A Semantic Trajectory Mining System Based on Deep Neural Network

Xinyuan Zhang, Jiying Peng, Xinqi Zhang

School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051
E-mail: 1050679160@qq.com

Abstract: At the beginning of 2020, epidemic of COrona VIrus Disease 19 (COVID-19) broke out. During the epidemic prevention and control, artificial intelligence, big data and other technologies have become powerful weapons against the epidemic, and have been widely used in the fields of epidemic tracing, confirming virus transmission path, resource allocation and so on. In this study, BiLSTM-CRF model, Bootstrap and Tornado frameworks are used to implement a neural network-based semantic trajectory mining system for the COVID-19 patients. On the basis of collecting the data published by the health committees of various provinces and cities, the semantic trajectories of the patients are extracted to ensure the accuracy of the data and then establish mapping relationship between the real space and the text description of the trajectories of the patients, while taking the time and space factors into account and excavating the dynamic changes of the patients.

Key Words: Bidirectional neural network; Named entity recognition; BiLSTM-CRF; API

2021 33rd Chinese Control and Decision Conference (CCDC) | 978-1-6654-4089-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/CCDC52312.2021.9602180

1 Background

At the beginning of 2020, epidemic of COVID-19 broke out. How to use artificial intelligence, big data, cloud computing and other advanced technologies to help prevent and control the epidemic has become what many organizations are working on together. For example, China Unicom Epidemic Prevention and Control Population Big Data Platform establish large database based on the Intelligent Footprint, and the independently developed Intelligent Footprint location trajectory space-time search algorithm, adjoint relationship algorithm together with the multi-epidemic areas SEIR model, provide epidemic crowd flow monitoring, epidemic risk warning, epidemic situation prediction, work resumption crowd flow monitoring and other functions to support the data requirement for epidemic prevention and control, judgment, analysis and decision-making. However, the research on big data and artificial intelligence technology applied to the prevention and control of sudden acute infectious diseases is still in its infancy in our country. The application prospect in this field can bring breakthrough progress to the prevention and control of sudden human infectious diseases, so it is very worthy of in-depth research.

At present, the existing approaches on the network that can be used by the masses are hard to satisfy public's need of obtaining exact and integrated information of the trajectories of the patients and the risk levels of different areas. For example, the peer inquiry tool of the COVID-19 patients on Sina Weibo can only inquire information about patients who travel by different means of transportation like airplane or train. The trajectory information of the COVID-19 patients provided in the public information of the provincial and municipal health committees cannot be directly searched and obtained, people need to consult the public data links of the provincial and municipal health committees one by one to identify high-risk areas. Besides, the data sources of some WeChat public accounts cannot guarantee their accuracy.

This study aims at making the masses obtain the correct travel information of the patients in quickly and effectively. The main contributions of this study are as follows:

(1) A semantic trajectory mining system based on neural network is designed, and a simple search tool is realized by combining the confirmed and suspected cases' trajectory data and the close contact group trajectory data disclosed by the National Health Commission.

(2) A named entity recognition model is built using BiLSTM-CRF model, and this model is applied to the task of patient trajectory analysis for the first time. In some scenes, the system built in this study can obtain accurate trajectory information and realize precise map display.

(3) The system designed in this study is helpful to the control and prevention of the epidemic situation, and can help the masses understand the development of the epidemic situation scientifically, track the source, analyze the areas with higher risk, and realize the risk early warning for cities, counties, streets and communities. On the basis of the existing data, Baidu Map API is used for epidemic risk assessment, which can provide intuitive reference for epidemic prevention and control.

2 Introduction to Related Technology

2.1 Recurrent Neural Network

When processing sequence type data, the most commonly used neural network is RNN (Recurrent Neural Network). Compared with the traditional neural network where the neurons are only connected between the layers, in RNN, the neurons in the same hidden layer are also connected to each other. Each RNN unit is connected in a chain form of the same neural network module[1]. This repeating module is

This work was supported by University of Science and Technology Liaoning College Student Innovation and Entrepreneurship Training Program.

composed of very simple modules in the standard RNN. However, due to the problem of gradient disappearance during the back propagation of RNN, the gradient gradually shrinks during the back propagation process, and if the gradient signal becomes very small, the learning cannot be sufficient. Moreover, RNN only has short-term memory, so it is difficult for the model to transmit long sequence information from the earlier time step to the later time step.

