International Journal of Geo-Information

MDPI

Article

CSMNER: A Toponym Entity Recognition Model for Chinese Social Media

Yuyang Qi [1], Renjian Zhai [1,*], Fang Wu [1], Jichong Yin [1], Xianyong Gong [1], Li Zhu [1] and Haikun Yu [2]

[1] Institute of Geospatial Information, Information Engineering University, Zhengzhou 450001, China; qi_yuyang@163.com (Y.Q.); wufang_630@126.com (F.W.); jichongy@whu.edu.cn (J.Y.); gongxygis@whu.edu.cn (X.G.); 18838226197@163.com (L.Z.)

[2] Henan Institute of Remote Sensing Surveying and Mapping, Zhengzhou 450003, China; hnwxfzx@163.com

* Correspondence: chxy_zrj@163.com

Abstract: In the era of information explosion, Chinese social media has become a repository for massive geographic information; however, its unique unstructured nature and diverse expressions are challenging to toponym entity recognition. To address this problem, we propose a Chinese social media named entity recognition (CSMNER) model to improve the accuracy and robustness of toponym recognition in Chinese social media texts. By combining the BERT (Bidirectional Encoder Representations from Transformers) pre-trained model with an improved IDCNN-BiLSTM-CRF (Iterated Dilated Convolutional Neural Network- Bidirectional Long Short-Term Memory- Conditional Random Field) architecture, this study innovatively incorporates a boundary extension module to effectively extract the local boundary features and contextual semantic features of the toponym, successfully addressing the recognition challenges posed by noise interference and language expression variability. To verify the effectiveness of the model, experiments were carried out on three datasets: WeiboNER, MSRA, and the Chinese social named entity recognition (CSNER) dataset, a self-built named entity recognition dataset. Compared with the existing models, CSMNER achieves significant performance improvement in toponym recognition tasks.

Keywords: Chinese social media; toponym recognition; named entity recognition

check for updates

Citation: Qi, Y.; Zhai, R.; Wu, F.; Yin, J.; Gong, X.; Zhu, L.; Yu, H. CSMNER: A Toponym Entity Recognition Model for Chinese Social Media. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 311. https://doi.org/10.3390/ijgi13090311

Academic Editors: Wolfgang Kainz, Mara Nikolaidou, Christos Chalkias, Marinos Kavouras and Margarita Kokla

Received: 15 June 2024
Revised: 25 August 2024
Accepted: 27 August 2024
Published: 29 August 2024

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

1. Introduction

In the context of today's exponential growth in information, social media platforms, as core channels for information exchange and sharing, have become an indispensable part of people's daily lives. The widespread use of social media such as Sina Weibo and X has led to the rapid growth of user-generated content (UGC), which, in turn, has increased the speed and breadth of information dissemination [1]. These rich UGC data are not only instantaneous and easy to access but also, more importantly, contain rich geographic information. This information is hidden in multimedia carriers such as text, images, and videos in unstructured or semi-structured forms [2–4]. These data not only intuitively record user locations in real-time but also map the behavior patterns, spatial preferences, and regional cultural characteristics of the individuals at a deeper level. These data are of important academic research value for in-depth analysis of population distribution characteristics, urban spatial patterns, and geographic connections of social networks, as well as provide valuable practical guidance for real-world applications such as urban planning and socioeconomic strategy formulation [5].

In the broad category of geographic information science and systems, accurately extracting toponym entities from text descriptions and locating them in a specific spatial reference system is a basic and important step in the automated semantic parsing of natural language texts. This process is of far-reaching significance for constructing intelligent information processing systems and promoting the deep integration of geospatial cognition and data analysis [6]. Compared with traditional structured text environments, social

media texts exhibit significant differences in their characteristics, posing unique challenges to toponym entity recognition (NER). Social media texts are usually presented in a short, concise, and information-dense way, quickly conveying rich information within a limited number of characters, which increases the complexity of place NER. This type of expression limits the high information density and leads to various information expression methods, often mixing formal language, internet slang, abbreviations, and regional dialects. Fusing these nonstandard language elements further aggravates the difficulty of information extraction. The internet slang words in social media texts, such as "吃瓜 (Observe the excitement)" and "酱紫 (Just like this)" as well as abbreviations such as "YYDS" and "U", both enrich the forms of expression and make the interpretation of the text information more ambiguous. The involvement of regional dialects, such as "咱这儿 (Our place)" and "中 (Okay)", adds local flavor to the text but also requires the recognition algorithm to have cross-regional adaptability. These characteristics, taken together, require that toponym recognition algorithms have excellent language adaptability, flexibility, and strong noise filtering capabilities, enabling them to accurately recognize toponym entities in texts with high information density and severe noise interference.

Compared with the research on entity recognition for English toponyms, the development of entity recognition technology for Chinese toponyms is still at a relatively early stage. High-quality research models are scarce, and sufficient and standardized training corpora are lacking, further increasing the difficulty in this research direction. In the academic exploration of NER, toponym recognition, as a key subtask of NER, has evolved from a preliminary to a mature stage. Early studies of toponym recognition were based on frameworks of rules and matching techniques, relying on sets of rules and dictionaries developed by experts. Subsequently, statistical models were introduced to capture the contextual features and patterns of toponyms. Recently, deep learning models have been introduced to drive toponym recognition to new levels. The challenge of toponym recognition also lies in how to construct corpora and optimize the model performance in noisy and diverse text expressions. When processing long texts, although their internal structure may be more complex, the richness of contextual information often allows toponym entities to be embedded in coherent contexts, thus reducing recognition difficulty. In contrast, in short texts on social media, the high fragmentation of information, casual expression, and lack of context pose significant obstacles to toponym recognition. Incorporating noise data such as emojis, hashtags, links, and other nonverbal information further blurs the boundary of toponym entities. Therefore, when constructing corpora for social media, balancing the proportion of noisy data is key to successful model learning [7,8].

The objective of this study is to explore and construct an efficient toponym recognition method for Chinese social media texts. With the help of the powerful representation ability of pre-trained models, a general and reliable toponym extractor is developed to solve the problem of toponym NER in Chinese social media. This research focuses on two core challenges: first, how to construct a rich corpus for Chinese social media recognition; and second, how to optimize the model performance in social media texts with considerable noise and diverse expressions to improve the accuracy and robustness of toponyms recognition. The main contributions are as follows:

1. Based on the bidirectional encoder representations from the transformers (BERT) pre-trained model and the improved IDCNN-BiLSTM-CRF (Iterated Dilated Convolutional Neural Network- Bidirectional Long Short-Term Memory- Condi-tional Random Field) model, the CSMNER (Chinese social media named entity recognition) model is proposed. The model uses the improved IDCNN and BiLSTM dual-channel joint feature extraction modules to extract local boundary features and contextual semantic features of toponyms, respectively and introduces the boundary extension (BE) module to enhance the perception of toponym boundary information, ultimately improving the overall performance of the model.

2. A Chinese social named entity recognition (CSNER) dataset is constructed. The dataset is sourced from the Sina Weibo platform, containing three entity categories with a total of 68,864 annotated samples. The dataset alleviates the scarcity of corpora for NER in Chinese social media, providing richer and more diverse data support for Chinese NER tasks.
3. To verify the superiority of the proposed method, specific modules such as the improved IDCNN, BiLSTM, and BE are investigated and discussed. A series of evaluations are conducted on the MSRA, WeiboNER, and CSNER datasets. Through a comprehensive comparison with other advanced models, comprehensive experimental results on general datasets in NER confirm the effectiveness of the proposed model.

The rest of this paper is organized as follows. Section 2 describes the current research work in toponym entity recognition. In Section 3 the corpus creation and processing procedure is introduced, focusing on the proposed method. Section 4 presents the experimental data and results. Conclusions and suggestions for further research are provided in Section 5.

2. Related Work

As a key component of geographic information, toponyms must be accurately collected and processed from various data sources before their effective integration and application [9–11]. As a subtask of named entity recognition, toponym named entity recognition can be divided into four categories: rule-based methods, gazetteer-based methods, statistical methods, and deep learning-based methods [12].

2.1. Rule-Based Methods

The toponym recognition strategies based on rules and matching techniques rely on the sets of rules carefully designed by experts and the comprehensive gazetteers to design a series of rules and patterns to match and extract the toponym information from the text. These methods usually rely on predefined rules and gazetteers. These rules may include typical toponym features, such as specific suffixes, prefixes, or common terms within toponyms. Giridhar et al. detected and located incident points by analyzing road traffic-related Twitter posts, using a part-of-speech (POS) tagger to mark the original tweets and identify toponym information based on nouns, determiners, and other items [13]. Dutt et al. first used an available POS tagger to identify proper nouns based on a heuristic algorithm; then, they employed a variety of regular expression matches to reduce the ambiguity of proper nouns appearing in different sentence positions [14]. Although rule-based methods exhibit certain accuracy and interpretability when dealing with clear structures and known toponyms, they have a limited ability to address complex contexts and toponym ambiguities. They heavily rely on manual work and require considerable time to establish a relevant knowledge base and vocabulary, which are difficult to update and maintain. Therefore, they may not be suitable for processing the diverse and non-standardized language found in social media.

