



Article

ChineseCTRE: A Model for Geographical Named Entity Recognition and Correction Based on Deep Neural Networks and the BERT Model

Wei Zhang ^{1,2,3}, Jingtao Meng ^{2,3}, Jianhua Wan ¹, Chengkun Zhang ⁴ , Jiajun Zhang ⁵, Yuanyuan Wang ^{4,6},
Liuchang Xu ^{5,7,8}  and Fei Li ^{2,3,*}

- ¹ College of Oceanography and Space Informatics, China University of Petroleum, Qingdao 266580, China; zhangweitgy@shandong.cn (W.Z.)
- ² Land Surveying and Mapping Institute of Shandong Province, Jinan 250102, China
- ³ Shandong Province Engineering Technology Research Center for Spatial Information and Big Data Applications, Jinan 250102, China
- ⁴ School of Earth Sciences, Zhejiang University, Hangzhou 310058, China
- ⁵ College of Mathematics and Computer Science, Zhejiang Agriculture and Forestry University, Hangzhou 311300, China; xuliuchang@zafu.edu.cn (L.X.)
- ⁶ Ocean Academy, Zhejiang University, Zhoushan 316021, China
- ⁷ College of Computer Science and Technology, Zhejiang University, Hangzhou 310063, China
- ⁸ Financial Big Data Research Institute, Sunyard Technology Co., Ltd., Hangzhou 310053, China
- * Correspondence: lifeigtgy@shandong.cn

Abstract: Social media is widely used to share real-time information and report accidents during natural disasters. Named entity recognition (NER) is a fundamental task of geospatial information applications that aims to extract location names from natural language text. As a result, the identification of location names from social media information has gradually become a demand. Named entity correction (NEC), as a complementary task of NER, plays a crucial role in ensuring the accuracy of location names and further improving the accuracy of NER. Despite numerous methods having been adopted for NER, including text statistics-based and deep learning-based methods, there has been limited research on NEC. To address this gap, we propose the CTRE model, which is a geospatial named entity recognition and correction model based on the BERT model framework. Our approach enhances the BERT model by introducing incremental pre-training in the pre-training phase, significantly improving the model's recognition accuracy. Subsequently, we adopt the pre-training fine-tuning mode of the BERT base model and extend the fine-tuning process, incorporating a neural network framework to construct the geospatial named entity recognition model and geospatial named entity correction model, respectively. The BERT model utilizes data augmentation of VGI (volunteered geographic information) data and social media data for incremental pre-training, leading to an enhancement in the model accuracy from 85% to 87%. The F1 score of the geospatial named entity recognition model reaches an impressive 0.9045, while the precision of the geospatial named entity correction model achieves 0.9765. The experimental results robustly demonstrate the effectiveness of our proposed CTRE model, providing a reference for subsequent research on location names.

Keywords: social media information; named entity recognition; named entity correction; VGI; BERT



Citation: Zhang, W.; Meng, J.; Wan, J.; Zhang, C.; Zhang, J.; Wang, Y.; Xu, L.; Li, F. ChineseCTRE: A Model for Geographical Named Entity Recognition and Correction Based on Deep Neural Networks and the BERT Model. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 394. <https://doi.org/10.3390/ijgi12100394>

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

Received: 9 August 2023

Revised: 16 September 2023

Accepted: 24 September 2023

Published: 27 September 2023



Copyright: © 2023 by the authors. 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

With the continuous promotion of the construction of smart cities, the methods used for obtaining geographic spatial data have become more diverse and widespread. Emerging technologies such as smartphones, the Internet of Things, and social media have led to explosive growth in geographic spatial data [1–3]. As important basic data for urban construction, geographic spatial data contain information on urban spatial structures, transportation networks, public facilities, and many other aspects, providing important support

for urban planning, transportation management, and other fields. Among geographic spatial data, geographic named entity data are very important, including place names, addresses, spatial locations, and other components, which play important roles in building the urban geographic spatial environment. Therefore, the geographic named entity recognition of geographic named entity data is crucial for these data to be used more effectively. Currently, the most advanced research in geographic named entities mainly applies relevant theories in natural language processing to the semantic understanding of the text for geographic named entities. This approach is more suitable for geographic named entity data with a single source and a simple structure, and it can achieve good results. However, due to the complexity of geographic named entity data and their diverse data sources, this method performs poorly in processing the data. Previous research roughly divided geographic named entity recognition into two categories: spatial statistical-based geographic named entity recognition [4] and deep neural network-based geographic named entity recognition [5]. Many previous studies [6] have achieved high recognition accuracy in geographic named entity recognition, ignoring the correctness of identified geographic entities; thus, further standardization and precision improvement are needed. Therefore, this study uses the BERT model and introduces the incremental pre-training method to improve the accuracy of the model in the pre-training phase. And we also enhance the corpus by enriching the geographic named entity database after completing the incremental pre-training. Additionally, we compare and verify the number of semantic feature extraction modules of the model to further improve its accuracy [7]. In addition, many of the geographic named entity recognition data used in previous research come from VGI (volunteered geographic information) data [8–12]. Despite the extraction of place names, errors may still exist in the resulting geographic named entity data. Therefore, place name correction work is also very important, as it plays an important role in expanding the standard place name library for the subsequent efficient use of geographic named entity data. Previous research roughly divided Chinese text correction into three categories: rule-based Chinese spelling correction methods, machine learning-based Chinese spelling methods, and deep learning-based Chinese text correction [13–15]. The existing research mainly focuses on correcting general natural language texts, while geographic named entities are a special type of text that contains potential spatial information. Therefore, this study conducts geographic named entity text correction based on BERT and provides an understanding of the semantic features of geographic named entities.

The innovative points of this article are as follows:

- (1) We introduce a transfer learning approach and use multiple sources of VGI data to achieve a highly accurate expansion of the geographic named entity database.
- (2) Incremental pre-training is introduced in the pre-training stage of the geographic named entity recognition model and the geographic named entity correction model, further improving the accuracy of geographic named entity text recognition and correction.
- (3) After performing geographic named entity recognition, we conduct further experiments and attempt to achieve geographic named entity text correction based on the existing geographic named entity recognition results, further improving the accuracy of geographic named entity recognition.

In this paper, we propose two models for Chinese geographical named entity recognition and geographical named entity correction, collectively known as CTRE, where “C” stands for Chinese context, “T” represents toponym, “R” denotes recognition, and “E” stands for error (the full term is error correction, which corresponds to text error correction in Chinese, both of which are based on the BERT framework and are extended and constructed). Previous studies have demonstrated the effectiveness of the CRF [16,17] and BiLSTM [18,19] frameworks in text sequence labeling and named entities, so our proposed model framework extends the structure of CRF and BiLSTM on the basis of BERT. Based on the two proposed models, this study aims to address two major challenges. Firstly, the proposed geographical named entity recognition model is utilized to extract and iden-

tify geographical named entities from social media data. However, due to the inherent fallibility of language, such as spelling errors and missing spellings in English, as well as homophonic and typo words in Chinese, and the spatial heterogeneity, which refers to different names for the same place and different places having the same name, the identified entities may contain errors. Therefore, the second objective is to employ the proposed geographical named entity error correction model to rectify the errors that are present in the extracted geographical named entities. For example, when identifying a geographic named entity in a Twitter text, the content is “Taylor Swift performs ‘Our Song’ from Taylor Swift (Debut) as the first surprise song for Day 2 of ‘The Eras Tour’ in Los Angeles, California!” We can identify the geo-named entity “Los Angeles, California” through the geographical named entity recognition model. Unlike English texts, Chinese vocabulary is based on a Chinese character, and each Chinese character has a specific meaning. In English, words are typically composed of letters or numbers, and each letter has no meaning in itself and needs to be combined to form a vocabulary. For example, in Chinese, Beijing is “Beijing”, where the three characters of “Bei” are spelled as a whole to form a word and are recognized in the process of geographical named entity recognition. In English, the letter a in California can be recognized separately. In addition, due to the diversity of online media data, online text is mostly user-generated, so there are errors such as misspellings, missing spelling, etc. For example, “Los Angeles” is misspelled as “Los Angelas”, so we need to correct it. Both the geographical named entity recognition and text correction models contribute to the advancement of natural language processing, providing more efficient, accurate, and intelligent solutions for real-world applications. This study’s contribution lies in the further correction research based on the original geographic named entity recognition, offering more efficient, accurate, and intelligent solutions for real-world applications. In addition, by correcting the identified geographic named entity data, it provides assistance in expanding the existing standard address dataset.

The remaining parts of our paper are as follows: The second part introduces the concepts related to geographical named entity recognition and the method for correcting Chinese addresses. The BERT model, the models of geographic named entity recognition and text correction, as well as the related content are introduced in the third part. The datasets, pre-processing methods, experiments for named entity recognition and address correction, and the resulting experimental outcomes are presented in the fourth part. The fifth part presents the study’s conclusions and prospects for future research.

