Research Article

The Journal of Clinical Microbiology

Constrained Random theory for Rapid Identification of Epidemic-Related Websites in Covid-19 Media Reports

Guiyun Zuang, Zinming Hao

School of Science, Hubei University of Technology, Wuhan, Hubei, China

***Corresponding Author :** Guiyun Zuang, School of Science, Hubei University of Technology, Wuhan, Hubei, China

Received : July 08, 2023 ;
Accepted : July 09, 2023 ;
Published : August 11, 2023 ;

Abstract

Background: Following the early December 2019 COVID-19 outbreak in Wuhan, China, the Chinese government established a system for information disclosure. In relation to newly diagnosed cases of novel coronavirus pneumonia, more than 400 cities have released precise location details, including residential areas and places of stay. Based on elements of Chinese geographical names, we have established a rule-dependent model and a conditional random field model. The named entity identification and the automatic extraction of sites related to the epidemic are done, using Guangdong province as an example. This approach will help identify the epidemic's spread, stop and manage it, and buy more time for vaccine clinical trials.

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

Findings : According to the findings of the analysis using the conditional random field model and the rule-dependent model, the major cities in Guangdong Province where new cases of coronavirus pneumonia were confirmed in mid-February were Guangzhou, Shenzhen, Zhuhai, and Shantou. There are more epidemic sites in Guangzhou's Futian district compared to Huangpu and Conghua district. Futian District is something that Guangzhou City government representatives ought to be aware of.

Interpretation : In the middle of February, Guangzhou Province's governments at all levels took action to contain the outbreak in a number of ways. The model analysis's results lead us to conclude that, in order to prevent the disease from spreading to other nearby administrative regions, the administrative regions with a higher number of diagnosed locations should concentrate on controlling it there. This can be done by implementing measures like blockades and personnel flow control.

Keywords : *COVID-19; Place Name Recognition; Web Crawler; Part-Of-Speech Tagging; Conditional Random Field; Elements of Chinese Place Name*

Introduction

The COVID-19 virus first surfaced in Wuhan, Hubei, China, in early December 2019, and in just a few short months, it spread throughout the world [1]. China has been experiencing a crisis over the last few months, but official government data indicates that China has essentially stopped local transmission. While some nations have moved into China's early stages, conditions in other nations are currently worse than they were during the height of the epidemic in China. We believe that one of the reasons China is able to contain the epidemic is that both the national and local governments have high levels of information transparency, and the appropriate agencies promptly disseminate the most recent information regarding the outbreak.

Finding possible patients can be greatly aided by examining data such as the residential area or activity location of officially confirmed COVID-19 diagnosed cases, according to epidemiological research. Viral experts can use these data to create models of epidemic transmission that allow them to assess and forecast the infection source, transmission speed, transmission path, and propagation risk. This is because the community can use these data to implement targeted prevention and control, granting people the right to know and better personal protection. The residential area or activity location of the diagnosed case is typically expressed in the epidemic report of web pages in a variety of ways, including the page body, embedded text, and screenshots. The distribution of epidemic-related sites from these information sources must first be timely analysed. To do this, pertinent information must be taken out of unstructured data and organised into structured data. This task was primarily completed by hand in the past by looking for and categorising the relevant information in the text. There is a lot of work to be done, not much efficiency, and no punctuality. As named entity recognition technology has advanced over the past few years, the focus of this work has gradually shifted from manual to automatic extraction, which not only uses less money and human resources but also processes tasks more quickly. The process of locating named entities in text and classifying them into related entity types is known as named entity recognition [2]. Names of people, places, dates, and organisations are among the general entity types. Our primary goal is to recognise place names in Chinese. The close arrangement of Chinese characters in Chinese text, the use of multiple characters in sentences, and the lack of spaces between words make it more challenging to identify named entities. From the original rule and dictionary technique to the conventional statistical learning method to the present deep learning method, named entity recognition technology has evolved to increase recognition accuracy [3–7]. In terms of accuracy for several common entities, current technology has essentially advanced. This article's goal is to process the text data on the webpage in order to present the entities and relationships.

