure 3. ERNIE embedding layer. ERNIE represents sentence as Embeddings constructed by summing Token Embeddings, Segment Embeddings and Position Embeddings through the embedding layer.

3.2. Bidirectional LSTM Network

The Long Short Terms Memory (LSTM) unit contains three specially designed gates (input gate, forgetting gate and output gate) for controlling the information transmission in a sequence. It can well solve the gradient disappearance and gradient explosion problems in the training process of traditional recurrent neural networks (RNN); at the same time, it has good sequence modelling ability and can better model long-range dependencies.

The Bidirectional Long Short Terms Memory (BiLSTM) network is a modification of RNN. It contains two sub-networks, forward and reverses LSTM, which can process contextual information simultaneously. The BiLSTM layer acts as a sequence encoder in NER tagging and inputs the word vector sequence of the input $x = \{x_1, x_2, \cdots, x_t\}$ sentence into the BiLSTM network for feature extraction. The probability of each token corresponding to the tag sequence $y = \{y_1, y_2, \cdots, y_n\}$ is output, and n is the number of tags. Specifically, ERNIE's word vector sequence output is encoded using the BiLSTM network. The forward LSTM network obtains the hidden forward state (historical features), and the reverse LSTM network obtains the backward hidden state (future features). The output of the hidden layer of the BiLSTM network is represented as:

$$h = \overrightarrow{LSTM(x)} + \overleftarrow{LSTM(x)} \tag{1}$$

The corresponding values between the forward and backward hidden states obtained by BiLSTM are summed to obtain h, where $h = \{h_1, h_2, \cdots, h_t\}$ is the hidden representation of the character. Input h into the Attention layer and use self-attention to obtain further the features that have the most significant impact on entity classification.

3.3. Attention Mechanism

The attention mechanism is a selection mechanism used to allocate limited information processing power, which is essentially a weighted summation, i.e., assigning higher weights to essential characters and smaller weights to other characters. The pre-trained model ERNIE uses a multi-layer Transformer as the primary encoder. The multi-headed attention mechanism in the Transformer structure can extract features of the text itself from multiple perspectives and levels. However, the "degree of influence" between the output information obtained from each time point of the LSTM is the same. In order to highlight the most critical part of the output information for entity classification, this paper adds self-attention after the BiLSTM network to capture the most critical semantic-level information in the sentence and automatically focus on the features that have a decisive impact on entity classification.

The matrix H is composed of the hidden state vector h output by the BiLSTM. Assume that w is the matrix parameter to be trained and w^T is the transpose of w, satisfying the following equation:

$$M = \text{relu}(H) \tag{2}$$

$$\alpha = \text{so}ftmax\left(w^T M\right) \tag{3}$$

$$r = H\alpha^T \tag{4}$$

where $H \in R^{d^w \times T}$. d^w is the word vector dimension in the sentence. α is the attention weight coefficient. The output h of the BiLSTM is weighted and summed to obtain r. The dimensions of w, α, and r are d^w, T, and d^w, respectively. After self-attention, we obtain the sentence representation vector containing the most critical information:

$$h^* = relu(r) \tag{5}$$

3.4. CRF Layer

Conditional Random Field (CRF) is a class of discriminative models best suited for prediction tasks. It is widely used in sequence labelling problems. Although the BiLSTM-Att network can handle long-range textual information and obtain more critical features for entity classification, the dependencies between neighboring tags are not effectively handled. The CRF layer is used as a sequence decoder for the NER tagger. The standard Viterbi algorithm obtains a globally optimal labeled sequence in the final decoding stage.

All outputs of the BiLSTM-Self-Attention (BiLSTM-Att) network are input to the CRF layer as a score matrix P. For a sequence of predicted tags $y = \{y_0, y_1, \cdots, y_{n+1}\}$, and the score is defined as:

$$x = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \tag{6}$$

where P is a matrix of size $t * n$, $P_{i,j}$ corresponds to the score of the i^{th} word in sen tence $x = \{x_1, x_2, \cdots, x_t\}$ corresponding to the j^{th} label. A is a matrix of transferred scores, and $A_{i,j}$ represents the scores transferred from label i to label j. The prediction sequence of y_0 and y_{n+1} are the two tags start and end that mark the beginning and end of the sentence, so A is a square matrix of size $n + 2$.