2.2. Gazetteer-Based Methods

A gazetteer is a geospatial dictionary that contains toponyms and their corresponding geographic coordinates. Currently, two representative geospatial gazetteers in use are OpenStreetMap and GeoNames [15]. To capture the location of plane crashes, Milusheva et al. used OpenStreetMap, GeoNames, and Google Places to establish a gazetteer and develop an improved address parsing algorithm [16]. De Bruijn et al. matched tweet texts with a gazetteer to achieve geographic location inference from tweets [1]. Middleton constructed an OpenStreetMap-based geographic parsing algorithm and proposed a geotagging algorithm combining many social media tags and multiple gazetteers [17]. Although geographic entity recognition assisted by gazetteers has achieved high accuracy, the rapid acceleration of global urbanization leads to the constant emergence of new toponyms, resulting in many geographic indications not being included in existing gazetteers. This limits the effectiveness and generality of static gazetteer-based recognition

systems, especially in fast-changing urban areas, where the names of new streets, communities, and public facilities are frequently updated, making the lagging issue of gazetteers particularly prominent.

2.3. Statistical Methods

With the maturation of statistical learning theory, a series of algorithms based on probabilistic models, such as hidden Markov models (HMMs) [18], conditional random fields (CRFs) [19,20], and maximum entropy models [21,22], have been applied to the sequence labeling tasks of toponym recognition. These models capture the contextual features and distribution patterns of toponym entities by constructing feature functions. Compared with the rule-based matching methods, they improve the recognition accuracy and generalization ability to a certain extent. For example, Wang et al. used a deep belief network (DBN) for toponym recognition, simultaneously using word representation and model explanation. The experimental results showed that this method outperforms the CRF model on a small Chinese corpus [23]. To solve the problem of unknown multiword toponyms for which machine learning methods lack sufficient data, Hu et al. combined rules, gazetteers, and deep learning methods to achieve good performance [24]. The advantage of statistical methods is that they can automatically capture patterns in the data, reducing manual intervention. However, when faced with the complexity and diversity of language expressions, such as the variability of syntactic structures and the emergence of new types of toponym entities, the performance improvement of these models has encountered a bottleneck. Additionally, these models heavily depend on high-quality annotated data and may not be flexible enough to deal with low-frequency toponyms and language variations.

2.4. Deep Learning-Based Methods

In recent years, the rise of deep learning methods has significantly improved the performance of toponym recognition tasks. In particular, deep neural network models such as recurrent neural networks (RNNs), long short-term memory networks (LSTMs) [25], attention mechanisms [26], and the transformer architecture can both capture word-level information and effectively integrate contextual semantics [27–29]. The accuracy and robustness of toponym recognition in complex text environments are significantly improved through multilevel feature learning mechanisms. For the task of Chinese NER, researchers have proposed a variety of deep learning models that incorporate lexical information, significantly enhancing the recognition accuracy and robustness by fusing external lexical knowledge. A typical method is Lattice-LSTM, which constructs a word-level lattice structure on character sequences to effectively use lexical information, thereby improving the model performance [30]. Inspired by such methods, Xu et al., proposed a Chinese NER model with character-level word embeddings, which further enhances the performance of the NER model by using multigranular semantic information at the radical, character, and word levels [31]. Pretrained language models such as BERT [32] learn rich language representations through unsupervised pretraining on large-scale corpora, demonstrating remarkable contextual understanding capabilities. BERT and its subsequent variants can efficiently and accurately recognize various types of toponym entities solely through a deep contextual understanding of the input text, without the need for manual feature engineering, pushing toponym recognition technology to an unprecedented level. Berragan et al. used pre-trained models such as BERT, RoBERTa, and DistilBERT for geographic NER and demonstrated excellent performance compared to previous work [33]. These advancements both improved toponym recognition and provided powerful technical support for fields such as geographic information science and geocomputation for social sciences.

3. Materials and Methods

3.1. Corpus Collection and Settings

To effectively alleviate the scarcity of corpus resources in Chinese social named entity recognition, a new corpus was collected and created in this paper—CSNER. This corpus aims to provide richer and more diverse data support for Chinese NER tasks, thus promoting the development and improvement of related technologies. The CSNER covers three main entity categories, person, location, and organization, with a total of 68,864 samples.

3.1.1. Corpus Sources

The corpus of CSNER was sourced from Sina Weibo, one of the largest social media platforms in China. Self-designed web crawler technology was used to obtain a large quantity of targeted social text data from the internet. After systematic screening and data cleaning, the Weibo content of real users was sorted to construct a corpus covering a total of 68,864 information-rich Weibo posts published by 3449 users from 2015 to the present. After analyzing the word frequency statistics of these microblogs, the top 2000 words with the highest frequency of occurrence were taken to build a word cloud, as shown in Figure 1. From the word cloud, it can be found that the CSNER dataset can more accurately reflect the language style, expression habits, and toponym usage on the microblogging platform because it comes directly from the microblogging platform. Compared with other datasets, other public datasets do not contain the latest online phrases, hot topics, or emerging toponyms because they were created earlier. The CSNER dataset was constructed more recently, which enables the CSMNER model to capture the latest trends and be more up-to-date. The self-constructed dataset allows for strict quality control to ensure annotation quality and consistency and reduce noise. The microblog text in the CSNER dataset usually contains a large amount of informal language, abbreviations, network terms, and emoticons, and contains many names of people, toponyms, organizations, and other words of different natures. They not only reflect the users' daily life and interests but also contain a lot of rich geographic information. It more fully reflects the linguistic characteristics of the Chinese social media environment. By collecting these social media text data, we have constructed a comprehensive and authentic Chinese social media corpus, which provides a solid foundation for the task of Chinese toponym entity recognition.



Figure 1. Word frequency statistics of the CSNER (Chinese social named entity recognition) dataset generated using the word cloud. The image lists the top 2000 words with the highest frequency of occurrence, including a wide range of words such as personal names, toponyms, names of institutions and organizations, locatives, adverbs of place, nouns, but excluding many meaningless verbs, adjectives, and intonational auxiliaries.

3.1.2. Corpus Annotation

Among many annotation methods, we chose the widely adopted begin-inside-outside (BIO) annotation method, which divides entities into three types of tags, namely, B, I, and O to locate and classify entities. Specifically, "B" represents the beginning of an entity and is used to mark the starting position of the entity, "I" represents the inside of an entity and is used to mark the continuation of the entity, and "O" represents the outside of an entity and is used to mark the nonentity part. The advantage of this annotation method lies in its simplicity and clarity, which can clearly define the boundary of the entity and make it easier for models to learn the rules of entity recognition.

When constructing the CSNER corpus, we used a simplified entity classification system to ensure that the model can accurately identify and process the key information frequently mentioned by users in social interactions. Specifically, we divided the entities in the corpus into three categories: location (LOC), organization (ORG), and person (PER). This classification system aims to capture the most common and information-rich entity types in social texts. The annotation of location entities involves the recognition of geographic names, which include specific geographic locations such as cities, countries, and landmarks and references to abstract geographic areas or cultural areas. Organization entities involve the recognition of companies, nongovernmental organizations, educational institutions, and various other groups or institutions; these entities are often associated with specific events, activities, or topics in social texts. Person entities include recognizing individual names, which may include public figures, historical figures, cultural figures, and other individuals mentioned by users. Table 1 shows some of the annotated information from the CSNER corpus.

Table 1. Examples of annotated corpus information. B represents the beginning word of an entity, I represents the middle or ending word of an entity, LOC stands for location entity, ORG for organization entity, and PER for person entity.

No.	Type	Number of Annotated Entities	Examples (English)	Example (Chinese)	Example of Labeling
1	Location	9897	Tiananmen Gate	天安门	{"天安门", [B-LOC, I-LOC, I-LOC]}
2	Organization	9119	European Union	欧盟	{"欧盟", [B-ORG, I-ORG]}
3	Person	13,426	Liu Yang	刘洋	{"刘洋", [B-PER, I-PER]}

The data annotation session is shown in Figure 2, and the study first uses the Baidu UIE (universal information extraction) interface for automatic annotation [34]. This step greatly improves annotation efficiency and ensures the rapid completion of basic annotation. However, errors may exist in automatic annotation. Therefore, Doccano, a collaborative annotation platform, was further used for manual correction and refined management. The use of the Doccano platform both improves the consistency of annotations and ensures annotation accuracy through multi-person collaboration. This ensures that the final generated corpus meets the high standard requirements in natural language processing and provides reliable data support for various downstream tasks.

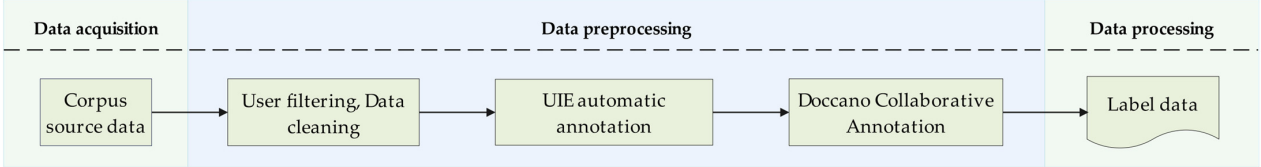


Figure 2. Corpus annotation process.

3.2. Proposed Method

3.2.1. Overall Framework and Workflow of the Model

The overall structure of the proposed CSMNER model is shown in Figure 3, which consists of three main components, input representation layer, feature coding layer, and label decoding layer. The input representation layer is the unit that encodes the input corpus features using the BERT pre-training model; the feature encoding layer is composed of three parts, namely, the IDCNN network, the dual-channel feature extraction unit composed of the BiLSTM network, and the BE module proposed in this paper; and the label decoding layer consists of the CRF module.

