Huang K, et al., Archiv Zool Stud 2020, 3: 014

DOI: 10.24966/AZS-7779/100014

HSOA Archives of Zoological Studies

Research Article

Automatic Extraction of Epidemic-Related Sites in Covid-19 Media Reports of Webpages Based on Conditional Random Field Model

Kuiyun Huang[1]**, Jinming Cao**[2] **and Bin Zhao**[1]*

[1]*School of Science, Hubei University of Technology, Wuhan, Hubei, China*

[2]*School of Information and Mathematics, Yangtze University, Jingzhou, Hubei, China*

Abstract

Background: Since the outbreak of the COVID-19 in Wuhan, China, in early December 2019, the Chinese government has formed a mode of information disclosure. More than 400 cities have announced specific location information for newly diagnosed cases of novel coronaviruspneumonia, including residential areas or places of stay. We have established a conditional random field model and a rule-dependent model based on Chinese geographical name elements. Taking Guangdong province as an example, the identification of named entities and the automatic extraction of epidemic-related sites are carried out. This method will help locate the spread of the epidemic, prevent and control the spread of the epidemic and gain more time for vaccine clinical trials.

Methods: Based on the presentation form of the habitual place or place of stay of the diagnosed cases in the text of the web page, a conditional random field model is established, and a rule-dependent model is established according to the combination rule of the elements of the place words and the place name dictionary composed of provinces, cities and administrative regions.

Findings: The results of the analysis based on the conditional random field model and the rule-dependent model show that the location of confirmed cases of new coronavirus pneumonia in Guangdong Province in mid-February is mainly concentrated inGuangzhou, Shenzhen, Zhuhai and Shantou cities.In Guangzhou, Futian district has more epidemic sites and Huangpu and Conghuadistricthas fewer epidemic sites. Government officials in Guangzhou City should pay attention to Futian District.

Interpretation: Governments at all levels in Guangzhou Province have intervened to control the epidemic through various means in mid-February. According to the results of the model analysis, we believe that the administrative regions with more diagnosed locations should focus on and take measures such as blockades and control of personnel flow to control the disease in those administrative regions to avoid affecting other adjacent administrative regions.

Keywords: Conditional random field; COVID-19; Elements of Chinese place name; Part-of-speechtagging; Place name recognition; Web crawler

Introduction

In early December 2019, the COVID-19 virus began to erupt from Wuhan, Hubei, China, and it has only become a global epidemic in just a few months [1]. In the past few months, China has been in crisis, but according to official government data, China has basically blocked local transmission. While some countries have entered the early stage of China, the current situation in some countries is even worse than when China's epidemic was the worst. From our point of view, one of the reasons why China can control the epidemic is that the central government and local governments have high information transparency and relevant agencies release the latest information on the epidemic on time.

From the epidemiological point of view, analyzing information such as the residential area or activity location of officially confirmed COVID-19 diagnosed cases is of great help in finding potential patients. Because these data provide a basis for the community to implement targeted prevention and control; make the people have the right to know and better personal protection; virus experts can build epidemic transmission models based on this, to the infection source, transmission speed, transmission Path and propagation risk to evaluate and predict. In the epidemic report of web pages, the residential area or activity location of the diagnosed case is usually expressed in various forms, such as the body of the page, embedded text on the pageand screenshots. To timely analyze the distribution of epidemic-related locations from these sources of information, first of all, relevant information must be extracted from unstructured data and converted into structured data. In the past, this work was mainly achieved by manually searching and classifying the specific information in the text. The workload is large and the efficiency is lowand it lacks timeliness.

In recent years, as the technology of named entity recognition has gradually matured, this work has gradually shifted from manual extraction to automatic extraction, which not only reduces human and financial resources, but also speeds up the processing of tasks. Named entity recognition refers to identifying named entities in

***Corresponding author:** Bin Zhao, School of Science, Hubei University of Technology, Wuhan, Hubei, China, Tel: +86 130 2851 7572; E-mail: zhaobin835@nwsuaf.edu.cn

Citation: Huang K, Cao J, Zhao B (2020) Automatic Extraction of Epidemic-Related Sites in Covid-19 Media Reports of Webpages Based on Conditional Random Field Model. Archiv Zool Stud 3: 014.

Received: May 21, 2020; **Accepted:** June 13, 2020; **Published:** June 19, 2020

text and dividing them into corresponding entity types [2]. The general entity types include person names, place names, dates, and organization names. Our main aim is to identify Chinese place names. Chinese characters are closely arranged in Chinese text, sentences are composed of multiple characters, and there are no spaces between words, which increases the difficulty of identifying named entities. In order to improve the accuracy of recognition, the named entity recognition technology ranges from the initial rule and dictionary method [3] to the traditional statistical learning method [4-6] to the current deep learning method [7]. The current technology has basically reached a higher level for some common entities Accuracy.