2.2 Long Short-term Memory Neural Network

In order to solve the problem of gradient disappearance, gradient explosion and long-term dependencies of hyperbolic tangent(tanh) function in RNN, Guillaume Lample et al. propose long short-term memory neural network (LSTM)[2] by changing the propagation structure of RNN. As a variant of RNN, LSTM network brings addition operation into the network through exquisite gate control, which effectively alleviates the problem of gradient disappearance. In order to have long-term memory, on the basis of RNN network, LSTM introduces a pathway called cell state which contains an input and a output. The introduction of cell state makes LSTM not only considers the latest state, its cell state will also judge the state that should be left or forgotten. LSTM has three gates to protect and control cell state, namely forget gate, input gate and output gate.

2.3 BiLSTM

Different from LSTM, bidirectional LSTM considers both past information and future information, obtains long-term memory of the past through the process of forward extraction, and predicts the future through the process of backward extraction. "Bidirectional" means the BiLSTM model has a forward input sequence and a backward input sequence, and take the conbination of the outputs of the two directions as the final output result[3]. Therefore, the basic idea of BiLSTM is two LSTMs that one has a forward input sequence and the other has a backward input sequence, both of which are connected to the same output layer, and taking the conbination of two outputs as the final output result.

2.4 Conditional Random Field

Conditional Random Field (CRF)[4] is an undirected graph learning model proposed by Laferty et al. in 2001, on the basis of Maximum Entropy Model and Hidden Markov Model. It is a conditional probability model used to label and segment ordered data. Let $G = (V,E)$ be an undirected graph, and $Y = \{Y_v | v \in V\}$ be a set of random variables Y_v indexed by nodes in G. When X is determined, if every random variable Y_v follows Markov property, that is, $P(Y_V|X,Y_U,U \neq V) = P(Y_V|X,Y_U,U\sim V)$, where $U\sim V$ indicates that U and V are adjacent edges, then (X,Y) constitutes a conditional random field. CRF is to calculate the joint probability of the whole tag sequence under the condition of giving observation sequence needed to be marked, i.e. to find the conditional distribution: $P(Y|O)$, rather than defining the distribution of the next state under the condition of giving the current state(HMM), i.e. to find the joint distribution: $P(Y,O)$.

2.5 Visualization

This study uses the JavaScript API of Baidu Map for visualization. It is a set of application program interfaces written in JavaScript. It can build mapping programs with strong interactions and various functions in websites, supports developing browser-based map application on PC and mobile terminals, and supports map development with HTML5 features[5]. Baidu Map uses BD09 Baidu coordinate system and are encrypted again based on GCJ02 coordinate system. In Baidu Map, longitude and latitude coordinates are represented by BD09ll, Mercator metric coordinates are represented by BD09MC, and WGS84 coordinates are used to uniformly represent maps of non-China regions. The coordinate system contains longitude and latitude spherical coordinate system, Mercator plane coordinate system and block numbering system.

By using Baidu Map API, the location data uploaded to LBS cloud by users can be rendered into layers in real time, and then superimposed on the map through CustomLayer object. LBS cloud supports users to store POI data, in which the storage fields include attribute information such as name and address besides longitude and latitude coordinates. The interface for reading LBS cloud data is provided by the CustomLayer class, and can automatically render user data to generate data layers, while providing the function of clicking the overlay layer to return POI data.

3 Module Implementation

3.1 Data Processing

There is no complete ready-made dataset of COVID-19 Re on the Internet, and the data of provincial and municipal health committees are difficult to obtain through crawler and other technologies, so this paper manually collects text data of COVID-19 patients' trajectory information published by provincial and municipal health committees to ensure the accuracy and comprehensiveness of the data. Considering the characteristics of the dataset, this paper proposes to combine regular expression and BiLSTM-CRF model to process the data. Firstly, this paper uses functions from regular expression processing library to cut the raw text data into sentences and separates them with a tab character. Because the expected output of this system is a <date-location> sequence, while BiLSTM-CRF model has poor recognition effect on time entity, this paper proposes to use regular expression to screen out sentences beginning with date according to the characteristics of the raw data set, then use jieba, a Chinese word segmentation module, to cut each sentence into words.

