applied sciences

MDPI

Article

Non-Standard Address Parsing in Chinese Based on Integrated CHTopoNER Model and Dynamic Finite State Machine

Mengwei Zhang, Xingui Liu *, Jingzhen Ma, Zheng Zhang, Yue Qiu and Zhipeng Jiang

Institute of Geospatial Information, Information Engineering University, Zhengzhou 450001, China; zmweiwei1999@163.com (M.Z.); zb50mjz@163.com (J.M.); giser_zzy@163.com (Z.Z.); qiuyue@whu.edu.cn (Y.Q.); jiangzp0803@163.com (Z.J.)

* Correspondence: a9812260211@126.com; Tel.: +86-13-0176-63935

Abstract: Information in non-standard address texts in Chinese is usually presented with rough content, complex and diverse presentation forms, and inconsistent hierarchical granularity, causing low accuracy in Chinese address parsing. Therefore, we propose a method for parsing non-standard address text in Chinese that integrates the Chinese Toponym Named Entity Recognition (CHTopoNER) model and a dynamic finite state machine (FSM). First, named entity recognition is performed by the CHTopoNER model. Sets of dynamic FSMs are then constructed based on the address hierarchical characteristics to sort and combine the Chinese address elements, thereby achieving address parsing on the Chinese internet. This method showed excellent accuracy in parsing both standard and non-standard placename addresses. In particular, this method performed better in address parsing for disordered or missing hierarchical elements than traditional methods using an FSM. Specifically, this method achieved accuracies of 96.6% and 96.8% for standard and non-standard placenames, respectively. These accuracies increased by 8.0% and 57.1%, respectively, compared with the integrated CHTopoNER model and traditional FSM, and by 7.4% and 19.8%, respectively, compared with the integrated CHTopoNER model and bidirectional FSM. After analysis, the address-parsing method showed good scalability and adaptability, which could be applied to various types of address-parsing tasks.

Keywords: Chinese address parsing; CHTopoNER model; finite state machine; dynamic finite state machine; Chinese internet text

Citation: Zhang, M.; Liu, X.; Ma, J.; Zhang, Z.; Qiu, Y.; Jiang, Z. Non-Standard Address Parsing in Chinese Based on Integrated CHTopoNER Model and Dynamic Finite State Machine. *Appl. Sci.* **2023**, *13*, 9855. https://doi.org/10.3390/app13179855

Academic Editor: Douglas O'Shaughnessy

Received: 31 July 2023
Revised: 29 August 2023
Accepted: 30 August 2023
Published: 31 August 2023

1. Introduction

With the rapid development of computer technology, services based on location data (such as express and food delivery) are becoming increasingly important [1]. Address matching is an integral part of these services. Problems, such as word segmentation, ambiguity, and unregistered words in Chinese, pose more demanding requirements and significant challenges to Chinese address parsing. Owing to its importance in address-matching accuracy, address parsing is a key part of the process of address matching [2]. However, addressing information usually includes difficulties, such as disordered structure, high degree of information fragmentation, complex and diverse presentation forms, and diversified information hierarchical granularity, in daily Chinese texts on the internet [3–5]. This increases the difficulty of parsing addresses in the text. Therefore, address parsing is the first issue to be considered when studying high-accuracy address matching.

Some studies have demonstrated results in Chinese address parsing. For example, Wu et al. [6] proposed a Chinese address parsing method based on the BERT-BiLSTM-CRF deep learning model. This method uses the BERT [7] pretrained language model to obtain character vectors rich in semantic information, thereby improving the ability to extract elements from complex addresses. By targeting the differences in local and racial cultures and the urban development in various regions in China, Ma et al. [8] expanded and refined

the classification of Chinese address elements considering their characteristics. They also summarized the common vocabulary of spatial relationships in the addresses and the combination patterns among Chinese address elements. Cheng et al. [9] built a labeling system for word segmentation of Chinese addresses using Bidirectional Long Short-Term Memory (BiLSTM) to extract contextual features. The optimal label sequence was determined by integrating conditional random fields, thereby improving address entity recognition.

In the field of Chinese address parsing, some significant research achievements have been made; nevertheless, several issues remain to be addressed. Specifically, the following challenges persist:

1. The existing methods for Chinese address element recognition segregate at the character level, without considering traditional Chinese word segmentation techniques commonly used in natural language processing. Consequently, the acquired semantic representations are at the character level and fail to capture the word-level features inherent in the Chinese language.
2. Previous studies on Chinese address element recognition rarely consider the integration of local and global features during feature extraction. Hence, during feature extraction, problems such as missing global or local semantic information arise.
3. Furthermore, after the address parsing process, the traditional finite state machines heavily rely on pre-recorded keywords of address elements or require threshold settings when ordering and combining these elements. While finite state machines (FSMs) and bidirectional FSMs perform well in handling standardized addresses, they struggle to effectively process address information descriptions present in web texts, characterized by hierarchical element disorder and omissions.

The innovations introduced in this paper to address the aforementioned issues are summarized as follows:

1. This study proposes the CHTopoNER model for identifying hierarchical address elements in online text. The innovation of this model is derived from the improved SoftLexicon approach and the integration of BiLSTM and Iterated Dilated Convolutional Neural Network (IDCNN) models to form a Two Channel Neural Network (TCNN) layer. Specifically, the enhanced SoftLexicon approach is utilized to acquire word-level semantic information while avoiding potential out-of-vocabulary issues, resulting in more accurate identification of Chinese toponyms' word boundaries. The TCNN layer comprehensively considers both character- and word-level local semantic features as well as global semantic features from the input text, thus minimizing the loss of semantic information and effectively addressing the ambiguity of Chinese address entity elements.
2. This study introduces a dynamic FSM. In comparison to the traditional FSM, the dynamic FSM is capable of adjusting its state set based on the types of address elements. This adaptation avoids the limitations of depending heavily on the collected keywords of address elements and the threshold settings inherent to the FSM. Consequently, it can more effectively handle address information descriptions in network text that exhibit issues such as hierarchical element disorder and omission.

The remainder of this paper is organized as follows. Section 2 presents the related work on address parsing and the application of FSM in address parsing. Section 3 presents the detailed deep learning architecture method using a dynamic FSM and the experimental data discussed in this study, and Section 4 presents and analyzes the experimental results. Finally, Section 5 summarizes the study and discusses future research directions.

2. Related Work

Address parsing is the process of decomposing address strings into address elements and determining their types [10]. This constitutes a pivotal undertaking within the domain of natural language processing [11]. To proficiently address this challenge, researchers have proposed diverse methodologies and techniques. In this section, we shall present a

comprehensive overview of pertinent research in the context of Chinese address parsing, encompassing methodologies rooted in dictionaries, rules, statistics, and deep learning. As FSMs can be harnessed to systematically arrange and amalgamate address components, they lay the cornerstone for ensuing endeavors in high-precision address matching. Accordingly, this section will also delve into the utilization of FSMs in the context of address-parsing tasks.

2.1. Relevant Research on Chinese Address Parsing

2.1.1. Dictionary-Based Address Parsing

The dictionary-based approach for addressing parsing stands as one of the earliest instances of applying mechanical text segmentation techniques to this task. This methodology involves breaking down addresses through string matching. Generally, the process initiates by first splitting the address text into individual characters and then sequentially comparing these characters with the placenames listed in an address dictionary. Should a match be identified, the term is retained; conversely, the method attempts to establish a match by either adding or removing a single character until only one character remains. If the string still resists segmentation, it is treated as an out-of-vocabulary term [12]. The development of the dictionary predominantly relies on a substantial repository of pre-existing toponym data, and the parsing procedure deliberately sidesteps the integration of address rule knowledge or statistical insights [13]. Currently, there is an ongoing refinement of the dictionary-based address-parsing technique. As an illustration, Ye et al. [14] augmented this approach by forming sets of potential placenames based on shared character features among toponyms and constructing a single-character index to elevate query efficiency and parsing precision. In another vein, Li et al. [15] adopted a forward adaptive length matching algorithm grounded in address indicator words to curtail redundant data input and heighten parsing efficiency.

The efficacy of these dictionary-based methods heavily hinges on the comprehensiveness of the dictionary, leaving them incapable of recognizing out-of-vocabulary terms within addresses. In the face of the incessant emergence of novel toponyms and addresses, traditional toponym dictionaries grapple with staying aligned with the evolving landscape [16].

2.1.2. Rule-Based Address Parsing

Chinese addresses possess specific rules and distinctive features. By systematically summarizing the attributes and regulations governing Chinese addresses, it becomes feasible to devise address-parsing techniques founded on the characteristics of address components and address structures. This approach facilitates the achievement of a structured address-parsing outcome [17,18]. Zhang et al. [19] undertook the task of consolidating the traits of address elements and dissecting the parsing regulations that govern these attributes. Consequently, they devised a rule-based address-parsing algorithm that integrates feature characters and regulations. Tan et al. [20] introduced innovative mechanisms encompassing rule trees and ambiguity storage, thereby achieving a rule-based approach for address segmentation and matching. This method tangibly enhanced the efficacy of matching addresses that are afflicted by omissions and ambiguities.

In practical scenarios, rule-based methodologies face challenges akin to those encountered by dictionary-based approaches. Both methodologies are susceptible to complications arising from incomplete dictionaries or deficient address element repositories. Consequently, an established approach is to amalgamate these two strategies [21,22]. Nevertheless, diverse geographical regions exhibit significant disparities in address nomenclature and usage conventions, rendering the generalized application of dictionaries and regulations across various locales a formidable task.

