Research on Chinese Address Resolution Model Based on Conditional Random Field

Zhu Feng  Zhao Tingting  Liu Yang  Zhao Ying

College of Geomatics of Shandong University of Science and Technology, Qingdao 266590, China
19116652@qq.com

Abstract. In this paper, a Chinese address resolution method based on Conditional Random Fields is proposed. In this method, the address resolution is divided into the address segmentation and the address component annotation issues, and moreover, the segmentation is combined with the address component annotation to form the annotated data set of output sequences. Meanwhile, taking a full consideration of the composition and use habits of Chinese address, the helpful characteristics to address resolution is set. Then, the address corpus and the corresponding characteristic template are constructed, and the conditional model which is suitable for Chinese non-standard address is obtained with model training of conditional random fields. In the end, the performance test is conducted on the conditional model of Chinese address on test set gained through training with experiments. Results show that the annotation accuracy of conditional random fields conforms to the requirements of address matching basically, and the accuracy is over 80%, with a certain practical value.

1. Introduction
Geographic Information Systems (GIS) has been integrated into human economic and political activities and has become an important service tool for guiding human behaviour decisions [1]. At present, there are a large number of location-based information in the domestic postal services, express delivery, taxation, planning, public security, and other areas. However, such information does not record the spatial location information corresponding to it, it can only be stored in the form of text. It unable to conduct spatial analysis and greatly limited the extraction and decision-making of business information [2][3]. Geocoding technology is to visualize the text information of the place name address on the map, then the location information will be better applied to data analysis, decision making, location services, and the like.

Geocoding is divided into address resolution and address matching. Different from western languages such as English, Chinese has no splitting symbols, multiple combinations, and arbitrary description. Although it increases the flexibility of expression, it also causes the study of geocoding great troubles [4][5]. Address resolution technology is developed on the basis of natural language segmentation technology. With the development of Chinese information processing technology, the concept of Chinese word segmentation was proposed in the 1980s [6]. KeJiang He introduced the expert system into the Chinese word segmentation. The reasoning mechanism and knowledge base of the expert system can transform the Chinese word segmentation process into a knowledge reasoning process. However, due to the complex knowledge base, difficulty in constructing, and the reasoning mechanism is too artificial, etc. The system has not been widely used [7]. Chinese word segmentation is mainly divided into statistical learning and rule-based word segmentation methods. Compared with
dictionary-based and rule-based Chinese word segmentation methods, Chinese word segmentation based on statistical learning can break through the dependence on the dictionary and better resolve the problem of unknown word recognition. So it has been widely used. In 2002, Zhang Huaping and Liu Qun developed ICTCLAS segmentation system based on Cascading Hidden Markov Models. The system obtained good word segmentation results. In terms of statistical learning methods, Song Zihui [8] dealt with non-canonical Chinese addresses based on the principles of Chinese word segmentation and semantic reasoning in natural language understanding and integrates natural language understanding into Chinese address resolution.

Conditional Random Field (CRF) as a discriminative model-based statistical learning method does not need to satisfy strict conditional independent assumptions and can adopt more flexible feature templates to accommodate more contextual information and is therefore more widely used in Chinese word segmentation processing [9]. As the conditional random field is used to mature the Chinese word segmentation technology, the place name address data is used as the training corpus, and an annotation feature set corresponding to the address tagging is established, and a corresponding address tagging feature template is constructed to train the address recognition and address element recognition. The conditional random field model has become a practical way to solve the problem of non-standard address resolution in Chinese. This paper draws on the Chinese parsing approach of natural language understanding and applies statistical annotation ideas to Chinese addresses. The CRF model is used as a statistical model to annotate the address, build an address-labeling corpus, select an appropriate feature template, train the conditional random field model for processing Chinese addresses, and finally obtain the address resolution result, thereby making the non-canonical address match and improve the matching accuracy.

2. Address resolution and CRF theory

2.1. Address resolution

Geocoding is divided into geo-geographic-based geocoding and geo-entity-based geocoding. The former is mainly used to make location classification by using the grid form, including the standard grid geocoding [10][11] and non-standard grid geocoding. The latter uses latitude and longitude coordinates as the conversion result, including encoding of address elements and encoding of address names. Address encoding based on geographical entities is also called address matching. Address matching refers to the process of parsing the address information of the place name through the parsing algorithm, and then matching with the standard address database to obtain the space coordinate corresponding to the information. The parsing process is as shown in the figure. 1.