CRF uses potential functions to estimate the conditional distribution probability $P(x|w)$ of the output tag sequence y for sequence x. The formula is shown below:

$$P(y|x, w) = \frac{\exp\left(w^T \varphi(x, y)\right)}{Z(w, x)} \tag{7}$$

where $\varphi(x, y)$ is the feature vector and w is the parameter vector. $Z(w, x)$ is the cumulative sum of the conditional distribution probabilities $P(y|x, w)$ for all possible tags y.

Training the model utilizing the maximum conditional likelihood function:

$$x = argmax_w P(y|x, w) \tag{8}$$

The CRF layer learns some constraints from the training dataset, which reduces the sequence of invalid predicted tags and ensures that the final output tag sequence is valid. When decoding the sequences, the tag sequence with the highest prediction score is selected as the best answer:

$$x = argmax_y P(y|x, w) \tag{9}$$

3.5. Regularization

In NLP tasks, adversarial training is no longer used to defend against gradient-based malicious attacks but more to strengthen the regularization of classification models. Adversarial training can be generalized to the following maximum minimization formulation:

$$\min_{\theta} E_{(x,y)} \sim D\left[\max_{\Delta_x \in \Omega} L(x + \Delta_x, y; \theta)\right] \tag{10}$$

where inside the middle brackets is a maximization. D, x and y denote the training set, input samples and sample labels, respectively; θ, $L(x, y; \theta)$ and Δ_x are the model parameters, the loss of individual samples and the adversarial perturbation superimposed on the input, respectively; and Ω is the perturbation space. The perturbation space is generally small to avoid damage to the original input samples. $max(L)$ is the optimization objective, i.e., finding the perturbation that maximizes the loss of a single sample. Meanwhile, the model parameters θ of the neural network are optimized using the outer layer $E_{(x,y)}$ to minimize them. When the perturbation is fixed, the model has a minor loss of the training data. In simple terms, the sample loss should be as significant as possible after adding the perturbation. In contrast, the model loss should be as small as possible, thus making the model more robust and avoiding the bias of the model inference results caused by small perturbations on the samples.

In this paper, we borrow the Fast Gradient Method (FGM) from [37] for the text classification task and add an adversarial perturbation Δ_x to the word embedding to train the model adversarially, with Δ_x defined as follows:

$$x = \epsilon \cdot (g / \parallel g \parallel_2) \tag{11}$$

where $g = \Delta_x L(x, y, \theta)$ is the gradient of the input sample and $\parallel g \parallel_2$ is the L_2 parametrization of g.

Compute Δ_x from the gradient of the word embedding matrix and add it to the current embedding, which is equivalent to:

$$x_{adv} = x + \Delta_x \tag{12}$$

Compute its forward loss, backpropagate to obtain the adversarial gradient, accumulate to the original gradient, recover the embedding, and update the parameters based on the gradient with the accumulated adversarial gradient.

4. Experiments

In Sections 4.1 and 4.2, the data set used for the experiments, the settings of the relevant parameters, and the evaluation standard settings are presented. The proposed model in this paper is compared with some SOTA models, and the main experimental results are presented in Section 4.3. To verify the effectiveness of the components in our model, a series of ablation experiments are performed in Section 4.4. Our model is tested three times on each dataset, and this is used to calculate the average values of Precision (P), Recall (R), and F-score (F1). The bolded numbers in tables represent the better performance of our model over the comparison model.

4.1. Dataset and Experimental Setup

- **Datasets:** This paper constructed a food safety domain dataset for Food to conduct experiments. To ensure the fairness of the experiments, a widely used public dataset, Weibo NER, is chosen to validate the validity and reasonableness of our model. Both datasets use standard BIO annotation to represent the named entity tags of tokens in the input sentences.

The Weibo NER dataset is generated based on historical data filtered from Sina Weibo between November 2013 and December 2014. It contains 1890 microblog messages, annotated based on the annotation standard of DEFT ERE of LDC2014. The entity categories in this dataset are divided into four categories: Person, Organization, Address, and Geopolitical entity; and each category can be subdivided into specific (NAM, e.g., " 张三" labelled as "PER.NAM") and generic (NOM, e.g., "NOM" for " 男人").