The purpose of this article is to process the text information in the web page until the entities and relationships are displayed.

Methods

Data

Since the outbreak of COVID-19, there are many news reports of newly diagnosed patients on the Internet, so it provides us with raw unstructured data to collect diagnostic locations. We collect official notifications and media reports on the itinerary of the activities of the diagnosed patients associated with Guangdong province from January 29, 2020 to February 19, 2020. The data comes from 366 webpages in 152 media. The original data is obtained through web crawler technology, and the data includes the main content of the webpage, the media of the news release and the corresponding URL. The basic principle of a web crawler is to simulate a browser making HTTP requests. The crawler client sends a request to the web server through the HTTP request, downloads the web page after obtaining the corresponding server, and completes the crawling work of the crawler system[8]. Part of the data is shown in Table 1.

NO.	Content	Media	URL
1	According to the latest news from the Guangdong Provincial Health Commission today (4th) ...	Southern Metropolis Daily	https://mp.weixin.qq.com/s/N76fZ6R-4zeiB6Zqa3TYaxg
2	From 0:00 to 24:00 on February 3, 2020, Guangzhou City reported a new model ...	Guangzhou Municipal Health Commission	http://wjw.gz.gov.cn/ztzl/xxfyyqfk/yqtb/content/post_5650395.html
3	According to the official website of the Guangdong Health Commission, as of 12:00 on January 31, the province ...	Southern Metropolis Daily	https://mp.weixin.qq.com/s/KkYpk-D4oV997_ZPNRa0vHg
...	...	...	...
366	Prevention and Control of Pneumonia Epidemic Infected by New Coronavirus in Haibei ...	Haibei new media	https://www.tibet3.com/news/zangqu/qh/2020-02-04/149341.html

Table 1: Part of the diagnosis locations data.

Text data preprocessing

The preprocessing of text data means that before performing named entity recognition, we must first determine whether there are missing values in the data. After checking, there are no missing values, and then we need to convert the data into a format that the model can easily handle. For example: delete some unnecessary character strings, split each press release in sentence units and remove the same sentence, etc. The part-of-speech feature is used in Peking University's [9].

The model

Based on the collected unstructured text data, we tried to identify and extract the place words in the text through named entity recognition technology, and then classify the recognized place words according to a certain rule and divide them into provinces, cities, administrative regions and detailed locations. Finally, we conduct statistical analysis of the location information to provide accurate data for the epidemic development model constructed by researchers in the future, as well as to assess and predict the source of infection, the speed of transmission and the route of transmission.

There are three main methods for studying Chinese place name recognition: rule-based methods, statistics-based methodsand deep learning-based methods. The rule-based method is intuitive and natural, and is easy to understand and expand by humans. However, rule writing depends on specific language knowledge and domain knowledge. The rules are more complicated, it is difficult to cover all the modes, and the portability is also poor [10,11]. Statistics-based methods do not require excessive language knowledge and domain knowledge and are highly portable, but require manual annotation of the corpus and selection of appropriate statistical learning models and parameters [12-14]. Deep learning-based methods do not require overly complex feature engineeringand can automatically discover information from the input to form an end-to-end model [13]. Considering that the collected text data is limited by time and data, the content of this paper is mainly based on the first two methods.

Recently, named entity recognition has achieved good results in some limited entity types. For example, the recognition effect on the names of people, places and organizations in news corpus is remarkable. Chinese place name recognition can be regarded as a sequence labeling problem. The place name is a combination of multiple words arranged in a certain order, and the place name entity recognition is the combination of marking the correct names from these word sequences. The conditional random field model combines the advantages of the maximum entropy model and the hidden Markov model, and can be used for the labeling and segmentation of sequence data [6]. Therefore, the effective solution to the sequence labeling problem is the conditional random field model [15]. We chose the conditional random field model as the identification model for epidemic locations.

Maximum likelihood method for estimation

Suppose there is a sample set $D=\{x^j, y^j\}$, $\forall j=1,2,\ldots,N$ for the training data, where the samples are independent of each other, and $p(x,y)$ is the empirical probability of (x,y) in the training samples. For a certain conditional model $p(y \mid x,\theta)$, the maximum likelihood function formula of the training data sample set can be expressed as:

$$L(\theta)=\prod_{x,y} p(y \mid x,\theta)^{\tilde{p}(x,y)}$$

Take the logarithmic form:

$$L(\theta)=\prod_{x,y} \tilde{p}(x,y)\log p(y \mid x,\theta)$$

Since we choose traditional conditional random field model, its conditional probability can be expressed as:

$$p(y|x,\theta)=\frac{1}{Z(x)}\exp\left(\sum_{i=1}^{n+1}\sum_{k}\lambda_k f_k(y_{i-1},y_i,x)+\sum_{i=1}^{n}\sum_{k}\mu_k g_k(y_i,x)\right)$$

Where λ and μ are the parameters to be estimated, and f and g are the vectors of the eigenfunctions.