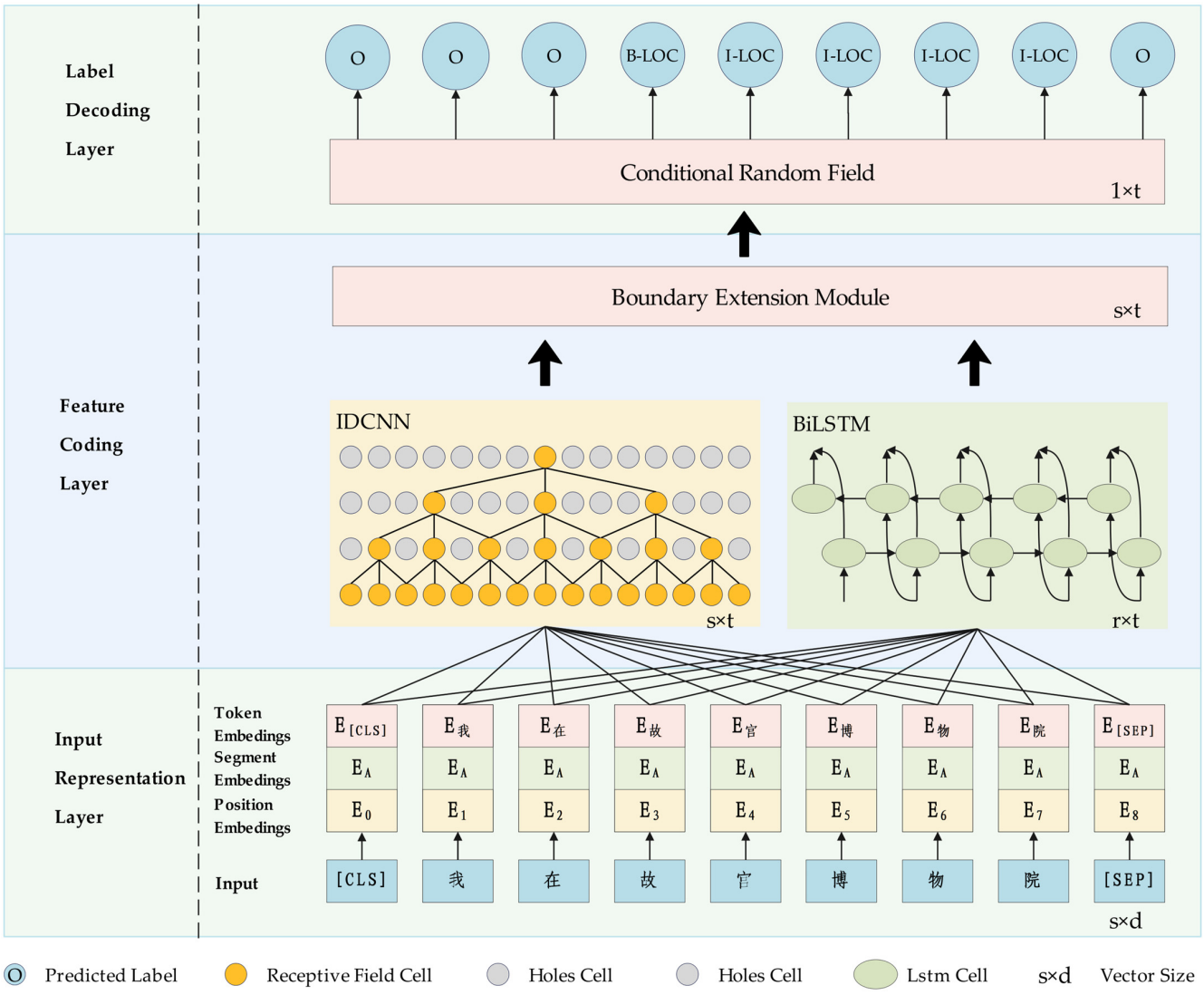


Figure 3. The architecture of the CSMNER (Chinese social media named entity recognition) model. Inputting the sentence "[CSL] I am at the Palace Museum [SEP]" demonstrates the overall structure of the model.

3.2.2. Input Presentation Layer

The method used in this paper employs the BERT pre-trained model to represent the input text. The BERT pre-trained model, with its innovative bi-directional Transformer encoding architecture, demonstrates an unrivaled advantage over traditional manual feature engineering, being able to more accurately understand and capture the subtle semantics of words in context [26,32]. This ability is particularly evident when fine-tuned to specific downstream tasks, making it ideal for processing short Chinese social texts.

Unlike traditional static word vector embeddings, BERT performs two core tasks—the masked language model (MLM) and next sentence prediction (NSP) —in the pretraining stage to generate different vector representations for each word according to the context. This dynamic method enables BERT to better consider word contexts, especially in Chinese, where a word can have multiple meanings. Thus, the embedding vector output of the BERT pretrained model can significantly improve its performance.

The input text is decomposed into tokens by the WordPiece word segmentation strategy, and the strategy is shown in Table 2. A greedy algorithm is used to segment words. It starts from the longest known word and tries to match the text fragments in the sentence. Once a match is found, it is treated as a token, and then it continues to search for the next longest match in the remaining text. If the remaining part no longer fully matches any token in the vocabulary, it is split into smaller parts until all the text is decomposed. If a text segment cannot match any token in the vocabulary, WordPiece then uses a special UNK (unknown) token to represent the unknown word, which helps to handle new or rare words not included in the vocabulary. This strategy endows the model with a superior ability to capture semantic information of fragmented text, enabling it to cope with complex and trivial Chinese social short texts containing many abbreviations, dialects, and Internet slang.

Table 2. Examples of WordPiece word segmentation strategy.

No.	Original Sentence	Translate Sentences	Word Segmentation Result
1	明天我们将进行一场mindstorming 会议。	Tomorrow we will have a mindstorming meeting.	[‘明天’, ‘我们’, ‘将’, ‘进行’, ‘一场’, ‘mi’, ‘##nd’, ‘##sto’, ‘##rming’, ‘会议’, ‘。’]
2	我在北京故宫博物院	I’m at the Palace Museum in Beijing.	[‘我’, ‘在’, ‘北京’, ‘故宫’, ‘博物院’]

“Mindstorming” is not included in the vocabulary, so it is split into a sequence of multiple tokens, each preceded by the ## symbol to indicate that they are contiguous parts of the same original word. In this way, even if the model has never seen the word “mindstorming”, it can understand the meaning of this word by learning the contextual meaning of the tokens.

BERT uses MLM to learn bidirectional contexts. It randomly selects a portion (usually 15%) of the vocabulary in the input text and replaces it with a special token [MASK]. The model then predicts the masked original word based on the context information around the word. This process forces the model to understand the complex relationships between words, including dependencies, co-occurrence patterns, and syntactic and semantic roles. As a result, this rich contextual information is encoded in the vectors while the character-level and word-level features of Chinese are obtained.

Although in some subsequent models (such as RoBERTa), the next sentence prediction (NSP) task is considered possibly unnecessary and is removed [35]; however, in the BERT model, it exists as an auxiliary task. By judging whether a pair of sentences are adjacent in the original text, the model can learn the coherence and contextual relationships at the sentence level, thus providing more effective features for short information-poor social texts. By combining the MLM and NSP tasks, the BERT model can capture contextual information at both word and sentence levels. These pieces of information are encoded into a 768-dimensional vector, with each dimension representing a certain language feature or pattern learned by the model. These feature vectors fully capture the Chinese POS, syntactic structure, and more complex linguistic phenomena so that the model can be adapted to NER tasks during fine-tuning.

3.2.3. Feature Encoding Layer

In English, a word can convey complete semantic information, while in Chinese, a phrase is needed to express the complete meaning. Moreover, Chinese does not have obvious word boundary characters or capitalization features, which makes entity boundary recognition difficult. Therefore, word boundary information is an important factor in

Chinese NER [36]. The model proposed in this paper uses two neural network structures, the improved IDCNN and BiLSTM, to extract semantic features. Figure 4 shows that the IDCNN is an iterative dilated convolutional neural network. By introducing holes in the convolution kernels, it can expand the receptive field without increasing the number of parameters and computational complexity, better capturing long-range dependence information in the text. By stacking convolutional layers with different dilation rates, features at different scales can be extracted, providing multiscale feature representations for the text sequence. This multilevel feature extraction helps capture the subtle differences in entity boundaries and the complex features of entity types, thus improving the recognition accuracy of NER. In text, the convolutional operation can capture local patterns, such as phrases or word groups, which are especially useful for recognizing short toponyms. When sliding over the entire text sequence, the weights of the convolutional operation are shared, which helps the model learn common features across text positions.

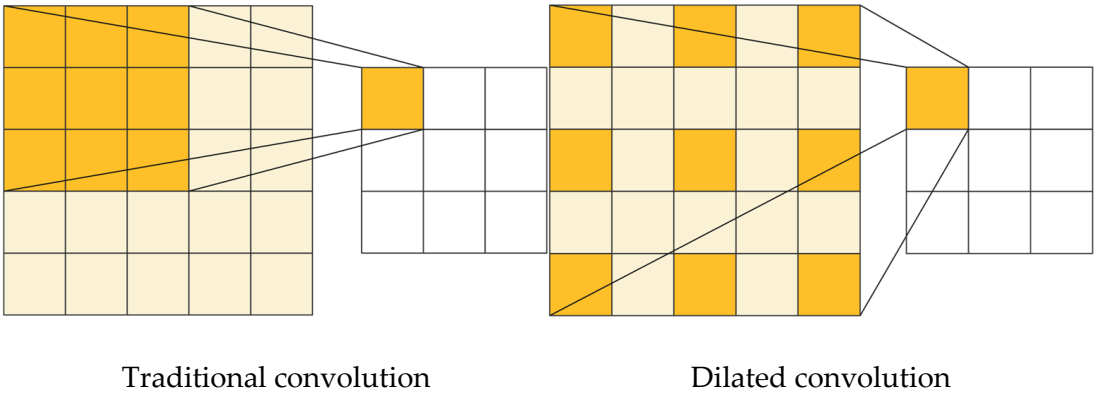


Figure 4. Comparison between dilated convolution and traditional convolution.