[Diagram of geocoding process]

The result of address resolution has a great influence on the accuracy of address matching. Address resolution is the basis and premise of address matching. Chinese address resolution methods are mainly divided into three methods based on rules, based on understanding, based on statistics. The rule-based method is also called a dictionary-based method. The principle of parsing is to use the established address element dictionary and address matching rule base. The model is shown in the figure2. Rule-based address resolution algorithms can only be used to describe very standard addresses, and are not suitable for solving problems with Chinese addresses.
Fig 2. Rule based address parsing model

The idea of the statistical address-based address resolution method mainly comes from the Chinese natural language understanding. The Chinese address is a branch of Chinese natural language, and its statistical labeling idea is also suitable for address resolution. This method has achieved good results by tagging the address with corpora training and obtaining an address statistical model to perform address segmentation and address element identification.

2.2. The mathematical principle of conditional random fields

The conditional random field model integrates the advantages of Hidden Markov Models and Maximum Entropy Models, can make good use of contextual features, and performs global normalization of all features [12]. Combining the difference with MEMM gives the formula (1) for the conditional random field model

$$P(Y|X) = \frac{\exp(\omega \phi(Y, X))}{\sum_{Y'} \exp(\phi(Y', X))}$$ (1)

Where X is the input sequence, Y is the output sequence, $\omega$ is the weight vector, and $\phi(Y', X)$ is the eigenfunction. It maps the input sequence and the output sequence to a real number vector, so that the eigenfunction has more state information. $Y^n$ is the number of state sequences corresponding to the input sequence. $\phi(Y', X)$ characterizes the eigenfunction of the input sequence and a state.

In the sequence labeling task, since the Chinese sentence satisfies the characteristics of the linear chain conditional random field, the linear chain conditional random field is used as the model for the sequence labeling. The current state and the previous state of the linear chain conditional random field is a maximum group, and the formula (2) can be obtained.

$$\phi(Y, X) = \sum_i \phi(y_{i-1}, y_i, X)$$ (2)

for the j-th dimension feature function can be defined as Formula (3)

$$\phi_j(Y, X) = \sum_i \phi_j(y_{i-1}, y_i, X)$$ (3)

This gives the first-order linear CRF formula (4)

$$P(Y|X) = \frac{\exp(\sum_j \omega_j \phi_j(y_{i-1}, y_i, X))}{\sum_{Y'} \exp(\sum_j \omega_j \phi_j(Y', X))}$$ (4)

3. Chinese address resolution and annotation construction

3.1. Chinese address resolution principle

The Conditional Random Field Chinese address resolution is a kind of statistical learning method. The basic thinking is: Firstly, the tagged address corpus is used as the training corpus. Through the corresponding learning algorithm, an appropriate labeling set is set up to train the conditional random
field model. The model then is used to annotate similar linguistic addresses and eventually translates into address resolution results. Chinese address resolution is essentially the recognition of word segmentation and address component types in Chinese addresses. The conditional random field model can solve the problem of word segmentation and address component recognition at the same time.

3.2. Chinese word segmentation labeling principle
CRF translates the process of address word segmentation into character-based sequence labeling in address strings. The idea is to train the CRF model using a segmentation tagged corpus. The Viterbi decoding algorithm is used to mark the probability of the untagged address based on the word formation. The labeling result is the segmentation result, and finally the possible sequence with the largest probability of word formation among the words in the output address. It comprehensively considers the frequency of words appearing in the corpus and the context of the words. It has strong learning ability and can solve the problems of ambiguity and unknown words that cannot be handled by the rules. The labeling of address segmentation by CRF generally takes the form of "B, M, E, S", where "B" represents the beginning of a word, "M" represents the middle of a word, "E" represents the end of a word, and "S" represents an individual Word formation. For the final decoded annotation result, the segmentation sequence of the address can be easily obtained according to the annotation form shown in Figure 3.