Differentiate the parameter λ_k:

$$\frac{\partial L(\theta)}{\partial \lambda_k}=\sum_{x,y}\widetilde{p}(x,y)\sum_{i=1}^{n+1}f_k(y_{i-1},y_i,x)-\sum_{x,y}\widetilde{p}(y|x,\theta)\sum_{i=1}^{n+1}f_k(y_{i-1},y_i,x)$$

Let the value of the above formula be 0 to meet some of the constraints and find λ_k. In the same way, μ_k can also be obtained, so that θ can be parsed, but this method may not always be able to directly find the solution value.

Feature selection for Chinese location recognition

The training set used in this article is the annotated corpus of the People's Daily in January 1998. It uses the part-of-speech annotation set of Peking University. The label set has marked all the names of people, places and organizations. Since these corpora are sentences that have been roughly segmented, in order to make the features of entity extraction more obvious, these sentences need to be finely segmented.

The 5n-gram template can fully express the lexeme information of the word. Among them B (the beginning word of the named entity), M (the middle word of the named entity), E (the tail of the named entity), S (the single word constitutes the named entity), N (unnamed entity). The three types and combinations of entities form various labels. The template can effectively mark the position of the word, so that the system can use the position feature to identify the boundary of the word. Table 2 shows the labeled labels.

Output tags	Explanations for different tags
Per-B	The start word(s) of the name entity
Per-M	The middle word(s) of name entities
Per-E	The ending word(s) of name entity
Per-S	The individual word(s) constitute name entity
Loc-B	The start word(s) of place name entity
Loc-M	The middle word(s) of place name entity
Loc-E	The ending word(s) of place name entity
Loc-S	The individual word(s) constitute a place name entity
Org-B	Start word(s) of organization name entity
Org-M	The middle word(s) of organization name entity
Org-E	The ending word(s) of organization name entity
Org-S	The individual word(s) constitute organization name entity
N	Unnamed entity word

Table 2: Corpus conversion tag.

Feature template selection

The conditional random field is very dependent on the selection of features, which has a great influence on the accuracy of the final recognition. Theoretically, if more context information is collected around the current word, that is, the larger the value of the observation window, the richer the information obtained and the more accurate the judgment of the current word [16].

But once the observation window is too large, the calculation of too much information will make the model identification efficiency inefficient and affect the operation efficiency. If the value of the window is too small, the relevant dependency information cannot be fully utilized, affecting the accuracy of recognition [16]. So, choosing the right window size is the premise of choosing the right feature template. The window size selected in this article is 2.

The base feature is the most basic feature that is stronger than the character itself, including the current character or key, the position of the first character in the pre-word and part of speech. Part-of-speech features can often improve the degree of discrimination.

For example, named entities are often noun part-of-speech words, while verb part-of-speech words are rarely used as named entities. For a word-based entity tagging system, the part-of-speech feature is to use the part-of-speech of the word in which the word is located. Based on these, we establish the basic features of the template as shown in Table 3.

Transfer features	Word features	Part of speech features
y_t	$word_{t-2}$	tag_{t-2}
	$word_{t-1}$	tag_{t-1}
	$word_t$	tag_t
	$word_{t+1}$	tag_{t+1}
	$word_{t+2}$	tag_{t+1}

Table 3: The basic features of the template.

Where t represents the position where the feature is currently being extracted, y represents the label, word represents the word, and tag represents the part of speech. Considering that there are too many ways to combine words, many binary grammatical features are not used.

Entity relationships

The seven types of relationships between entities are: partial overall relationship, geographical location relationship, generic relationship, metaphor relationship, manufacturing use relationship, organizational structure dependency relationship and character relationship [17]. For this work, what is studied is the relationship between locations in an outbreak press release, so more attention is paid to the identification of geographic location relationships.

According to the characteristics of the recognized text itself, the geographical location relationship is classified in detail (Table 4):

The geographical location relationships that may be involved in the entry are divided into four categories: provinces, cities, administrative regions and geographical locations. The head word of a relationship can be not only a noun but also a verb. The system needs to classify the identified entity relationships. The specific division is shown in Table 5.