2. Related Work

2.1. Geographic Named Entity Recognition Based on Social Media Platforms

Place names are proper names assigned by people to geographical entities in physical space. In addition to denoting specific geographic locations, place names may also include natural or human features. Place names are widely used in people’s daily lives and are the basic resources of geographic information [20]. Place name data can be collected through empirical data such as interviews and social surveys [21,22], but these approaches are difficult to popularize and apply on a large scale due to problems such as a high cost, low efficiency, and weak generalization. As geographic spatial information services become more widespread, the data are growing exponentially. Social media, location-based travel blogs, and housing advertisements are becoming more prevalent in people’s daily lives, and geographic named entity information is widely present in these different types of text [23–25]. However, this geographic named entity information contained in text is often not effectively utilized. With the development and progress of natural language processing technology, geographic named entity recognition driven by big data has become a hot research topic. Currently, research on methods for geographic named entity recognition in ubiquitous network text can be broadly categorized into two types, traditional spatial statistical methods and deep neural network-based methods.

2.1.1. Methods Based on Spatial Statistics

The spatial statistical method involves researchers creating rules for recognizing geographic named entities in specific research areas or extracting them from text by analyzing their spatial distribution patterns. For instance, de Bruijn et al. proposed a method for extracting place names by matching them with existing databases and OpenStreetMap [26]. However, in ubiquitous network text data, particularly in crowdsourced data, the place name information contributed by users usually has a certain degree of arbitrariness. In addition, there are some local conventions for place names, which often contain irregular language and some abbreviations that cannot be identified using methods based on specific place name databases. Furthermore, McKenzie et al. combined multiple spatial statistical metrics and random forest ensemble learning methods to extract neighborhood names from rental property listings [27]. Lai et al. used a spatial point pattern analysis method to extract place names from geotagged tweets [28]. Nevertheless, these spatial statistical methods may encounter issues such as high dimensionality, computational complexity, and overfitting despite their effectiveness in certain research domains.

2.1.2. Methods Based on Deep Neural Networks

Geographic named entity recognition (NER) based on deep neural networks is an approach that utilizes deep learning models to automatically identify geographic entities from textual data. This method leverages deep neural networks to extract features and classify input text, enabling the automatic identification of geographic named entities such as countries, cities, rivers, etc. In this method, the geographic named entity recognition task is often treated as a sequence labeling problem, where each word in the text sequence is tagged as a geographic entity or non-geographic entity. To achieve this, deep learning models such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and attention mechanisms are commonly used to model the text sequence, and classifiers such as multilayer perceptrons (MLPs) are used to predict the tag of each word. These deep learning models learn the language features and contextual information of geographic entities in the text by training on large amounts of annotated geographic named entity text data, which improves the recognition effectiveness of geographic named entities. Hu et al. proposed a deep learning architecture called C-LSTM that combines geographic dictionaries and rules and is applied to the process of geotagging Weibo data [29]. Additionally, NER tools have been utilized for extracting historical corpora, but this approach may lack generalization. Using artificial neural networks is an effective way to address generality and scalability. However, this method cannot solve the problem of toponym disambiguation because it does not utilize the contextual features of the text. To solve the problem of toponym disambiguation, Wang et al. proposed a neural network geotagging model based on the BiLSTM-CRF architecture to extract locations from social media messages and trained the model to manually annotate tweet data and a dataset from Wikipedia [30]. With the development of natural language processing technology, pre-training and fine-tuning language models represented by BERT have made it possible to efficiently and accurately extract geographic named entities from ubiquitous network texts [31]. Liu et al. implemented the Geo-NER for geological reports based on the BERT model [32]. However, directly applying BERT to geographic named entity recognition may ignore the inter-label constraint relationship between label sequences, which can significantly affect the model's performance in the task. To address this, Ma et al. proposed a BERT-BiLSTM-CRF deep neural network architecture for Chinese text geotagging tasks [33]. Qiu et al. proposed a ChineseTR architecture based on weakly supervised BERT + BiLSTM + CRF for Chinese geotagging and trained the Chinese geotagging model on a training dataset generated from the People's Daily corpus [34]. Tao et al. proposed an improved BERT model method for geographical named entity recognition and verified the effectiveness of the method [35]. The proposed method employed ALBERT + BiLSTM + CRF. The biggest difference from the model framework in this study is the BERT of the base. Compared with BERT, ALBERT has made certain improvements. It adopts the methods of parameter

sharing and embedding layer sharing, which reduce the number of parameters in BERT and can improve the training efficiency and reasoning speed. However, this method has limited resource scenarios. For example, the experimental dataset used by Tao et al. is TPCNER, and the dataset size is at the 600,000 level. The resources in this study are abundant, and the data magnitude reaches the 3,000,000 level. Models trained with different scales of data have different semantic feature extraction abilities. Our data size is larger, so the model has better semantic feature extraction ability and the evaluation scale is relatively large. Therefore, we use BERT to build the model on the base to obtain the semantic entities in the training model. The research conducted by Tao et al. was biased towards the improvement of the accuracy of identifying entities in geographic naming recognition. After identification, the recognition effects of different models and methods on the same dataset were compared, and the quality of the dataset was verified. However, our study further carried out the related tasks of text error correction as an in-depth study and supplement to the identification research, which Tao et al. did not have in the previous study. As a result, the scalability of our methodology is better. Deep learning-based geographic named entity recognition models often divide pre-training and fine-tuning into two stages, allowing them to better learn the semantic features of the text by pre-training on massive unsupervised data.

2.2. Chinese Address Correction

Text correction plays an important role in the field of natural language processing, and a good correction model is crucial for improving downstream task performance [36]. However, Chinese text correction is a challenging task due to its complexity. Research in related fields has proposed various methods for Chinese text correction, which can be categorized into three main groups: rule-based text spelling correction methods, machine learning-based text spelling correction methods, and deep learning-based text spelling correction methods.

2.2.1. Ruled-Based Chinese Text Spelling Correction Method

Rule-based Chinese text spelling correction methods rely on knowledge resources such as dictionaries. They identify a character as a spelling error if it does not comply with the predefined rules, such as not being present in the dictionary, and provide candidate characters as correction options. Early Chinese text correction methods first detected the misspelled position, generated candidate characters for these positions, and then selected a suitable one to replace the misspelling [37–40]. Another early study on Chinese text spelling correction proposed by Chang [41] used a Chinese character dictionary that accounted for the similarities in shape, pronunciation, meaning, and input method code to handle the spelling correction task. Each Chinese character in the sentence was replaced with a similar one from the dictionary, and the probability of all modified sentences was calculated based on a language model. Zhang et al. proposed another rule-based Chinese text correction method that differentiated between Chinese and English matching methods. This method could handle not only Chinese character replacement errors, but also insertion and deletion errors, greatly improving the correcting performance compared to Chang's method [42]. In addition, Huang et al. used a segmentation tool (CKIP) to generate correction candidates for detecting Chinese spelling errors [43]. Hung et al. corrected Chinese text spelling errors based on manually edited error templates [44]. Similarly, Jiang et al. designed a new grammar rule system for correcting Chinese grammar and spelling errors [45]. In rule-based Chinese text spelling correction methods, if a character does not comply with predefined rules, the method identifies it as a spelling error and provides candidate characters as correction options. However, in practical applications, rule-based Chinese text spelling correction methods heavily rely on linguistic knowledge and rules, and building language knowledge and rule libraries requires significant human and time costs and cannot cover the complex linguistic phenomena of Chinese. In addition, due to the complexity of Chinese, there are many unknown and ambiguous words that rule-

based methods cannot effectively handle. More importantly, rule-based methods require full-text scanning and matching for each input text, resulting in a low processing efficiency, and thus are not suitable for large-scale text processing. Therefore, the limitations of rule-based Chinese text spelling correction methods restrict their widespread use in practical scenarios. With the continuous development of artificial intelligence and natural language processing technologies, machine learning and deep learning-based Chinese text spelling correction methods have gradually become mainstream.

2.2.2. Machine Learning-Based Spelling Method

The machine learning-based Chinese spelling correction method [46–48] is a technique that utilizes machine learning algorithms and language models to automatically detect and correct spelling errors in Chinese text. This method requires a large amount of Chinese language corpus to train the language model and teach it the patterns of spelling errors. When a spelling error is detected, the method uses the trained model to analyze and correct it, providing more accurate spelling correction suggestions. This method can be widely used in areas such as Chinese input methods, word processing software, search engines, and social media to improve the accuracy and efficiency of text processing. Compared with traditional rule-based spell checking methods, machine learning-based spell checking methods can better handle complex language patterns and spelling errors and can provide more accurate corrections based on the user's input history and contextual information. In the research on machine learning-based Chinese text spelling correction methods, unsupervised n-gram language models are often used for error detection [49,50]. This approach introduces a confusion set of similar characters after error detection to limit the candidate options. Xie et al. replaced characters with confusion sets and evaluated the modified sentence using a joint bi-gram and tri-gram language model [49]. In the studies by Jia et al. and Xin et al., graph models were used to represent sentences, and the single-source shortest path (SSSP) algorithm was performed on the graph to correct spelling errors [51,52]. In addition, transforming text spelling correction tasks into sequence labeling problems and using machine learning methods such as conditional random fields or hidden Markov models is also a solution [50,53]. Xiong et al. proposed a Chinese spelling correction framework called HANSpeller, which uses an extended hidden Markov model and ranking model to correct spelling errors in Chinese articles, and achieves further refinement using a rule-based model [54]. Unlike rule-based Chinese text spelling correction methods, machine learning-based methods can automatically learn correction rules according to the data, making them more adaptable to different domains and types of text, with relatively good adaptability and scalability. However, these methods require a large amount of annotated data in practical applications, and obtaining and standardizing text data require a lot of labor and time costs. Moreover, these methods can easily produce erroneous prediction results when processing long texts due to inherent issues with the model.