2.1.3. Statistical-Based Address Parsing

The fundamental concept of statistical segmentation is primarily derived from the understanding of the Chinese language. It suggests that characters combine to form words, and the greater the frequency of adjacent characters appearing in the same order, the higher the likelihood of constituting a word [23]. The statistical-based approach to address parsing treats the address string as an observational sequence and address component type annotations as sequences of states. Through the training of an address-parsing model on annotated address datasets, this method automatically annotates untagged addresses, achieving the division and identification of address elements [24]. Presently, the prevailing statistical models for Chinese address segmentation are predominantly founded on traditional probability statistics, such as Hidden Markov Models (HMMs), Maximum Entropy Markov Models (MEMMs), and Conditional Random Fields (CRFs) [25,26].

Song [10] employed the HMM to annotate address components and incorporated constraints between the address components using the Viterbi algorithm, proposing a Chinese address-matching method grounded in natural language comprehension. Zhu et al. [27] synthesized the composition and features of Chinese addresses by segmenting address components and assembling them into an output sequence with feature annotations. They trained a CRF-based model for Chinese address parsing. Tang et al. [28] applied CRF to Chinese address parsing, establishing a fuzzy matching method for local address information and a standardization approach for placenames, thus bolstering parsing precision. Wei et al. [29] devised composite geographical name features based on geographical name element attributes, part-of-speech attributes, and syntactic attributes, utilizing a CRF model for recognition, which enhanced the accuracy and recall for recognizing Chinese geographical names. Yuan [30] amalgamated statistical techniques with rules, devising and implementing a Chinese address segmentation system grounded in both statistics and rules. This system employs rule algorithms to remove ambiguities and validate the deduced results, thereby elevating accuracy considerably.

This approach shares core principles with statistical-based Chinese word segmentation. Its central attribute includes its autonomy from dictionaries but reliance on corpora. Nevertheless, its efficacy is curbed by the configuration of attributes, often leading to overfitting when attributes are overly abundant. Furthermore, relying on probability conditions alone presents challenges in serving as the foundation for address parsing [31].

2.1.4. Deep Learning-Based Address Parsing

The sequence-labeling model is a typically used deep learning Chinese address-parsing method, where the address text is regarded as a sequence whose characters at each position are labeled using the address components to which they correspond.

Many Chinese address-parsing methods based on deep learning have been proposed. Certain research results have been obtained using these methods [32,33]. For example, Zhang et al. [34] proposed a Chinese address-parsing method using the RoBERTa-BiLSTM-CRF deep learning model to alleviate the heavy reliance on word segmentation dictionaries and the inability to effectively recognize address elements and their types when using existing address-matching methods. Parsed addresses were standardized, and their composition was analyzed to improve the address-matching results. Zhang [35] improved the accuracy of Chinese address parsing by constructing an address-parsing model based on BERT-BiLSTM-CRF. Liu et al. [36] proposed a Chinese address-parsing method that integrated neural networks and spatial relationships by targeting the word segmentation and randomness, diversity, and ambiguity of Chinese address elements. The address model that adopted spatial relationships could inherit the address elements of the hierarchical relationship model and accurately locate the address according to the spatial relationships among its elements. Cheng et al. [9] devised a labeling system for segmenting Chinese addresses. They utilized BiLSTM networks to capture contextual features and integrated CRF to ascertain the optimal labeling sequence. This approach led to an improved identification of placename entities.

Although Chinese address-parsing methods based on deep learning have achieved good results, most deep learning models often ignore the fusion of local and global features during feature extraction from address elements. This results in the loss of semantic feature information, thereby reducing the accuracy of address element recognition. Finally, the error is propagated in the next processing link of address parsing. Therefore, in this study, a method for fusing local and global features was introduced into the deep learning model to improve the accuracy of Chinese address parsing.

2.2. Sorting and Combination of Chinese Addresses Based on FSM

FSMs [37] are models for directed graphs used to study the computation processes of finite states. They consist of a finite state set, input set, and state transition rule set. The finite state set describes the system state, the input set represents the input information of the system, and the state transition rule set describes the transition conditions between states. The FSM is widely used in fields such as computer science, linguistics, logic, and mathematics. Owing to its good flexibility and scalability, the FSM can adapt to different types and forms of address elements and quickly process large amounts of text data because they only require a small amount of calculation. Therefore, FSMs are widely used for parsing Chinese address elements.

Currently, many Chinese address-sorting and combination methods based on the FSM have been implemented. Certain results have been obtained using these methods. For example, because elements in address descriptions that comply with regular patterns may appear disordered, incomplete, or inconsistent with cognitive habits, Gu [38] used a method based on pattern matching integrated with address element vocabulary to establish a rule set for address description patterns. Integrated with the bidirectional FSM, the recognized address information was sorted and combined according to the hierarchical elements; therefore, the obtained address was standardized. Luo et al. [39] proposed a new method based on an address hierarchy classification driven by FSM. First, this method removed redundant noise words, such as punctuation marks, notes, and localizers, from the original address. Subsequently, a general word segmentation software was used to preliminarily segment the words. Then, an algorithm based on feature word recognition using a classification model driven by an FSM for address hierarchy classification, recognition, labeling, and standardized coding was developed. This method could effectively solve the difficult problem of Chinese address standardization. Wang et al. [40] proposed the T-FA model, which uses natural language processing to address administrative regions based on the Trie model and the finite state automaton model to extract elements of non-standard addresses. This method performed better at processing addresses in batches. Tan [41] adopted a unique strategy that deviates from the conventional method of assigning individual characters as weights to arcs within an FSM; on the contrary, their method utilizes administrative unit names from addresses as the arc weights. This innovative approach not only diminished the space complexity of the algorithm's implementation but also enhanced its execution speed. As a result, it successfully facilitated the recognition of Chinese addresses.

Although the methods of sorting and combining Chinese addresses based on FSM have achieved certain results, they cannot process scenarios where the address elements are completely disordered or missing. In addition, traditional FSM relies heavily on the included vocabulary or requires a threshold setting, which reduces its generalizability for address sorting and combination. Therefore, this study proposes a dynamic FSM algorithm to improve element sorting and the combination of non-standard addresses in Chinese.

Considering the classification system of address elements and the characteristics of Chinese address descriptions, the method of combining address elements becomes more complex with the gradual refinement of the granular address descriptions, and the order of elements may be omitted or reversed. Therefore, after the address elements are recognized using the deep learning model, the dynamic FSM is further needed to sort and parse the address elements to address the disorder and missing hierarchical elements in the descrip-

tion of address information in the text. Consequently, addresses with correct hierarchical structures can be obtained for subsequent high-accuracy address matching. Considering the relevant research status, this study proposes a method to parse Chinese address text on the internet by integrating the CHTopoNER model and dynamic FSM to effectively solve the abovementioned problems.

3. Data and Methodology

3.1. Data

In this study, we used text from the Zhengzhou Epidemiological Survey Data (COVID-19 ESD) released on social media. A total of 77,451 pieces of data were available. The dataset was divided into training and test sets at a ratio of 4:1. To label these data by referring to the Chinese national rules for the geocode of addresses and the standard specification on industrial address classification and considering the universality, uniformity, and extensibility of the address, this study used BMES labeling [42] to categorize the Chinese address elements into the provincial, municipal, county (district), and town (township) administrative divisions, villages, roads, local regions, doorplates, building addresses, and unit numbers. The specific classifications are listed in Table 1.

Table 1. Labeling system of Chinese address elements.

Label	Meaning	Example
XZQHS	Provincial administrative divisions	河南省 (in English: Henan Province)
XZQHCS	Municipal administrative divisions	郑州市 (in English: Zhengzhou City)
XZQHQX	County (district) administrative divisions	中牟县 (in English: Zhongmu County)
XZQHZ	Town (township) administrative divisions	白沙镇 (in English: Baisha Town)
XZQHC	Village	高庄村 (in English: Gaozhuang Village)
JD1	Road 1	商都路 (in English: Shangdu Road)
JD2	Road 2	万山公里 (in English: Wansan Highway)
JXK	Intersection	交叉路 (in English: Intersection)
DIR	Direction	西北方向 (in English: Northwest)
DIS	Distance	50 m (in English: 50 m)
BES	Blur shift	附近 (in English: Nearby)
MP1	Local area 1	489号 (in English: No. 489)
MP2	Local area 2	博士嘉园 (in English: Boshi Jiayuan)
MP3	Local area 3	25幢 (in English: Building 25)
POI1	Point of interest 1	白沙商贸城 (in English: Baisha Trade City)
POI2	Point of interest 2	茶百道 (in English: Chabaidao)

Moreover, to validate the model's ability to generalize, this study introduced two distinct datasets, namely People's Daily Annotated Corpus (PFR) and Microsoft Research Asia (MSRA), in addition to the COVID-19 ESD dataset created for this research. These datasets differ in that the PFR and MSRA datasets consist of a sole entity type that is labeled as "LOC", referring to geographical placenames. The partitioning of training and testing sets for both the PFR and MSRA datasets adhered to a 4:1 ratio. The PFR dataset, extracted from the January 1998 edition of the People's Daily newspaper's annotated corpus (https://www.heywhale.com/mw/dataset/5ce7983cd10470002b334de3/content (accessed on 15 January 2023)), encompasses over six million bytes of text. It has found extensive use in international competitions focused on named entity recognition tasks for geographical places and personal names. Owing to its proven reliability and gradual adoption, it has evolved into the most widely utilized standard corpus in this domain. The MSRA corpus [43] (https://tianchi.aliyun.com/dataset/144307 (accessed on 15 January 2023)), originating from MSRA, comprises approximately 45,000 sentences, with more than 30,000 instances of geographical placenames. This corpus serves as a significant dataset extensively employed in named entity recognition tasks.

3.2. Methodology

The proposed method for parsing Chinese address text on the internet integrated the CHTopoNER model and dynamic FSM. First, the CHTopoNER model was used for named entity recognition (NER) of the multi-hierarchical address elements in texts from the internet. Subsequently, the hierarchical elements of each address were obtained. Furthermore, according to the hierarchical features of each address, sets of dynamic FSMs were created to parse Chinese address texts on the internet through state transitions. This method is shown in Figure 1.