The Food dataset was generated by filtering negative reviews about food (fruit category) from the sentiment/opinion/review propensity analysis dataset online_shopping_10_cats of the Chinese Natural Language Processing Language Resources Project (https://github.com/liuhuanyong/ChineseNLPCorpus (accessed on 20 August 2021)) and manually annotated using the NER annotation tool YEDDA [38]. First, food-safe named entities were extracted from the sentences, as shown in Figure 4; then, as shown in Figure 5, the sentences were processed into sequence labels to feed into the NER model for training. These reviews come from the Jingdong e-commerce platform and contain 1914 messages. Under the guidance of food safety-related experts, based on the content of the review, we assessed consumer evaluations about a particular type of fruit sold by a store on an e-commerce platform with a particular problem and the appearance of a specific symptom by consumers after consumption; the entity categories of this dataset are divided into five categories: type of fruit (FRU), description of the risk present (SIG), e-commerce platform sold (ECP), store sold (MER), and symptoms appeared (SYM).

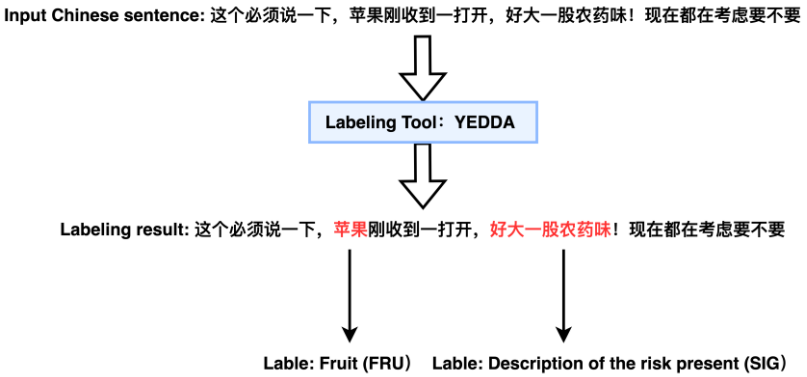

Figure 4. Manual annotation instance. The term " 苹果" represents the apple, indicating the presence of risky fruit categories, labeled "FRU", and " 好大一股农药味" indicates that the apple strongly smells of pesticides, which is labeled as food safety risk description "SIG".

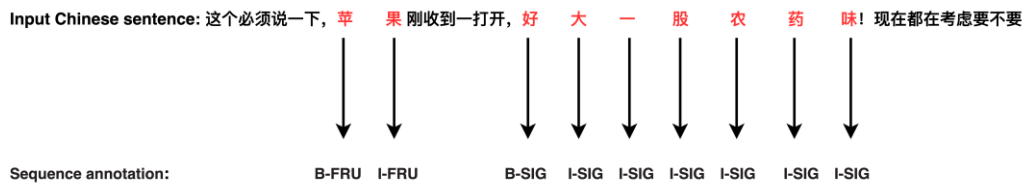

Figure 5. Sequence annotation instance. According to the BIO labeling scheme, B indicates the beginning of an entity, I indicates an intermediate entity, and O indicates a non-entity. The Chinese word " 苹果" indicates the apple, labeled as "B-FRU, I-FRU," and the Chinese characters " 好大一股农药味" describing apples with a strong pesticide smell is a food safety risk description and is labeled as "B-SIG, I-SIG, I-SIG, I-SIG, I-SIG, I-SIG, I-SIG". The rest of the Chinese characters are marked as "O".

The training, validation and test sets of the two datasets segmented according to the number of sentences are shown in Table 1. The Weibo NER dataset is its initial segmentation [33] and has not been changed in this paper.

Table 1. Dataset segmentation statistics.

Dataset	Train Set	Dev Set	Test Set
Weibo NER	1350	270	270
Food	1500	200	214

Furthermore, the named entity labels for both datasets are shown in Table 2.

Table 2. Dataset label statistics.