Dirty data should be removed first when reading data. In the training dataset, dirty data mainly includes line break, Chinese punctuation, tab characters and continuous spaces. This article uses the sub function in regular expressions processing library to delete dirty data from the dataset. Then save the date information in an individual file corresponding to the output file of the result of the named entity recognition model one by one. At the same time, the cleaned dataset is fed into the BiLSTM-CRF model trained by the People's Daily corpus for named entity recognition.

The results are converted from labels to Chinese and then filtered to reserve the named entities of location and organization. The files of date and the files of organization and location are fused into json format files, which are used as input data of the web end.

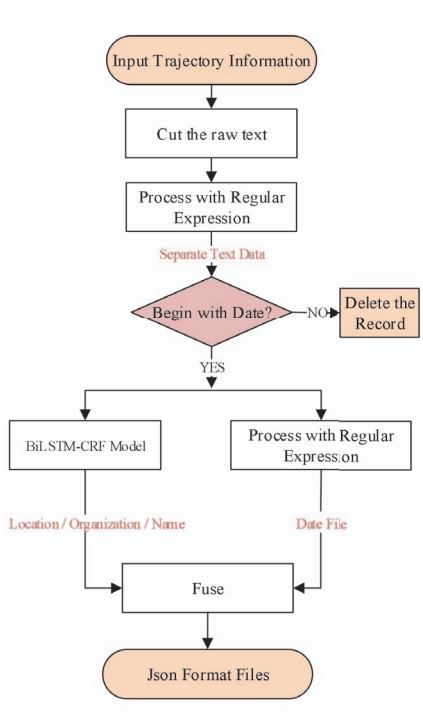

Fig 3.1. Data Processing Module Flow Chart

3.2 BiLSTM-CRF Model Framework

This study applies BiLSTM-CRF model to the location mining model for the first time, combining a BiLSTM network with a CRF network, as shown in Figure 3.2. The network can effectively use past input features and future input features through Bi-LSTM layer and construct sentence-level location tag information through CRF layer. The CRF layer is presented by a linear chain which continuously connecting the results of the output layer. At the same time, the CRF layer uses the state transition matrix as the parameter, and the objective function also considers the state characteristic function of the input.

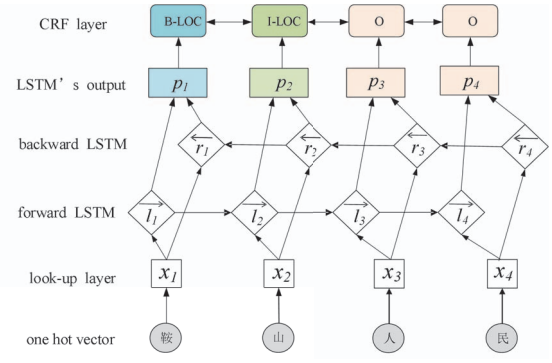

Fig 3.2. BiLSTM-CRF Structure Diagram

CRF assumes that each input node forms a first-order Markov random field. The model uses SGD to update the model parameters during the training process. Through such layer, past and future tags can be efficiently used to predict current location tags.

3.3 Principle of BiLSTM-CRF Model

This study assumes that a text message contains n words, and takes sentence X as the unit to process the text message of the patient's trajectory information:

$$X = (x_1, x_2, \ldots\ldots, x_n) \qquad (3.1)$$

where x_i represents the id of the i-th word of the current trajectory information sequence in the dictionary, and then one-hot vector with dictionary size can be obtained for each word in the current trajectory information sequence. The semantic trajectory mining model is constructed with following three layers: Word embeddings layer, Bi-LSTM layer and CRF layer. The first layer is the Word embeddings layer, which is the input data processing layer in the model. The one-hot vector of each word x_i in the current trajectory information sequence maps the word in high-dimensional sparse space to character embedding in low-dimensional continuous space by using pre-trained or randomly initialized embedding matrix, $x_i \in R^d$, where d refers to the dimension of embedding. Each word is represented by an embedding vector of the same dimension and passed into the bidirectional LSTM layer as input data for the following operations. The second layer of the model is Bi-LSTM layer, which automatically extracts sentence features by combining past features (extracted by forward process) and future features (extracted by backward process). The general process is to extract high-dimensional features from the i-th word of the current trajectory information sequence first, and then obtain the probabilities of the features to each label in the label set by learning the maps of features to labeling results. Finally, the CRF layer is used to obtain a globally optimal chain of labels for the current given sequence. Specifically, in this study the char embedding sequence corresponding to each word in the current trajectory information sequence is fed as input data at each time step of bidirectional LSTM, the output hidden state sequence $(\overrightarrow{l_1}, \overrightarrow{l_2}, \ldots\ldots, \overrightarrow{l_n})$ is extracted by the forward process of LSTM, and the output hidden state sequence $(\overleftarrow{r_1}, \overleftarrow{r_2}, \ldots\ldots, \overleftarrow{r_n})$ is extracted by the backward process of LSTM. The output at each time $h_t = [\overrightarrow{l_t}; \overleftarrow{r_t}] \in R^m$ is formed by concatenating these two hidden states, the complete hidden state sequence representation of the concatenating sequence is shown in formula (3.2).