Methods

Data

We can gather diagnostic locations thanks to the abundance of news reports of newly diagnosed patients that have appeared on the Internet since the COVID-19 outbreak. This raw, unstructured data is available to us. We gather official announcements and news articles about the schedule of the diagnosed patients' activities related to Guangdong province between January 29, 2020, and February 19, 2020. 366 webpages across 152 media contain the data. Web crawler technology is used to obtain the original data, which consists of the news release's media, the main content of the webpage, and the relevant URL. A web crawler's fundamental function is to mimic a browser sending HTTP requests. After locating the relevant server, the crawler client uses the HTTP request protocol to send a request to the web server, downloads the web page, and concludes the crawling task of the crawler system [8]. Table 1 displays a portion of the data.

Text Data Pre-processing

Due to the pre-processing of text data, we must first ascertain whether any values in the data are missing before executing named entity recognition. Once we have verified that no values are missing, we must transform the data into a format that the model can process with ease. As an illustration, eliminate a few extraneous character strings, divide each press release into separate sentence units, remove the same sentence, etc. The Peking University uses the part-of-speech feature [9].

The Model

Using named entity recognition technology, we attempted to identify and extract place words from the collected unstructured text data. We then classified the identified place words based on a set of rules, dividing them into administrative regions, cities, provinces, and specific locations. In order to provide precise data for the epidemic development model built by researchers in the future, as well as to evaluate and forecast the source of infection, the rate of transmission, and the route of transmission, we lastly perform statistical analysis on the location data. Chinese place name recognition can be studied using three main approaches: rule-based, statistics-based, and deep learning-based. The rule-based approach is natural and intuitive, making it simple for people to comprehend and use. Rule writing, however, requires domain and language specific knowledge. The portability is also poor, covering all the modes is challenging, and the rules are more complex [10,11]. While statistics-based approaches are very portable and do not necessitate extensive language or domain knowledge, they do require manual corpus annotation and the selection of suitable statistical learning models and parameters [12–14]. In order to create an end-to-end model, deep learning-based techniques can automatically extract information from the input without the need for unduly complicated feature engineering [13]. This paper's content is mostly based on the first two methods because the amount of text data that may be gathered is limited in terms of both time and data.Named entity recognition has shown promise recently for a few restricted entity kinds. For instance, there is a notable recognition effect on the names of individuals, locations, and organisations in news corpuses. One could consider Chinese place name recognition to be a sequence labelling challenge. The place name entity identification process consists of identifying the right names from these word sequences. The place name is made up of several words arranged in a specific order. The hidden Markov model and the maximum entropy model are combined in the conditional random field model. and are applicable to the segmentation and labelling of sequence data [6]. Thus, the conditional random field model is an efficient way to solve the sequence labelling problem [15]. As the identification model for epidemic locations, we went with the conditional random field model.

Feature Selection for Chinese Location Recognition

The annotated corpus of the People's Daily from January 1998 served as the training set for this article. It makes use of Peking University's part-of-speech annotation set. All of the names of the people, places, and organisations are marked on the label set. These corpora consist of sentences that have been coarsely segmented; fine segmentation of these sentences is necessary to highlight the entity extraction features. The word's lexeme information can be fully expressed by the 5n-gram template. They are B (the named entity's beginning word), M (the named entity's middle word), E (the named entity's tail), S (the single word that makes up the named entity), and N (unnamed entity). Different labels are formed by the three sorts and combinations of entities. The word's location can be efficiently marked by the template, allowing the system to use the position feature to determine the word's boundary. The labelled labels can be seen in Table 2.

Feature Template Selection

The final recognition accuracy greatly depends on feature selection, which in turn affects the conditional random field. Theoretically, richer information can be obtained and a more accurate judgement of the current word can be made if more context information is gathered around it, that is, the larger the value of the observation window[16]. However, if the observation window is too big, the computation of too much data will impact the operation efficiency and inefficiently identify the model. The accuracy of recognition may suffer if the window value is too small because it will not be possible to fully utilise the relevant dependency information [16]. Selecting the appropriate feature template therefore starts with selecting the appropriate window size. In this post, the chosen window size is 2.The most fundamental characteristic that is more powerful than the character itself is called the base feature. Examples of base features include the current character or key, the first character's location in the pre-word, and the part of speech. The degree of discriminating can frequently be increased by part-of-speech features. Verb part-of-speech words are seldom utilised as named entities, whereas noun part-of-speech words are frequently employed as named things. Using the part-of-speech of the word in which it appears is the part-of-speech feature for a word-based entity tagging system. We determine the fundamental characteristics of the template based on these, as indicated in Table 3.