Dataset	Label
Weibo NER	PER.NOM
	PER.NAM
	LOC.NOM
	LOC.NAM
	GPE.NOM
	GPE.NAM
	ORG.NOM
	ORG.NAM
Food	FRU
	SIG
	ECP
	MER
	SYM

- **Hyper-Parameter Setting:** The experiments use ERNIE1.0, a knowledge-based augmented Chinese pre-training model released by Baidu, to train the word vectors. This pre-training model is improved on the Chinese direction of Google BERT and can be handled in the same way as BERT when used in downstream tasks and model transformation. According to the default configuration, the output vector size of each character is set to 768, and the dropout rate of ERNIE is 0.1. The LSTM is set as a bidirectional network, the hidden layer size is 768, the number of layers is 1, and the dropout rate of LSTM is set to 0.5. The initial learning rate is a critical parameter that needs to be adjusted according to the target task. AdamW is used as an optimizer for pre-training model fine-tuning and NER model training, both of which have different learning rates. For pre-training model fine-tuning, the initial learning rate is 3×10^{-5} for the Weibo NER dataset and 8×10^{-5} for the Food dataset; for NER model training, the LSTM learning rate is 2×10^{-5} for both datasets 2×10^{-2} for CRF. The difference between the optimal learning rates of ERNIE and BERT is extensive and requires a higher initial learning rate. Since the models' weights are randomly initialized at the beginning of training, choosing a more extensive learning rate at this time may bring instability (oscillation) to the model. In order to stabilize the model, the warm-up learning rate is chosen to make the model converge faster and better. The initial warm-up step number is set to 80.

4.2. Evaluation Standard Setting

In this paper, three experimental results of precision, recall, and F1 value are used as performance measurement criteria. Their calculation equations are as follows:

$$P_i = \frac{TP_i}{TP_i + FP_i} \tag{13}$$

In the precision (P) calculation Equation (13), TP_i denotes the number of positive classes correctly predicted by the model and FP_i denotes the number of positive classes predicted by the model from the negative classes.

$$R_i = \frac{TP_i}{TP_i + FN_i} \tag{14}$$

In the recall (R) calculation Equation (14), TP_i is the same as the above-mentioned equation and FN_i denotes the number of negative classes predicted by the model from the positive classes.

$$F1 = \frac{P * R * 2}{P + R} \tag{15}$$

Since precision and recall are a pair of contradictory metrics, in order to better evaluate the performance of the classifier, the harmonic mean $F1$ score of precision and recall is used as an evaluation standard to evaluate the comprehensive performance of the model.

4.3. Experimental Results

The results of the experiments on the datasets Weibo NER and Food are shown in Tables 3 and 4. Comparisons are made with other SOTA models on the Weibo NER dataset and some baseline models on the Food dataset.

Table 3. Detailed statistics of our model on Weibo NER.

Model	Weibo NER		
	P	R	F1
FLAT	N/A	N/A	60.32
SLK-NER	61.8	66.3	64.0
Locate and Label	70.11	68.12	69.16
AESINER	N/A	N/A	69.78
SA-NER	N/A	N/A	69.80
BERT-LMCRF	66.88	67.33	67.12
FLAT + BERT	N/A	N/A	68.55
FGN	69.02	**73.65**	71.25
Our Model	**72.82**	72.46	**72.64**

Table 4. Detailed statistics of our model on Food.

Model	Food		
	P	R	F1
ERNIE + Softmax	64.47	67.74	66.07
ERNIE + BiLSTM + CRF	66.37	69.12	67.72
Our Model	**68.44**	**70.97**	**69.68**

The ERNIE-Adv-BiLSTM-Att-CRF model proposed in this paper takes ERNIE-BiLSTM-CRF as the baseline and adds adversarial training and self-attention. In the SOTA model shown in Table 3:

1. Locate and Label uses a two-stage entity recognizer that locates entities and labels boundaries for nested NER;
2. SA-NER uses semantic expansion to improve the performance of NER;
3. AESINER improves the named entity recognition ability of the model by introducing extra knowledge;
4. SLK-NER uses an attention mechanism to fuse lexical knowledge into character-based models;
5. FLAT without the pre-trained model BERT uses Lattice information;
6. BERT-LMCRF is a BERT model that uses BiLSTM-CRF as a NER tagger;
7. FLAT + BERT is a SOTA model based on BERT;
8. FGN is a fused glyph network based on BERT.