$$(h_1, h_2, \ldots\ldots, h_n) \in R^{n \times m} \qquad (3.2)$$

After this operation, the hidden state vector obtained by BiLSTM layer will be mapped from m dimension to k dimension by connecting to a full connection layer, that is, a score matrix $P = (p_1, p_2, \ldots\ldots, p_n) \in R^{n \times k}$ will be automatically extracted by bidirectional LSTM network, where k represents the number of labels of different classes in the label set, and each dimension p_{ij} of $p_i \in R^k$ is regarded as the score of the word x_i under the j-th label in

the current trajectory information sequence, as shown in the yellow area in Figure 3.3.

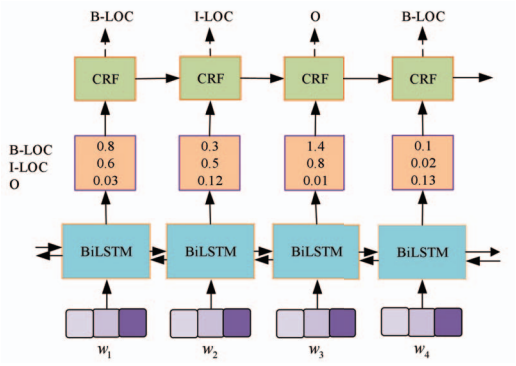

Fig 3.3. BiLSTM Prediction Schematic

As shown in the above figure, the output (yellow area) of Bi-LSTM layer is the predicted score of each tag corresponding to each word x_i at each time step. The third layer of the model is CRF layer. CRF combines the characteristics of maximum entropy model (MEMM) and hidden Markov model (HMM), selects all features for global normalization, defines a global score, considers the transistion between tag results, and considers contextual information. That is, adjacent tags are used in predicting the current tags, so the global optimal result can be obtained, as shown in Figure 3.4.

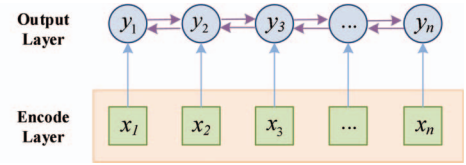

Fig 3.4. CRF Layer Prediction Schematic

CRF predicts the current tag by using contextual information, so that tags the location sequence at sentence level, and obtain the globally optimal location tag sequence. For a location tag sequence $Y = (y_1,y_2,.....,y_n)$, the length of Y is equal to the length of the original sentence sequence $X = (x_1,x_2,.....,x_n)$. The calculative process that obtaining the label sequence Y by the model with sentence X is shown in formula (3.3).

$$Score(X,Y) = \sum_{i=0}^{n} A_{y_i,y_{i+1}} + \sum_{i=1}^{n} P_{i,y_i} \quad (3.3)$$

where A is a transition matrix of size $(k+2) \times (k+2)$, A_{ij} represents the transition score from the i-th tag to the j-th tag. y_0 and y_n represent the starting state and ending state of the sentence, which are added to the set of possible tags. The purpose of adding two here is to add the starting state and ending state to the beginning and ending state of the sentence. The score of the whole text sequence is equal to the sum of the scores of each position, that is, the transition matrix A of CRF and P_i output by LSTM determine the score of each position together. By using the activation function softmax for normalization operation, The calculative process of obtaining probability of generating sequence Y on all possible tag sequences is shown in formula (3.4).

$$p(Y|X) = \frac{exp\,(Score(X,Y))}{\sum_{\tilde{y} \in Y_X} exp\,(Score(X,\tilde{y}))} \quad (3.4)$$