Compared with the traditional IDCNN module, the proposed model replaces the rectified linear unit (ReLU) activation function with the Gaussian error linear unit (GELU). The GELU is a smooth function, and its nonlinear nature is more complex than that of the ReLU, as shown in Figure 5. This means that the GELU generates a more stable gradient flow during optimization, which contributes to convergence stability during model training and reduces the oscillation during training. Its design is inspired by the cumulative distribution function (CDF) of the Gaussian distribution, which in theory can better simulate the probability distribution characteristics of natural language when processing natural language data, thereby improving the model's ability to learn language features.

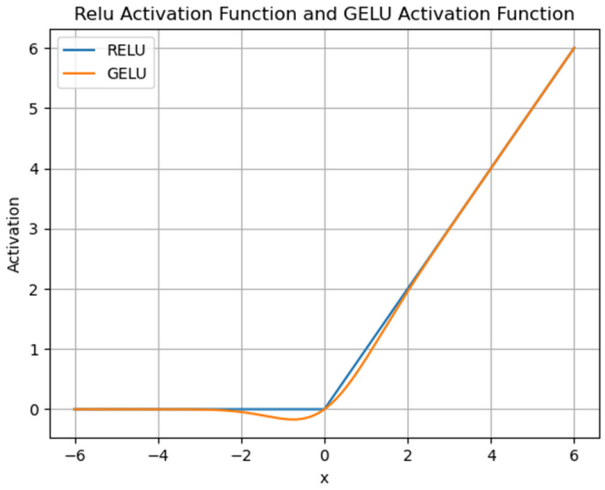


Figure 5. Comparison of ReLU and GELU activation functions.

BiLSTM is a neural network structure widely used in NER tasks. Through its unique gating mechanism, this structure can effectively capture long-distance dependencies so that the model can better consider the forward and backward information of the input sequence. In tasks such as toponym recognition, there is often a long distance between an entity and its modifiers or explanations. For example, in the "Nanjing Changjiang Bridge", such an entity and modifier may be separated by numerals, adjectives, or other nouns, making such contextual information crucial for accurately recognizing toponyms. In this context, BiLSTM has unique advantages. Its bidirectional structure allows the model to simultaneously consider information from before and after the input sequence. Here, the forward LSTM is taken as an example, which processes the input sequence from left to right. The hidden state h_t and cell state c_t of each unit are calculated by the following formulas:

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \tag{1}$$

$$f_t = \sigma\left(W_{fx}x_t + W_{fh}h_{t-1} + b_f\right) \tag{2}$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \tag{3}$$

$$\widetilde{c}_t = \tan\mathrm{h}(W_{cx}x_t + W_{ch}h_{t-1} + b_c) \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \widetilde{c}_t \tag{5}$$

$$h_t = o_t \odot \tan\mathrm{h}(c_t) \tag{6}$$

where i_t, f_t, o_t and $\widetilde{c}_t$ are the input gate, forget gate, output gate, and candidate state representing current information, respectively; W and h are the weight matrix and bias term, respectively; σ and $\tan\mathrm{h}$ are the sigmoid and hyperbolic tangent functions, respectively; and $\odot$ denotes elementwise multiplication. Similarly, the output of the backward LSTM can be obtained. The final output of the BiLSTM is obtained by concatenating the hidden states from both directions as $y_t = [h_t; h'_t]$.

To effectively integrate lexical and syntactic features from various aspects in the Chinese NER model and to reduce the interference of noisy information, we introduce the BE module at the end of the feature encoding layer, the structure of which is shown in Figure 6. By integrating the attention mechanism and the improved IDCNN feature vectors, this module focuses on and weighs the input sequences to ensure that the recognition of each potential toponym can focus on the most critical information segments. It highlights the local features of toponyms and expands the contextual features of toponym boundaries. Specifically, after the linear transformation of the extracted X_{BiLSTM} feature vectors, the X_{IDCNN} vectors are utilized to generate the query matrix Q_i by multiplying with t W_i^Q matrices, respectively, and then the X_{BiLSTM} vectors are utilized to generate the key matrix K_i and value matrix V_i by multiplying with t W_i^K and W_i^V matrices, respectively, with the formula of Equation (7). After obtaining the Q, K, and V matrices for the t attention heads, respectively, the attention score A_i is computed in each attention head using the formula for Q_i and K_i (Equation (8)), and the attention score matrix O_1 is obtained by using the obtained t A_i weighted and summed with V_i, respectively. To reduce the loss of features to the IDCNN network, the O_1 vector is again spliced and fused with the X_{IDCNN} vector to obtain the O_2 vector. A linear variation was performed to obtain the O_3 vector, which was directly summed with the X_{IDCNN} vector to obtain the final feature vector V.

$$Q_i = X_{IDCNN}W_i^Q, K_i = X_{BiLSTM}W_i^K, V_i = X_{BiLSTM}W_i^V \tag{7}$$

$$A_i = \mathrm{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_k}}\right) \tag{8}$$

where W_i^Q, W_i^K, and V_i^V are the linear transformation matrices of the queries, keys, and values, respectively; X_{BiLSTM} and X_{IDCNN} are the feature vectors obtained by BiLSTM and

IDCNN, respectively; Q_i, K_i, and V_i are the query matrix, key matrix, and value matrix for each head, respectively; and $\sqrt{d_k}$ is the scaling factor.

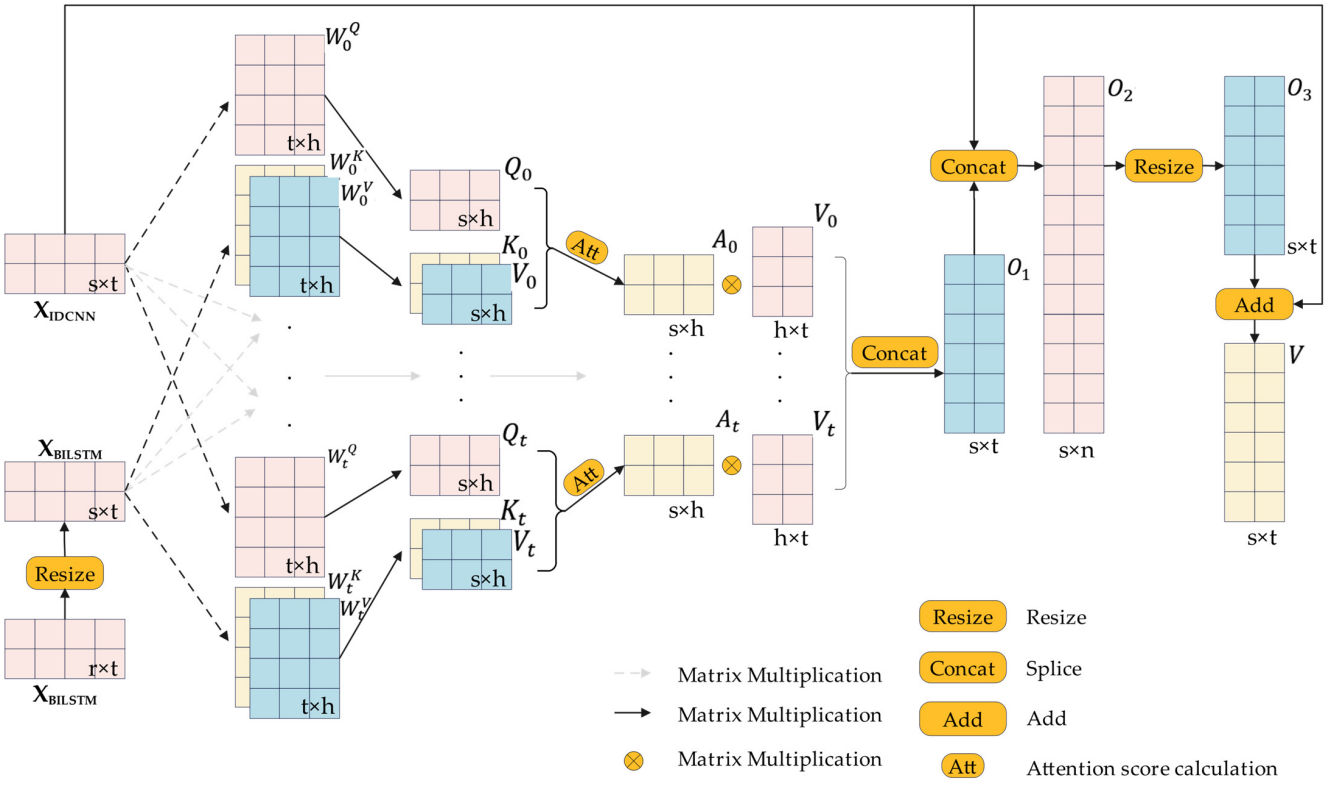


Figure 6. Structure diagram of the toponym BE module.

To enhance the model's ability to recognize toponym features in short social media texts, the BE module splices the IDCNN output feature vectors with other feature vectors by dimension to synthesize information from different feature sources. At the same time, the element-by-element summing of IDCNN output feature vectors with other feature vectors can help the model maintain a stable gradient propagation during feature fusion. Ultimately, the feature learning ability of the model is enhanced by diverse feature fusion methods, which improves the accuracy and robustness of the CSMNER model.