3.3. Chinese address decoding algorithm
The main idea of the Chinese address decoding algorithm uses the dynamic programming idea in the Viterbi decoding idea. The word segmentation is used as an example to illustrate the process of decoding Chinese addresses. Through this algorithm, the address "50 meters south of the university of Science and Technology" is decoded and word segmentation is marked, and the array is obtained by analyzing the possibility of labeling. As shown in Figure 4. Viterbe decoding is to find an optimal path in the tags of the following array, so that the probability of the path is maximized.

3.4. Construction of Chinese address corpus
In this paper, a method of establishing an address-labeling corpus is designed. In the case of confirming the annotation system, using the method of modeling and labeling at the same time, and the combination of model annotation and manual inspection is used to speed up the construction of corpus to reduce the enormous energy spent by the corpus through pure manual establishment. The
javascript API provided by GaoDeMap is used to retrieve and extract address data. The data extraction process is shown in the figure. The information extraction page is written in JavaScript language, and the data is saved in SVC format. The extraction program interface is shown in Figure 5 and Figure 6.

![Corpus construction flow chart](image1)

*Fig 5. Corpus construction flow chart*

Enter the city name into the city, enter the extraction type to determine the extraction keyword, select the city to limit the city to be extracted, click on latitude and longitude, and click to extract the output result to the "Extract Dst" in the root directory of drive C, the step length is used to set the minimum search range; Step length is not more than 0.2.

![Acquisition system page](image2)

*Fig 6. Acquisition system page*

3.5. Label address data

This paper adopts the method of hierarchical labeling and training. Manually label part of the address as corpus in advance and train the conditional random field model. Through the trained model, the possible types of address labels is predicted. Then the remaining untagged corpora of the address is marked, then the corrected results is added to the corpus and the new model is trained. The specific labeling process is shown in Figure 7.

![Process of label address data](image3)

*Fig 7. Process of label address data*
4. The construction of Chinese address resolution model
This chapter mainly introduces the address resolution process based on conditional random fields, and then studies the feature selection, feature template function and problem setting in the process of establishing a CRF model with Chinese addresses. By setting possible features and optimizing the combination of features to optimize the address resolution training model.

4.1. Address resolution process based on CRF
The address resolution process based on the conditional random field is divided into two phases: model establishment and model prediction. The model establishment stage is mainly to set the corpus to train CRF model through the conditional random field idea. The model prediction stage uses the trained model to label untagged addresses to determine address segmentation results and identify address components. The specific process is shown in Figure 8.

4.2. Address feature template
In the process of establishing an address model for address resolution, the address characteristics that may affect the establishment of the address model are fully considered. The address features used in this paper include: atomic features, part-of-speech features, syntax features, address word features, and output address component features. Among them, atomic features, part-of-speech features, and syntactic features belong to linguistic features. Word-use features refer to the use of words or expressions that are helpful for the recognition of address components, and the address component features are the output features of the solution. The design of the feature template is based on the annotation feature of the training corpora. The conditional template setting needs to fully consider the combination and use of various address features in order to obtain an efficient and accurate training model for the ultimate purpose. The essence of the condition template is a formal description of how address features are combined. Feature templates generally use the form of %X[row,column] to describe atomic features, and through its combination to describe the combination features.

5. Address resolution performance test based on CRF model
This paper adopts the CRF++ application program under Windows system. Through the DOS window, under the premise of setting appropriate training corpus data and feature templates for the program, an address model based on conditional random fields can be trained and the model can be evaluated and output the test results. The data preparation is mainly to transform the prepared address corpora into train.data data conforming to the CRF++ format. The format is shown in Table 1 below.
Table 1. Training data format sample

| Observation array | Feature of vocabulary | Word Features | Output array  |
|-------------------|-----------------------|--------------|--------------|
| Qing              | n                     | OW           | B-City       |
| Dao               | n                     | OW           | M-City       |
| Shi               | n                     | GW           | E-City       |
| Hai               | n                     | OW           | B-Town       |
| Qing              | n                     | OW           | M-Town       |
| Zhen              | n                     | GW           | E-Town       |
| Qi                | n                     | OW           | B-poi        |
| Che               | n                     | OW           | M-poi        |
| Zhan              | n                     | OW           | E-poi        |
| Nan               | f                     | TW           | S-Direct     |
| 200               | m                     | TW           | B-Distance   |
| Mi                | q                     | TW           | E-Distance   |

For the test of address resolution performance, this paper adopts a measurement standard similar to natural language understanding and uses 200,000 labeled data as a training corpus. The remaining 80,000 data are used as test corpus to evaluate the address resolution performance. It mainly evaluates the accuracy of the output sequence address component type. The evaluation index is the recall rate, accuracy rate, and comprehensive index of the address component type. The specific measurement description is shown in Table 2.