2.2.3. Deep Learning-Based Chinese Text Spelling Correction Method

The deep learning-based Chinese text spelling correction method is a technology that uses deep learning models to automatically detect and correct spelling errors in Chinese text. The purpose of this technology is to automatically identify spelling errors in text by training a model to provide correct suggestions or automatic corrections to improve the accuracy and readability of the text. Deep learning-based methods can typically handle various types of spelling errors, including typos, homophones, and look-alike characters, and can adaptively process various types and styles of text. This technology has been widely applied in various fields such as search engines, intelligent customer service, and natural language processing [35]. Some research is based on end-to-end networks (e.g., RNN), directly treating Chinese text spelling correction as a sequence labeling task [55,56]. Wang et al. proposed an end-to-end confusion-set-guided encoder–decoder model based on the sequence-to-sequence framework, treating Chinese text spelling correction as a sequence-to-sequence task and injecting confusion set information through

a copying mechanism [40]. However, this method also has a limited generalization ability and performs poorly when dealing with spelling errors that have not appeared in the dataset. Overall, machine learning-based Chinese text spelling correction methods are becoming increasingly popular.

3. Methods

The geographical named entity recognition model uses the BERT framework as the encoder, and further adds a decoder, which includes a fully connected feedforward neural network layer, a bidirectional LSTM layer, and a CRF layer. In addition, adding a CRF layer after BERT can increase the constraint of the data sequence order. The geographical named entity correction model adds a decoder on top of the BERT framework, which includes two sub-modules: error detection and correction. The input of the decoder is passed to the fully connected layer for linear transformation, and then binary classification is used to judge whether the character is correctly segmented or not. The correction module uses the maximum value selection method to convert the output of the encoder into the corresponding characters.

3.1. BERT Model

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language model proposed by Google AI Research [31]. It leverages the transformer's encoder structure, and its performance is enhanced through two pre-training tasks on a vast text corpus, Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), to further improve the performance of the model. The model's architecture is shown in Figure 1.

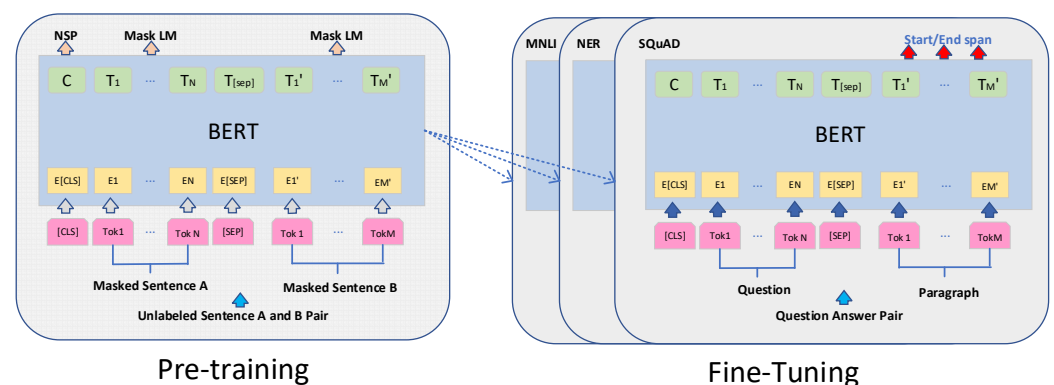


Figure 1. The overall framework for pre-training and fine-tuning the BERT language model.

In the fine-tuning stage, the BERT model is initialized with pre-trained parameters, and a neural network structure is designed for specific downstream tasks. It is then fine-tuned using labeled downstream task datasets.

The emergence of BERT has revolutionized natural language processing, introducing a new pre-training and fine-tuning method for language models. Compared to previous models, the BERT model has stronger semantic understanding capabilities [57]. By leveraging semantic information from the text corpus during pre-training, the self-supervised approach implicitly introduces linguistic knowledge for downstream tasks. In the pre-training of the BERT model, the MASK objective function is used to train the model's loss function, which is commonly referred to as the Masked Language Model (MLM) loss. The formula can be specifically represented as

$$L_{MLM} = -\sum_{(i \in m)} \log P(x_{i_true} | X_{\{<i\}}, X_{\{>i\}}) \quad (1)$$

where M is the set of tokens that are replaced with the MASK token, x_{i_true} is the true i -th token, and $P(x_{i_true} | X_{\{<i\}}, X_{\{>i\}})$ is the probability that the model predicts x_{i_true} , obtained through the softmax function. The objective of this loss function is to

minimize the difference between the predicted and true values. During training, the model will attempt to optimize this loss function to better understand the context and semantic information of the input sequence.

The computation formula of the transformer in the BERT model is as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

In this equation, Q is the query matrix, K is the content to be attended to, and V is the value matrix. The purpose of scaling by d_k is to avoid the dot product from being too large, as a large dot product can result in a very small gradient after passing through softmax.

3.2. Incremental Pre-Training

Similar to fine-tuning, the incremental pre-training of language models is also a transfer learning method. Incremental pre-training refers to further pre-training on a new dataset based on an existing pre-trained language model to enhance the model's performance in a specific domain. Incremental pre-training usually involves the following steps: first, select an existing pre-trained model; then, conduct additional pre-training on the new dataset, typically using the same or similar tasks as the original pre-training; finally, fine-tune the model that has undergone incremental pre-training to adapt to specific downstream tasks.

The advantage of incremental pre-training is that it can utilize information from the new dataset to enhance the model's representational power. Additionally, as the pre-trained model has already learned a significant amount of semantic information, the model can converge faster during incremental pre-training and require less training data. For incremental pre-training, we selected 139,255 Sina Weibo text data points from check-in locations in Jinan city from March to December 2022 after pre-processing and cleaning. We used these data to construct a general web text corpus. Some examples of Sina Weibo text data are shown in Table 1.

Table 1. An example of incrementally pre-trained Weibo text data.

| 景点名称 | Name of Scenic Spot | Longitude | Latitude | Weibo Text |
|-----------|-----------------------------------|-----------|----------|------------|
| 大明湖风景名胜 | Daming Lake Scenic Spot | 117.0244 | 36.6754 | 大明湖畔捡到夏雨荷 |
| 济南千佛山风景名胜 | Jinan Qianfo Mountain Scenic Spot | 117.0369 | 36.6389 | 记录在千佛山看的日落 |
| 趵突泉公园 | Baotu Spring Park | 117.01566 | 36.6615 | 今天的趵突泉有点雾气 |

3.3. Geographic Named Entity Recognition Model

Named entities, also known as name entities, are essential terms in the field of natural language processing, encompassing words and phrases that pertain to specific objects, tasks, places, and more, and are identifiable by their names. Among them, geographical named entities represent a subclass that primarily denotes geographic locations, organization names, institutions, and other relevant entities. For the purpose of this study, geographical named entities specifically refer to texts that are abundant in web texts, imbued with geographic semantics. Apart from merely denoting a geographic location, they encompass descriptions of spatial relationships with specific geographic points, including orientation, distance, and other related aspects.

Based on the BERT model, we constructed a geographic named entity recognition framework that can extract text semantics and perform task-oriented recognition. We first conducted incremental pre-training on the collected ubiquitous network text data, and then added new neural network structures for the geographic named entity recognition task. The resulting model predicts whether each character in the ubiquitous network text belongs to a geographic named entity. The incremental pre-training fine-tuning learning

model and training framework were built for the geographic named entity recognition task. The input data for the model are ubiquitous network text data, which are more natural-language-like and are therefore used by the geographic named entity semantic model for incremental pre-training. The final neural network structure of the geographic named entity recognition model is shown in Figure 2.

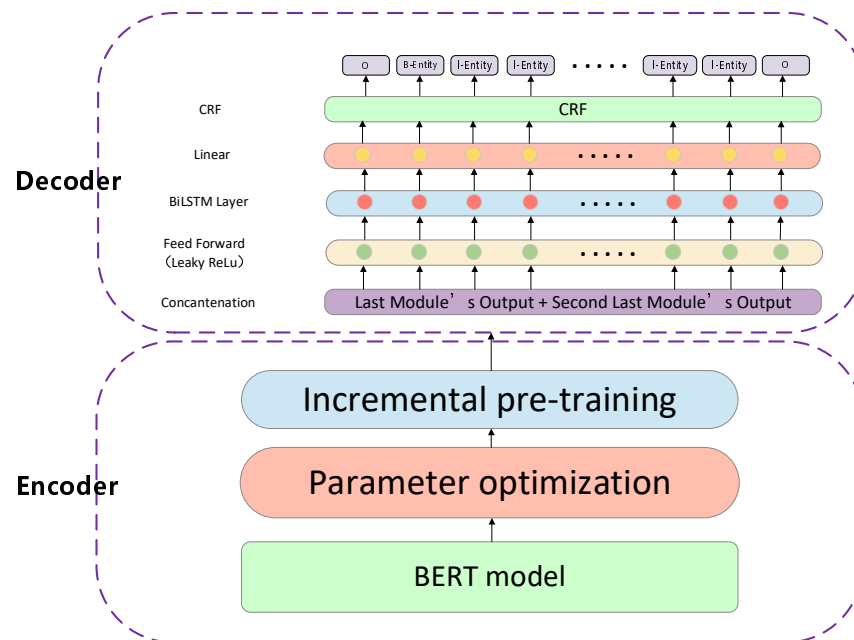


Figure 2. Network structure of the geographic named entity recognition model.