3.2.4. Label Decoding Layer

In NER tasks, the CRF layer plays a key role in label sequence decoding. Its design addresses the dependency challenge in sequence data by optimizing the label sequence configuration, which is crucial for improving recognition accuracy and model robustness. The unique advantage of the CRF layer is that it both examines the classification of individual words and fully considers the dependency between labels to ensure that the logical combination of adjacent labels is reasonable. For example, in recognizing toponyms (with B-LOC as the starting word label and I-LOC as the subsequent word label), the consistency of the internal structure of the entity can be effectively captured. Specifically, the CRF layer globally optimizes sequence labeling by learning the label transition probability, which is expressed as the conditional probability $P(y|x)$.

$$P(y|x) = \frac{\exp\left(\sum_{k=1}^{K}\sum_{i=1}^{T}\psi_k(y_i, x) + \sum_{l=1}^{L}\sum_{i=1}^{T-1}\phi_l(y_i, y_{i+1})\right)}{Z(x)} \tag{9}$$

where y represents the observed label sequence, x is the input text sequence, and ψ_k and ϕ_l represent the state feature function and transition feature function, respectively, which

jointly quantify the rationality of label selection based on the current input and context. $Z(x)$ is used as a normalization factor to normalize the probability distribution.

Through Equation (9), the CRF layer can overcome the local restrictions to achieve the global optimization of the entire sequence label configuration. This mechanism promotes the model to focus on the classification labels of individual words during prediction and comprehensively consider the context information of the entire sequence and the intrinsic structure of the label sequence to ensure that the output label sequence is in line with the expected logic syntactically and semantically, thereby improving performance on the NER tasks.

4. Experiment and Discussion

4.1. Experimental Data, Evaluation Metrics, and Experimental Settings

4.1.1. Experimental Datasets

To verify the effectiveness of the proposed method, the experiments were conducted on three Chinese datasets in Table 3, including two public datasets—WeiboNER and MSRA—and the CSNER dataset proposed in this paper. The WeiboNER dataset was collected from Sina Weibo, the largest social media platform in China. The dataset is specifically designed for Chinese NER tasks. The entity annotation follows the general BIO annotation system, and three entity types are annotated: person (PER), location (LOC), and organization (ORG). It contains a total of 1890 annotated data points, with high quality assured through multiple verifications to ensure the accuracy and consistency of the annotations. The corpus is short and diverse, covering a wealth of entity categories and language styles, making it suitable for training and evaluating Chinese NER tasks [37]. The MSRA dataset was released by Microsoft Research Asia. The dataset is tailored for texts in the news field and contains more than 50,000 annotated data points, with annotated entity types such as person names, toponyms, and organization names. Compared with other Chinese NER datasets, the MSRA dataset is large and, due to its focus on the news field, is especially suitable for dealing with formal texts on social media and NER in news reports. Many studies and algorithm models have used the MSRA dataset as a basis for evaluation and comparison.

Table 3. Dataset information statistics.

Dataset	Entity Type	Training Set Size	Validation Set Size	Test Set Size
WeiboNER	3	1.4 k	0.3 k	0.3 k
MSRA	3	46.4 k	4.4 k	4.4 k
CSNER	3	48.2 k	10.3 k	10.3 k

The three data sets are all Chinese data sets, and the entity types annotated in the data sets are person (PER), location (LOC), and organization (ORG).

In this paper, the three experimental datasets are each divided into three parts: a test set, a validation set, and a training set, with the ratio of the number of corpora among the three being 7:1.5:1.5. Figure 7 shows the proportions of annotated entities in each dataset, and the proportions of the three types of annotated entities in each dataset are relatively balanced. Table 4 shows the annotation examples of the corpus in the datasets. The CSNER and WeiboNER datasets have meaningless expressions such as punctuation marks and "哈哈 (haha)"; the corpus in the MSRA dataset is formalized and carefully annotated, and the news corpora contain many abbreviated toponym entities such as "港 (Hong Kong)" and "台 (Taiwan)", which are sufficiently difficult and representative to effectively test the named entity recognition model for Chinese social media and evaluate the robustness of the proposed model.

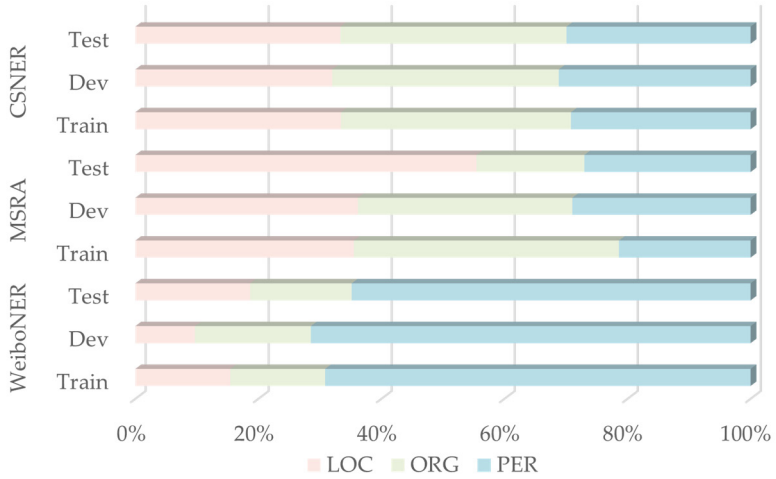


Figure 7. Statistical graphs of labeled entity information for each part of the experimental dataset, the three experimental datasets of WeiboNER, MSRA, and CSNER are all divided into three parts: Test, Dev, and Train.

Table 4. Examples of annotated information of the datasets.

Dataset	Annotation Example
WeiboNER	日/O 中/O 午/O ，/O 宋/B-PER 同/I-PER 志/I-PER 抵/O 达/O 汉/B-LOC 口/I-LOC 站/I-LOC 转/O 动/O 车/O 回/O 家/O 。/O 哈/O 哈/O
MSRA	把/O 欧/B-LOC 美/B-LOC 、/O 港/B-LOC 台/B-LOC 的/O 图/O 书/O 汇/O 集/O 起/O 来/O
CSNER	又/O 到/O 了/O 可/O 以/O 出/O 门/O 踏/O 青/O 的/O 日/O 子/O 了/O 我/O 的/O 家/O 乡/O 就/O 是/O 宁/B-LOC 波/I-LOC 象/I-LOC 山/I-LOC

4.1.2. Evaluation Metrics

In the experiments, the *Precision*, *Recall*, comprehensive evaluation metric (*F*1), and accuracy were used to evaluate the model. The F1 value provides a comprehensive evaluation of the model performance. The *Precision* represents the degree of preciseness, which is calculated by dividing the number of entities correctly recognized by the model by the total number of entities recognized, as shown in Equation (10).

$$Precision = TP/(TP + FP) \tag{10}$$

Recall is calculated by dividing the number of correctly recognized entities by the total number of true entities, as shown in Equation (11).

$$Recall = TP/(TP + FN) \tag{11}$$

*F*1 is a comprehensive evaluation metric, which is the harmonic mean of P and R, as shown in Equation (12).

$$F1 = 2 \times Precision \times Recall/(Precision + Recall) \tag{12}$$

The *Accuracy* represents the proportion of the total number of samples correctly classified by the model, as shown in Equation (13).

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \tag{13}$$

where TP (true positive) represents the number of entities correctly recognized, TN (true negative) represents the number of nonentities correctly recognized, FP (false positive) represents the number of nonentities incorrectly recognized, and FN (false negative) represents the number of true entities not recognized.

4.1.3. Experimental Settings

1. Hyperparameter settings: To illustrate how the experiments were conducted, Table 5 lists some important hyperparameters used in the experiments.

Table 5. Hyperparameter settings.

No.	Parameters	Value
1	Embedding Dimension	768
2	Max Length	128
3	Batch Size	64
4	Learning Rate	0.00003
5	Hidden Layer Size	512
6	Dropout	0.5

2. Model training strategy: The model used the BERT-based-Chinese pre-trained model to map text features to 768-dimensional vectors to obtain rich representations at the character level, sentence level, and context. The model selected the Adam optimizer to adaptively adjust the learning rate based on the historical gradient of each parameter and a parameter that set the strength of weight decay was used to control model complexity and prevent overfitting [38]. The model avoided overfitting by using a regularization technique and early stopping strategy [39], with the dropout and patience set to 0.5 and 5, respectively. When the model did not significantly improve within five consecutive epochs, the current best performance result is output, and the training stops, thus improving the training efficiency and avoiding overfitting.

4.2. Performance Comparison

The performance of the proposed method is compared with the Chinese toponym entity recognition models emerging in recent years in WeiboNER and MSRA datasets, respectively, and the results are shown in Table 6. Table 7 represents the performance comparison between the proposed CSMNER model and the latest Chinese NER models in recent years.

Table 6. Comparison of previous research work on publicly available datasets, comparing toponym entity recognition performance only.

Model	WeiboNER			MSRA		
	Precision	Recall	F1	Precision	Recall	F1
Qiu et al., 2022 [15]	59.00	56.00	57.00	86.00	85.00	86.00
Ma et al., 2022 [40]	-	-	-	99.00	90.00	94.00
Zhao et al., 2023 [41]				96.98	95.85	96.41
Ours (CSMNER)	82.35	71.62	76.61	97.49	97.17	97.33

Table 7. Comparison of previous research work on publicly available datasets to compare the combined performance of named entity models for PER, LOC, and ORG entity recognition.