Whereas y denotes the label, word denotes the word, tag denotes the part of speech, and t indicates the position from which the feature is currently being extracted. Given the abundance of word combinations, many binary grammatical features go unused.

Entity Relationships

The following are the seven categories of relationships between entities: overall partial relationship relationships based on character, organisational structure dependency, manufacturing use, generic relationships, geographic location relationships, and metaphors [17]. More focus is placed on the identification of geographic location relationships because the relationship between locations in an outbreak press release is the subject of this work.The geographic location relationship is categorised in detail based on the features of the identified text itself (Table 4). Provinces, cities, administrative regions, and geographical locations are the four categories into which the geographical location relationships that may be involved in the entry are separated. A relationship's head word can be either a verb or a noun.

The identified entity relationships need to be categorised by the system. The specific division is shown in Table 5.

Entity Relationships Extraction Method

Relational semantics recognition is always changing and can be categorised into two categories: machine learning and rule matching methods. During relationship identification, the rule template is compared to the statement using the rule template matching method, which is predefined. The entity in the statement has the relationship indicated in the template Attributes if the statement satisfies the characteristics of the characteristic template [18]. The drawback is that it has poor portability and takes a long time to write a large number of feature templates, requiring more experienced linguists [19]. Using a variety of pattern recognition feature models, the machine learning method computes entity relationship features and weight values in sentences using associated algorithms. For handling entity relationships, there are currently two widely used categories of machine learning techniques: kernel-based techniques and feature vector-based techniques [20, 21]. Our research aims to carry out location extraction. The geographic location relationship's feature template is highly portable and comparatively fixed. Therefore, in order to extract the identified place words for relationship, we will employ the rule-based matching method. Rule-making and corpus pre-processing have three facets.

Corpus Pre-processing

Word segmentation and entity recognition are the primary pre-processing steps used on the corpus, converting the sentences into a stream of words with entity identification. Sentences with fewer than two place name entities in the text are filtered out, and sentences with two or more place name entities are used as the recognition corpus. This is because entity relationship extraction involves a relationship between two entities. Journalist Certain sentences might only contain information about the administrative region and not the province or city, for example, without a complete place name. Currently, we have to compile all of the Chinese province, city, and district names and create a dictionary structure. Figure 1 presents some of the data.Since Guangzhou, China is the focus of our study, we won't be taking into account the COVID-19 outbreak scenario overseas. As a result, we must compile the names of important international nations and locations from the Internet and create a dictionary foundation. Figure 2 presents some of the data.

Rule Making

By examining the structural features of epidemic location words, developing regular expression-based rules to extract location words identified by named entity recognition word-by-word, and organising the relevant location data that has been extracted into a suitable data structure for further processing.

Model Evaluation Criteria

The model's evaluation is conducted using the F1-score evaluation index. These three indicators have definitions for every kind of named entity and relationship extraction.

Overall Framework for Automatic Extraction of Chinese Place Names

This article's goal is to process the text data on the webpage in order to present the entities and relationships. The entity and relationship recognition module and the web page processing module make up the two core modules of the implementation process. Figure 3 displays the frame diagram for the automatic extraction of Chinese localities.

Results

Place Name Entity Recognition Result

This paper makes use of the corpus that People's Daily marked in January 1998, of which 80% is chosen as the training set, 20% is used as the closed test set, and the open test set is the news release about the COVID-19 outbreak that was crawled through the Internet. Table 6 displays the entity recognition results.

The experimental data shows that place name recognition yields results with higher accuracy. Both the open and closed training sets have potential F values of 0.771 and 0.870, respectively. The entity recognition results show that the following categories of incorrect place entity recognition predominate:

a. The text contains abbreviations for cities and provinces, and place names in unclear forms can be recognised. For instance, "Zhongshan" can refer to both a city in the province of Guangdong and an administrative district in the province of Liaoning;

b. Certain place names are used in more than one city. For instance, several cities' roads go by the name "Baojian Road." It can be challenging to figure out which city this route name belongs to when there are several cities mentioned in a phrase;

b. Words have distinct meanings depending on where they are placed. For instance, a town or building may be known by the name "Bajiao Tower";

c. The incorrect words appear in the recognition location when the entity label is incorrectly labelled. For instance, the algorithm classifies the words "Sputum," "when getting on the train," "from the day," "and," and "more" as place names even though they are not.