Table 3 shows the data of each SOTA model are the experimental results in their original papers. Two baseline models are selected in Table 4 for comparison with the model in this paper:

1. ERNIE-based and using softmax for entity classification;
2. ERNIE-based and using BiLSTM-CRF as the NER sequence coder-decoder.

As shown in Table 3, our model outperforms other SOTA models on the Weibo NER dataset. Compared with the FLAT, SLK-NER, Locate and Label, and AESINER models, the F1 values obtained notable improvements of 12.32%, 8.64%, 3.48%, and 2.86%, respectively; compared with the BERT-LMCRF, FLAT + BERT, SA-NER, and FGN, they also obtained remarkable improvements in the F1 values, with 5.52%, 4.09%, 2.84%, and 1.39% improvement, respectively.

On the Food dataset, it can be seen from Table 4 that our model is significantly higher than the other two baseline models in terms of the F1 value, which is 69.68%. The improvement is 3.61% and 1.96%, respectively. It can be seen that adding the BiLSTM-CRF network after ERNIE is better than directly classifying the output of ERNIE for prediction, with an F1 value improvement of 1.65%. After adding adversarial training to the model training process and self-attention in BiLSTM-CRF, the model is further improved with another F1 value improvement of 1.96%.

From this, we can see that our model can alleviate the impact of noise on NER performance in Weibo NER and Food, two small-scale datasets with noise confounding. After self-attention, the features with a high impact on entity classification among all the features output by BiLSTM get greater weights, making entity recognition better.

4.4. Ablation Experiments

To demonstrate that adding self-attention and adversarial training can effectively improve the NER performance of small-scale datasets with noisy interference based on the ERNIE-BiLSTM-CRF as the baseline model, a series of ablation experiments are conducted in this paper using the Weibo NER dataset as an example. As shown in Table 5, the experimental results illustrate the effects of self-attention and adversarial training on NER performance.

Table 5. Performances of the various component on Weibo NER dataset.

Model	P	R	F1
Baseline	67.49	72.71	70.00
Baseline + Attention	69.93	70.77	70.35
Baseline + Adversarial	69.05	72.22	70.60
Baseline + both	**72.82**	**72.46**	**72.64**

The experimental results show that the dataset Weibo NER has an F1 value of 70.00% on the baseline model ERNIE-BiLSTM-CRF; only self-attention is added to the baseline model, and the F1 value is 70.35%, which is a mere 0.35% improvement. From this, it can be concluded that even if self-attention is added to the BiLSTM-CRF model to assign

greater weights to those features that impact entity classification, the noise still dramatically influences the model. Similarly, adding adversarial training to the baseline model only, the F1 value is 70.60%, which is just a 0.60% improvement. Although adversarial training can alleviate the effect of noise on entity recognition during model training, it does not capture the most important features without adding self-attention for the BiLSTM-CRF model. Ultimately, it does not obtain good entity labeling performance. While adding both simultaneously, the model has a remarkable improvement with an F1 value of 72.64%. Compared with the three models mentioned above, the F1 values are improved by 2.64%, 2.29%, and 2.04%, respectively. This shows that adding both self-attention and adversarial training can effectively improve the performance of small-scale nonstandard Chinese NER with noise by assigning more weight to the features that help entity recognition while reducing the effect of noise. The validity of our model for unstandardized Chinese NER is demonstrated.

Based on the above experiments, it can be seen that adding adversarial perturbation to the original samples and adding the self-attention mechanism to the BiLSTM-CRF network can both alleviate the effect of noise on the model and capture the features that are beneficial to entity classification. In addition, this paper further analyzes the effects of different pre-training models and adversarial training methods on entity recognition, as detailed in Tables 6 and 7.

Table 6. Performances of various Pre-trained Language models on the Weibo NER dataset.

Pre-Trained Model-Type	P	R	F1
BERT-base	67.76	70.05	68.88
RoBERTa-wwm-ext	69.21	72.22	70.69
ERNIE	**72.82**	**72.46**	**72.64**

Table 7. Performances of various Adversarial Training on Weibo NER dataset.