Entity relationships extraction method

Recognition of relational semantics is constantly evolving at any time and is divided into methods based on rule matching and methods based on machine learning. The method based on rule template matching is to define the rule template beforehand, and compare the statement with the rule template during the relationship identification. If the statement matches the characteristics of the characteristic template, it means that the entity in the statement has the relationship specified in the template Attributes [18]. The disadvantage is that it requires more professional linguists to write a large number of feature templates, which takes a long time and has poor portability [19].

Category		Symbol	Word
Administrative Area	Province	S1	Provinces, Municipalities, Autonomous regions, Special administrative regions
	City	S2	City, Region, League, Autonomous prefecture
	County	S3	County, Flag, Qu, District
	Town	S4	Township, Town, Street office
	Village	S5	Village, Zhuang, Tun, Li
Street		S6	Road, Avenue, Road, Street, Lane, Alley, Strip
Residential area		S7	Garden, Square, Residence, Apartment
Landmarks		S8	Building, Square, Hotel, Center, Building, Floor, Field, Square, Pavilion, Residence
House number		S9	NO., #

Table 4: Relationship category.

Category	Symbol
province	S1
city	S2
administrative region	S3
geographical location	S4, S5, S6, S7, S8, S9

Table 5: Specific geographical name division.

The method based on machine learning is a method that uses various pattern recognition feature models to calculate the entity relationship features and weight values in sentences through related algorithms. There are currently two popular types of machine learning methods for dealing with entity relationships, namely kernel-based methods and feature vector-based methods [20,21].

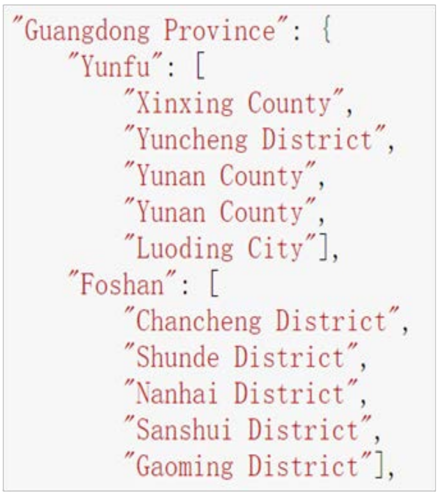
The purpose of our research is to perform location extraction. The feature template of geographic location relationship is relatively fixed and the portability is high. Therefore, we will use the rule-based matching method to extract the identified place words for relationship. There are three aspects of corpus preprocessing and rulemaking.

Corpus preprocessing

The preprocessing of corpus is mainly through the steps of word segmentation and entity recognition, which transform the sentences in the corpus into a stream of words with entity identification. Since entity relationship extraction is a relationship between two entities, the sentences with less than two place name entities in the text are filtered out, and the sentences containing two or more place name entities are used as recognition corpus.

Gazetteer

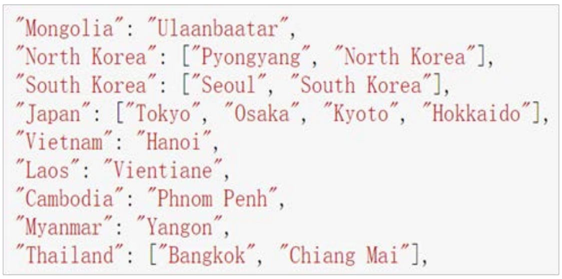
There may not be a complete place name in some sentences, for example: only the information of the administrative region, and no information of the province or city. At this time, we need to collect all the names of provinces, cities and districts in China and establish a dictionary framework. Part of the data is shown in Figure 1.



```
"Guangdong Province": {
    "Yunfu": [
        "Xinxing County",
        "Yuncheng District",
        "Yunan County",
        "Yunan County",
        "Luoding City"],
    "Foshan": [
        "Chancheng District",
        "Shunde District",
        "Nanhai District",
        "Sanshui District",
        "Gaoming District"],
```

Figure 1: Structural framework of some Chinese provinces, cities and administrative regions.

Since the object of our research is Guangzhou, China, we will not consider the COVID-19 epidemic situation of abroad. Therefore, we need to collect major foreign countries and place names from the Internet and establish a dictionary framework. Part of the data is shown in Figure 2.



```
"Mongolia": "Ulaanbaatar",
"North Korea": ["Pyongyang", "North Korea"],
"South Korea": ["Seoul", "South Korea"],
"Japan": ["Tokyo", "Osaka", "Kyoto", "Hokkaido"],
"Vietnam": "Hanoi",
"Laos": "Vientiane",
"Cambodia": "Phnom Penh",
"Myanmar": "Yangon",
"Thailand": ["Bangkok", "Chiang Mai"],
```

Figure 2: Structural framework of some foreign countries place name.

