A Novel Chinese Address Segmentation Method with Self-growth Feature

Yong zhang [1, a], Yingqiu Li [1], Fengkun Li[1], Yujun Shen[2] and Yanxin Xu[2]

[1] computer and software department, Dalian Neusoft University of Information, Dalian, China 116023, China;

[2] Hangzhou Hikvision Digital Technology Co., Ltd, 310051 Hangzhou, China

[a] zhangyong@neusoft.edu.cn

Abstract. Chinese Address Segmentation (CAS) is a crucial step that can greatly enhance the performance, accuracy, and reliability of geo-coding technology. However, it presents a tremendous challenge due to the inherent lack of obvious word boundaries, complex grammatical and semantic features. To address this challenge, we propose a novel CAS method or model that starts from scratch, without relying on any pre-installed knowledge about Chinese addresses. Instead, it dynamically evolves and grows its knowledge library by leveraging contextual information and comparing addresses during the process of dividing them into address elements. Our approach does not rely on Chinese language or address-element dictionaries, nor does it depend on address statistics. The knowledge library is automatically extracted and organized in a tree data structure. This unique approach allows our method to effectively segment addresses from any area of China, including regions with intricate address expressions, such as the Inner Mongolia Autonomous Region. Experimental results demonstrate that our method achieves high precision in address segmentation.

Keywords: List the; keywords covered; in your paper.

1. Introduction

Presently, there are tens of thousands of textual address information stored on the Internet. However, this valuable information is often disconnected from its corresponding spatial location. As a result, these web pages are unable to provide location-based services, hindering comprehensive big data analysis [1, 2] and disabling a wide range of applications and services related to urban spatial positioning. To address this challenge, geocoding has emerged as a coding technique that enhances the value of information with text addresses by mapping non-spatial information to spatial information, such as geographic or GIS coordinates. Geocoding involves several primary steps, including address element segmentation, address rectification or standardization, and address positioning [3]. Address element segmentation, in particular, is a fundamental and critical step in geocoding as it involves dividing Chinese addresses into the smallest address-semantic units.

Many authors of existing Chinese Address Segmentation (CAS) papers consider CAS to be a highly complex task. It's not entirely accurate to consider Chinese address segmentation as a Natural Language Processing (NLP) task. First, it is not advisable to solve a simpler problem using a complex method, as NLP tasks are generally more intricate than address segmentation tasks. Second, even a proficient Chinese speaker may not be knowledgeable about Chinese address segmentation, particularly for addresses unfamiliar to them. Third, Chinese addresses do not conform to normal Chinese sentences. Chinese addresses are simply sequences of nouns without any verbs [4,5].

By considering these factors, it becomes evident that Chinese address segmentation requires specialized methods tailored specifically for the task, distinct from general NLP approaches.

2. Related Works

The topic of CAS has garnered significant attention, with many researchers exploring its various aspects. Existing papers related to CAS can be roughly categorized into two distinct groups.

2.1 Artificial tagging and manual feature extraction based methods

These methods involve the extraction of features or the discovery of rules by analyzing Chinese language, Chinese addresses, and address structures. For instance, researchers examine word formation, part-of-speech features, semantic rules, and other linguistic aspects. Based on these identified features or rules, certain address information libraries are constructed and stored in data structures such as double-character-hash-indexing (DCHI) [6], double hash structures[7], or Double-Array Trie [8], among others. Research papers falling under this category were among the earliest to emerge, and they primarily focus on matching Chinese textual addresses with a dictionary composed of address elements, following a specific method.

Different approaches can be identified depending on the direction of string matching, such as forward maximum matching, backward maximum matching, and bidirectional maximum matching. Notably, the paper [9] proposed a Chinese word segmentation technique based on maximum matching and word binding force, primarily intended for general Chinese text rather than specifically targeting Chinese address segmentation.

2.2 Automatic feature extraction based methods

In contrast to the previous approaches, these methods typically rely on neural networks and do not require explicit feature extraction or predefined rules as input. They leverage self-learning and training techniques, enabling the neural networks to determine the optimal weights for achieving accurate address segmentation. For instance, paper [10] introduces a long short-term memory neural network that serves as a foundation for developing address segmentation methods [11]. Similarly, paper [12] presents a gated recursive neural network that utilizes contextual character features and a supervised layered training method.

3. Problem Discussion

In this section, we will illustrate our comprehension of Chinese addresses, which forms the foundation for our address segmentation approach.