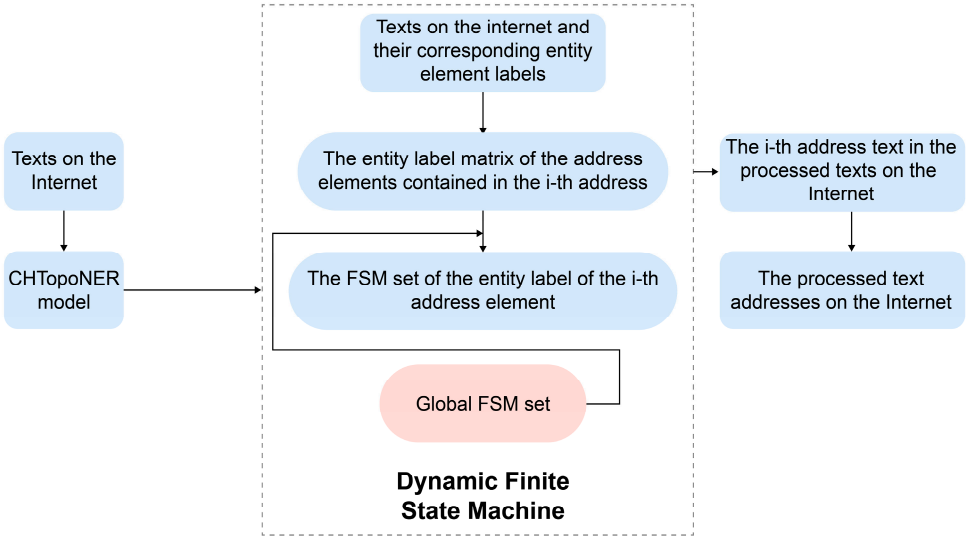

Figure 1. Method for parsing Chinese address text on the internet based on the integration of the CHTopoNER model and dynamic finite state machine (DFSM).

3.2.1. CHTopoNER Model

The CHTopoNER model includes a Chinese-roberta-wwm-ext layer [44], an improved SoftLexicon layer, and a TCNN-CRF layer composed of an IDCNN [45], a BiLSTM [46], and a CRF layer [47]. The model is shown in Figure 2.

First, the Chinese text data were input. The model first converted the input text into character vectors using the Chinese-roberta-wwm-ext layer. Subsequently, improved SoftLexicon was used to obtain word-level semantic information. Next, the character- and word-level vectors were concatenated to fuse the semantic information on both levels. In addition, the model extracted and concatenated the semantic features of the vectors through a TCNN layer to obtain the forward and backward contextual dependency relationships of each character and word in the Chinese text, thereby exploring their potential semantic associations. Finally, the output of the TCNN layer was mapped to a predefined category space through a fully connected layer. The CRF layer was ultimately used to model and decode the label sequences because of its ability to consider the dependencies among labels, thereby improving the accuracy of sequence labeling.

To avoid the out-of-vocabulary problem, possibly caused by SoftLexicon, the improved SoftLexicon improves the embedding weights of the corresponding words from the number of words matched in the training and verification sets into the calculated weights from the corresponding embedding obtained during pre-training using GLoVe [48].

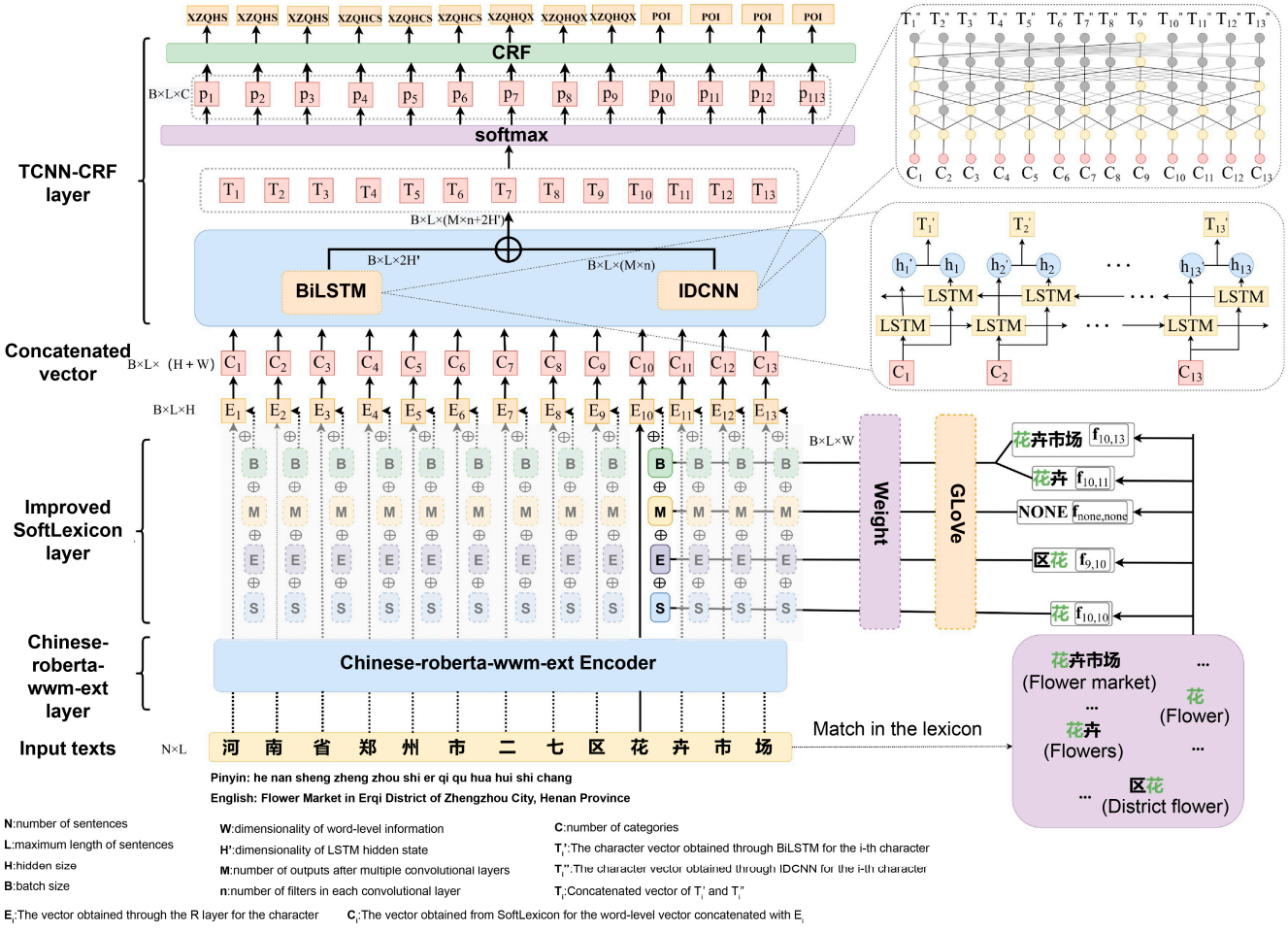

Figure 2. CHTopoNER model.

Compared with BiLSTM, which can only retain contextual information but cannot efficiently process long sequences, the IDCNN uses a separable convolutional approach to reduce the number of model parameters, thereby reducing the computational complexity and adapting it for processing long sequences. Therefore, the CHTopoNER model integrates BiLSTM and IDCNN to form a TCNN network layer. The respective advantages of BiLSTM and IDCNN were fully exploited while avoiding their disadvantages. In the TCNN model, BiLSTM is primarily used to capture the contextual information of the sequence, whereas IDCNN is used to extract its local features. This integration method can be used for the accurate recognition of Chinese address elements.

3.2.2. Dynamic FSM

The dynamic FSM adds a global set to the quintuple FSM and attributes the following elements to the original quintuple: the *i*th FSM set, the *i*th entity label set of the hierarchical elements of the input address, the *i*th state transition function, and the *i*th initial state set and final state set. Notably, the generation process of the dynamic FSM is to generate a corresponding set of FSMs based on the elemental level features of each address, which achieves the dynamization of FSM. Unlike the quintuple FSM, the dynamic FSM is a sextuple. The *i*th dynamic FSM M_i is shown in Equation (1).

$$M = (W, Q_i, \Sigma_i, \delta_i, q_{i0}, F_i) \quad (1)$$

where W is an already defined global FSM set, Q_i is the ith state set, $\sum_i$ is the entity label set of the input address-level elements, δ_i is the state transition function, and F_i and q_i are the ith initial and final state sets, respectively.

Dynamic FSM is an algorithm that dynamically generates a model of a directed graph. It can ameliorate the excessive reliance of FSM on keywords in the address-level elements and avoid the uncertain threshold setting of the bidirectional FSM. Where the transition relationship between each state s and the next state s' is added as a key–value pair into the state transition function δ_i, with s being the key and s' being the value. Here, the set is used to store values because state s may have multiple successor states. The state transition function is constructed using range(len(sorted_E2)-1) in the loop to traverse every position of state s. Each time, the two adjacent positions are taken out as one state and the next state. Finally, the state transition function δ_i is added to the sextuple of the dynamic FSM $E3$, which is returned. Algorithm 1 is as follows:

Algorithm 1: Dynamic FSM algorithm.

```
Input: entity labels E1 of address elements, global FSM set W
Output: dynamic FSM E3

function generateDynamicFSM(E1, W):
    // initialize the sextuple of dynamic FSM E3
    Q_i = {}
    Σ_i = set of all entity labels in E1
    δ_i = {} // state transition function, initialized as an empty set
    q_i = {} // initial state set, initialized as an empty set
    F_i = {} // final state set, initialized as an empty set
    E3 = (W, Q_i, Σ_i, δ_i, q_i, F_i)

    // on each path P in W, do
    for each P_i in W:
        // obtain the entity labels E2 of the address elements of the path P
        E2 = entity labels in path P_i
        // determine whether every element in E1 is contained in E2
        if all entity labels in E1 are in E2:
            // select this path P for the sorting and organizing the entity labels
            // traverse E1 in E2 and record the positions of the entity elements of E1 with
respect to E2
            indices = []
            for each e in E1:
                index = index of e in E2
                append index to indices
            // sort the entity elements of E1 according to the index in indices
            sorted_E1 = [E1[i] for i in sorted(indices)]
            // sort the entity elements of E2 according to the index in indices
            sorted_E2 = [E2[i] for i in sorted(indices)]
            // add the sorted entity elements to the state set Q_i as new states
            q_i = (sorted_E2, P)
            add q_i to Q_i

            // update the state transition function δ
            for i in range(len(sorted_E2)−1):
                s = (sorted_E2[i], sorted_E2[i + 1])
                s_next = (sorted_E2[i + 1],)
                if s not in δ:
                    δ_i [s] = set()
                δ_i [s].add(s_next)

    return E3
```

In this study, the Jin Rong Yue Hui Cheng Convenience Store (the northeast corner of the intersection between Yongzhou Road and Xunhang Road) was taken as an example to visualize the generation of dynamic FSM, as shown in Figure 3.