Table 2. Evaluation method of address tagging

| Measurement indicators | Evaluation formula |
|------------------------|--------------------|
| Recall rate (R)        | \( R = \frac{tp}{tp + fp} \) |
| Accuracy rate(P)       | \( P = \frac{tp}{tp + fn} \) |
| Comprehensive indicators (F) | \( F = \frac{R \times P \times 2}{R + P} \) |

In the table, \( tp \) denotes the correct label that has been identified, \( fp \) denotes the incorrect label that has been identified, \( fn \) denotes the correct label that was not identified.

On the basis of determining the best window for selecting context features, different address feature templates need to be designed for different address features to verify the applicability of the selected features to address resolution. When the most suitable feature template is selected, the test set address corpus is used for testing and the results are analyzed. The test results are shown in the figure 9. From the analysis of the following figure, we can see that the accuracy rate of using only general features is 87.30%, which basically meets the requirements of general address recognition elements, and the performance of address analysis is improved by adding part-of-speech features, and the accuracy of the collection of part-of-speech features and word features is achieved 89.02%, the correct rate in the address resolution phase is relatively important, and its F value is also the largest of the four tests, so this paper believes that adding the part-of-speech feature and the word feature into the address feature template is conducive to the establishment of the address model, basically satisfying the requirements for address resolution tasks. It proved the feasibility of address resolution based on Conditional Random Fields.
6. summary
In this paper, under the influence of Chinese text segmentation and the idea of Chinese named entity recognition, the possibility of applying statistical ideas to address resolution is studied. Through the study of address segmentation and address component type recognition, the current mainstream statistical conditional model CRF is used as a Chinese address training model to improve the matching accuracy.

References
[1] Hai Zhou. Geocoding method based on conditional random fields and spatial reasoning[D]. PLA University of Information Science and Technology, 2015.
[2] Xuri Tang, Xiaohe Chen, Xueying Zhang. Research on geographical names resolution in Chinese texts[J]. Journal of Wuhan University (Information Science Edition), 2010, 35(08): 930-935+982.
[3] Yingfei Yan. Research on address matching based on semantic analysis[D]. Chongqing Jiaotong University, 2015.
[4] Ying Hong. Research and experiment on address matching of geographical names in cities[D]. Liaoning Technical University of Technology, 2008.
[5] Haitao Wu, Li Yu , Guijun Zhang. Urban Chinese Address Coding System Based on Fuzzy Matching Strategy[J]. Computer Engineering, 2011, 37(02): 194-196+199.
[6] Wei Wang. A Study of Chinese Word Segmentation Based on Statistical Learning [D]. Northeastern University, 2015.
[7] Rongrong Xu. Research on Automatic Segmentation Technology and Algorithm of Chinese Text [D]. Zhengzhou University, 2010.
[8] Zihui Song. Chinese address matching algorithm for natural language understanding[J]. Journal of Remote Sensing, 2013, 17(04):788-801.
[9] Jiangde Yu, Yanqiang Ge, Zhengtao Yu. Chinese part of speech tagging based on conditional random fields[J]. Microelectronics & Computer,2011,28(10):63-66.
[10] Cheng Chengqi, Guan Li. Research on Global Decomposition Model and Its Address Coding Based on Scale Extension of Map[J]. Acta Geodaetica et Cartographica Sinica, 2010, 39(3): 295-302.
[11] Jia Cheng. Population census technology experiment and application based on geocoding [D]. Chinese Academy of Surveying and Mapping, 2011.
[12] Jiayi Zhao, Xipeng Qiu, Shu Zhang. Part-of-Speech Tagging fo Chinese-English Mixed Texts with Dynamic Featurues[j]. Journal of Computational Information System(JCIS),2012:1379-1388.