Model	WeiboNER			MSRA		
	Precision	Recall	F1	Precision	Recall	F1
Zhang et al., 2022 [36]	-	-	67.33	-	-	94.83
Wu et al., 2022 [42]	-	-	61.94	94.92	94.19	94.55
Song et al., 2023 [43]	51.24	41.47	45.84	65.28	69.32	67.24
Deng et al., 2023 [44]	-	-	70.12	-	-	96.51
Qin et al., 2023 [45]	74.90	72.70	73.80	96.50	96.60	96.50
Ours (CSMNER)	76.43	75.81	76.12	96.71	95.65	96.17

- Qiu et al., proposed a weakly supervised learning model for toponym recognition, which mainly uses the bidirectional LSTM and CRF model and extends them to enhance the model's ability to recognize multiword toponyms.
- Ma et al. proposed a BERT-BiLSTM-CRF deep learning model by adding a pre-trained BERT representation.
- Zhao et al. proposed a multi-layer deep learning model ERNIE-Gram-IDCNN-BiLSTM-CRF to capture toponym features through dynamic vectors, and a new deep learning framework EIBC.
- Zhang et al., proposed an NER model, LSF-CNER, that fuses lexical and syntactic information.
- Wu et al. proposed an InterFormer module, which simultaneously models character and word sequences of different lengths through a nonplanar grid structure and constructs the NFlat model that decouples lexical fusion and context encoding.
- Song et al. proposed a Chinese NER model with fused graph embedding. The model utilizes the phonetic relations of Chinese characters to construct an undirected graph, represents each Chinese character through the fusion of graph embedding and semantic embedding, and implements prediction using a BiLSTM-CRF network model.
- Deng et al. proposed a Kcr-FLAT Chinese NER model that extracts and encodes three types of syntactic information and fuses them with lexical information using an attention mechanism to address word segmentation errors introduced by lexical information.
- Qin et al. proposed a multitask learning-based model for Chinese NER, MTL-BERT. This model decomposes the NER task into two subtasks—entity boundary annotation and type annotation—and dynamically adjusts the task weights according to the real-time learning effect of the task, thus improving model learning efficiency.

In recent years, there has been a scarcity of research utilizing the same dataset for toponym entity recognition. To ensure the rigor and scientific validity of comparative experiments, we not only compared our results with studies solely focused on toponym recognition performance testing but also contrasted them with recent Chinese-named entity recognition research. Table 6 shows that the CSMNER model proposed in this paper not only exhibits significant advantages in the task of toponym entity recognition but also has a good overall performance in the domain of named entity recognition. In comparison to the model of Qin et al., ours achieved notable performance improvements on both WeiboNER and MSRA datasets. Due to the relatively small size and high noise complexity of the WeiboNER dataset, named entity recognition models tend to perform poorly in predicting it. However, on the WeiboNER dataset, the model structure proposed in this paper identifies named entities with much better F1 scores than Qiu et al.'s study, indicating that the use of the GELU function, which is more suitable for the task of natural language processing, and the BE module designed in this paper, can improve the acronym toponym and long toponym by enhancing the filtering of the contextual and localized information of the toponym recognition correctness.

Table 7 shows that the CSMNER model can effectively recognize different named entity types in Chinese social media short texts, and the comprehensive performance has reached a good level. However, there are still some performance gaps and recognition errors when facing texts of different domains and language styles. Compared with the studies of Ma et al., Zhao et al., Deng et al., and Qin et al., the performance improvement of our model is not obvious, and it does not perform as well as their models in some metrics. The main reason is that the training corpus of our model only comes from short text data of a single social media platform, which has certain limitations. In summary, although their research model shows greater efficacy on the standardized MSRA news dataset, the model proposed in this paper can better capture the features of toponym entities in non-standardized text. Meanwhile, limited by the size and type of training data, the noise difference of the data leads to significant differences in the toponym features and patterns therein, which will limit the recognition accuracy of the CSMNER model in different scenarios, affecting the further improvement of the model robustness.

4.3. Ablation Experiment

To validate the effectiveness of the proposed method, the BERT-BiLSTM-CRF model is used as the experimental baseline, and the model is improved by various strategies. The test results on three datasets are shown in Table 8.

Table 8. Performance test results of the ablation experiment.

Model	Attention	IDCNN (ReLU)	IDCNN (GELU)	WeiboNER	MSRA	CSNER
Baseline				68.32	92.07	76.39
+ BE	✓			70.12	92.21	78.52
+ IDCNN (ReLU)		✓		70.93	96.01	80.70
+ IDCNN (GELU)			✓	73.64	96.19	81.67
+ IDCNN(ReLU)&BE	✓	✓		75.35	95.71	81.62
Ours (CSMNER)	✓		✓	76.12	96.17	82.14

Where Baseline is the BERT-BiLSTM-CRF model, BE is the boundary extension module proposed in this paper, IDCNN (ReLU) is the basic IDCNN model, IDCNN (GELU) is the IDCNN module that changes the activation function to GELU, and CSMNER is the model proposed in this paper. "✓" indicates that the model contains this module.

Table 8 shows that the improvement strategies in the experiments result in varying degrees of improvement in the performance of the original model. Among them, adding the IDCNN (GELU) structure improves the model performance most significantly, by approximately 7.8% on the WeiboNER dataset and by nearly 4.1% and 5.8% on the MSRA and CSNER datasets, respectively. Moreover, adding the BE module improves the accuracy by 1.8%, 0.14%, and 2.13% on the WeiboNER, MSRA and CSNER datasets, respectively. Since the MSRA dataset is mainly derived from standard news texts, its language style is more formal and standardized, and the introduction of the BE module does not have a significant effect on improving the performance of the baseline model. On the contrary, the WeiboNER and CSNER datasets are derived from Sina Weibo, whose text content is more colloquial and networked, containing many non-standard terms and complex noise. In this case, after fine-grained feature screening and trade-offs, the BE module significantly improves the toponym recognition ability of the CSMNER model, which improves the accuracy of Chinesenamed entity recognition in different semantic scenarios.

4.4. Experimental Analysis

To analyze the comprehensive performance of the models and the experimental effect in detail, the training performance and recognition effect of the models are analyzed in this section. Figure 8 shows the F1 score performance of different models after training iterations, and it can be found that the Baseline + IDCNN (ReLU) and BE and CSMNER models converge faster and achieve better performance with the same number of iteration rounds after adding the BE module. The combination of Baseline + IDCNN (ReLU)

and BE is less different from the CSMNER model, but the learning process is not very stable and there is a large fluctuation, which is obviously reduced after the adoption of IDCNN (GELU), which makes the model learning process more stable. learning process more stable.

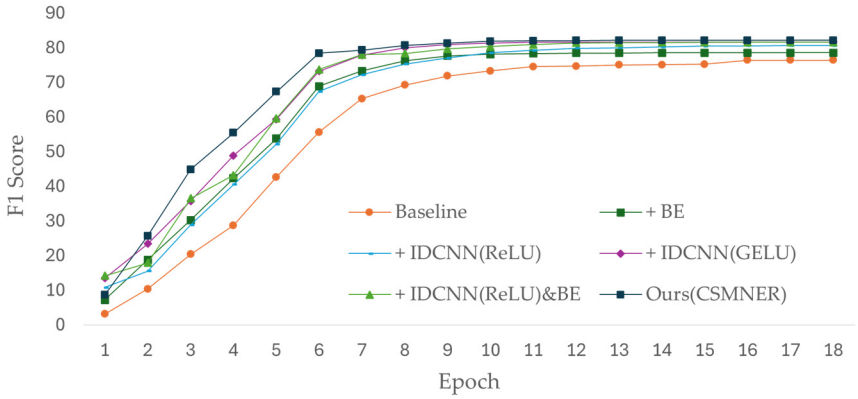


Figure 8. F1 score performance of different models on CSNER dataset.

Figure 9 further shows that the model CSMNER proposed in this paper can learn useful features as soon as possible, as the Loss value can decrease rapidly during the learning process and then level off. Compared with the Baseline model, the added BE module and IDCNN (GELU) network play different degrees of positive roles in the feature learning process of the model.

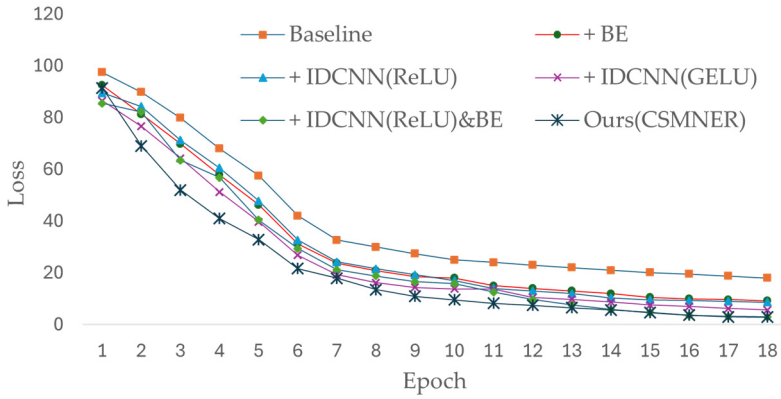


Figure 9. Loss performance of different models on CSNER dataset.

We have listed some typical example sentences for analysis, one is an example of a toponym abbreviation that requires cultural background and the other is an example of a toponym that requires geographical background.