During the training process, the maximum logarithmic likelihood function is used to calculate the correct tag sequence, as shown in formula (3.5) (3.6).

$$\log p(Y|X) = Score(X,Y) - \log(\sum_{\tilde{y} \in Y_X} \exp\,(Score(X,\tilde{y}))) \quad (3.5)$$

$$\log p(Y|X) = Score(X,Y) - \log_{\tilde{y} \in Y_X} add\,Score(X,\tilde{y}) \quad (3.6)$$

where Y_X represents all possible tag sequences for a sentence X, even if this sequence does not meet the tag format. Viterbi algorithm is used to predict the optimal location tag sequence. On the basis of the score matrix P and the transition matrix A, the dynamic programming Viterbi algorithm uses formula 3.3 to obtain the tag sequence Y with sentence X, then calculating the tag sequence with the largest probability as the final tag result of the model.

$$y^* = arg_{\tilde{y} \in Y_X} maxScore(X,\tilde{y}) \quad (3.7)$$

According to formula (3.7), the model can obatin the location tag sequence in a more effective way. During the training process, CRF layer can automatically learn the constraint rules of labels from the training data to ensure that the predicted labels are legal. Based on some constraint rules, such as the first word in each trajectory information always starts with the label "B-loc" or "O" instead of "I-", the probability of illegal sequences appearing in the semantic trajectory mining model is greatly reduced.

4 Experiments

4.1 Environment Configuration

Windows 10 is used as the main development environment and testing environment. Python is used as the development language, PyTorch framework is used to build BiLSTM-CRF model, Anaconda is used as the package manager, Bootstrap is used as the front-end framework and Tornado is used as the back-end framework on the web end. MySQL database is used to store the text and data obtained by named entity recognition process.

4.2 Experimental Results

In order to test the accuracy of the model and have an intuitive display of the whole training process, this paper uses Python's Matplotlib library to draw the change of the return value of the loss function during training and validation process. People's Daily Chinese corpus is used as the training set to train BiLSTM-CRF model which contains more than two million results of Chinese named entity recognition tags. In this paper, we have trained a total of 20 epochs, each poch has 67 outputs.

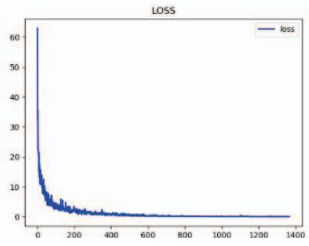

Fig 4.1. Change of Loss

It is shown that the training loss was as high as over 60 at the beginning, which shows that the dispersion degree of the data is very large. After 120 iterations, the gradient descent of loss value slowed down considerably, almost became stable around 0.1 after the 1000th iteration, and finally became stable at 0.05. The model built in this paper has good convergence after several iterations, which shows that the selection of parameters and the principle of the model are correct.

This paper uses Accuracy, recall rate and F1-score as evaluation indexes. These three indexes apply to not only the two-classification problem, but also the multi-classification problem in this paper. Since there is no obvious change in evaluation indexes after 10 epochs in the training process, only the changes of indexes in the first 10 epochs are shown.