Our model consists of two main parts: the encoder and the decoder. The encoder is mainly used for the pre-training and fine-tuning stages of the BERT model. In the pre-training stage, we adopt an incremental pre-training method to further improve the accuracy of semantic recognition. In the fine-tuning stage, we mainly train the module quantity of the semantic feature extraction module to ensure that BERT's semantic feature extraction achieves the best results. The output of the encoder is used as the input of the decoder, and its dimensionality is kept consistent.

For the decoder part of the model, the output of the encoder is used as the input. It first passes through a fully connected feedforward neural network layer and is transformed nonlinearly via an activation function to maintain the same dimensionality as the input. Then, the output of this layer is used as the input to the bidirectional LSTM layer. The bi-directional LSTM layer contains two LSTM layers, which read the input sequence from both the forward and backward directions and generate forward and backward hidden state sequences, respectively. Finally, these two sequences are concatenated into a new sequence, and the concatenated output is fed into a fully connected layer for sequence labeling classification. The last layer is a CRF layer, which is added after BERT and can add constraints on the sequence data relationships. The CRF layer calculates the CRF loss function based on the sequence labeling classification output of the fully connected layer, thereby adding constraints on the order of generated labels to ensure that the output results are legal. Compared with the BERT model, the previous modules (i.e., the model's input part and the semantic feature extraction module) are exactly the same.

Next, the objective function and loss function of the geographical named entity recognition model will be introduced. When introducing the model structure, it was mentioned that the role of the CRF layer is to calculate the loss function based on the output of the fully connected layer. Specifically, in the geographical named entity recognition task, the loss

function of the CRF layer of the model contains two types of scores: emission score and transition score. In the CRF layer, Equation (3) can be used to represent the probabilities.

$$P(y|x) = \frac{\exp\left(\sum_{i=1}^n \left(\sum_{y_i} X_{iy_i}\right) + \sum_{y_i, y_{i+1}} t_{y_i, y_{i+1}}\right)}{Z(x)} \quad (3)$$

In Equation (3), n is the sequence length, and $\sum_{y_i} X_{iy_i}$ represents the summation of emission scores, which is used to denote the scores of all labels observed at position i with the given feature X_{iy_i} . $\sum_{y_i, y_{i+1}} t_{y_i, y_{i+1}}$ represents the summation of transition scores, which is used to denote the scores of all transitions from one label to another. $Z(x)$ represents the normalization factor used to calculate the sum of possible state sequences.

The emission score x_{iy_i} comes from the output of the BiLSTM layer, which focuses on expressing which geographical named entity label the current character can be mapped to. The transition score t_{y_i, y_j} represents the transition score between categories in the text sequence. For example, $t_{B-Entity, I-Entity} = 0.9$ means the score from the $B-Entity$ category to the $I-Entity$ category is 0.9. The transition score matrix composed of all category transitions is a parameter of the geographical named entity recognition model, and these scores are updated during the iterative process of training. When calculating the CRF loss function, the true path score (emission score and transition score) and the total score of all paths (emission score and transition score) are calculated. The goal of model training is to make the score of the true path the highest among all paths. Therefore, the objective function of the model can be obtained:

$$Objective\ Function = \frac{P_{RealPath}}{P_1 + P_2 + \dots + P_N} \quad (4)$$

Taking the logarithm of the above equation yields the following:

$$Log\ Objective\ Function = \log \frac{P_{RealPath}}{P_1 + P_2 + \dots + P_N} \quad (5)$$

By taking the negative of the objective function above, since the goal of the model training is to minimize the loss function, we obtain the following loss function:

$$Loss\ Function = -\log \frac{P_{RealPath}}{P_1 + P_2 + \dots + P_N} \quad (6)$$

In the equation above, $P_{realpath}$ represents the true path fraction when calculating the CRF loss function, which is composed of the transmit fraction and the transfer fraction. P_i represents the path fraction of the i -th path, so the denominator in the fraction represents the total fraction of all paths.

To make the fine-tuning task more efficient, we propose several optimization strategies. The early stopping strategy is used to prevent overfitting of the neural network. The hierarchical fine-tuning strategy allows for different layers to use different learning rates for model fine-tuning. The layer-by-layer unfreezing strategy, similar to the hierarchical fine-tuning strategy, only trains the parameters related to the target task in higher layers while freezing the general knowledge in lower layers, ensuring that the model's general knowledge is not forgotten. In addition, to evaluate the accuracy of the geographic named entity recognition model, we introduce the evaluation metrics of precision, recall, and F1 score. To introduce these metrics, we first explain several commonly used concepts:

True Positive (TP): The number of positive samples that the model correctly classifies as positive, that is, the number of correctly identified geographic named entities in the sample.

False Positive (FP): The number of negative samples that the model incorrectly classifies as positive, that is, the number of non-geographic named entities identified as geographic named entities in the sample.

True Negative (*TN*): The number of negative samples that the model correctly classifies as negative, that is, the number of correctly identified non-geographic named entities in the sample.

False Negative (*FN*): The number of positive samples that the model incorrectly classifies as negative, that is, the number of geographic named entities identified as non-geographic named entities in the sample.

The formulas for calculating precision, recall, and F1 score are as follows:

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (9)$$

As the above formulas show, precision refers to the samples predicted as positive by the model that are actually positive. Recall refers to the proportion of actual positive samples that are correctly predicted by the model. The F1 score is the harmonic mean of precision and recall, which can comprehensively evaluate the performance of both indicators.

3.4. Geographic Named Entity Correction

Chinese geographical named entity text error correction refers to detecting and correcting errors in spelling, words, and other aspects of geographical named entity text to improve the accuracy and readability of the text. Unlike English text, Chinese text has certain particularities in the form of language. Chinese vocabulary is composed of one or more Chinese characters, and each Chinese character has its own meaning, while English words usually represent a letter or number, where each letter has no meaning in itself, so letters need to be combined to form a vocabulary. Therefore, for English, entity error correction is used to correct the letters in the word, while for Chinese, entity error correction is used to correct a Chinese character. Types of errors generally include spelling errors and word errors. Geographic named entity correction provides researchers with a new perspective for studying geographical named entities and expands the limitations of the original geographic named entity recognition, which was only focused on recognition accuracy. Examples of geographical named entity of different error types are shown in Table 2.

Table 2. Geographical named entity of different error types.

| Error Type | Geographical Named Entity | Translation | Error Entities |
|-------------------|---------------------------|-----------------|----------------|
| Spelling mistakes | 九如山 | Jiuru Mountain | 久如山 |
| Missing spelling | 千佛山 | Qianfo Mountain | 千山 |
| Wrong word used | 大明山 | Daming Mountain | 大名山 |

The geographical named entity text correction model is based on the BERT framework and is also divided into two parts: the encoder and the decoder. The resulting model structure is shown in Figure 3.

The encoder part of the model is consistent with the geographic named entity recognition model. The encoder part of the model is identical to that of the geographic named entity recognition model. As for the decoder part, two modules were designed for error detection and correction. For the error detection module, the input of the decoder first passes through a fully connected layer for linear transformation. Since error detection is a binary classification task that classifies each character as either correct or incorrect, the output dimension of this layer needs to be converted to two dimensions, making the final output dimension suitable for the target task's classification dimension. The softmax function is

applied to the linearly transformed output to compute the probabilities of correctness and incorrectness for each character. These probabilities are normalized, resulting in output values of either 0 or 1. In the output vector, the value on the first dimension represents the probability of being incorrect, while the value on the second dimension represents the probability of being correct. For the error correction module, the output of the encoder is converted to the corresponding character using the maximum value selection method. The ID of the dimension with the maximum output value is used to predict the character and query the lookup table to obtain the corresponding model predicted character.

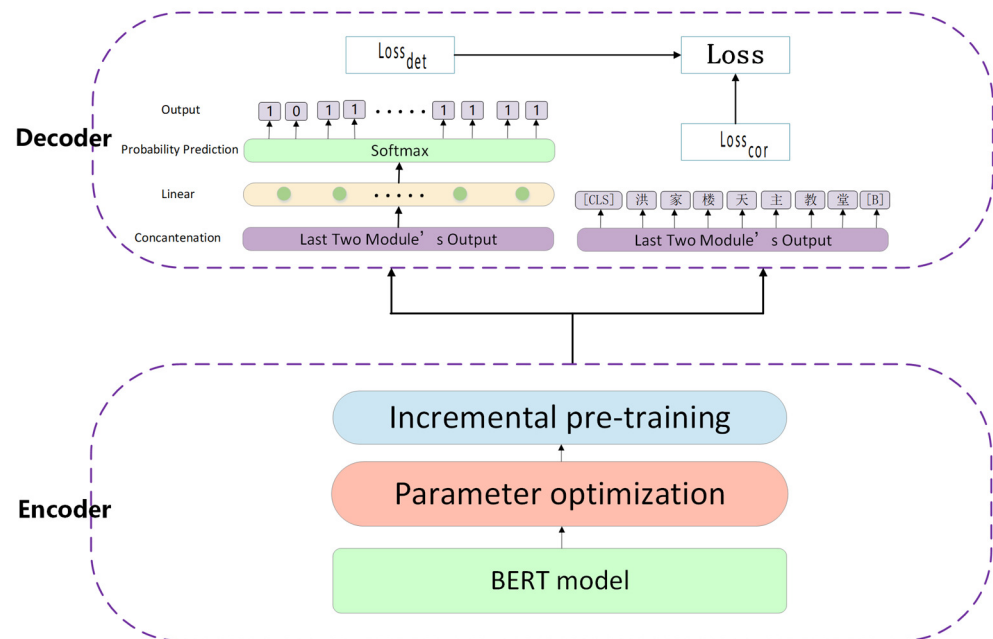


Figure 3. Network structure of the geographic named entity correction model.