Table 9 shows the recognition ability of the proposed model for abbreviated toponyms in the MSRA news corpus. It is found that the baseline model and the simple introduction of the IDCNN are not effective for toponym extraction. In this corpus, the toponyms are all region abbreviations. As adjectives in Chinese expressions, the expression of contextual information of toponyms is limited, thus increasing the difficulty in extracting toponym features. After adding the BE module, the features between the two abbreviated toponyms "欧美 (Europe and America)" and "港台 (Hong Kong and Taiwan)" are enlarged, while the influence of non-noun information in the context of the toponyms is suppressed, thereby achieving correct prediction. Compared with the IDCNN (ReLU), the difference in the prediction results is very small, but the error of misrecognizing the two regions "Hong Kong and Taiwan" as the same toponym is effectively avoided.

Table 9. Entity recognition results using different models for the sentence "把欧美、港台流行的食品汇集 (Bringing together popular foods from Europe, America, Hong Kong, and Taiwan)".

Original Sentence	把	欧	美	、	港	台	流	行	的	食	品	汇	集
Translate sentences	Bringing together popular foods from Europe, America, Hong Kong, and Taiwan.												
Correct prediction	O	B-LOC	B-LOC	O	B-LOC	B-LOC	O	O	O	O	O	O	O
Baseline prediction	O	O	O	O	O	O	O	O	O	O	O	O	O
+ IDCNN (ReLU)	O	O	O	O	O	O	O	O	O	O	O	O	O
+ IDCNN (GELU)	O	O	O	O	O	O	O	O	O	O	O	O	O
+ IDCNN(ReLU)&BE	O	B-LOC	B-LOC	O	B-LOC	I-LOC	O	O	O	O	O	O	O
Our (CSMNER)	O	B-LOC	B-LOC	O	B-LOC	B-LOC	O	O	O	O	O	O	O

To further test the model performance, Table 10 shows the prediction results of the proposed model for long toponyms in the CSNER dataset. The results show that the BE module performs better in recognizing the long toponym entity "意大利风景区 (Italian scenic area)". The main reason is that the BERT-BILSTM-CRF model and the model with the IDCNN network do not highlight the boundary information of the entities, which makes it impossible to locate the toponym boundaries accurately. The added BE module improves the textual information of the toponym entity "Italy Scenic Area" by considering the semantic association between the two nouns "意大利 (Italy)" and "风景区 (scenic area)" and focusing on the use of the hollow convolutional features extracted by IDCNN. The textual features of the toponym entity "Italy Scenic Spot" are improved, and a better prediction performance is achieved.

Table 10. Analysis of the entity recognition results using different models for the sentence "走走浪漫的五大道感受浪漫的意大利风景区 (Walk on the romantic Fifth Avenue and feel the romantic Italian scenic area)".

Original Sentence	走	走	浪	漫	的	五	大	道	感	受	浪	漫	的	意	大	利	风	景	区
Translate sentences	Walk on the romantic Fifth Avenue and feel the romantic Italian scenic area																		
Correct prediction	O	O	O	O	O	B-L	I-L	I-L	O	O	O	O	O	B-L	I-L	I-L	I-L	I-L	I-L
Baseline prediction	O	O	O	O	O	B-L	I-L	I-L	O	O	O	O	O	B-L	I-L	I-L	O	O	O
+ IDCNN (ReLU)	O	O	O	O	O	B-L	I-L	I-L	O	O	O	O	O	B-L	I-L	I-L	O	O	O
+ IDCNN (GELU)	O	O	O	O	O	B-L	I-L	I-L	O	O	O	O	O	B-L	I-L	I-L	O	O	O
+ IDCNN(ReLU)&BE	O	O	O	O	O	B-L	I-L	I-L	O	O	O	O	O	B-L	I-L	I-L	I-L	I-L	I-L
Our (CSMNER)	O	O	O	O	O	B-L	I-L	I-L	O	O	O	O	O	B-L	I-L	I-L	I-L	I-L	I-L

B-L in the table is B-LOC, representing the beginning of the location entity, and I-L is I-LOC, representing the inside of the location entity.

In Table 11, we provide many example sentences for model error identification and indicate the types of errors caused by the model. The given examples can be used to analyze the shortcomings and limitations of the proposed CSMNER model and suggest further research. Example 1 shows the limitations of the collected dataset size and geographic background knowledge, resulting in complete toponyms being partially recognized. Example 2 shows that although the toponym can be correctly recognized during the process of capturing semantic features, there are still some errors in the process of processing connected toponyms. The toponym in Example 3 has strong toponym features but is affected by the irrelevant noise interspersed in the short text, resulting in the model not being able to recognize it correctly. In Example 4, under the influence of semantic information in the context, the model incorrectly associates the two sentences and recognizes them as "No. 1 bus" due to the confusion of the combination of numbers and nouns.

Table 11. Some example sentences for CSMNER model misidentification.

No.	Example Sentence	Sentence Translation	Predicted Results	Annotated Result	Error Type
1	这是道观你信吗太原东社街区	This is a Taoist temple. Can you believe it? Taiyuan Dongshe neighborhood.	Taiyuan	Taiyuan Dongshe neighborhood	Lack of geographical background knowledge, toponym recognition is incomplete
2	我的家乡就是平顶山郏县	My hometown is Pingdingshan Jia County	Pingdingshan, Jia County	Pingdingshan Jia County	Lack of geographical background, identified two toponyms
3	中秋豪礼相送快来抢购大连金石滩国家旅游度假区	Mid-Autumn Gifts for Dalian Jinshitan National Tourist Resort	None	Dalian Jinshitan National Tourist Resort	The context-valid information is confused
4	小家伙跟了我一路车子也不要了	The little guy followed me all the way to the car and didn't want it.	Number 1 bus	None	No toponyms, semantic error

In the Chinese example in Example 4, because of the removal of punctuation marks, the three words "一路车" are linked together, resulting in a semantic misunderstanding, and "一路车" in Chinese can be customarily interpreted as a No. 1 bus.

Through the above performance analysis, it can be found that although the CSMNER model proposed in this paper can achieve better recognition results for most of the toponyms, it still inevitably generates certain errors under the influence of the lack of exogenous background knowledge and contextual information with a lot of irrelevant noise. This also points out that further research can be conducted to integrate the named entity recognition model into multi-domain background knowledge to further optimize the model performance.

5. Conclusions

This paper proposes a toponym entity recognition model, CSMNER, for Chinese social media. This model effectively extracts local boundary features and contextual semantic features of toponyms through the improved IDCNN and BiLSTM dual-channel feature extraction module. Meanwhile, the BE module is introduced to evaluate the influence of contextual information on local feature information and expand toponym entity boundary detection, which ultimately improves the efficiency, accuracy, and robustness of toponym entity recognition in Chinese social media texts. Toponym entity recognition models usually face problems such as difficulty in identifying toponym boundaries, ambiguous or polysemous toponym entities, context dependency, and ambiguous or incorrect toponyms when facing the short and highly fragmented expressions of social media texts, as well as many noises such as spelling, abbreviations, and slang. The experimental results on three datasets, WeiboNER, MSRA, and CSNER, show that the CSMNER model performs well in coping with various problems, and can effectively recognize complex toponyms such as toponym abbreviations and long toponyms in Chinese social texts. In this study, limited by the size of the training data, noise, and domain differences, the recognition results of the model will have some errors. Facing texts from different domains, different language habits, and different standards, the model needs to deal with more linguistic style variations and noise, which requires that the robustness of the model needs to be further improved. In future research, we will focus on how to incorporate external toponymic information data and social media linguistic features more effectively to enhance the model's ability to recognize toponym entities in complex contexts. We will also apply the model to geographic information-related tasks such as location services, travel recommendation systems, geographic intelligence analysis, and disaster monitoring and response.

Author Contributions: Conceptualization, Yuyang Qi and Renjian Zhai; methodology design and implementation, Renjian Zhai; validation and data verification, Fang Wu and Jichong Yin; formal data analysis, Xianyong Gong; investigation and primary data collection, Li Zhu; resource management, Haikun Yu; original draft preparation, Yuyang Qi; manuscript review and editing, Renjian Zhai. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by National Natural Science Foundation of China (No. 42371461) and the project of cyberspace information intelligence generalization technology.

Data Availability Statement: All original data and codes can be found in the figshare (https://figshare.com/s/77cb3523143208578619 (accessed on 1 August 2024)).