3.1 Addresses in China

The Chinese address framework is structured around five administrative divisions: province, city, county (district), town, and village (community) [1]. These divisions serve as the basis for coding addresses in China, incorporating feature words such as province, city, district, street, and road. Furthermore, according to Chinese language semantics, addresses typically start with high-level address elements (e.g., province-level) and conclude with low-level address elements (e.g., street, road, building) as illustrated in Fig. 1.

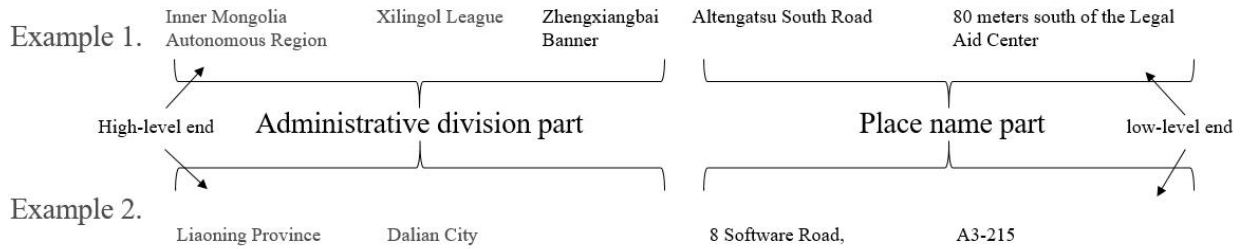

Fig. 1 Structure of Chinese addresses

3.2 Definitions

In this section, we provide a set of definitions to establish a common understanding of the key entities associated with address segmentation.

Definition 1. Geographic Area (GA) refers to a specific physical region on the surface of the Earth.

Definition 2. Address Element (AE) is a text-based representation that carries spatial semantics and refers to a specific Geographic Area (GA). It can be thought of as the label or name assigned to the corresponding GA.

AE can be represented as an ordered list of Chinese characters, as shown in Eq. (1), where c_i indicates a Chinese character and "n" is an unknown integer and only can be determined based on the specific context.

$$\mathrm{AE} = c_1 \cdots c_i \cdots c_n. \tag{1}$$

Definition 3. Chinese Address (CA for short) is a sequential arrangement of Chinese Address Elements (CAEs) and can be considered as an ordered list of Chinese characters. Therefore, we can represent a CA using the following Eq. (2).

$$\begin{aligned} \mathrm{CA} &= \mathrm{AE}_1 \cdots \mathrm{AE}_j \cdots \mathrm{AE}_m \\ &= c_{11} \cdots c_{1i} \cdots c_{1n_1} \cdots c_{j1} \cdots c_{ji} \cdots c_{jn_j} \cdots \\ &= c_1 c_2 \cdots c_{k-1} c_k c_{k+1} \cdots c_l \end{aligned} \tag{2}$$

Where c also indicates a Chinese character, i, j, l, k, n_1 and n_j are unknown integers and $l = n_1 + n_j + \cdots$. Those integers only can be decided in a specific situation.

The operators used in set theory, such as, $\supset$, $\subset$, $\cap$, $\cup$, $\supseteq$ and $\subseteq$, play a role in expressing the hierarchical structure of Chinese addresses.

4. Segmentation Methodology

In this section, we employ a tree data structure called the Address Knowledge Tree (AKT) to store the address context information.

4.1 Address Knowledge Tree

An Address Knowledge Tree is an un-directed graph G that satisfies the following conditions：

G is connected and acyclic; The root of G is "Zhong Guo"(China) and is virtual that it does not appear in all the Chinese addresses; The vertices of G are address elements, denoted by $V_{i,j}^{G}$, where i indicates the level of this vertex on the tree and j states the number of this vertex in the range of its parent vertex; The vertices of G near the virtual root have higher level than vertices far from the root; The corresponding GA of AE_G^i has two children at least.