Next, we will introduce the objective function and loss function of the geographical named entity correction model. The decoder structure of the geographical named entity text correction model consists of two parts, which include error detection and error correction subtask modules. Therefore, we will discuss the loss functions of these two subtasks separately.

For the error detection subtask, there is a problem of class imbalance in the dataset, which means that the number of samples with incorrect characters is much smaller than that of correct characters. This may lead to poor performance of the model when predicting incorrect characters, which are the focus of the correction task. Therefore, in this section, we use the improved cross-entropy loss function, Focal Loss, to calculate the loss value of the error detection subtask, which has the following mathematical form:

$$Loss_{det} = Focal\ Loss(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (10)$$

The core idea of the Focal Loss function is to balance the importance of positive and negative samples, reducing the weight of easily classified positive samples while focusing on the weight of difficult-to-classify negative samples. In Equation (10), p_t represents the predicted probability of the model for the correct character, and $-\log(p_t)$ is consistent with the cross-entropy loss function. That is, the degree of punishment for correctly classified characters is proportional to the logarithm of the predicted probability, meaning that the smaller the predicted probability, the greater the punishment. $(1 - p_t)^\gamma$ is the unique aspect of Focal Loss, controlling the degree of punishment for difficult-to-classify samples through an adjustable parameter γ . α_t represents the weight of the sample class. In this

study, difficult-to-classify negative samples refer to erroneous characters that need to be detected, and we set the value of α_t to 0.25 and the value of γ to 2.

For the error correction subtask, it is similar to the target task of confusion word correction. Therefore, we only list the mathematical form of the loss value without further explanation.

$$Loss_{cor} = - \frac{\sum_{x \in true_prob} x}{\sum_{x \in true_prob} \delta(x)} \quad (11)$$

In the training process of the model, the loss value $Loss_{det}$ of the error detection subtask and the loss value $Loss_{cor}$ of the error correction subtask need to be weighted and summed to obtain the final loss function. x represents the input text sequence, while $\delta(x)$ represents the text sequence obtained by making small modifications to the input text, which can include operations such as replacing, inserting, or deleting a single character, as shown in Equation (12).

$$Loss = detection_weight \cdot Loss_{det} + (1 - detection_weight) \cdot Loss_{cor} \quad (12)$$

In Equation (12), $detection_weight$ is a hyperparameter for model training, representing the proportion of the loss value of the error detection subtask to the total loss value, which indicates the importance of the error detection subtask in the task of geographical named entity text correction.

To make the task of geographical named entity text correction more effective, we use a gradient optimization strategy during training. This strategy is a way to increase the batch size to improve the training efficiency without increasing the memory usage when training deep neural networks.

4. Experiments and Discussion

4.1. Datasets and Pre-Training

The original data used in this study come from two sources, one of which being geographical named entity data that were manually collected by government departments, and the other being geographical named entity data that were extracted from ubiquitous web text data. Jinan is located in the eastern coastal region of China and is the capital city of Shandong Province. It covers an area of approximately 10,244.45 square kilometers and has 10 districts and two counties under its jurisdiction, as shown in the specific administrative map in Figure 4.

The general preprocessing of geographical named addresses mainly involves methods such as correcting geographical named addresses, filling geographical named elements, and identifying and filling geographical named address elements to preprocess geographical named address data. However, in this study, semantic representation learning was used to obtain the semantic features of geographical named entities. The geographical named entity corpus used in this study not only includes standardized geographical named address data, but also includes geographical named entity data obtained from ubiquitous networks such as social media. Ubiquitous network text data, taking Weibo as an example of the original data source, are shown in Table 3. During the preprocessing of the geographical named entity data, we mainly corrected and cleaned the erroneous data and eliminated the data noise by removing duplicate geographical named entity records, useless characters, and stop words and unifying full-width/half-width characters. Example of the cleaned and preprocessed geographic named entity data are shown in Table 4.

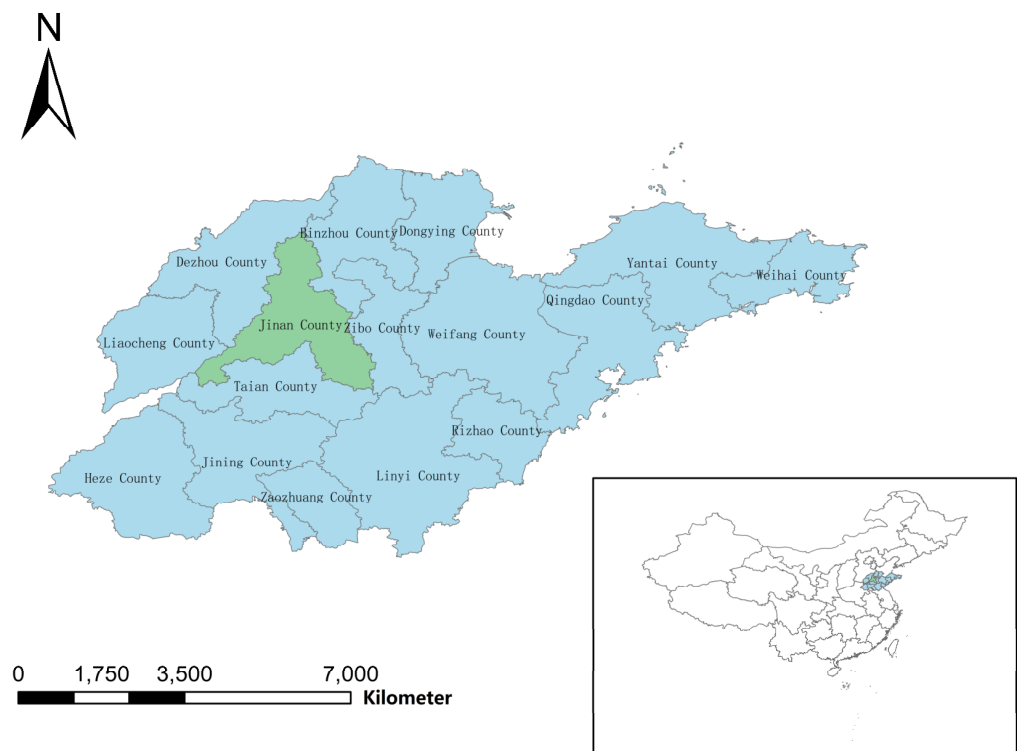


Figure 4. Jinan City location map.

Table 3. Examples of Weibo text data.

| Weibo Original Data Source | Translation |
|----------------------------|---|
| 济南租房独立卫浴春华居可短租到1214800 | Jinan rental independent bathroom Chunhua residence can be short-term rent to 1,214,800 |
| 初冬济南疫情 | Jinan epidemic in early winter |
| 当代济南华山的封闭生活 | The Closed Life of Huashan in Contemporary Jinan |
| 大明湖没有夏雨荷 只有废阳阳的池某灿 | There is no Xia Yuhe in Daming Lake, only Chi Moucan in Abandoned Yangyang |

Table 4. Examples of cleaned and preprocessed geographical named entity data.

| 中文地理命名实体 | Chinese Geographic Named Entity | Longitude | Latitude |
|-------------------|---|------------|-----------|
| 顺河街66号银座晶都国际3号楼2层 | 2nd Floor, Building 3, Ginza Jingdu International, 66 Shunhe Street | 117.001696 | 36.660089 |
| 净水大世界 | Clean water world | 117.044367 | 36.690995 |
| 鑫苑世家公馆正南方向30米 | Xinyuan Family Mansion, due south 30 m | 116.907415 | 36.689018 |

This study uses a total of 3,530,611 geographic named entity text data points located in Jinan City, Shandong Province in 2022, which have been preprocessed and cleaned as the experimental corpus for constructing model instances of geographic named entities.

4.2. Pre-Training and Validation of Place Names Based on the BERT Model

In order to verify the performance and training cost of the proposed BERT model, we calculated various indicators of model training and validation, including the training time,

loss value, perplexity of the validation set, and accuracy of the validation set. These results are shown in Table 5.

Table 5. Training and validation metrics of semantic models for geographical named entities using character encoding strategy in BERT model.

| Metric | Character Encoder Scheme | Training Time | Final Training Loss | Validation Set Perplexity | Validation Set Accuracy |
|--------------|--|----------------|---------------------|---------------------------|-------------------------|
| Metric value | Character Encoding Scheme Used in BERT Model | 24 h 21 m 39 s | 0.6615 | 1.8511 | 86.52% |

Unlike supervised learning in downstream tasks, the pre-training phase of the geographic named entity semantic model is a self-supervised learning task, and there are no standard labeled data. Therefore, this study added perplexity as the main indicator to evaluate the performance of the model. The calculation formula is as follows:

$$perplexity = e^{eval_loss} \quad (13)$$

where e is the natural logarithm, and $eval_loss$ is the average loss value of the model on the validation set.