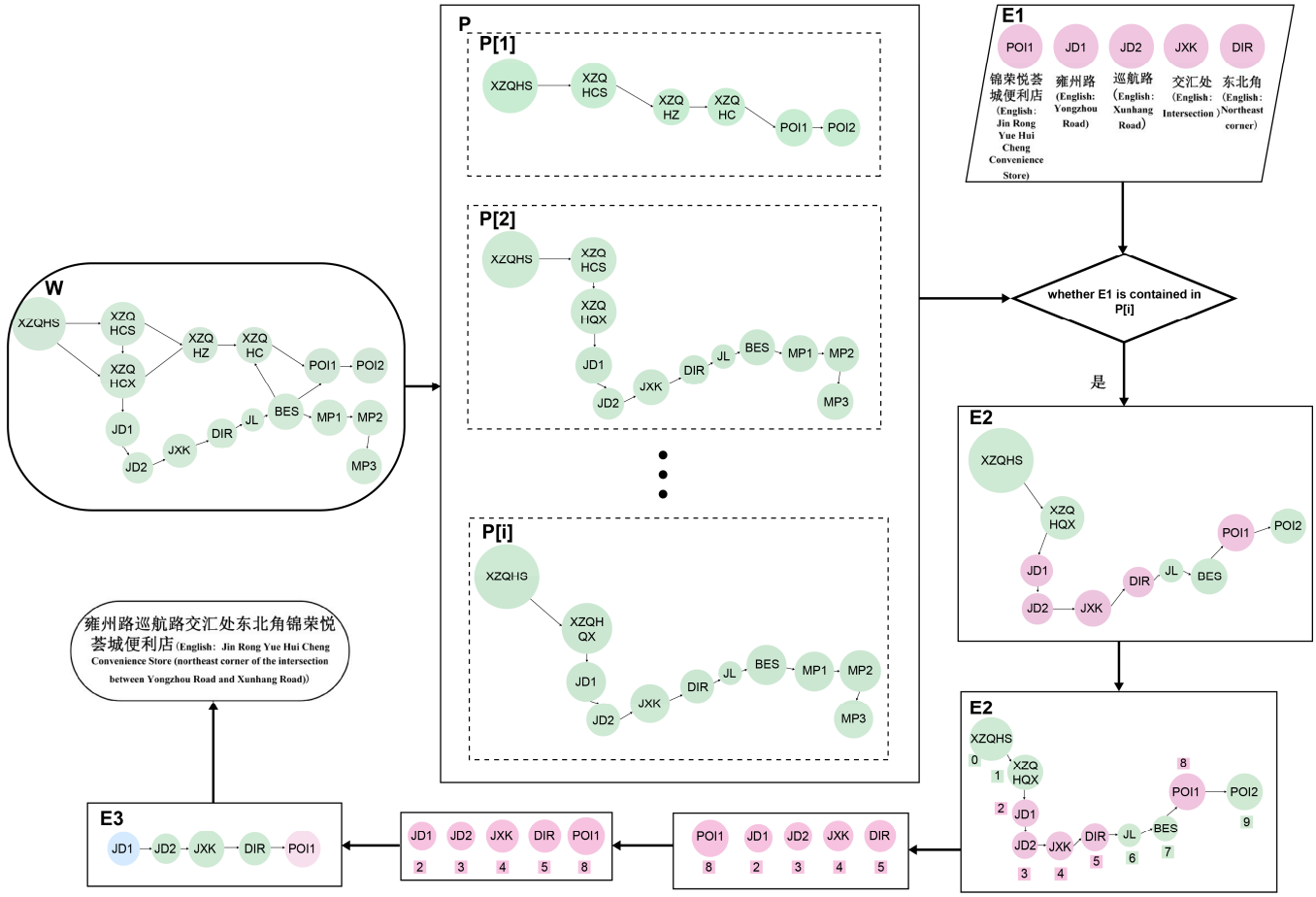

Figure 3. Visualization of the dynamic FSM for address instance acquisition.

There are missing address elements in the Jin Rong Yue Hui Cheng Convenience Store (northeast corner of the intersection between Yongzhou Road and Xunhang Road), and the structure is disordered. As shown in Figure 3, the entity labels of the address elements (E1 for short) were obtained first, including POI1, JD1, JD2, JXK, and DIR. Subsequently, the entity labels of the obtained address elements were organized by considering the global FSM set (W). The specific steps are as follows:

(1) Traverse path P in W to obtain the entity label of the address element (P[i]) of each path.
(2) Determine whether each element in E1 is contained in P[i]. If yes, the path was selected as E2.
(3) Traverse from E1 to E2. If the label element in E1 was consistent with that in E2, the index of the element was recorded.
(4) Finally, the indices were sorted from small to large to obtain a new dynamic FSM (E3; blue represents the initial state of the dynamic FSM, and pink represents the final state).

Finally, the text of the address information was obtained based on the text index corresponding to each entity label of the address element.

When implementing the dynamic FSM, there is no need to include keywords or set thresholds. However, by setting all entity label types of address elements as a global FSM set, the corresponding dynamic FSM set could be obtained based on the entity label of the *i*th address element in the text from the internet. This method is suitable for parsing

standard and non-standard addresses, which cannot be accomplished using FSM and bidirectional FSM.

Owing to the limited types of entity labels of address elements and good compatibility with uncommon keywords of address elements or ambiguously oriented keywords, the dynamic FSM abandons the collection of keywords and adopts the collection of entity labels of address elements, forming the global FSM set. Consequently, the dynamic FSM could largely solve the problem of an incomplete collection of keywords.

In addition, because of the uncertainty of the threshold setting of the bidirectional FSM, the dynamic FSM designed in this study discarded the threshold setting and sorted all types of entity labels in an address according to the global FSM set, even when the address was missing or disordered. Therefore, an FSM set corresponding to this address was obtained.

4. Evaluation Metrics and Experimental Results

In this study, we compared the CHTopoNER model with current advanced deep learning models using specified evaluation metrics to explore its advantages in addressing element recognition tasks in Chinese texts from the internet. We integrated the CHTopoNER model with FSM, bidirectional FSM, and dynamic FSM to parse addresses to verify the effectiveness of dynamic FSM parsing texts from the internet.

4.1. Evaluation Metrics

This study employs the evaluation metrics proposed during the MUC assessment conference for named entity recognition (NER). Specifically, the initial evaluation metrics for NER—namely F1 score, precision (P), and recall (R)—introduced by MUC-2 [49], are utilized to assess the efficacy of the model in extracting Chinese geographical placename entities.

4.2. Experimental Setup and Parameters

The parameter settings are as follows: batch size is set to 32, number of training epochs is 100, optimizer is Adam, loss function is CRF Loss, base learning rate is 5×10^{-6}, dropout rate is 0.5, and the maximum length of input text is 50. To prevent overfitting during the training process, enhance training efficiency, and avoid excessive training, we implemented the EarlyStop mechanism with an early_stop_ratio set to 20.

The software and hardware facilities mainly used in this study are listed in Table 2.

Table 2. Software and hardware used in our investigation.

Component	Details
Central Processing Unit (CPU)	Intel(R) Core(TM) i9-12900H
Graphics Card (GPU)	NVIDIA GeForce RTX 3080 Ti
Operating System	Ubuntu 18.04
Programming Language	Python 3.7
Deep Learning Framework	TensorFlow1.14.0

4.3. Experimental Results and Analysis

4.3.1. Experimental Results

To verify the effectiveness of the method proposed in this study, the BiLSTM-CRF [50], IDCNN-CRF [51], BiLSTM-attention-CRF [52], BERT-BiLSTM-CRF [53], and CHTopoNER models were applied to the chosen dataset. Furthermore, FSM, bidirectional FSM, and dynamic FSM were added to the experiments based on the CHTopoNER model.

Experimental results of geographical placename recognition using different models on the PFR dataset are presented in Table 3.

Table 3. Results of the PFR dataset across different models.

	Precision	Recall	F1
BiLSTM-CRF (baseline)	0.912	0.774	0.837
IDCNN-CRF1	0.992	0.773	0.869
IDCNN-CRF2	0.994	0.790	0.883
BiLSTM-Attention-CRF	0.938	0.740	0.827
BERT-BiLSTM-CRF	0.996	0.891	0.940
CHTopoNER	**0.997**	**0.953**	**0.975**

Results of the geographical placename recognition experiments on the MSRA dataset for different models are shown in Table 4.

Table 4. Results of the MSRA dataset across different models.