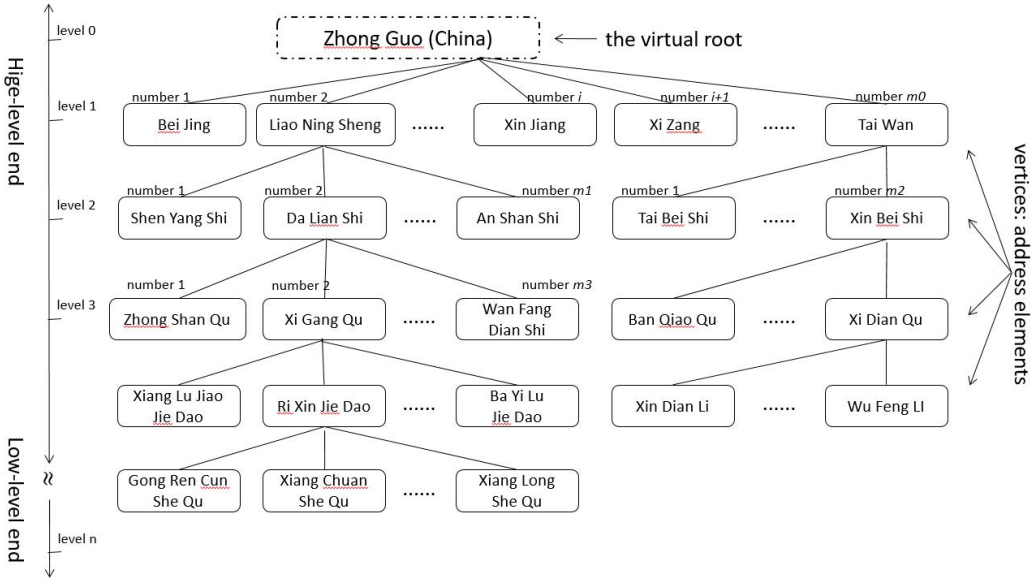

Fig. 2 Example of an Address Knowledge Tree

An example of AKT is shown in Fig.2. Every branch of AKT is an address context which includes two kinds of important information: address elements and parent-child relationship between them.

4.2 Chinese Address Segmentation Algorithm (CASA) based on AKT

Suppose $A = c_1c_2\cdots c_{k-1}c_kc_{k+1}\cdots c_l$ is a CA with l characters totally, G is a full-trained and qualified AKT for A, $V_{i,j}^G$ is the vertex of G and c_x^G is a character of $V_{i,j}^G$.

The following steps will be taken to do Chinese address segmentation:

Step 1.get the first character c_1 of the CA;

Step 2.get the set of the first level vertices of G, $V_1^G = \{V_{1,j}^G\}$;

Step 3.for every $V_{1,j}^G$, test if the first character $c_1^{'}$ of $V_{1,j}^G$ is equal to c_1, that is, $c_1^{'} = c_1$.There must be value for j which makes $c_1^{'} = c_1$true because the AKT G is a qualified one for A. Let's suppose $j = k$,that is, $V_{1,k}^G$, the test pass and we get that $V_{1,k}^G$ is the candidate first address element of CA. Test if the left characters of $V_{1,k}^G$ are equal to the corresponding characters of A. The test pass and we get the final first address element.

Step 4.get the set of the next level or children vertices of $V_{n,k}^G$, where $n > 1$, each one of them can be denoted by $V_{n,k}^G$;$l_s = \sum_1^{n-1} \left|V_{i,k}^G\right|$ indicates the sum of the length of all the segmented AE.

Step 5. for every $V_{n,j}^G$, test if every character $c_i^{'}$ of $V_{n,j}^G$ is equal to c_{l_s+i}, that is, $c_i^{'} = c_{l_s+i}$ where i begin with 1 and end with $\left|V_{n,j}^G\right|$. Suppose $j = k$, the test pass and we get that $V_{n,k}^G$ is the next address element of CA. Step 4 and step 5 repeat before the address A has been segmented completely.

4.3 Address Knowledge Tree (AKT) Building Algorithm (AKT-BA)

As shown in section 4.2, CAs can be segmented correctly based on a full-trained and qualified AKT. In this section, we show how to build an AKT from scratch based on a raw address corpus.

Suppose $A_i = c_1^ic_2^ic_3^i\cdots c_{k-1}^ic_k^ic_{k+1}^i\cdots$ is a raw address from the prepared address corpus. G is an empty AKT.

Step 1. Get a raw address $A_i = c_1^ic_2^ic_3^i\cdots c_{k-1}^ic_k^ic_{k+1}^i\cdots$;

Step 2. Get the first level vertices of G, denoted by $V_{1,j}^G$;

Step 3. for every $V_{1,j}^G$, test if the first character $c_1^{'}$ of $V_{1,j}^G$ is equal to c_1, that is, $c_1^{'} = c_1^i$.

Branch 1. Let's suppose when $j = k$, the test pass. That is, $V_{1,k}^{G}$ is the vertex which support the equation $c_1^{'} = c_1^{i}$. And then, test if the left characters of $V_{1,k}^{G}$ are equal to the corresponding characters of A_i.