Place Entity Extraction Results

The central indicator has a significant influence on the location extraction link. Different kinds of central words will cause the entity relationship pairs to form different semantic relationships when other features are the same (Table 7). Consequently, the incorrect entity centre must be identified by the central indicator thesaurus. To increase the recognised features, the centre is extracted and added to the central indicator thesaurus.

Errors that commonly occur in entity relationship extraction are as follows:

a. Certain place name words use pronouns in place of entity words. For instance, the terms "province" and "city" have been dropped from a large number of provinces and cities;

b. The text contains two identical entities with distinct connection classes that are unable to determine precedence. As an illustration, "Jilin" is the name of both a province and a city;

c. A single sentence mentions more than one place. There are several place words in a sentence, identifying the link between the two is difficult, and judging the subordinate relationship is limited to the word's placement within the sentence.

Final Results

Guangdong province's locations are chosen and arranged based on the data in the city column because certain provinces on the form are left blank. Table 8 presents the findings.

Table 7 suggests that the major cities within the COVID-19 epidemic area in Guangdong Province are Guangzhou, Shenzhen, Zhuhai, and Shantou. Among them, the epidemic areas in Shenzhen are in Futian, Luohu, and Longgang; in Zhuhai, they are in Xiangzhou; in Guangzhou, they are centred in Yuexiu, Tianhe, and Qiewan; and in Shantou, they are in Chaoyang. While Heyuan's epidemic area is relatively small, Guangzhou has the largest.

Discussion

Without a doubt, there are a variety of perspectives from which we can examine the evolution of COVID-19 in a particular region. This article primarily examines the degree of epidemic spread. We think that the number of epidemic outbreak locations in a region can be used to quantify the spread of epidemics. Research by other academics has revealed that cities with greater levels of economic development and a greater number of migrant populations experience a higher number of imported cases compared to other cities [22]. Greater levels of economic development are primarily found in the Pearl River Delta's core urban agglomeration, which is located in the northern region of the province of Guangdong. Guangzhou is one of these cities. hence there has been a greater Guangzhou pandemic spread. There aren't many imported cases in Heyuan City because it's in a rural, mountainous location with less transportation. If the epidemic is thought to be spreading swiftly, action should be taken as soon as the COVID-19 infectious disease reaches its early stages. Owing to COVID-19's protracted incubation period, infectious diseases might have spread before a case's symptoms manifested [23]. The traditional method of gathering data takes longer than the method of extracting data, so it can be used to identify the area that requires attention and identify the source of the disease.

Limitations

In this work, the entity relationship of the COVID-19 pneumonia patient's itinerary is extracted using the rule-dependent model, even though the conditional random field model is utilised to achieve the entity recognition of the epidemic location. However, the following areas still require improvement:

a. Long-range reliance on the information in this article can enhance recognition accuracy for the conditional random field model, but at the expense of increased model cost and inefficient recognition efficiency. Future improvements to the conditional random field model should be made in a way that maintains accuracy while also increasing efficiency.

b. The entity pairs are identified in sentence units during the entity recognition and relationship extraction processes. This means that errors in relationship extraction may occur if two related entity pairs are found in two sentences or if a whole pronoun is used in a sentence. We should have to do more research on the pronoun entity in the future.

Conclusion

The fundamental task of text processing, known as natural language processing, has many applications, one of which is the recognition of named entities. In order to accomplish the task of place name recognition, extraction, and classification, this paper proposes a named entity recognition method based on conditional random field model and a relationship extraction method based on rule matching.

Mainly, this article finished the following tasks:

a. To start, we download 366 epidemic websites using web crawler technology in order to collect unstructured data. Because different websites on the Internet have different organisational structures, the fixed search mode is unable to efficiently crawl data. It is still unclear how to incorporate crawling rules to enhance crawler performance;

b. Next, we test the epidemic text using the learned conditional random field model. Conditional random fields are a rather good machine learning technique that have shown promise in entity recognition. In order to provide a strong theoretical framework for future research, this article begins with the theoretical side of things. It then elaborates on the model derivation, training algorithm, and labelling techniques of the conditional random field model; Ultimately, we extract place terms using rule-based techniques, categorising them into four groups, and obtaining structured data on epidemic sites. To increase the classification accuracy, we must incorporate additional features into the relationship extraction rules in further work. In the future, relationship extraction and named entity identification can be combined to provide.