By studying the structural characteristics of epidemic location words, establishing rules based on regular expressions to extract word-by-word the location words identified by the named entity recognition, and then putting the extracted relevant location information into an appropriate data structure for subsequent deal with.

By analyzing location words, a total of 20 matching rules were established for 4 categories of location information. Each rule is a regular expression composed of keyword(s), of which there are 9 categories of keyword(s). The rule form is as follows:

Rule 1: (.*?)(S1)?

S1 = Provinces | Municipalities | Autonomous regions | Special administrative regions

Rule 2: (.*?)(S2)?

S2 = City | Region, League | Autonomous prefecture

Rule 3: (.*?)(S3)

S3 = County | Flag | Qu | District

Rule 4: (.*?)(S4)

S4 = Township | Town | Street office

Rule 5: (.*?)(S5)

S5 = Village | Zhuang | Tun, Li

Rule 6: (.*?)(S6)

S6 = Road | Avenue | Road | Street | Lane | Alley | Strip

Rule 7: (.*?)(S7)

S7 = Garden | Square | Residence | Apartment

Rule 8: (.*?)(S8)

S8 = Building | Square | Hotel | Center | Building | Floor | Field | Square | Pavilion | Residence

Rule 9: (.*?) (S9)

S9 = NO. | #

Rule 10: "Province" = Rule 1

Rule 11: "City" = Rule 2

Rule 12: "administrative region" = Rule 3

Rule 13: "geographical location" = [Rule 4, Rule 5, Rule 6, Rule 7, Rule 8, Rule 9]

Rule 14: Delete words that meet the Rule 1 to Rule 9 conditions, but do not belong to any place names. For example: "Epidemic Area", "Common Youth City", "Community", "Outer Province", "Inner Province", "You Province", "Reprinted City", "Ministry of Labor", etc.

Rule 15: Delete words that are incorrectly marked in the recognition of named entities. For example: "Sputum", "when getting on the train", "from the day", "and", "more", etc.

Rule 16: Delete words that belong to place names but are not related to this study. For example: "People's Republic of China", "China", "People's Hospital", "Chest Hospital", etc.

Rule 17: Use foreign gazetteer to delete foreign place names

Rule 17: If "administrative region"! = "" and "City" == "", find the corresponding place name from the Chinese gazetteer and fill it in "City"

Rule 18: If "administrative region"! = "" and "Province" == "", find the corresponding place name from the Gazetteer and fill it in "Province"

Rule 19: If "City"! = "" and "Province" == "", find the corresponding place name from the Gazetteer and fill it in "Province"

Rule 20: According to the principle of proximity, re-sort "administrative region" and "geographical location", and stitch the sorted words together. For example: after processing "Bailudong Street / Shili Bay / Suxian District", it becomes "Shili Bay, Bailudong Street,Suxian District"

Model evaluation criteria

For the evaluation of the model, the F1-score evaluation index is used for evaluation. For each type of named entity and relationship extraction, these three indicators are defined:

$$P = \frac{\text{Correctly identify the number of entities in this category}}{\text{Total number of entities identifying this category}}$$

$$R = \frac{\text{Correctly identify the number of entities in this category}}{\text{Total number of named entities of this category}}$$

$$F_1 = \frac{2 \times P \times R}{P + R}$$

Overall framework for automatic extraction of Chinese place names

The purpose of this article is to process the text information in the web page until the entities and relationships are displayed. The implementation process is mainly divided into two modules: web page processing module and entity and relationship recognition module. The frame diagram of the automatic extraction of Chinese locations is shown in Figure 3.