	Precision	Recall	F1
BiLSTM-CRF (baseline)	0.990	0.890	0.864
IDCNN-CRF1	0.994	0.889	0.864
IDCNN-CRF2	0.995	0.916	0.875
BiLSTM-Attention-CRF	0.989	0.840	0.840
BERT-BiLSTM-CRF	0.989	0.895	0.940
CHTopoNER	**0.999**	**0.965**	**0.981**

Table 5 lists the experimental results of address element recognition by different models on the text data of the Zhengzhou COVID ESD (2020–2022) released on social media.

Table 5. Address element recognition by different models in the COVID-19 ESD social media dataset.

	Precision	Recall	F1
BiLSTM-CRF (baseline)	0.961	0.710	0.817
IDCNN-CRF1	0.976	0.736	0.839
IDCNN-CRF2	0.988	0.729	0.839
BiLSTM-Attention-CRF	0.936	0.703	0.809
BERT-BiLSTM-CRF	0.989	0.735	0.843
CHTopoNER	**0.991**	**0.743**	**0.849**

Experimental results of different types of geographical placename entities and location orientation entities extracted by the CHTopoNER model on the COVID-19 ESD social media dataset are presented in Figure 4.

The experimental results after adding the FSM, bidirectional FSM, and dynamic FSM to the CHTopoNER model are listed in Tables 6–10.

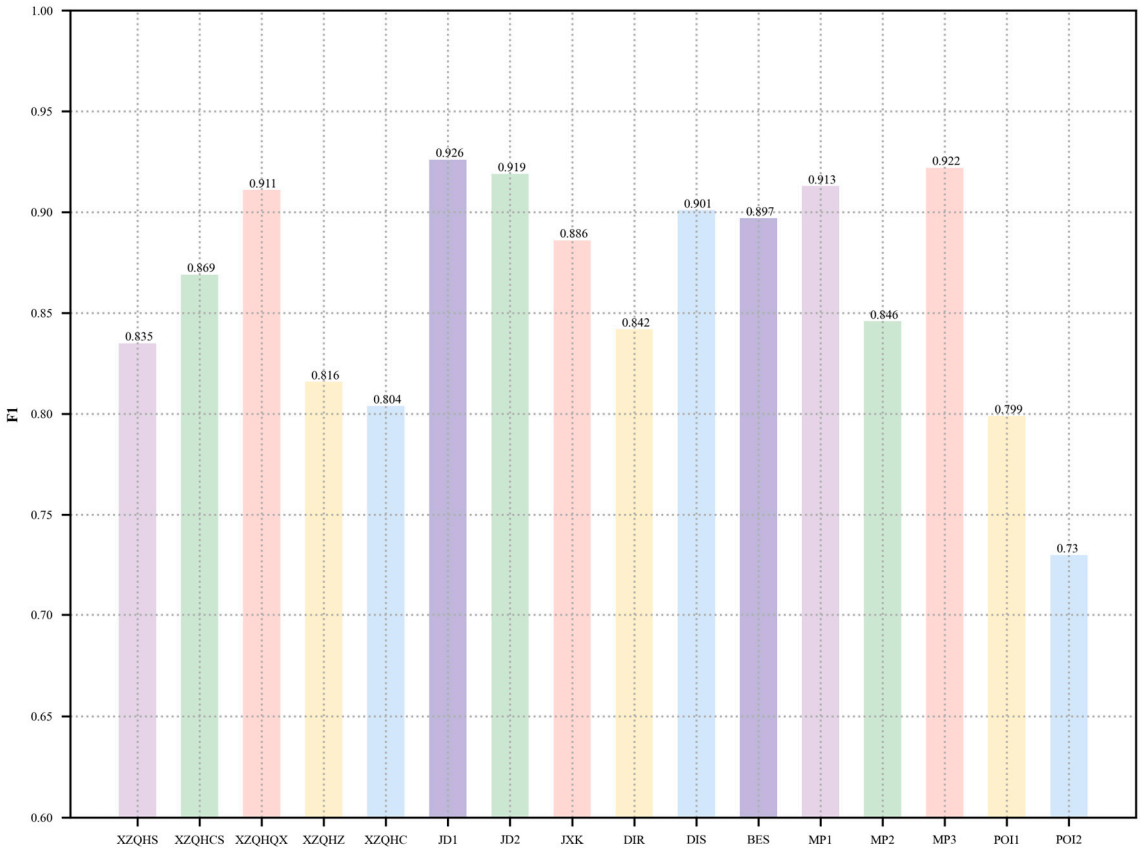

Figure 4. Experimental results of different types of address element entities extracted by the CHTopoNER model on the COVID-19 ESD social media dataset.

Table 6. Accuracies of the CHTopoNER models with added FSM, bidirectional FSM, and Algorithm 1 (dynamic FSM) to standard address data.

Model	Accuracy
CHTopoNER + FSM	0.777
CHTopoNER + bidirectional FSM	0.781
CHTopoNER + Algorithm 1 (dynamic FSM)	**0.839**

Table 7. Accuracies of the CHTopoNER models with added FSM, bidirectional FSM, and Algorithm 1 (dynamic FSM) to non-standard address data.

Model	Accuracy
CHTopoNER + FSM	0.532
CHTopoNER + bidirectional FSM	0.698
CHTopoNER + Algorithm 1 (dynamic FSM)	**0.836**

Table 8. Specific experimental results of the CHTopoNER model with added FSM, bidirectional FSM, and Algorithm 1 (dynamic FSM) to standard address data.

Model	Example	Processing of Address Elements
CHTopoNER + FSM	Xinyuan Modern Cheng (Cheng has the same meaning as city but uses a different Chinese character. In this study, Cheng is used to differentiate from city), No. 17 Qingfeng Street, Erqi District, Zhengzhou City	Zhengzhou City (XZQHCS)/Erqi District (XZQHQX)/Qingfeng Street (JD1)/No.17 (MP1)/Xinyuan Modern Cheng
CHTopoNER + bidirectional FSM	Xinyuan Modern Cheng, No. 17 Qingfeng Street, Erqi District, Zhengzhou City	Zhengzhou City (XZQHCS)/Erqi District (XZQHQX)/Qingfeng Street (JD1)/No.17 (MP1)/Xinyuan Modern Cheng (MP2)
CHTopoNER + Algorithm 1 (dynamic FSM)	**Xinyuan Modern Cheng, No. 17 Qingfeng Street, Erqi District, Zhengzhou City**	**Zhengzhou City (XZQHCS)/Erqi District (XZQHQX)/Qingfeng Street (JD1)/No.17 (MP1)/Xinyuan Modern Cheng (MP2)**

Table 9. Specific experimental results of the CHTopoNER model with added FSM, bidirectional FSM, and Algorithm 1 (dynamic FSM) to non-standard address data.

Model	Example	Processing of Address Elements
CHTopoNER + FSM	No. 132 Wangwu Road, Zheng Shang Ming Zuan	Invalid address
CHTopoNER + bidirectional FSM	No. 132 Wangwu Road, Zheng Shang Ming Zuan	Wangwu Road, Zheng Shang Ming Zuan (JD1)/No. 132 (MP1)
CHTopoNER + Algorithm 1 (dynamic FSM)	**No. 132 Wangwu Road, Zheng Shang Ming Zuan**	**Wangwu Road (JD1)/No. 132 (MP1)/Zheng Shang Ming Zuan (MP2)**

Table 10. Specific experimental results of the CHTopoNER model with the addition of FSM, bidirectional FSM, and Algorithm 1 (dynamic FSM), respectively, on non-standard address data.

Model	Example	Processing of Address Elements
CHTopoNER + FSM	Jin Rong Yue Hui Cheng Convenience Store (northeast corner of the intersection of Yongzhou Road and Xunhang Road)	Invalid address
CHTopoNER + bidirectional FSM	Jin Rong Yue Hui Cheng Convenience Store (northeast corner of the intersection of Yongzhou Road and Xunhang Road)	Invalid address
CHTopoNER + Algorithm 1 (dynamic FSM)	**Jin Rong Yue Hui Cheng Convenience Store (northeast corner of the intersection of Yongzhou Road and Xunhang Road)**	**Yongzhou Road (JD1)/Xunhang Road (JD2)/Intersection (JXK)/Northeast Corner (DIR)/Jin Rong Yue Hui Cheng Convenience Store (POI1)**

4.3.2. Analysis of Experimental Results

Table 3 reveals that on the PFR dataset, our proposed CHTopoNER model performs the best across all three evaluation metrics: Precision (P), Recall (R), and F1 score (F1). Compared to the baseline model BiLSTM-CRF, the CHTopoNER model exhibits improvements of 8.5% in P, 17.9% in R, and 12.4% in F1 values. Compared to the state-of-the-art BERT-BiLSTM-CRF model, the CHTopoNER model achieves improvements of 0.1% in P, 6.2% in R, and 2.1% in F1 values. Similarly, from Table 4, on the MSRA dataset, our proposed CHTopoNER model demonstrates superior performance in all the three evaluation

metrics (P, R, and F1). In comparison with the baseline model BiLSTM-CRF, the CHTopoNER model exhibits improvements of 0.9% in P, 7.5% in R, and 10.2% in F1 values. Compared to the state-of-the-art BERT-BiLSTM-CRF model, the CHTopoNER model achieves improvements of 1.0% in P, 7.0% in R, and 2.6% in F1 values. The CHTopoNER model proposed here exhibited the best evaluation metrics in the COVID-19 ESD dataset: P, R, and F1 (Table 2). Compared with the baseline model BiLSTM-CRF, P, R, and F1 improved by 3.1%, 4.6%, and 3.9%, respectively. Compared with the state-of-the-art BERT-BiLSTM-CRF model, the three evaluation metrics improved by 0.2%, 1.1%, and 0.7%, respectively.

To further compare and analyze the experimental results of our CHTopoNER model with other models on the PFR, MSRA, and COVID-19 ESD datasets, we present the experimental outcomes graphically in Figure 5.