Adversarial-Type	P	R	F1
FreeLB	68.47	**73.43**	70.86
PGD	71.50	71.50	71.50
FGM	**72.82**	72.46	**72.64**

- **Pre-trained Language Model:** As shown in Table 6, the NER performance of different pre-trained models is analyzed with the other settings of our model held constant. BERT-base (https://github.com/google-research/bert (accessed on 12 September 2021)) and RoBERTa-wwm-ext (https://github.com/ymcui/Chinese-BERT-wwm (accessed on 21 September 2021)) are Chinese pre-trained models. Google publishes the former, and the latter is published by Xunfei Joint Lab of Harbin Institute of Technology. It should be noted that RoBERTa-wwm-ext is not the original RoBERTa model, but only a BERT model trained in a similar way to Roberta training, i.e., RoBERTa-like BERT.

It can be seen that the BERT-base-based NER model is a minor performance, with an F1 value of 68.88%, because the Chinese in BERT-base is sliced at character granularity, which does not consider Chinese word separation (CWS) in traditional NLP tasks. Instead, RoBERTa-wwm-ext uses Chinese Wikipedia (both simplified and traditional) for training and applies the Whole Word Masking (WWM) technique to Chinese. At the same time, the LTP of Harbin Institute of Technology is used as a word-splitting tool to Mask all Chinese characters that form the same word instead of being limited to Masking a single Chinese character in BERT-base. The F1 value is 70.69%, 1.81% higher than the F1 value of the BERT-base.

The F1 value of the ERNIE-based NER model is 72.64%, which is 3.76% and 1.95% higher than the BERT-base and RoBERTa-wwm-ext, respectively. Because the experimen-

tal datasets, Weibo and Food, are annotated from the unstandardized text generated by user comments on social media and e-commerce platforms. However, BERT-base and RoBERTa-wwm-ext use Wikipedia data for training, and they are more effective in modeling canonical text. ERNIE adds web data such as Baidu Encyclopedia, Baidu News, and Baidu Post, and it has advantages in modeling such unstandardized text. Therefore, BERT-base and RoBERTa-wwm-ext are not as effective as ERNIE for entity recognition in our datasets.

- **Adversarial Training:** The Fast Gradient Method (FGM), the Project Gradient Descent (PGD) [39], and the Free Large Batch Adversarial Training (FreeLB) [40] are three adversarial training methods, i.e., three different adversarial perturbation generation methods.

Table 7 shows the different performances of the three adversarial training modalities. The experimental results show that FGM outperforms the remaining two on the NER task with an F1 value of 72.64%. 1.78% and 1.14% higher than FreeLB and PGD, respectively. That is to say, adding the perturbation generated by FGM to word embedding can obtain higher accuracy of entity classification. FGM is more suitable for small sample NER models.

5. Conclusions

In order to obtain as many practical features as possible from the noisy mixed small-scale corpus to improve the performance of named entity recognition of unstandardized Chinese text, we propose the ERNIE-Adv-BiLSTM-Att-CRF model:

- ERNIE, a pre-trained model with advantages for modelling unstandardized Chinese text, is chosen to generate context-based word vectors that retain rich implicit information;
- Adversarial training is added to the model training as a regularization tool to alleviate the effect of dataset noise on the NER model;
- Self-attention is added to the BiLSTM network to automatically focus on the features that have a decisive impact on entity classification and encode them in sequence;
- Sequence decoding is performed in the CRF layer to obtain the best label corresponding to each token.

The experimental results show that our approach obtains SOTA performance on the public dataset Weibo NER with an F1 value of 72.64%. As shown in Table 3, our model outperforms other SOTA models on the microblog NER dataset. Compared with the FLAT, SLK-NER, Locate and Label, and AESINER models, the F1 values obtained significant improvements of 12.32%, 8.64%, 3.48%, and 2.86%, respectively; compared with the BERT-LMCRF, FLAT + BERT, SA-NER, and FGN models, our F1 values also obtained significant improvements of improved by 5. 52%, 4.09%, 2.84%, and 1.39%, respectively. Good performance was also obtained on the self-built in-domain dataset Food, with an F1 value of 69.68%. As can be seen from Table 4, our model is significantly higher than the other two baseline models in terms of F1 values, with improvements of 3.61% and 1.96%, respectively. The effectiveness of our method in the non-standardized NER task is fully demonstrated. Moreover, the performance of different pre-trained models and adversarial training methods are discussed in the ablation experiments.