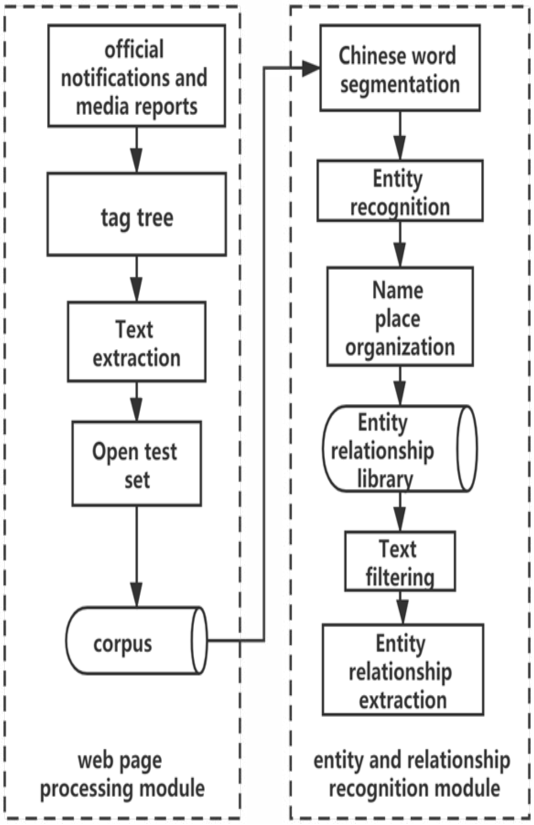

official notifications and media reports
tag tree
Text extraction
Open test set
corpus
web page processing module
Chinese word segmentation
Entity recognition
Name place organization
Entity relationship library
Text filtering
Entity relationship extraction
entity and relationship recognition module

Figure 2: Overall framework for automatic extraction of Chinese place names.

Results

Place name entity recognition result

This article uses the corpus marked by People's Daily in January 1998, of which 80% is selected as the training set, the remaining 20% is used as the closed test set, and the COVID-19 outbreak news release crawled through the Internet will be used as the open test set. The results of entity recognition are shown in Table 6.

Closed test corpus				Open test corpus			
Entity type	**P**	**R**	**F**	**Entity type**	**P**	**R**	**F**
Place entity	0.854	0.831	0.870	Place entity	0.781	0.763	0.771
Loc-B	0.881	0.852	0.872	Loc-B	0.732	0.718	0.720
Loc-M	0.873	0.852	0.870	Loc-M	0.729	0.706	0.711
Loc-E	0.869	0.841	0.857	Loc-E	0.703	0.690	0.692
Loc-S	0.896	0.853	0.878	Loc-S	0.693	0.752	0.764

Table 6: Place entity results.

It can be seen from the experimental data that the results of place name recognition have higher accuracy. The closed training set can reach the F value of 0.870, and the open training set can reach the F value of 0.771. As can be seen from the results of entity recognition, there are mainly the following types of incorrect recognition of place entities:

a. There are abbreviations for cities and provinces in the text, and it is possible to identify place names in ambiguous forms. For example: "Zhongshan" can be either a city in Guangdong Province or an administrative district in Dalian, Liaoning Province;

b. Some place names appear in multiple cities. For example: "Baojian Road" is the road name of many cities. When there are multiple cities in one sentence, it is not easy to determine which city this road name belongs to;

c. Place words in different places have different meanings. For example: The "Bajiao Tower" can be the name of a building or a town;

d. The wrong labeling of the entity label itself leads to the wrong words in the recognition location. For example: The words "Sputum", "when getting on the train", "from the day", "and", "more" are not place names, but they are classified as place names according to the algorithm.

Place entity extraction results

The location extraction link is very dependent on the central indicator (Table 7). In the case of other features that are the same, different types of central words will make the entity relationship pairs form different semantic relationships. Therefore, the central indicator thesaurus needs to identify the wrong entity center. The center of is extracted and added to the central indicator thesaurus to increase the recognized features. The main types of errors in entity relationship extraction are:

a. Some place name words omit entity words or substitute pronouns. For example: many provinces and cities have omitted the entity words "Province" or "City";

b. Two identical entities in the text have different relationship classifications and cannot determine priority. For example: "Jilin" is both a province name and a city name;

c. Multiple locations are involved in one sentence. There are multiple place words in a sentence, it is not easy to judge when recognizing the relationship, and the subordinate relationship can only be judged based on the position of the word in the sentence.

Category	P	R	F
Province	0.776	0.741	0.765
City	0.753	0.712	0.732
Administrative region	0.702	0.653	0.682
Geographical location	0.521	0.494	0.501

Table 7: Place entity relationship recognition result.

Final results

As some provinces in the form are not filled in, the locations of Guangdong province are selected and sorted according to the information in the city column. The results are shown in Table 8.

It can be inferred from Table 7 that the cities of the COVID-9 epidemic area in Guangdong Province are mainly concentrated in Guangzhou, Shenzhen, Zhuhai and Shantou. Among them, the epidemic areas in Guangzhou are concentrated in Yuexiu, Tianhe and Qiewan, the epidemic areas in Shenzhen are in Futian, Luohu and Longgang, the epidemic areas in Zhuhai are in Xiangzhou, and the epidemic areas in Shantou are in Chaoyang. Guangzhou has the widest epidemic area, the epidemic area of Heyuan is relatively small.