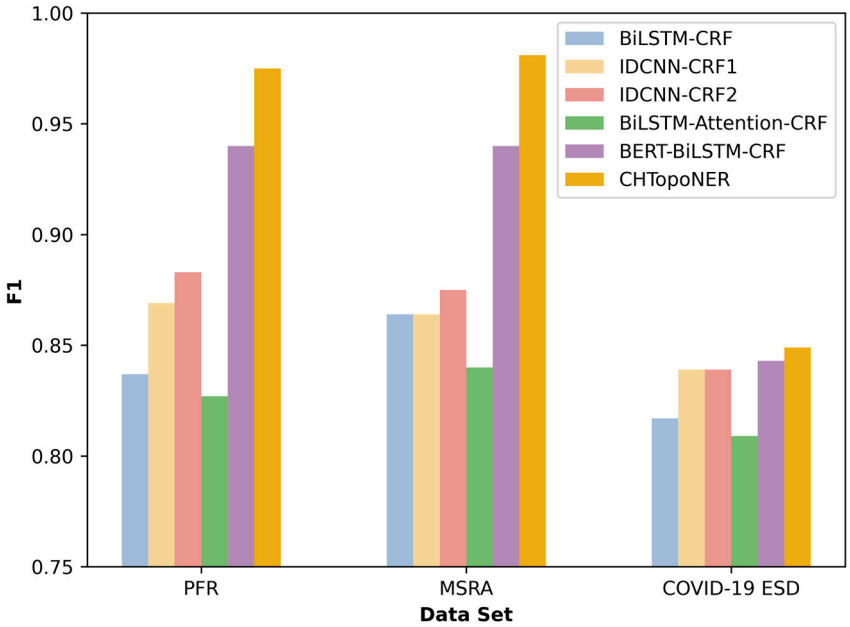

Figure 5. Experimental results of the six models on the three different datasets.

Figure 5 shows that the proposed CHTopoNER model, compared to the other models, demonstrates superior experimental performance across the three distinct datasets. This can be attributed to the utilization of the Chinese-roberta-wwm-ext pretraining with a Chinese full-word masking strategy and the improved SoftLexicon approach to capture character-level and word-level information from input texts. Furthermore, the integration of the TCNN layer in semantic feature extraction enables the consideration of both local and global semantic information, thus minimizing the loss of semantic features. However, the performance of the CHTopoNER model on the COVID-19 ESD dataset is slightly inferior to that on the PFR and MSRA datasets. This discrepancy arises from the fact that while the PFR and MSRA datasets focus on recognizing a single entity type, namely geographical places, the COVID-19 ESD dataset involves the recognition of 16 different types of address element entities. Typically, the presence of multiple entity types may increase data sparsity and context ambiguity, rendering it challenging for the model to capture subtle features of each entity type and accurately determine the correct entity category. Consequently, the precision of the CHTopoNER model's experimental results on the COVID-19 ESD dataset is marginally lower.

The F1 values of different types of address element entities are shown in Figure 4. Among them, the JD1 type exhibits the highest F1 value, while the POI2 type has the lowest F1 value. This discrepancy can be attributed to the higher frequency and distinct features of JD1 entities in the text, such as terms like "XX Road" or "XX Avenue". Conversely, the lower F1 value for the POI2 type stems from the less prominent textual features associated with this type and its infrequent appearance in the text, with examples like

"Cha Bai Dao" or "Wen Xuan Art Studio". As the dataset is based on epidemic-related text data from Zhengzhou city, administrative regions, such as provinces, cities, towns, and villages, have relatively low occurrences. Consequently, the recognition performance of the corresponding entity labels "XZQHS", "XZQHCS", "XZQHZ", and "XZQHX" is comparatively lower.

The parsing method for non-standard address text in Chinese exhibited an accuracy of 83.9% for standard placenames and 83.6% for non-standard placenames by integrating the CHTopoNER model with dynamic FSM (Tables 6 and 7). Compared with the approach integrating the CHTopoNER model with traditional FSM, the improvements were 8.0% and 57.1%, respectively; compared with the approach integrating the CHTopoNER model with bidirectional FSM, the improvements were 7.4% and 19.8%, respectively. Notably, the CHTopoNER model could only recognize entity labels starting with B or I but not those starting with E. Therefore, it was impossible to obtain a complete entity label starting with B and ending with E, so the dynamic FSM failed to correctly sort and combine all address elements recognized by the CHTopoNER model.

Dynamic FSM was better than FSM and bidirectional FSM in parsing and sorting both standard and non-standard placenames (Tables 8–10). This is because FSM relies on the included keywords (such as province, city, county (district), and road), whereas text from the internet often includes incomplete descriptions of keywords and other information due to its non-standardized description.

For standard addresses, the corresponding address element cannot be recognized if a keyword is not included in the FSM. For example, for Xinyuan Modern Cheng, No. 17 Qingfeng Street, Erqi District, Zhengzhou City, the keyword city was included in the state set of the FSM (Figure 6A). Therefore, the address elements of Zhengzhou were obtained. Subsequently, the next state was triggered. If the next state is District (County), Street can be triggered next, and so on. However, the keyword Cheng was not included during this process. Therefore, the FSM could only obtain the address elements on three levels: Zhengzhou City/Erqi District/Qingfeng Street. When processing standard addresses, the bidirectional FSM set (Figure 6B) was similar to the FSM set. For the dynamic FSM set, the entity label types of the address elements were obtained first. Subsequently, a new FSM set was obtained according to the initially defined general state set (Figure 6C). Finally, the text corresponding to the address element was parsed and sorted using the new FSM set.

For non-standard addresses, such as No. 132 Wangwu Road, Zheng Shang Ming Zuan, with missing address element levels and a disordered structure, the initial state of the FSM state set (Figure 7A) was City. However, because City is not present in No. 132 Wangwu Road, Zheng Shang Ming Zuan, the next state could be triggered. Therefore, when the FSM processed No. 132 Wangwu Road, Zheng Shang Ming Zuan, it was regarded as an invalid address. In contrast, when the bidirectional FSM (Figure 7B) processed this address, the forward trigger could not happen due to the lack of the initial trigger keywords in this address. However, reverse triggering was enacted as the number was in the end state of the bidirectional FSM; therefore, No. 132 was obtained. Subsequently, the next state—Road—was triggered, thereby obtaining Wangwu Road, Zheng Shang Ming Zuan (JD1)/No. 132 (MP1). The working process of the dynamic finite state set was as follows. First, the entity label types of the address elements in "No. 132 Wangwu Road, Zheng Shang Ming Zuan" were obtained, including MP2, JD1, and MP1. Next, according to the initially defined general state set, a new finite state set was obtained (Figure 7C). Finally, according to the entity label indices of the address elements, the corresponding address text of the new finite state set Wangwu Road (JD1)/No. 132 (MP1)/Zheng Shang Ming Zuan (MP2) was obtained.

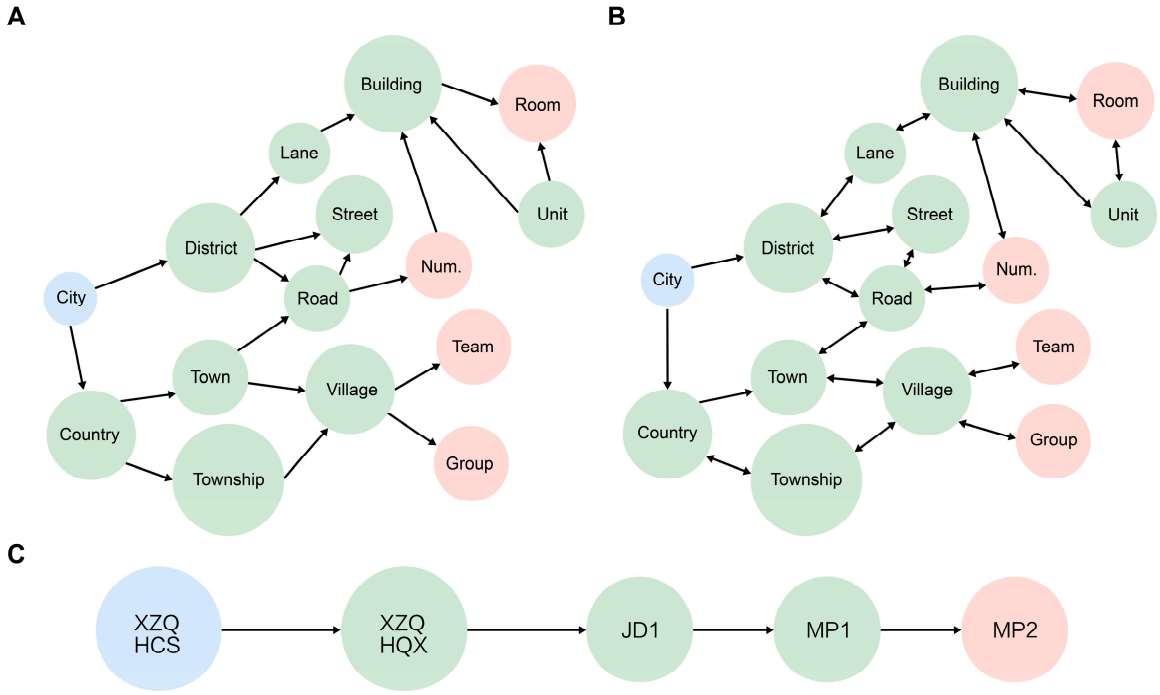

Figure 6. Three types of FSMs corresponding to addresses in Table 8. (**A**): The included keywords, initial state (blue), intermediate state (green) and end state (pink) of the FSM. (**B**): The included keywords, initial state (blue), intermediate state (green) and end state (pink) of the bidirectional FSM. (**C**): The state machine set, initial state (blue), intermediate state (green) and end state (pink) of dynamic FSM.