Based on the experiments in this paper, it is demonstrated that our model can extract specific entity information from non-standardized web texts. This is useful for NER tasks in corpus-poor domains, such as the food safety domain. To address the problem that there is no publicly available NER dataset in the food safety domain, our approach can extract named entities related to food safety from short non-standardized Chinese texts generated from web users' comments. In this way, a regulatory knowledge graph in the food safety domain can be constructed to help relevant authorities to regulate food safety issues and mitigate the harm caused by food safety problems.

Author Contributions: Conceptualization and methodology, T.Y.; writing—original draft preparation, T.Y.; investigation, T.Y., C.W. and X.Q.; resources and supervision X.Q.; data curation, T.Y. and C.W.; project administration, X.Q.; All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the Major science and technology special projects of Xinjiang Uygur Autonomous Region (2020A03001) and its sub-program Key technology development and application demonstration of integrated food data supervision platform in Xinjiang region (2020A03001-2).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We would like to express our special thanks to the director of the Xinjiang Uygur Autonomous Region Institute of Product Quality Supervision and Inspection, Wensheng Ran, for his advice on the self-built food safety NER dataset in this paper and his participation in the project management of this paper's support project, the Xinjiang Region Food Big Data Comprehensive Supervision Platform Key Technology Development and Application Demonstration (2020A03001-2).

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

References

1. Yu, Q.; Jing, S.; Tai, Y.; Wu, S.; Chang, J.; He, P. Multidimensional Analysis of the Healthcare Big Data Policy Documents in China. *Chin. Gen. Pract.* **2019**, *22*, 3209.
2. Yu, K.; Tan, L.; Lin, L.; Cheng, X.; Yi, Z.; Sato, T. Deep-learning-empowered breast cancer auxiliary diagnosis for 5GB remote E-health. *IEEE Wirel. Commun.* **2021**, *28*, 54–61. [CrossRef]
3. Zhou, B.; Yang, G.; Shi, Z.; Ma, S. Natural Language Processing for Smart Healthcare. *arXiv* **2021**, arXiv:2110.15803. [CrossRef] [PubMed]
4. Yu, K.; Tan, L.; Mumtaz, S.; Al-Rubaye, S.; Al-Dulaimi, A.; Bashir, A.K.; Khan, F.A. Securing critical infrastructures: Deep-Learning-Based threat detection in IIoT. *IEEE Commun. Mag.* **2021**, *59*, 76–82. [CrossRef]
5. Nadeau, D.; Sekine, S. A survey of named entity recognition and classification. *Lingvist Investig.* **2007**, *30*, 3–26. [CrossRef]
6. Li, J.; Sun, A.; Han, J.; Li, C. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 50–70. [CrossRef]
7. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.
8. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv* **2015**, arXiv:1508.01991.
9. Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; Dyer, C. Neural architectures for named entity recognition. *arXiv* **2016**, arXiv:1603.01360.
10. Chiu, J.P.C.; Nichols, E. Named entity recognition with bidirectional LSTM-CNNs. *Trans. Assoc. Comput. Linguist.* **2016**, *4*, 357–370. [CrossRef]
11. Peters, M.E.; Ammar, W.; Bhagavatula, C.; Power, R. Semi-supervised sequence tagging with bidirectional language models. *arXiv* **2017**, arXiv:1705.00108.
12. Baevski, A.; Edunov, S.; Liu, Y.; Zettlemoyer, L.; Auli, M. Cloze-driven pretraining of self-attention networks. *arXiv* **2019**, arXiv:1903.07785.
13. Li, X.; Sun, X.; Meng, Y.; Liang, J.; Wu, F.; Li, J. Dice loss for data-imbalanced NLP tasks. *arXiv* **2019**, arXiv:1911.02855.
14. Cao, P.; Chen, Y.; Liu, K.; Zhao, J.; Liu, S. Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 182–192.
15. Zhou, J.T.; Zhang, H.; Jin, D.; Zhu, H.; Fang, M.; Goh, R.S.M.; Kwok, K. Dual adversarial neural transfer for low-resource named entity recognition. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 3461–3471.
16. Lu, Y.; Zhang, Y.; Ji, D. Multi-prototype Chinese character embedding. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, 23–28 May 2016; pp. 855–859.
17. 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. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 2742–2753.
18. Li, H.; Hagiwara, M.; Li, Q.; Ji, H. Comparison of the impact of word segmentation on name tagging for Chinese and Japanese. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, 26–31 May 2014; pp. 2532–2536.