Sub-branch1. If the test pass, it means the AKT has mastered the knowledge about $V_{1,k}^{G}$ and A_i has nothing useful knowledge for the AKT until now. Now, get the children vertices of $V_{1,k}^{G}$, and repeat the Step 3 to process the left part of A_i

Sub-branch2. If the test does not pass, that is, if we suppose $n < \left|V_{1,j}^{G}\right|$, we get $c_n^{'} \neq c_n^{i}$. So the vertex $V_{1,k}^{G}$ should be split into two parts. The first part take place of $V_{1,k}^{G}$ and the latter part is inserted into the AKT as a new vertex and all the descendant vertices of $V_{1,k}^{G}$ become the descendant vertices of the new inserted vertex. It means A_i supply the AKT with new and more precise address element knowledge. At the same time, for A_i, all the left characters index equal to or bigger than n are new for G and are inserted into G as a child vertex of $V_{1,k}^{G}$. The process stops and quits.

Branch 2. The test does not pass, it means A_i is new for the AKT and AKT does not know A_i at all. So A_i is inserted into the AKT as a new vertex. The process stops and quits.

For example, now we have a small address corpus consisted of 4 raw addresses. They are listed in the order as follows:

A_1 ="Liao Ning Sheng Da Lian Shi Gan Jing Zi Qu Yi Xin Jie"

A_2 ="Liao Ning Sheng Shen Yang Shi Tie Xi Qu Hong Qi Lu 8 Hao"

A_3 =" Liao Ning Sheng Da Lian Shi Zhong Shan Qu Fu Qiang Luv 307 Hao"

A_4 =" Liao Ning Sheng Da Lian Shi Gan Jing Zi Qu Ling Xiu Jie 81 Hao 2 Dan yuan 301"

Fig.3 demonstrates how to build an AKT G from scratch using the above small address corpus.

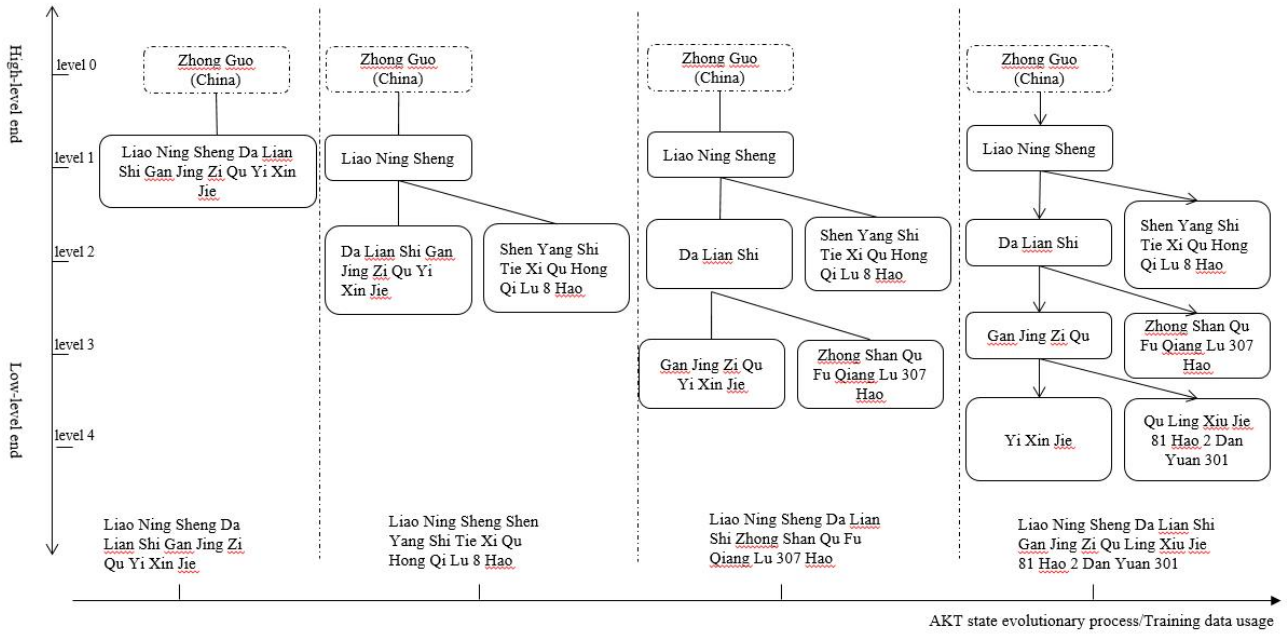

Fig. 3 The growing process of AKT

5. Result and Discussion

5.1 Results

Definition 4 A **breakpoint** in an address refers to the specific location where two address elements intersect, indicating the point of segmentation between them.

For a better evaluation, the following indexes are introduced.

Breakpoint Precision (B.K.P. for short) is defined by formula (3) and is the ratio of the breakpoints segmented correctly to the real breakpoints.

$$B.K.P = \text{Count}(\text{segmentd} - \text{correctly breakpoints})/\text{Count}(\text{the real breakpoints}) \quad (3)$$

Where Count() is a function used to count.