Acknowledgments: The authors thank the seven anonymous reviewers for their positive, constructive, and valuable comments and suggestions.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Purves, R.; Hollenstein, L. Exploring Place through User-Generated Content: Using Flickr to Describe City Cores. *J. Spat. Inf. Sci.* **2010**, *1*, 21–48. [CrossRef]
2. Xu, L.; Du, Z.; Mao, R.; Zhang, F.; Liu, R. GSAM: A Deep Neural Network Model for Extracting Computational Representations of Chinese Addresses Fused with Geospatial Feature. *Comput. Environ. Urban Syst.* **2020**, *81*, 101473. [CrossRef]
3. Lai, J.; Lansley, G.; Haworth, J.; Cheng, T. A Name-led Approach to Profile Urban Places Based on Geotagged Twitter Data. *Trans. GIS* **2020**, *24*, 858–879. [CrossRef]
4. Gelernter, J.; Zhang, W. Geocoding Location Expressions in Twitter Messages: A Preference Learning Method. *J. Spat. Inf. Sci.* **2014**, *9*, 37–70. [CrossRef]
5. McDonough, K.; Moncla, L.; van de Camp, M. Named Entity Recognition Goes to Old Regime France: Geographic Text Analysis for Early Modern French Corpora. *Int. J. Geogr. Inf. Sci.* **2019**, *33*, 2498–2522. [CrossRef]
6. Hu, Y.; Mao, H.; McKenzie, G. A Natural Language Processing and Geospatial Clustering Framework for Harvesting Local Place Names from Geotagged Housing Advertisements. *Int. J. Geogr. Inf. Sci.* **2019**, *33*, 714–738. [CrossRef]
7. Wallgrün, J.O.; Karimzadeh, M.; MacEachren, A.M.; Pezanowski, S. GeoCorpora: Building A Corpus to Test and Train Microblog Geoparsers. *Int. J. Geogr. Inf. Sci.* **2018**, *32*, 1–29. [CrossRef]
8. Wang, J.; Hu, Y.; Joseph, K. NeuroTPR: A Neuro-Net Toponym Recognition Model For Extracting Locations From Social Media Messages. *Trans. GIS* **2020**, *24*, 719–735. [CrossRef]
9. Paul, C. Robert Pasley Images and Perceptions of Neighbourhood Extents. In Proceedings of the 6th Workshop on Geographic Information Retrieval, Zurich, Switzerland, 18 February 2010; ACM: Zurich, Switzerland, 2010; pp. 1–2.
10. Jones, C.B.; Purves, R.S.; Clough, P.D.; Joho, H. Modelling Vague Places with Knowledge From the Web. *Int. J. Geogr. Inf. Sci.* **2008**, *22*, 1045–1065. [CrossRef]
11. Montello, D.R.; Goodchild, M.F.; Gottsegen, J.; Fohl, P. Where's Downtown?: Behavioral Methods for Determining Referents of Vague Spatial Queries. *Spat. Cogn. Comput.* **2003**, *3*, 185–204. [CrossRef]
12. Leidner, J.L.; Lieberman, M.D. Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language. *SIGSPATIAL Spec.* **2011**, *3*, 5–11. [CrossRef]
13. Giridhar, P.; Abdelzaher, T.; George, J.; Kaplan, L. On Quality of Event Localization from Social Network Feeds. In Proceedings of the 2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops), St. Louis, MO, USA, 23–27 March 2015; IEEE: St. Louis, MO, USA, 2015; pp. 75–80.
14. Dutt, R.; Hiware, K.; Ghosh, A.; Bhaskaran, R. SAVITR: A System for Real-Time Location Extraction from Microblogs during Emergencies. In Proceedings of the Web Conference 2018, Lyon, France, 23–27 April 2018.
15. Qiu, Q.; Xie, Z.; Wang, S.; Zhu, Y.; Lv, H.; Sun, K. ChineseTR: A weakly Supervised Toponym Recognition Architecture Based on Automatic Training Data Generator and Deep Neural Network. *Trans. GIS* **2022**, *26*, 1256–1279. [CrossRef]
16. Milusheva, S.; Marty, R.; Bedoya, G.; Williams, S.; Resor, E.; Legovini, A. Applying Machine Learning and Geolocation Techniques to Social Media Data (Twitter) to Develop a Resource for Urban Planning. *PLoS ONE* **2021**, *16*, e0244317. [CrossRef] [PubMed]
17. Middleton, S.E.; Kordopatis-Zilos, G.; Papadopoulos, S.; Kompatsiaris, Y. Location Extraction from Social Media: Geoparsing, Location Disambiguation, and Geotagging. *ACM Trans. Inf. Syst.* **2018**, *36*, 1–27. [CrossRef]
18. Habib, M.B.; van Keulen, M. A Hybrid Approach for Robust Multilingual Toponym Extraction and Disambiguation. In *Language Processing and Intelligent Information Systems*; Mieczysław, A., Kłopotek, J.K., Małgorzata, M., Agnieszka, M., Sławomir, T., Wierzchoń, Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2013; Volume 7912, pp. 1–15. ISBN 978-3-642-38633-6.
19. Sharma, P.; Samal, A.; Soh, L.-K.; Joshi, D. A Spatially-Aware Algorithm for Location Extraction from Structured Documents. *GeoInformatica* **2023**, *27*, 645–679. [CrossRef]

20. Sobhana, N.; Mitra, P.; Ghosh, S. Conditional Random Field Based Named Entity Recognition in Geological text. *Int. J. Comput. Appl.* **2010**, *1*, 143–147. [CrossRef]
21. Curran, J.R.; Clark, S. Language Independent NER Using a Maximum Entropy Tagger. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*; Edmonton, AB, Canada, 31 May 2003, Association for Computational Linguistics: Edmonton, AB, Canada, 2003; Volume 4, pp. 164–167.
22. Lingad, J.; Karimi, S.; Yin, J. Location Extraction from Disaster-Related Microblogs. In Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, 13 May 2013; ACM: Rio de Janeiro, Brazil, 2013; pp. 1017–1020.
23. Santos, R.; Murrieta-Flores, P.; Calado, P.; Martins, B. Toponym Matching through Deep Neural Networks. *Int. J. Geogr. Inf. Sci.* **2018**, *32*, 324–348. [CrossRef]
24. Hu, X.; Al-Olimat, H.S.; Kersten, J.; Wiegmann, M.; Klan, F.; Sun, Y.; Fan, H. GazPNE: Annotation-Free Deep Learning for Place Name Extraction from Microblogs Leveraging Gazetteer and Synthetic Data by Rules. *Int. J. Geogr. Inf. Sci.* **2022**, *36*, 310–337. [CrossRef]
25. Xu, C.; Li, J.; Luo, X.; Pei, J.; Li, C.; Ji, D. DLocRL: A Deep Learning Pipeline for Fine-Grained Location Recognition and Linking in Tweets. In Proceedings of the The World Wide Web Conference, San Francisco, CA, USA, 13 May 2019; ACM: New York, NY, USA, 2019; pp. 3391–3397.
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2023**, arXiv:1706.03762.
27. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv* **2015**, arXiv:1508.01991.
28. Tao, L.; Xie, Z.; Xu, D.; Ma, K.; Qiu, Q.; Pan, S.; Huang, B. Geographic Named Entity Recognition by Employing Natural Language Processing and an Improved BERT Model. *ISPRS Int. J. Geo Inf.* **2022**, *11*, 598. [CrossRef]
29. Ma, X.; Hovy, E. End-to-End Sequence Labeling via Bi-Directional LSTM-CNNs-CRF. *arXiv* **2016**, arXiv:1603.01354.
30. Zhang, Y.; Yang, J. Chinese NER Using Lattice LSTM. *arXiv* **2018**, arXiv:1805.02023.
31. Xu, C.; Wang, F.; Han, J.; Li, C. Exploiting Multiple Embeddings for Chinese Named Entity Recognition. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, China, 3 November 2019; ACM: New York, NY, USA, 2019; pp. 2269–2272.
32. Devlin, J.; Chang, M.; Lee, K. Kristina Toutanova BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
33. Berragan, C.; Singleton, A.; Calafiore, A.; Morley, J. Transformer Based Named Entity Recognition for Place Name Extraction from Unstructured Text. *Int. J. Geogr. Inf. Sci.* **2023**, *37*, 747–766. [CrossRef]
34. Lu, Y.; Liu, Q.; Dai, D.; Xiao, X.; Lin, H.; Han, X.; Sun, L.; Wu, H. Unified Structure Generation for Universal Information Extraction. *arXiv* **2022**, arXiv:2203.12277.
35. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
36. Zhang, M.; Li, B.; Liu, Q.; Wu, J. Chinese Named Entity Recognition Fusing Lexical and Syntactic Information. In Proceedings of the 2022 the 6th International Conference on Innovation in Artificial Intelligence (ICIAI), Guangzhou, China, 4 March 2022; ACM: Guangzhou, China, 2022; pp. 69–77.
37. 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; Association for Computational Linguistics: Lisbon, Portugal, 2015; pp. 548–554.
38. Kingma, D.P. Adam: A Method for Stochastic Optimization. *arXiv* **2017**, arXiv:1412.6980.
39. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. *arXiv* **2019**, arXiv:1711.05101.
40. Ma, K.; Tan, Y.; Xie, Z.; Qiu, Q.; Chen, S. Chinese Toponym Recognition with Variant Neural Structures from Social Media Messages Based on BERT Methods. *J. Geogr. Syst.* **2022**, *24*, 143–169. [CrossRef]
41. Zhao, Y.; Zhang, D.; Jiang, L.; Liu, Q.; Liu, Y.; Liao, Z. EIBC: A Deep Learning Framework for Chinese Toponym Recognition with Multiple Layers. *J. Geogr. Syst.* **2024**, *26*, 407–425. [CrossRef]
42. Wu, S.; Song, X.; Feng, Z.; Wu, X.J. NFLAT: Non-Flat-Lattice Transformer for Chinese Named Entity Recognition. *arXiv*, 2022; arXiv:2205.05832.
43. Song, X.; Yu, H.; Li, S.; Wang, H. Robust Chinese Named Entity Recognition Based on Fusion Graph Embedding. *Electronics* **2023**, *12*, 569. [CrossRef]
44. Deng, Z.; Tao, Y.; Lan, R.; Yang, R.; Wang, X. Kcr-FLAT: A Chinese-Named Entity Recognition Model with Enhanced Semantic Information. *Sensors* **2023**, *23*, 1771. [CrossRef] [PubMed]
45. Fang, Q.; Li, Y.; Feng, H.; Ruan, Y. Chinese Named Entity Recognition Model Based on Multi-Task Learning. *Appl. Sci.* **2023**, *13*, 4770. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.