19. Liu, Z.; Zhu, C.; Zhao, T. Chinese named entity recognition with a sequence labeling approach: Based on characters, or based on words? In *International Conference on Intelligent Computing, Xiamen, China, 29–31 October 2010*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 634–640.
20. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
22. Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Chen, X.; Zhang, H.; Tian, X.; Zhu, D.; Tian, H.; Wu, H. Ernie: Enhanced representation through knowledge integration. *arXiv* **2019**, arXiv:1904.09223.
23. Lu, Y.; Stathopoulou, T.; Vasiloglou, M.F.; Christodoulidis, S.; Stanga, Z.; Mougiakakou, S. An artificial intelligence-based system to assess nutrient intake for hospitalised patients. *IEEE Trans. Multimed.* **2020**, *23*, 1136–1147. [CrossRef]
24. Ritter, A.; Clark, S.; Etzioni, O. Named entity recognition in tweets: An experimental study. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK, 27–31 July 2011; pp. 1524–1534.
25. Ma, X.; Hovy, E. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; pp. 1064–1074.
26. Dong, C.; Zhang, J.; Zong, C.; Hattori, M.; Di, H. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition. In *Natural Language Understanding and Intelligent Applications*; Springer: Cham, Switzerland, 2016; pp. 239–250.
27. Zhang, Y.; Yang, J. Chinese NER using lattice LSTM. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 19–20 July 2018; pp. 1554–1564.
28. Liu, W.; Xu, T.; Xu, Q.; Song, J.; Zu, Y. An encoding strategy based word-character LSTM for Chinese NER. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 2379–2389.
29. Xuan, Z.; Bao, R.; Jiang, S. Fgn: Fusion glyph network for chinese named entity recognition. In *China Conference on Knowledge Graph and Semantic Computing*; Springer: Singapore, 2020; pp. 28–40.
30. Li, X.; Yan, H.; Qiu, X.; Huang, X. FLAT: Chinese NER using flat-lattice transformer. *arXiv* **2020**, arXiv:2004.11795.
31. Hu, D.; Wei, L. Slk-ner: Exploiting second-order lexicon knowledge for chinese ner. *arXiv* **2020**, arXiv:2007.08416.
32. Nie, Y.; Tian, Y.; Song, Y.; Ao, X.; Wan, X. Improving named entity recognition with attentive ensemble of syntactic information. *arXiv* **2020**, arXiv:2010.15466.
33. 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.
34. Aguilar, G.; Maharjan, S.; López-Monroy, A.P.; Solorio, T. A multi-task approach for named entity recognition in social media data. *arXiv* **2019**, arXiv:1906.04135.
35. Nie, Y.; Tian, Y.; Wan, X.; Song, Y.; Dai, B. Named entity recognition for social media texts with semantic augmentation. *arXiv* **2020**, arXiv:2010.15458.
36. Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; Xu, B. Attention-based bidirectional long short-term memory networks for relation classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Berlin, Germany, 7–12 August 2016; pp. 207–212.
37. Miyato, T.; Dai, A.M.; Goodfellow, I. Adversarial training methods for semi-supervised text classification. *arXiv* **2016**, arXiv:1605.07725.
38. Yang, J.; Zhang, Y.; Li, L.; Li, X. YEDDA: A lightweight collaborative text span annotation tool. *arXiv* **2017**, arXiv:1711.03759.
39. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv* **2017**, arXiv:1706.06083.
40. Zhu, C.; Cheng, Y.; Gan, Z.; Sun, S.; Goldstein, T.; Liu, J. Freelb: Enhanced adversarial training for natural language understanding. *arXiv* **2019**, arXiv:1909.11764.