According to the experimental results, our validation accuracy reached 86.52%, which is at a high level and proves that the BERT model, after improvement, can better understand the semantic meaning of geographic named entity text. After discussing and analyzing the pre-training phase used in previous relevant studies, this study obtained similar conclusions to those of previous studies, proving that even if the research area and data objects are different, the model proposed in this study follows the rules defined by previous studies. This model provides a foundation for subsequent model construction and related application research.

4.3. Comparative Analysis of the Number of Semantic Modules in BERT Model

Previous research [48] has conducted comparative studies on the number of semantic modules in the BERT model. To verify whether the BERT-based geographic named entity recognition model conforms to the conclusions of previous research, we conducted comparative validation experiments on the number of modules and on whether digits are uniformly replaced. In the number comparison validation experiment, we compared the output results of the BERT-based geographic named entity recognition model with those of previous research to evaluate the differences and similarities between them. In the comparison validation of whether digits are uniformly replaced, we verified whether our model can correctly identify and process digits and replace them uniformly to improve the accuracy and consistency of geographic named entity recognition, addressing the issues found in previous research. Through these validation experiments, we can more accurately evaluate the performance and reliability of the BERT-based geographic named entity recognition model and provide guidance and reference for further improving and optimizing geographic named entity recognition technology.

Therefore, we conducted a comparative analysis of different numbers of semantic feature extraction modules. After training with the training set of the geographic named entity corpus, we set different numbers of semantic feature extraction modules (6, 8, 10, and 12) for four control models. The specific training and validation indicators are shown in Table 6.

Table 6. Training and validation metrics of semantic models for geographical named entities with different numbers of semantic feature extraction modules.

| Metric | Number of Semantic Feature Extraction Modules | Training Time | Final Training Loss | Validation Set Perplexity | Validation Set Accuracy |
|--------------|---|----------------|---------------------|---------------------------|-------------------------|
| Metric value | 6 | 29 h 45 m 43 s | 0.1813 | 1.2616 | 94.72% |
| | 8 | 35 h 52 m 41 s | 0.2567 | 1.2572 | 94.84% |
| | 10 | 41 h 45 m 51 s | 0.1745 | 1.2518 | 94.96% |
| | 12 | 47 h 40 m 07 s | 0.1736 | 1.2497 | 95.03% |

According to Table 6, as the number of semantic feature extraction modules increases, the training time also increases correspondingly, showing a positive correlation with the module quantity. The perplexity of the training set is roughly negatively correlated with the number of modules, and the loss of the training set and the perplexity of the validation set are relatively small, indicating that there is no overfitting phenomenon. This study ultimately used a model with 12 semantic feature extraction modules as the basic framework for further research to ensure the ultimate effectiveness of downstream tasks.

To further improve the accuracy of subsequent models based on the BERT framework, we conducted a digit replacement experiment after selecting the upper module features. In the geographical named entity recognition task, identifying digits is relatively difficult and less practical. Therefore, we conducted an experiment to replace Arabic numerals in the geographical named entity text corpus, and the experimental results are shown in Table 7.

Table 7. Training and validation metrics of semantic models for geographical named entities in text corpus where Arabic numerals are replaced with (NUM) and where they are not replaced.

| Metric | Whether Arabic Numerals Are Replaced by (NUM) | Training Time | Final Training Loss | Validation Set Perplexity | Validation Set Accuracy |
|--------------|---|----------------|---------------------|---------------------------|-------------------------|
| Metric value | Y | 50 h 16 m 00 s | 0.0331 | 1.0975 | 97.96% |
| | N | 47 h 40 m 07 s | 0.1736 | 1.2497 | 95.03% |

Based on Table 6, we can see that the Arabic numerals in the geographic named entity corpus affect the model's ability to extract semantic features from the text. At the same time, only a small amount of training time cost was added, and the loss value of the training set and the perplexity of the validation set decreased significantly. In order to improve the semantic feature extraction ability of the BERT model, we replaced the Arabic numerals in the corpus with unified codes. Finally, we decided to use the geographic named entity corpus with the replaced Arabic numerals as the basis for subsequent research.

4.4. Geographic Named Entity Recognition

In order to improve the robustness and resilience of the geographic named entity recognition model, we conducted incremental pre-training on a ubiquitous network text corpus constructed from 139,255 Sina Weibo texts that were checked-in within Jinan city from March to December 2022, after cleaning and preprocessing. Sina Weibo is one of the most popular social media platforms in China; it is a microblog-based social network where users can post short texts, pictures, videos, and audios and interact with other users. As of 2022, the number of active users on Sina Weibo exceeded 580 million, making it one of the most important platforms in the field of social media in China. The obtained Weibo text data can be identified using the named entity recognition model to obtain the corresponding entities. The example process of geographic named entity recognition is shown in Figure 5. Examples of recognized entities are shown in Table 8.

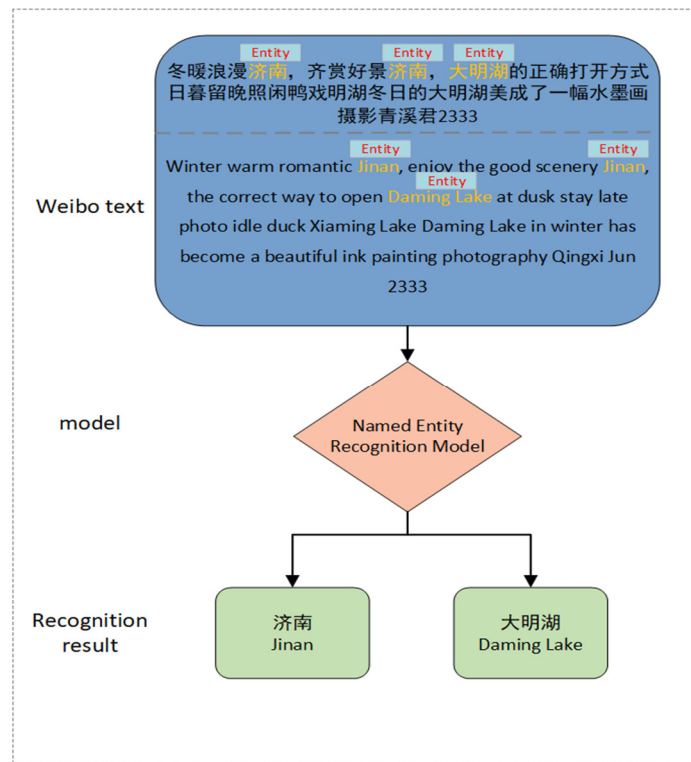


Figure 5. The example process of geographic named entity recognition.

Table 8. Training data examples for the geographical named entity recognition model.

| Chinese Text | Display in English | Recognition Result |
|---------------------|--|---------------------|
| 济南华山湖朝霞现天空之镜 | Jinan Huashan Lake sunrise mirror of the sky | Jinan; Huashan Lake |
| 第一次来济南和 大明湖合个影 | First time coming to Jinan and taking a photo with Daming Lake | Jinan; Daming lake |
| 趵突泉很好 就是体感47度已经肝不动了 | Baotu Spring is very good, but feeling 47 degrees is already so tired | Baotu Spring |
| 初冬大明湖济南身边事同城种草 | Early winter Daming Lake Jinan things around the same city recommended | Daming Lake; Jinan |

For the geographic named entity recognition model, this study developed a training framework with hyperparameters, as shown in Table 9.

Table 9. Training hyperparameters for the geographical named entity recognition model.

| Hyperparameters | Value |
|--------------------------|------------------------------|
| Batch_size | 64 |
| Initial_learning_rate | 0.00003 |
| epoch | 100 |
| num_labels | 3(B-entity, I-entity, and O) |
| Lstm_hidden_size | 1024 |
| Num_transformer_module | 12 |
| Patience_num | 0.0002 |
| Layer_wise_learning_rate | 15 |
| Min_epoch_num | 10 |

This study first manually constructed 14,728 labeled data points of geographic named entity recognition based on the collected Sina Weibo data, and integrated 2363 data points from the CLUENER2020 dataset to form the dataset for the geographical named entity recognition task. The model was fine-tuned based on the above incremental pre-training. The ratio of the training set, validation set, and testing set during the training process was 8:1:1, and the data selection method was random. In addition, to evaluate the recognition accuracy of the proposed method for geographic named entity recognition, we compared it with several traditional methods for geographic named entity recognition, including Fully Connected Neural Network + CRF, RNN + CRF, and BiLSTM + CRF. Three commonly used classification indicators were used to evaluate the effectiveness of geographic named entities, namely precision, recall, and F1-score.

Similar to the pre-training task of the geographical named entity recognition model, to verify the performance and training cost of the recognition model, we also calculated various indicators for model training and validation, as shown in Table 10.