Acknowledgement

The Hubei Education Department's Philosophical and Social Sciences Research Project (19Y049) and the Staring Research Foundation for the Ph.D. at Hubei University of Technology (BSQD2019054), located in Hubei Province, China, provided funding for this work.

References

1. Guo YR, Cao QD, Hong ZS, Tan YY, Chen SD, Jin HJ et al. The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak-an update on the status. Mil Med Res. 2020 ;7(1):1-10.
2. Chen S D, Ouyang X Y. Overview of named entity recognition technology [J/OL]. Radio Communications Technology: 1-11. http://kns.cnki.net/kcms/detail/13.1099.TN.20200414.1436.002.html [Last accessed on 13 June 2020]
3. Rau LF. Extracting company names from text [C]// Proceedings of the Seventh IEEE Conference on Artificial Intelligence Application. IEEE. 1991,1:29-32.
4. Rathaparkhi AA. Maximum entropy model for part-of-speech tagging [C]//. Conference on Empirical Methods in Natural Language Processing, 1996;133-142.
5. Mccallum A, Freitag D, Pereta FCN. Maximum entropy markov models for information extraction and segmentation [C]// ICML. 2000;17:591-8.
6. Lafferty J, Mccallum A, Pereira FCN. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]// Proceedings of the 18th International Conference on Machine Learning (ICML 2001). 2001:282-9.
7. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. J Machine Learn Res. 2011;12::2493-537.
8. Pan XY, Chen L, Yu HM. Survey on research of topic crawling technique [J]. Application Research of Computers. 2020;37(4):961-72.
9. Yu SW, Duan HM, Zhu XF. The basic processing of contemporary Chinese corpus at peking university specification [J]. J Chinese Info Processing. 2002(5):49-64.
10. Tan KK. Rule-based Chinese address segmentation and matching methods [d]. Qingdao: Shandong Uni Sci Technol. 2011.
11. Du P, Liu Y. Recognition of Chinese place names based on ontology [J]. J Northwest Normal Uni. (Natural Science). 2011:47(6):87-93.
12. Sha Q, Yuan A, Fuyan W, Hai-yan DI. Study on automatic recognition of Chinese location names based on statistical method. Comp Technol Develop. 2011;21(11):35-8.
13. Tang XR, Chen XH, Zhang XY. Research on toponym resolution in Chinese text [J]. Geomatics Info Sci Wuhan Uni. 2010;35(8):930-5.
14. Aaron LFH, Derek FW, Lidia SC. Chinese named entity recognition with conditional random fields in the light of Chinese characteristics [M]. LP&IIS2013, Warsaw: Springer. 2013;57-68.
15. Wei Y, Li HF, Hu DL. A method of Chinese place name recognition based on composite features [J]. Geomatics Info Sci Wuhan Uni. 2018;43(1):17-23.
16. Kan Q. Research and application of CRF on named entity and entity relationships based on recognition [D]. Beijing: Beijing Jiao Tong Uni. 2015.
17. Xu QY. Joint learning of named entity recognition and relation extraction based on CRF. Shanghai: Shanghai Jiao Tong Uni. 2012.
18. Xu J, Zhang ZX. The technical method analysis of typical relation extraction system [J]. Digital Library Forum. 2008;(9):13-18.
19. Aone C, Ramos-Santacruz M. REES: A large-scale relation and event extraction system [C]// Proceeding of 6th Applied Natural Language Processing Conference. 2000:76-83.
20. Zhang M, Zhang J, Su J. A Composite kernel to extract relations between entities with both flat and structured features [C]// Int Conference on Acl. DBLP. 2006.
21. Yi E, Lee GG, Song Y, Park SJ. SVM-based biological named entity recognition using minimum edit-distance feature boosted by virtual examples. Natural Language Processing-IJCNLP 2004. 2005;807-14.
22. Liu Y, Li Y, Li ZL. The diffusion characteristic of an outbreak of 2019 novel coronavirus disease (COVID-19) in Guangdong province [J/OL]. Trop Geography. 1-9.
23. Zhai PL, Liu XY, Duan R. Real-time regional spread analysis, prediction and early warning of COVID-19 epidemic [J]. Acta Mathematicae Applicatioe Sinica, 2020;43(2):295-309.