Discussion

There is no doubt that we can analyze the development of COVID-19 in a certain area from different angles. In this article, we mainly consider the extent of the spread of the epidemic. We believe that the spread of epidemics can be quantified by the number of epidemic outbreak locations in a region. It is learned from the studies of other scholars that cities with higher economic development level and larger floating population have more imported cases than other cities[22]. The cities with higher economic development levels in Guangdong Province are mainly concentrated in the core urban agglomeration of the Pearl River Delta, and Guangzhou is located in the northern part of the Pearl River Delta, so the spread of the Guangzhou epidemic is wider. Heyuan City is located in a mountainous area with poor traffic, so there are few imported cases.

If it is considered that the epidemic is spreading quickly, intervention should start from the early stage of the COVID-19 infectious disease. Due to the long incubation period of COVID-19, infectious diseases may have spread before the symptoms appear in the case[23]. Since the traditional method of collecting data is time-consuming, it takes less time than the method of extracting data, so this method can be used to find the source of the disease, so as to determine the area that needs attention.

City	Administrative region	The number of geographical location	City	Administrative region	The number of geographical location
Guangzhou	Yuexiu	11	Shenzhen	Futian	19
	Tianhe	10		Luohu	18
	Qiewan	10		Longgang	16
	Haizhu	7		Yantian	2
	Fanyu	1		Baoao	1
	Huangpu	Unknown	Huizhou	Huidong	5
	Conghua	Unknown		Boluo	4
Foshan	Shunde	4		Huicheng	1
	Nanhai	3		Huiyang	Unknown
	Chancheng	2	Zhuhai	Xiangzhou	15
	Sanshui	Unknown		Jinwan	1
Meizhou	Meijiang	2		Doumen	Unknown
	Meixian	Unknown	Jiangmen	Pengjiang	3
	Pingyuan	Unknown		Jianghai	1
	Dapu	Unknown		Heshan	1
	Fengshun	Unknown		Xinhui	Unknown
	Jiaoling	Unknown	Jieyang	Puning	Unknown
	Wuhua	Unknown		Jiexi	Unknown
Maoming	Binhai	Unknown	Qingyuan	Qingcheng	2
	Dianbai	Unknown		Qingxin	Unknown
	Huazhou	Unknown		Yangshan	Unknown
	Xinyi	Unknown		Yingde	Unknown
Shantou	Chaoyang	16		Lianshan	Unknown
	Nan’ao	2	Shanwei	Luhe	1
	Longhu	1		Lufeng	1
	Jingping	1	Yangjiang	Yangxi	1
	Chaonan	Unknown		Yangchun	Unknown
	Chenghai	Unknown		Yangdong	Unknown
Zhanjiang	Potou	2		Jiangcheng	Unknown
	Xiashan	2	Zhaoqing	Guangning	Unknown
	Chikan	1		Dinghu	Unknown
	Lianjiang	1		Duanzhou	Unknown
	Wuchuan	Unknown		Gaoyao	Unknown
	Leizhou	Unknown		Sihui	Unknown
	Development zone	Unknown	Zhongshan	Development zone	Unknown
Heyuan	Yuancheng	1		Shiqi	1

Table 8: Epidemic sites in Guangdong Province.

***Unknown:** The cities and administrative regions have appeared in the list, but there is no specific location.

Limitations

Although the conditional random field model is used in this paper to realize the entity recognition of the epidemic location, the rule-dependent model is used to extract the entity relationship of the COVID-19 pneumonia patient’s itinerary. But the following aspects still need to be improved:

a. The conditional random field model can rely on the information in this article for a long distance to improve the recognition accuracy, but this also increases the cost of the model and makes the recognition efficiency inefficient. The conditional random field model should be appropriately improved in the future, so that the efficiency can be improved while ensuring accuracy.

b. In the process of entity recognition and relationship extraction, the entity pairs are identified in sentence units, so that once two related entity pairs exist in two sentences, or after a complex pronoun is used in the sentence it will cause mistakes in relationship extraction. In the future, we should need to increase the research on the pronoun entity.

Conclusion

The recognition of named entities has a wide range of applications in various fields of natural language processing, which is the basic work of processing text. This paper proposes a named entity recognition based on conditional random field model and a relationship extraction method based on rule matching to achieve a task of recognition, extraction and classification of place names. This article mainly completed the following work:

a. Firstly, we use web crawler technology to download 366 epidemic websites to obtain unstructured data. In the face of the different webpage organization forms between different websites on the Internet, the fixed search mode cannot efficiently crawl data. How to improve crawler performance by integrating crawling rules remains to be studied;

b. Then, we use the trained conditional random field model to test the epidemic text. Conditional random field, as a relatively excellent method of machine learning, has achieved good results in entity recognition. This article firstly starts from the theoretical aspect, elaborate on the model derivation, training algorithm, and labeling methods of the conditional random field model, which lays a solid theoretical foundation for further research in the future;

c. Finally, we use rule-based methods to extract and classify place words into four categories, and obtained structured epidemic sites data. In future work, we need to add more features to the relationship extraction rules to improve the classification accuracy. In the future, we can also combine named entity recognition and relationship extraction to create a new model and simplify the process.