Table 10. Incremental pre-training task for geographical named entity recognition model on various metrics on the training and validation sets.

| Task Name | Training Time | Final Training Loss | Validation Set Perplexity | Validation Set Accuracy |
|---|---------------|---------------------|---------------------------|-------------------------|
| Geographic named entity recognition model incremental pre-training task | 4 h 42 m 56 s | 0.5199 | 1.9393 | 87.72% |

From Table 9, we can see that the model has relatively good performance on the ubiquitous web text corpus with a relatively small amount of data, indicating that the model has strong text semantic understanding ability. After verifying the effectiveness of the model, we compared it with several traditional models. The comparison results of the precision, recall, and F1 score for each method are shown in Table 11.

Table 11. Comparison results of precision, recall, and F1 score between the geographical named entity recognition model and other methods.

| Geographic Named Entity Recognition Method | Precision | Recall | F1 Score |
|--|---------------|---------------|---------------|
| FCNN + CRF | 0.7981 | 0.7533 | 0.7751 |
| RNN + CRF | 0.8535 | 0.8299 | 0.8415 |
| BiLSTM + CRF | 0.8524 | 0.8379 | 0.8451 |
| BERT + CRF | 0.9018 | 0.9074 | 0.9045 |

As shown in Table 11, our improved BERT-based geographic named entity recognition model has significant advantages over traditional methods (FCNN + CRF, RNN + CRF, and BiLSTM + CRF), with an F1 score of around 0.90, which proves that incorporating social media information into geographic named entity recognition with pre-trained language models can greatly improve the accuracy of geographic named entity recognition tasks. By comparing BiLSTM + CRF with our method, it can be concluded that the method based on the pre-trained model can enable the model to learn more language features and have better generalization ability, and thus improve the recognition effect. Our proposed geographic named entity recognition model outperforms previous methods in precision, recall, and F1 score, demonstrating its excellent identification performance. Overall, our proposed model performs better than previous traditional models and has better comprehensive performance, laying a foundation for further research on geographic named entity recognition text.

4.5. Geographic Named Entity Correction

This study used a total of 3,530,611 geographic named entity text data points located in Jinan City, Shandong Province in 2022, which were preprocessed and cleaned, as the experimental corpus for constructing model instances of geographic named entities.

To perform the correction of geographic named entities, we used a dataset of 779,924 text samples of geographic named entities obtained through manual annotation and data augmentation techniques. The data augmentation methods used included vocabulary replacement, back-translation, random insertion, and random deletion.

The dataset was randomly divided into training, validation, and testing sets with proportions of 85%, 10%, and 5%, respectively. The structure of the dataset included two fields, which were the original geographic named entities and the corrected geographic named entities. The training data accounted for 85% of the total dataset, which was used with the aim of training a geographic named entity correction model with a large amount of data, while the validation data accounted for 10%, and was mainly used to evaluate the trained correction model and adjust and optimize the model parameters based on the evaluation results. Some examples of geographic named entity correction are shown in Table 12. Additionally, the hyperparameters used to train the geographic named entity correction model are shown in Table 13.

Table 12. Examples of corrected geographical named entities.

| Geographical Named Entity | Display in English | Is Corrected | Error Character Position | After Correction |
|---------------------------|--------------------|--------------|--------------------------|------------------|
| 久如山 | Jiuru Mountain | N | 1 | 九如山 |
| 千佛山 | Qianfu Mountain | Y | | 千佛山 |
| 老山 | Lao Mountain | N | 1 | 崂山 |
| 大明湖 | Daming Lake | N | 2 | 大明湖 |

Table 13. Model training hyperparameters.

| Hyperparameters | Value |
|-------------------------|-------|
| Batch_size | 64 |
| Accumulate_grad_batches | 4 |
| Detection_weight | 0.3 |

After data augmentation and model training on the geographical text corpus, in order to evaluate the geographical named entity text correction model, we compared it with the RNN (seq2seq) method. The comparison results of our geographical named entity correction model and the traditional error correction method based on RNN and the seq2seq model architecture on the sequence metrics are shown in Table 14.

Table 14. Comparison results of sequence-level metrics between the geographical named entity text correction model and the method based on the seq2seq model structure.

| Geographic Named Entity Correction Method | Sequence-Level Precision | Sequence-Level Recall | Sequence-Level F1 Score | Sequence-Level Accuracy |
|---|--------------------------|-----------------------|-------------------------|-------------------------|
| RNN (seq2seq) | 0.8983 | 0.4369 | 0.5879 | 0.8051 |
| BERT | 0.9765 | 0.7647 | 0.8577 | 0.8595 |

From Table 14, compared with traditional RNN error correction methods, it can be seen that our proposed error correction method improves the precision rate from 89.8% to 97.6%, improves the recall rate from 43.7% to 76.5%, improves the F1 value from 58.8% to

85.8%, and improves the accuracy rate from 80.5% to 85.9%. We can see that our proposed model for geographic named entity correction outperforms the seq2seq model architecture on all metrics, demonstrating that the deep learning architecture using incremental pre-training and data augmentation can greatly improve the accuracy and effectiveness of geographic named entity text correction tasks. Other parameters of the geographic named entity correction model are shown in Table 15.

Table 15. Various error correction metrics of the geographical named entity correction model.

| Geographic Named Entity Correction Method | Precision of Error Detection | Recall of Error Detection | F1 Score of Error Detection | Error Correction Accuracy | Error Correction Recall Rate | Error Correction F1 Score | Sequence-Level Precision | Sequence-Level Recall | Sequence-Level F1 Score | Sequence-Level Accuracy |
|---|------------------------------|---------------------------|-----------------------------|---------------------------|------------------------------|---------------------------|--------------------------|-----------------------|-------------------------|-------------------------|
| BERT | 0.9388 | 0.8845 | 0.9108 | 0.9200 | 0.8679 | 0.8932 | 0.9765 | 0.7647 | 0.8577 | 0.8595 |

Based on Table 15, we can see that our proposed geographic named entity text correction model achieved good results in most of the detection, correction, and sequence-level metrics. Our model's F1 scores for detection, correction, and sequence level were 0.9108, 0.8932, and 0.8577, respectively, all of which reached a high level, demonstrating the excellent correction effect of our proposed geographic named entity correction model. This indirectly proves that the geographic named entity correction model has a strong understanding ability of geographic named entity text semantic features, which lays the foundation for future research on geographic named entities.

5. Conclusions

In this paper, we first introduced the BERT model framework in natural language processing tasks. Then, we fine-tuned the BERT framework and performed incremental pre-training, which improved the BERT framework and achieved good results. Based on the BERT framework, we extended and constructed a geographical named entity recognition (NER) model. By comparing it with other geographical NER methods, we proved that our pre-training fine-tuning approach performed better than the other recognition methods, laying the foundation for further research on geographical named entity recognition. Finally, our proposed geographical named entity correction (NEC) model was also extended and improved based on the BERT framework. Through a comparison with traditional sequence-to-sequence-based correction methods, our proposed geographical NEC model achieved much better correction results. With the continuous development and popularization of natural language processing technology, geographical named entity recognition and correction techniques will be further developed and applied and will become important tools for people to process geographical information and manage geographical knowledge. However, current geographical named entity recognition and correction technologies still have some shortcomings, such as high error rates in geographical named entity correction, incomplete coverage of geographical information data, and differences in processing geographical names in different languages. In summary, geographical named entity recognition and correction techniques have broad application prospects, but also have certain defects. In the future, we will focus on realizing more content related to geographical named entities and further expanding and upgrading these functions. We will also improve the BERT model to rebuild the geographical NER and NEC models and propose functions such as place name recommendation.

Author Contributions: Conceptualization: Wei Zhang, Liuchang Xu, Fei Li, and Jianhua Wan; methodology: Jingtao Meng, Jiajun Zhang, and Chengkun Zhang; writing—original draft: Wei Zhang, Jingtao Meng, and Fei Li; writing—review and editing: Yuanyuan Wang and Jianhua Wan; visualization: Jiajun Zhang and Liuchang Xu. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Major Science and Technology Innovation Project of Shandong Province, Grant No. 2019JZZY020103; the National Natural Science Foundation of China, Grant No. 42050103; and the Natural Science Foundation of Zhejiang Province, Grant No. LGG22D010001.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding authors upon request.