When parsing non-standard addresses or when the hierarchy of address elements is missing, the structure of address elements is disordered, and the address element contains ambiguous locator words, such as distance. For example, in "Jin Rong Yue Hui Cheng Convenience Store (the northeast corner of the intersection between Yongzhou Road and Xunhang Road)", the initial state of the FSM state set (Figure 8A) is "City". However, because "City" is not present in "Jin Rong Yue Hui Cheng Convenience Store (the northeast corner of the intersection between Yongzhou Road and Xunhang Road)", the next state could not be triggered. As a result, when the FSM processed "Jin Rong Yue Hui Cheng Convenience Store (the northeast corner of the intersection between Yongzhou Road and Xunhang Road)", it was regarded as an invalid address. However, when the bidirectional FSM (Figure 8B) processed the address "Jin Rong Yue Hui Cheng Convenience Store (the northeast corner of the intersection between Yongzhou Road and Xunhang Road)", state changes were not triggered in either direction owing to the lack of bidirectional initial trigger keywords. As a result, "Jin Rong Yue Hui Cheng Convenience Store (the northeast corner of the intersection between Yongzhou Road and Xunhang Road)" was regarded as an invalid address. The working process of the dynamic finite state set is as follows. First, the types of entity labels of address elements from "Jin Rong Yue Hui Cheng Convenience Store (the northeast corner of the intersection between Yongzhou Road and Xunhang Road)" were obtained, including "POI1", "JD1", " JD2", "JXK", and "DIR". Subsequently, a new finite state set was obtained according to the global FSM set (Figure 8C). Finally, according to the indices of entity labels of address elements, the corresponding address text of the new finite state set was obtained as follows: "Yongzhou Road (JD1)/Xunhang Road (JD2)/Intersection (JXK)/Northeast Corner (DIR)/Jin Rong Yue Hui Cheng Convenience Store (POI1)".

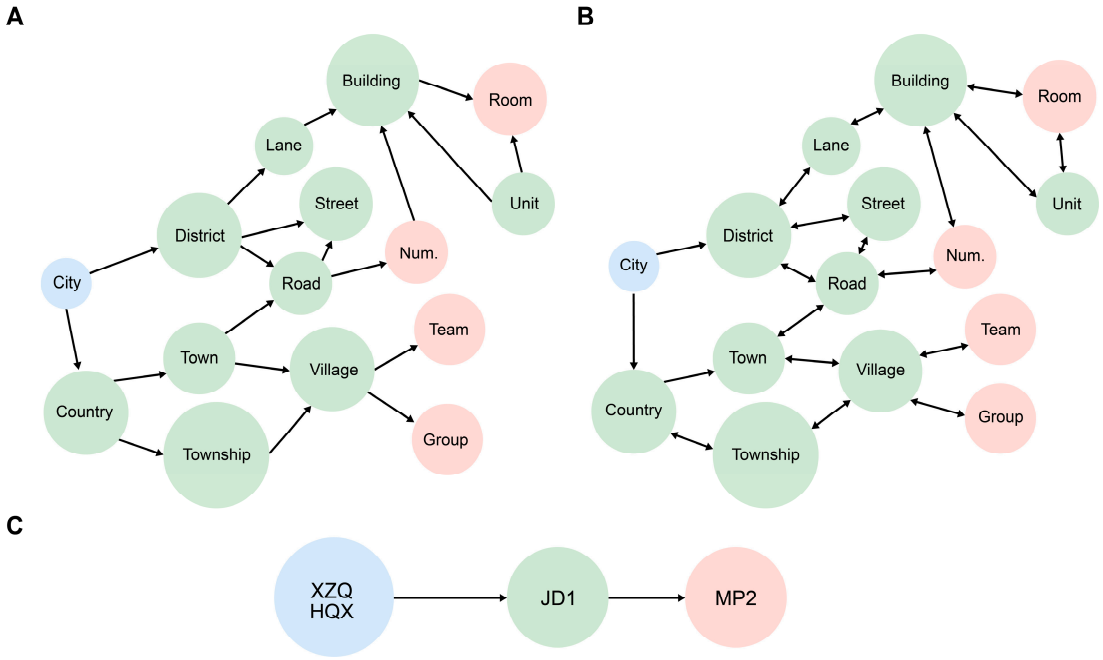

Figure 7. Three types of FSMs corresponding to addresses in Table 9. (**A**): The included keywords, initial state (blue), intermediate state (green) and end state (pink) of the FSM. (**B**): The included keywords, initial state (blue), intermediate state (green) and end state (pink) of the bidirectional FSM. (**C**): The state machine set, initial state (blue), intermediate state (green) and end state (pink) of dynamic FSM.

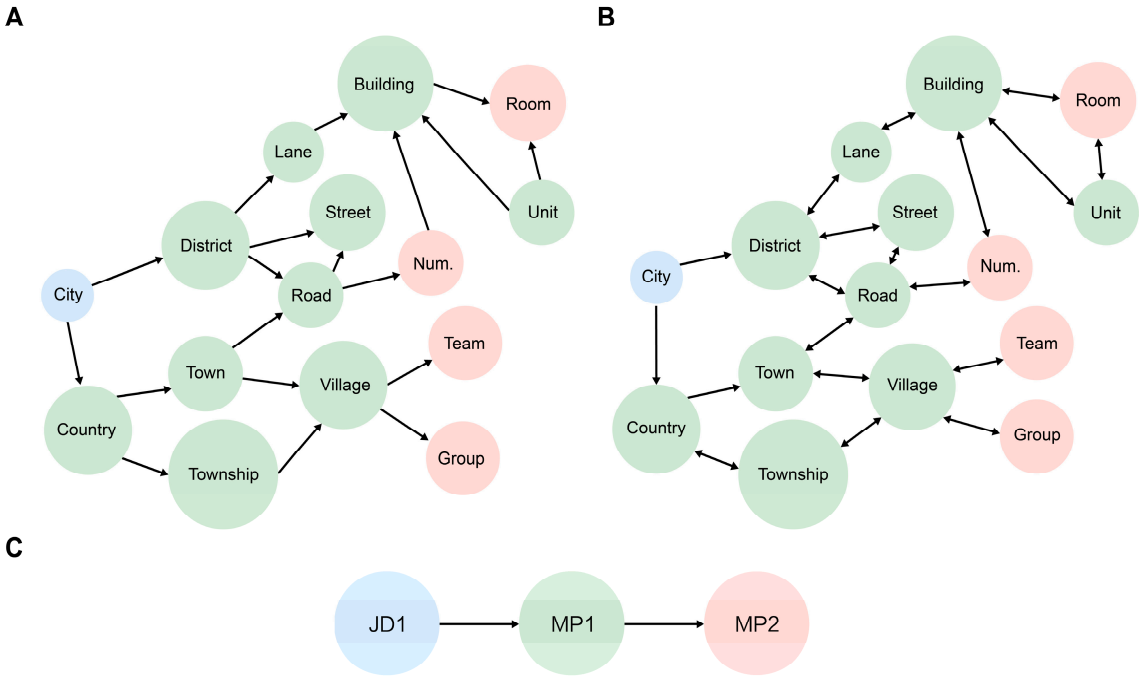

Figure 8. Three types of FSMs corresponding to addresses in Table 10. (**A**): The included keywords, initial state (blue), intermediate state (green) and end state (pink) of the FSM. (**B**): The included keywords, initial state (blue), intermediate state (green) and end state (pink) of the bidirectional FSM. (**C**): The state machine set, initial state (blue), intermediate state (green) and end state (pink) of dynamic FSM.

After the analysis, we concluded that the parsing results obtained using the three types of FSM were not significantly different for standard addresses. However, when the

hierarchical elements were disordered or missing in the address information in text from the internet, the CHTopoNER model and dynamic FSM showed improved address parsing.

5. Conclusions

This study integrated the CHTopoNER model and dynamic FSM to parse non-standard text addresses in Chinese and achieved good experimental results. The accuracy rate was 96.6% for standard placenames and 96.8% for non-standard placenames, which increased by 82.6% and 38.7%, respectively, compared to the integrated traditional and bidirectional FSMs. Integration with FSM supplemented the deficiencies of the CHTopoNER model, thereby improving the accuracy and robustness of address element parsing.

In the future, more natural language processing techniques integrated with the FSM can be explored to improve the accuracy and efficiency of Chinese address parsing. For example, deep learning methods can be integrated with FSM to improve their ability to recognize and sort address elements. In addition, knowledge graphs [54] can be applied to address parsing to improve semantic understanding and address combinations. Moreover, integrating multimodal information into address parsing could be considered. For example, image recognition techniques [55] can be integrated to extract address information from pictures, which can be considered together with text information, thereby improving the accuracy and robustness of address parsing. The application and promotion of these new ideas and methods are expected in Chinese address parsing.

Author Contributions: Conceptualization, Z.Z.; software, M.Z.; validation, M.Z.; formal analysis, Y.Q.; resources, Z.J.; data curation, Z.J.; writing—original draft preparation, M.Z.; writing—review and editing, M.Z., J.M., Z.Z. and Y.Q.; project administration, X.L.; funding acquisition, J.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Joint Fund of Collaborative Innovation Center of Geo-Information Technology for Smart Central Plains, Henan Province and Key Laboratory of Spatiotemporal Perception and Intelligent processing, Ministry of Natural Resources, grant number 212102; and National Natural Science Foundation of China, grant number 42101454.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data will be made available on request.