Conflict of Interest

We have no conflict of interests to disclose and the manuscript has been read and approved by all named authors.

Acknowledgement

This work was supported by the Philosophical and Social Sciences Research Project of Hubei Education Department (19Y049), and the Staring Research Foundation for the Ph.D. of Hubei University of Technology (BSQD2019054), Hubei Province, China.

References

1. Guo YR, Cao QD, Hong ZS, Tan YY, Chen SD, et al. (2020) The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak - an update on the status. Mil Med Res 7: 11.

2. Chen SD, Ouyang XY (2020) Overview of named entity recognition technology[J/OL]. Radio Commun Tech 1-11.

3. Rau LF (1991) Extracting company names from text[c]// proceedings of the seventh IEEE conference on artificial intelligence application. IEEE 1: 29-32.

4. Rathaparkhi A (1996) A maximum entropy model for part-of-speech tagging. Conference on Empirical Methods in Natural Language Processing pp. 133-142.

5. Mccallum A, Freitag D, Pereta FCN (2000) Maximum entropy markov models for information extraction and segmentation. ICML 17: 591-598.

6. Lafferty J, Mccallum A, Pereira FCN (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the 18th International Conference on Machine Learning pp. 282-289.

7. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, et al. (2011) Natural language proceeding(almost) from scratch. J Mach Learn Res 12: 2493-2537.

8. Pan XY, Chen L, Yu HM, et al. (2020) Survey on research of topic crawling technique. Appl Res Comput 37: 961-972.

9. Yu SW, Duan HM, Zhu XF, Bin S (2002) The Basic Processing of Contemporary Chinese Corpus at Peking University Specification. J Chinese Inform Process 16: 49-64.

10. Tan KK (2011) Rule-based Chinese address segmentation and matching methods. Qingdao: Shandong University of Science and Technology.

11. Du P, Liu Y (2011) Recognition of Chinese place names based on ontology. J. Northeast Univ Nat Sci 47: 87-93.

12. Qiu SAY, Wang FY (2011) Study on automatic recognition of Chinese location names based on statistical method. Comput Tech Develop 21: 35-38.

13. Tang XR, Chen XH, Zhang XY (2010) Research on toponym resolution in Chinese text. Geo Inform Sci Wuhan Univ 35: 930-935.

14. Aaron LFH, Derek FW, Lidia SC (2013) Chinese named entity recognition with conditional random fields in the light of Chinese characteristics. LP&IIS, Springer, Warsaw.

15. Wei Y, Li HF, Hu DL, LI X (2018) A method of Chinese place name recognition based on composite features. Geo Inform Sci Wuhan Univ 4: 17-23.

16. Kan Q (2015) Research and Application of CRF on Named Entity and Entity Relationships Based on Recognition. Beijing Jiaotong University, Beijing.

17. Xu QY (2012) Joint learning of named entity recognition and relation extraction based on CRF. Shanghai Jiao Tong University, Shanghai.

18. Xu J, Zhang ZX (2008) The technical method analysis of typical relation extraction system. Digit Lib Forum pp. 13-18.

19. Aone C, Ramos-Santacruz M (2000) REES: A large-scale relation and event extraction system. Proceeding of 6th Applied Natural Language Processing Conference pp. 76-83.

20. Zhang M, Zhang J, Su J, Zhou G (2006) A composite kernel to extract relations between entities with both flat and structured features. International Conference on ACL, Sydney, Australia pp. 825–832.

21. Yi E, Lee GG, Song Y, Park SJ (2005) SVM-based biological named entity recognition using minimum edit-distance feature boosted by virtual examples. Natural Language Processing-IJCNLP 2004 pp. 807-814.

22. Liu Y, Li Y, Li ZL (2020) The Diffusion Characteristic of An Outbreak of 2019 Novel Coronavirus Disease (COVID-19) in Guangdong Province. Tropical Geo pp. 1-9.

23. Zhai PL, Liu XY, Duan R (2020) Real-time Regional Spread Analysis, Prediction and Early Warning of COVID-19 Epidemic. Acta Mathematicae Applicatioe Sinica 43: 295-309.

Herald Scholarly Open Access, 2561 Cornelia Rd, #205, Herndon, VA 20171, USA.
Tel: +1 202-499-9679; E-mail: info@heraldsopenaccess.us
http://www.heraldopenaccess.us/