Address Precision (Addr. P. for short) is defined by formula (4) and is the ratio of the addresses segmented correctly to the total addresses. That an address is segmented correctly means all the breakpoints of the address are segmented correctly.

$$\text{Addr. P} = \text{Count}(\text{segmentd} - \text{correctly addresses})/\text{Count}(\text{the total addresses}) \quad (4)$$

Table 1. Evaluation statistics

Cities	Addr. Amount	Addr. P.	B.K.P.
Dalian	142012	98.86%	99.32%
Shenzhen	143227	97.10%	98.20%
Inner Mongolia	134620	96.56%	97.99%
Xinjiang	100168	96.31%	97.76%
Total/Average	520027	97.21%	98.54%

5.2 Result Analysis

Table2 shows the performance comparison of four algorithms. The test result of Bi-GRU, GRU and LSTM are from the paper [1]. From Table 2, The conclusion that can be drawn from Table 2 is: the address segmentation scheme does not utilize special data structures, such as neural networks, and does not require dedicated pre-training.

Table 2. Performance comparison of four algorithms.

Algorithms	Trained	Addr. P.
The proposed method	No, Self-growth	97.21%
Bi-GRU[1]	Yes	97.81%
GRU[1]	Yes	91.15%
LSTM[1]	Yes	90.65%

6. Conclusions

We present a novel CAS method/model that starts from scratch, without any pre-installed knowledge. It autonomously develops its knowledge library by comparing addresses and dividing them into address elements. It does not rely on Chinese language or address semantics, address-element dictionaries, or address statistics. Our approach achieves excellent address precision and breakpoint accuracy. However, our method also encounters an over-segmentation issue, resulting in an excessive number of breakpoints and the division of Chinese addresses into excessive elements.

Acknowledgment

This work received partial support from the National Key R&D Program of China (2021YFC3320302) and the Liaoning Province Education Science Plan "13th Five-Year" project (JG20DB037). The authors express their gratitude to the anonymous reviewers and the editor for their valuable feedback and suggestions on this paper.

References

[1] Li Pengpeng, Luo An, Liu Jiping, Wang Yong, Zhu Jun, Deng Yue, et al., Bidirectional Gated Recurrent Unit Neural Network for Chinese Address Element Segmentation, ISPRS International Journal of Geo-Information, Oct. 2020, 9(11)

[2] Dhar Subhankar, and Varshney Upkar, Challenges and business models for mobile location-based services and advertising, Communications of the ACM, 2011, 54(5)

[3] Song Zihui, Address matching algorithm based on chinese natural language understanding, Journal of Remote Sensing, 2013, 17(4)

[4] Li Lin, Wang Wei, He Biao, and Zhang Yu, A hybrid method for Chinese address segmentation, International Journal of Geographical Information Science, 2018, 32(1)

[5] Mengjun KANG, Qingyun DU, and Mingjun WANG, A new method of Chinese address extraction based on address tree model, Acta Geodaetica et Cartographica Sinica, 2015, 44(1)

[6] Jin Zhihui, and Tanaka-Ishii Kumiko, Unsupervised segmentation of Chinese text by use of branching entropy, in Proceedings of the COLING/ACL on Main conference poster sessions -, Sydney, Australia: Association for Computational Linguistics, 2006.

[7] Li Qing-hu, Chen Yu-jian, and Sun Jia-guang, A new dictionary mechanism for Chinese word segmentation, Journal of Chinese Information Processing, 2003, 4

[8] Mo Jian-Wen, Zheng Yang, Shou Zhao-Yu, and Zhang Shun-Lan, Improved Chinese word segmentation method based on dictionary, Computer Engineering and Design, 2013, 34(5)

[9] Wang Sili, Zhang Huaping, and Wang Bin, Research of optimization on double-array trie and its application, Journal of Chinese Information Processing, 2006, 20(5)

[10] Chen Xinchi, Qiu Xipeng, Zhu Chenxi, Liu Pengfei, and Huang Xuan-Jing, Long short-term memory neural networks for chinese word segmentation, in Proceedings of the 2015 conference on empirical methods in natural language processing, 2015.

[11] Yao Yushi, Huang Zheng, Bi-directional LSTM recurrent neural network for Chinese word segmentation, Neural Information Processing: 23rd International Conference, ICONIP 2016, Kyoto, Japan, October 16–21, 2016

[12] Chen Xinchi, Qiu Xipeng, Zhu Chenxi, and Huang Xuan-Jing, Gated recursive neural network for Chinese word segmentation, in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015.