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

References

1. Tian, Q.; Ren, F.; Hu, T.; Liu, J.; Li, R.; Du, Q. Using an optimized Chinese address matching method to develop a geocoding service: A case study of Shenzhen, China. *ISPRS Int. J. Geo Inf.* **2021**, *5*, 65. [CrossRef]
2. Kang, M.; Du, Q.; Wang, M. The Chinese address extraction method based on the address tree model. *J. Surv. Mapp.* **2015**, *44*, 99–107.
3. Melo, F.; Martins, B. Automated geocoding of textual documents: A survey of current approaches. *Trans. GIS* **2017**, *21*, 3–38. [CrossRef]
4. Lin, Y.; Kang, M.; He, B. Spatial pattern analysis of address quality: A study on the impact of rapid urban expansion in China. *Environ. Plan. B Urb. Anal. City Sci.* **2021**, *48*, 724–740. [CrossRef]
5. Qiu, Q.; Xie, Z.; Wu, L.; Li, W. Geoscience keyphrase extraction algorithm using enhanced word embedding. *Expert Syst. Appl.* **2019**, *125*, 157–169. [CrossRef]
6. Wu, K.; Zhang, X.; Ye, P.; Huai, A.; Zhang, H. The Chinese address parsing method based on BERT-BiLSTM-CRF. *Geo Geogr. Inf. Sci.* **2021**, *37*, 10–15.
7. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pretraining of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
8. Ma, M. *Study on Key Techniques of Data Organization for Spatiotemporal Information of Internet News*; National University of Defense Technology: Changsha, China, 2016.
9. Cheng, B.; Li, W.; Tong, H. Chinese hierarchical address segmentation based on BiLSTM-CRF. *J. Geo-Inf. Sci.* **2019**, *21*, 1143–1151.
10. Song, Z. Chinese address matching algorithm for natural language understanding. *J. Remote Sens.* **2013**, *17*, 788–801.

11. Hu, X.; Hu, Y.; Resch, B.; Kersten, J. Geographic Information Extraction from Texts (GeoExT). In Proceedings of the European Conference on Information Retrieval, Dublin, Ireland, 2–6 April 2023; Springer Nature: Cham, Switzerland, 2023; pp. 398–404.
12. Zhu, S.M. Research and Implementation of Chinese Word Segmentation Algorithms. Master's Thesis, University of Electronic Science and Technology of China, Chengdu, China, 2011.
13. Guo, Y.C. Research on Technology for Chinese Address Services. Master's Thesis, Wuhan University, Wuhan, China, 2017.
14. Ye, P.; Zhang, X.Y.; Du, M. Query Method of Chinese Gazetteer Based on the Character Features. *J. Geo-Inf. Sci.* **2018**, *20*, 880–886.
15. Li, J.; Zhu, G.X.; Zhou, L.; Zheng, X.C. Address Segmentation Algorithm Based on Forward Adaptive Length Matching by Mark Words and Supplementary Method of Missing Address Elements. *China Med. Devices* **2019**, *34*, 112–114+130.
16. Li, P.P. Research on Self-Learning Construction Method of Chinese Address Element Library Based on Internet POI. Master's Thesis, Lanzhou Jiaotong University, Lanzhou, China, 2019.
17. Zhu, J. Key Techniques for Chinese Standard Address Database Construction. Master's Thesis, Nanjing Normal University, Nanjing, China, 2013.
18. Zhuang, H.D.; Zhang, H.E. Rule-based Chinese Address Matching System. *J. Fujian Comput.* **2013**, *29*, 130–132+146.
19. Zhang, X.Y.; Lv, G.N.; Li, B.Q.; Chen, W. Rule-based Approach to Semantic Resolution of Chinese Addresses. *J. Geo-Inf. Sci.* **2010**, *12*, 9–16. [CrossRef]
20. Tan, K.K. Rule-Based Chinese Address Segmentation and Matching Methods. Master's Dissertation, Shandong University of Science and Technology, Jinan, China, 2011.
21. Zhao, Y.; Zhan, B.B.; Jia, P.Z.; Li, Y.H. Address Matching Algorithm Based on Rules and Dictionaries. *Beijing Surv. Mapp.* **2017**, *5*, 50–54. [CrossRef]
22. Hong, Y. Study and Experiments on Urban Geocoding Method. Master's Thesis, Liaoning Technical University, Dalian, China, 2008.
23. Mao, R.C. Research on Address Standardization and Semantic Model Construction Based on Deep Neural Network. Ph.D. Thesis, Zhejiang University, Hangzhou, China, 2019.
24. Jian, R.J. Building Standardization Model of Address Based on Statistical Methods. Master's Thesis, Yunnan University, Kunming, China, 2015.
25. Quan, Y.X. New Progress in Research on Chinese Word Segmentation Techniques in China. *J. Intell.* **2002**, *11*, 29–30.
26. Zhang, X.Y.; Wang, T.; Chen, H.W. Research on Named Entity Recognition. *Comput. Sci.* **2005**, *32*, 5. [CrossRef]
27. Zhu, F.; Zhao, T.; Liu, Y.; Zhao, Y. Research on Chinese Address Resolution Model Based on Conditional Random Field. *J. Phys. Conf. Ser.* **2018**, *1087*, 052040. [CrossRef]
28. Tang, X.R.; Chen, X.H.; Zhang, X.Y. Research on Toponym Resolution in Chinese Text. *Geomat. Inf. Sci. Wuhan Univ.* **2010**, *35*, 930–935+982.
29. Wei, Y.; Li, H.F.; Hu, D.L.; Li, X.; Ma, L. A Method of Chinese Place Name Recognition Based on Composite Features. *Geomat. Inf. Sci. Wuhan Univ.* **2018**, *43*, 17–23.
30. Yuan, X.D. Design and Implementation of Segmentation System for Chinese Address Based on Statistics and Rules. Master's Thesis, Southeast University, Nanjing, China, 2018.
31. Lilleberg, J.; Zhu, Y.; Zhang, Y. Support vector machines and word2vec for text classification with semantic features. In Proceedings of the IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC), Beijing, China, 6–8 July 2015; IEEE: Manhattan, NY, USA, 2015; pp. 136–140.
32. Li, H.; Lu, W.; Xie, P.; Li, L. Neural Chinese address parsing. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 3421–3431.
33. Ling, G.; Mu, X.; Wang, C.; Xu, A. Enhancing Chinese Address Parsing in Low-Resource Scenarios through In-Context Learning. *ISPRS Int. J. Geo Inf.* **2023**, *12*, 296. [CrossRef]
34. Zhang, H.; Du, Q.; Chen, Z.; Zhang, C. A Chinese address parsing method using RoBERTa-BiLSTM-CRF. *J. Wuhan Univ.* **2022**, *47*, 665–672.
35. Zhang, H. Study on the Parsing and Matching Methods of Chinese Addresses Based on BERT Pretrained Model. Ph.D. Thesis, Nanjing Normal University, Nanjing, China, 2021. [CrossRef]
36. Liu, X.; Li, Y.; Yin, B.; Tian, Q. Chinese address parsing integrating neural network with spatial relationship. *Sci. Surv.* **2021**, *46*, 165–171+212. [CrossRef]
37. Lee, D.; Yannakakis, M. Principles and methods of testing finite state machines-a survey. *Proc. IEEE* **1996**, *84*, 1090–1123. [CrossRef]
38. Gu, J. A Spatiotemporal Information Parsing Method for Cases and Events in Chinese. Ph.D. Dissertation, Nanjing Normal University, Nanjing, China, 2016.
39. Luo, M.; Huang, H. A Chinese address standardization method based on finite state machine. *Appl. Res. Comput.* **2016**, *33*, 3691–3695.
40. Wang, Y.; Liu, S.; Wang, Z. A Chinese address parsing model based on Trie and finite state automaton. *Comput. Mod.* **2016**, *7*, 60–67.
41. Tan, T.C. Finite State Machines and Its Application. Master's Thesis, South China University of Technology, Guangzhou, China, 2013.

42. Ma, R.; Peng, M.; Zhang, Q.; Wei, Z.; Huang, X. Simplify the usage of lexicon in Chinese NER. *arXiv* **2019**, arXiv:1908.05969. [CrossRef]
43. Levow, G.A. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In Proceedings of the 5th SIGHAN Workshop on Chinese Language Processing, Sydney, Australia, 22–23 July 2006; pp. 108–117.
44. 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.
45. Strubell, E.; Verga, P.; Belanger, D.; McCallum, A. Fast and accurate entity recognition with iterated dilated convolutions. *arXiv* **2017**, arXiv:1702.02098. [CrossRef]
46. Graves, A.; Mohamed, A.; Hinton, G. Speech recognition with deep recurrent neural networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; IEEE: Manhattan, NY, USA, 2013; pp. 6645–6649.
47. Lafferty, J.; McCallum, A.; Pereira, F.C.N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the 18th International Conference on Machine Learning, San Francisco, CA, USA, 28 June 2001.
48. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
49. Grishman, R.; Sundheim, B.M. Message understanding conference-6: A brief history. In Proceedings of the COLING 1996: The 16th International Conference on Computational Linguistics, Copenhagen, Denmark, 5–9 August 1996; Volume 1.
50. Di, L.; Ling, X.; Guangwen, W. Design of Chinese named entity recognition algorithm based on BiLSTM-CRF model. In Proceedings of the IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), Shenyang, China, 10–11 December 2021; IEEE: Manhattan, NY, USA, 2021; pp. 37–41.
51. Yu, B.; Wei, J. IDCNN-CRF-based domain named entity recognition method. In Proceedings of the IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT), Weihai, China, 14–16 October 2020; IEEE: Manhattan, NY, USA, 2020; pp. 542–546.
52. Zhang, S.; Zhu, H.; Xu, H.; Zhu, G.; Li, K.-C. A named entity recognition method towards product reviews based on BiLSTM-attention-CRF. *Int. J. Comput. Sci. Eng.* **2022**, *25*, 479–489.
53. Li, X.; Zhang, H.; Zhou, X.H. Chinese clinical named entity recognition with variant neural structures based on BERT methods. *J. Biomed. Inform.* **2020**, *107*, 103422. [CrossRef]
54. Ji, S.; Pan, S.; Cambria, E.; Marttinen, P.; Yu, P.S. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 494–514. [CrossRef] [PubMed]
55. Jaderberg, M.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep structured output learning for unconstrained text recognition. *arXiv* **2014**, arXiv:1412.5903.