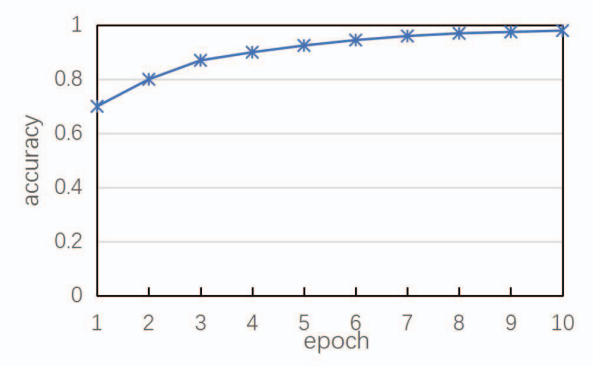

Fig 4.2. Accuracy of Validation Dataset

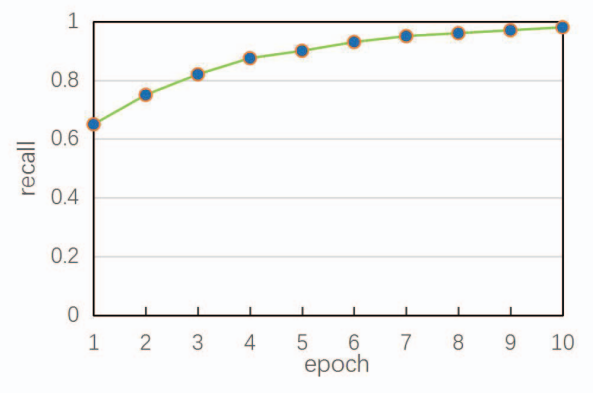

Fig 4.3. Recall Rate of Validation Dataset

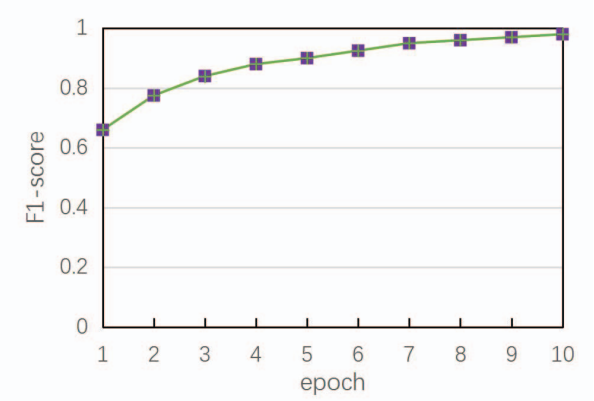

Fig 4.4. F1-score of Validation Dataset

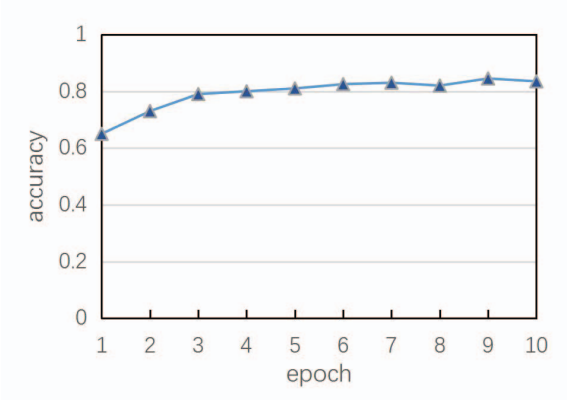

Fig4.5. Accuracy of Test Dataset

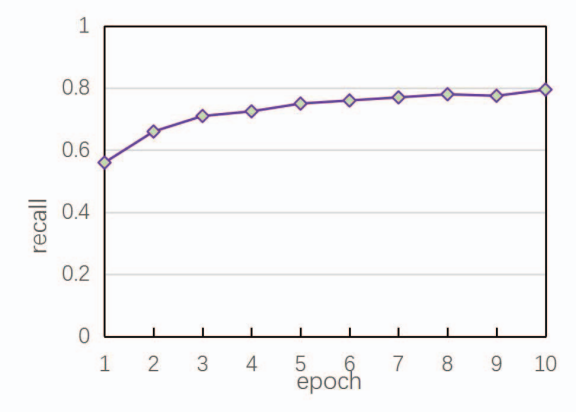

Fig 4.6. Recall Rate of the Test Dataset

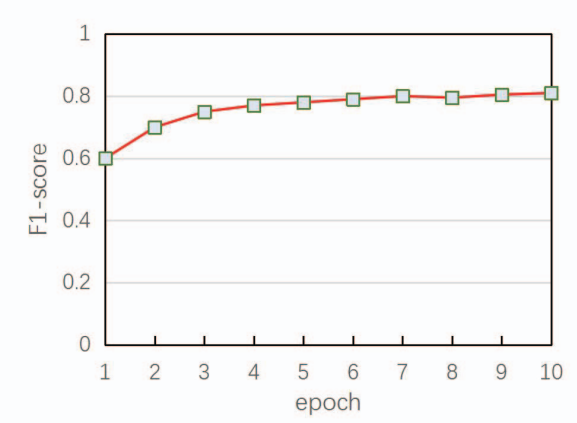

Fig 4.7. F1-score of Test Dataset

After 20 epochs, the scores of this model is:
Validation dataset:

Precision: 0.97

Recall: 0.99

F1: 0.98

Test dataset:

Precision: 0.84

Recall: 0.79

F1: 0.81

The scores are relatively high, indicating that the overall performance of the model is good.