Acknowledgments: The authors would like to thank the editor and the reviewers for their contributions.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

- Huang, H.; Gartner, G.; Krisp, J.M.; Raubal, M.; Van de Weghe, N. Location based services: Ongoing evolution and research agenda. *J. Locat. Based Serv.* **2018**, *12*, 63–93. [\[CrossRef\]](#)
- Yao, X.A.; Huang, H.; Jiang, B.; Krisp, J.M. Representation and analytical models for location-based big data. *Int. J. Geogr. Inf. Sci.* **2019**, *33*, 707–713. [\[CrossRef\]](#)
- Li, W. GeoAI: Where machine learning and big data converge in GIScience. *J. Spat. Inf. Sci.* **2020**, *20*, 71–77. [\[CrossRef\]](#)
- Mozharova, V.A.; Loukachevitch, N.V. Combining knowledge and CRF-based approach to named entity recognition in Russian. In Proceedings of the 5th International Conference on Analysis of Images, Social Networks and Texts, AIST 2016, Yekaterinburg, Russia, 7–9 April 2016; Revised Selected Papers 5. Springer International Publishing: New York, NY, USA, 2017; pp. 185–195.
- Yin, W.; Kann, K.; Yu, M.; Schütze, H. Comparative study of CNN and RNN for natural language processing. *arXiv* **2017**, arXiv:1702.01923.
- 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. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#)
- Sagcan, M.; Karagoz, P. Toponym recognition in social media for estimating the location of events. In Proceedings of the 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, USA, 14–17 November 2015; IEEE: New York, NY, USA, 2015; pp. 33–39.
- Bae, S.H.; Yun, H.J. Spatiotemporal distribution of visitors' geotagged landscape photos in rural areas. *Tour. Plan. Dev.* **2017**, *14*, 167–180. [\[CrossRef\]](#)
- Musaev, A.; Wang, D.; Shridhar, S.; Lai, C.A.; Pu, C. Toward a real-time service for landslide detection: Augmented explicit semantic analysis and clustering composition approaches. In Proceedings of the 2015 IEEE International Conference on Web Services, New York, NY, USA, 27 June–2 July 2015; IEEE: New York, NY, USA, 2015; pp. 511–518.
- Zhu, J.Q.; Lu, L.; Ma, C.M. From interest to location: Neighbor-based friend recommendation in social media. *J. Comput. Sci. Technol.* **2015**, *30*, 1188–1200. [\[CrossRef\]](#)
- Zhang, C.; Zhang, Y.; Zhang, J.; Yao, J.; Liu, H.; He, T.; Zheng, X.; Xue, X.; Xu, L.; Yang, J.; et al. A Deep Transfer Learning Toponym Extraction and Geospatial Clustering Framework for Investigating Scenic Spots as Cognitive Regions. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 196. [\[CrossRef\]](#)
- Zhang, S.; Huang, H.; Liu, J.; Li, H. Spelling error correction with soft-masked BERT. *arXiv* **2020**, arXiv:2005.07421.
- Liu, S.; Yang, T.; Yue, T.; Zhang, F.; Wang, D. PLOME: Pre-training with misspelled knowledge for Chinese spelling correction. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing 2021, Online, 1–6 August 2021; Volume 1: Long Papers. pp. 2991–3000.
- Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv* **2015**, arXiv:1508.01991.
- Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; Dyer, C. Neural architectures for named entity recognition. *arXiv* **2016**, arXiv:1603.01360.
- Ma, X.; Hovy, E. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv* **2016**, arXiv:1603.01354.
- Chiu, J.P.; Nichols, E. Named entity recognition with bidirectional LSTM-CNNs. *Trans. Assoc. Comput. Linguist.* **2016**, *4*, 357–370. [\[CrossRef\]](#)
- Zhang, R.; Pang, C.; Zhang, C.; Wang, S.; He, Z.; Sun, Y.; Wu, H.; Wang, H. Correcting Chinese spelling errors with phonetic pre-training. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 1–6 August 2021; pp. 2250–2261.
- 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. [\[CrossRef\]](#)
- 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.

22. Clough, P.; Pasley, R. Images and perceptions of neighbourhood extent. In Proceedings of the 6th Workshop on Geographic Information Retrieval, Zurich, Switzerland, 18–19 February 2010.
23. 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](#)]
24. Medway, D.; Warnaby, G. What's in a name? Place branding and toponymic commodification. *Environ. Plan. A* **2014**, *46*, 153–167. [[CrossRef](#)]
25. Zhang, W.; Gelernter, J. Geocoding location expressions in Twitter messages: A preference learning method. *J. Spat. Inf. Sci.* **2014**, *9*, 37–70.
26. de Bruijn, J.A.; de Moel, H.; Jongman, B.; de Ruiter, M.C.; Wagemaker, J.; Aerts, J.C. A global database of historic and real-time flood events based on social media. *Sci. Data* **2019**, *6*, 311. [[CrossRef](#)]
27. McKenzie, G.; Liu, Z.; Hu, Y.; Lee, M. Identifying urban neighborhood names through user-contributed online property listings. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 388. [[CrossRef](#)]
28. 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](#)]
29. 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](#)]
30. 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](#)]
31. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
32. Liu, H.; Qiu, Q.; Wu, L.; Li, W.; Wang, B.; Zhou, Y. Few-shot learning for name entity recognition in geological text based on GeoBERT. *Earth Sci. Inform.* **2022**, *15*, 979–991. [[CrossRef](#)]
33. 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](#)]
34. 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](#)]
35. 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](#)]
36. Guo, Z.; Ni, Y.; Wang, K.; Zhu, W.; Xie, G. Global attention decoder for Chinese spelling error correction. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 1–6 August 2021.
37. Yeh, J.F.; Li, S.F.; Wu, M.R.; Chen, W.Y.; Su, M.C. Chinese word spelling correction based on n-gram ranked inverted index list. In Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing 2013, Nagoya, Japan, 14–18 October 2013.
38. Yu, J.; Li, Z. Chinese spelling error detection and correction based on language model, pronunciation, and shape. In Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, Wuhan, China, 20–21 October 2014.
39. Xiong, J.; Zhang, Q.; Zhang, S.; Hou, J.; Cheng, X. HANSpeller: A unified framework for Chinese spelling correction. *Int. J. Comput. Linguist. Chin. Lang. Process.* **2015**, *20*, 1–22.
40. Wang, D.; Tay, Y.; Zhong, L. Confusionset-guided pointer networks for Chinese spelling check. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019.
41. Chang, C.H. A new approach for automatic Chinese spelling correction. *Proc. Nat. Lang. Process. Pac. Rim Symp.* **1995**, *95*, 278–283.
42. Zhang, L.; Zhou, M.; Huang, C.; Pan, H. Automatic detecting/correcting errors in Chinese text by an approximate word-matching algorithm. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong, China, 3–6 October 2000.
43. Huang, C.; Wu, M.; Chang, C. Error detection and correction based on Chinese phonemic alphabet in Chinese text. In *Modeling Decisions for Artificial Intelligence, Proceedings of the 4th International Conference, MDAI 2007, Kitakyushu, Japan, 16–18 August 2007*; Springer: Berlin/Heidelberg, Germany, 2007.
44. Hung, T.H.; Wu, S.H. Chinese essay error detection and suggestion system. In Proceedings of the Taiwan E-Learning Forum; 2008.
45. Jiang, Y.; Wang, T.; Lin, T.; Wang, F.; Cheng, W.; Liu, X.; Wang, C.; Zhang, W. A rule based Chinese spelling and grammar detection system utility. In Proceedings of the 2012 International Conference on System Science and Engineering (ICSSE), Dalian, China, 30 June–2 July 2012; IEEE: New York, NY, USA, 2012.
46. Hong, Y.; Yu, X.; He, N.; Liu, N.; Liu, J. FASPELL: A fast, adaptable, simple, powerful Chinese spell checker based on DAE-decoder paradigm. In Proceedings of the 5th Workshop on Noisy User-Generated Text (W-NUT 2019), Hong Kong, China, 4 November 2019.
47. Song, J.; Guo, Z.; Gao, L.; Liu, W.; Zhang, D.; Shen, H.T. Hierarchical LSTM with adjusted temporal attention for video captioning. *arXiv* **2017**, arXiv:1706.01231.
48. Guo, Z.; Gao, L.; Song, J.; Xu, X.; Shao, J.; Shen, H.T. Attention-based LSTM with semantic consistency for videos captioning. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016.

49. Xie, W.; Huang, P.; Zhang, X.; Hong, K.; Huang, Q.; Chen, B.; Huang, L. Chinese spelling check system based on n-gram model. In Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing, Beijing, China, 30–31 July 2015.
50. Tseng, Y.H.; Lee, L.H.; Chang, L.P.; Chen, H.H. Introduction to SIGHAN 2015 bake-off for Chinese spelling check. In Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing, Beijing, China, 30–31 July 2015.
51. Jia, Z.; Wang, P.; Zhao, H. Graph model for Chinese spell checking. In Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing (SIGHAN-7), Nagoya, Japan, 14 October 2013.
52. Xin, Y.; Zhao, H.; Wang, Y.; Jia, Z. An improved graph model for Chinese spell checking. In Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, Wuhan, China, 20–21 October 2014.
53. Wang, D.; Song, Y.; Li, J.; Han, J.; Zhang, H. A hybrid approach to automatic corpus generation for Chinese spelling check. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018.
54. Xiong, J.; Zhang, Q.; Hou, J.; Wang, Q.; Wang, Y.; Cheng, X. Extended HMM and ranking models for Chinese spelling correction. In Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, Wuhan, China, 20–21 October 2014.
55. Zheng, B.; Che, W.; Guo, J.; Liu, T. Chinese grammatical error diagnosis with long short-term memory networks. In Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications, Osaka, Japan, 12 December 2016.
56. Yang, Y.; Xie, P.; Tao, J.; Xu, G.; Li, L.; Si, L. Alibaba at IJCNLP-2017 task 1: Embedding grammatical features into LSTMs for Chinese grammatical error diagnosis task. In Proceedings of the IJCNLP 2017, Shared Tasks, Taipei, Taiwan, 27 November–1 December 2017.
57. Xu, L.; Mao, R.; Zhang, C.; Wang, Y.; Zheng, X.; Xue, X.; Xia, F. Deep Transfer Learning Model for Semantic Address Matching. *Appl. Sci.* **2022**, *12*, 10110. [[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.