4.3 Implementation of Visualization System

Bootstrap is used as the front-end framework and Tornado as the back-end framework in the web end of this system. The data in the front end is transferred to the 8001 port monitored by Tornado through POST, and the input text is processed into json format that can be recognized by the front end just like how the data is previously processed. Figure 4.8 shows the form submit page.

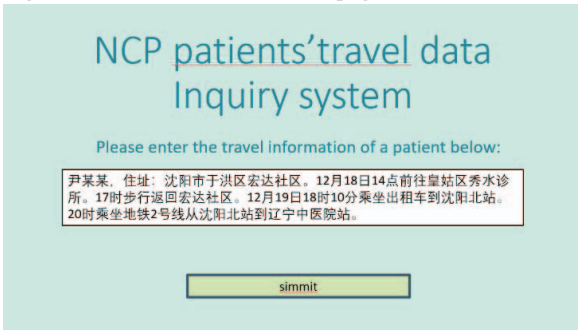

Fig 4.8. Submit Form

Since the coordinate values returned by the same geographical noun on Baidu map may not be uniform, this paper smooths the trajectories with different strategies by dividing the returned results into three categories:

(1) If the result contains one point, take the value directly;

(2) If the result contains two points, take the coordinates of the end point;

(3) If the result contains three or more points, take the first three points to form a triangle and calculate the midpoint of the triangle. The formula is as follows:

$$(x,y) = \left(\frac{x_1+x_2+x_3}{3}, \frac{y_1+y_2+y_3}{3}\right) \qquad (4.1)$$

Through the semantic trajectory mining model, location detection model and trajectory smoothing operation proposed in this paper, the semantic trajectory of users is mined together with the smoothed location and smoothed time sequence. Through the Baidu Map API, visualized trajectory can be displayed on the map.

On the basis of the trained BiLSTM-CRF model and saving the parameters of the model in file form, in this study, the manually collected data of patients' trajectory information are saved in json format through a series of processing steps, which can be easily read by the web end and displayed directly on the web page.

Taking the trajectory text message " Yin, female, Address: Hongda Community, Yuhong District, Shenyang City. She went to Xiushui Clinic in Huanggu District at 14:00 on December 18th and walked back to Hongda Community at 17:00. on December 19th, she took a taxi to Shenyang North Station at 18:10. At 20:00, She took Metro Line 2 from Shenyang North Station to Liaoning Traditional Chinese Medicine Hospital Station. " .

In the process of mapping high-risk areas, in this paper, Canvas is used to draw the trajectory on map surface in the form of broken lines, and each point of the polyline is shown on the map. To calculate the longtitude and latitude, this paper selects the first point and the second point by using the strategy of trajectory smoothing, and the green line between the two points is positioned as a high-risk zone, because many infected people appear between the two points. The mapping is shown in Figure 4.9, so far the inquiry is completely processed.

Fig 4.9. Mapping and Path drawing

5 Conclusion

This paper aims at making the masses obtain the correct trajectory information of the patients quickly and effectively. On the basis of collecting the data published by the health committees of various provinces and cities, the semantic trajectories of the patients are extracted to ensure the accuracy of the data and then establish mapping relationship between the real space and the text description of the trajectories of the patients while taking the time and space factors into account and excavating the dynamic changes of the patients. By using BiLSTM-CRF entity recognition model, a semantic trajectory mining system based on named entity recognition is built, and the analysis results are visualized. In addition, this paper also adds cloud database to the system, ensuring the input data can be persisted.

REFERENCES

[1] L Shuai,W Li. A Fully Trainable Network with RNN-based Pooling[J]. Neurocomputing. 2017, 338(21): 72-82.

[2] Xinchi Chen ,Xipeng Qiu ,Chenxi Zhu,Pengfei Liu,Xuanjing Huang .Long Short-Term Memory Neural Networks for Chinese Word Segmentation[C].Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.2015:1197-1206.

[3] E Kiperwasser ， Y Goldberg. Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations[J].Transactions of the Association for Computational Linguistics.2016,4(7):313-327.

[4] Lafferty J , Mccallum A , Pereira F C N . Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[J]. Proceedings of Icml, 2001, 3(2):282-289.

[5] Zhuoya Dong.The map display of communication sites and lines based on Baidu maps JavaScript API[J].Electronic Design Engineering.2013,21(18):73-76.