Extraction of Information from Public Health Emergency Web Documents

Li Wang [1, a],Yuanpeng Zhang[1,b] ,Danmin Qian[1,c] and Min Yao[2,d,*]

[1] Department of Medical Informatics, Medical School, Nantong University, 19 Qixiu Road, Nantong 226001, Jiangsu Province, China

[2] Department of Immunology, Medical School, Nantong University, 19 Qixiu Road, Nantong 226001, Jiangsu Province, China

[a]wangli@ntu.edu.cn, [b]maxbirdzhang@ntu.edu.cn, [c]Qiandm@ntu.edu.cn,

[d,*]erbei@ntu.edu.cn（Corresponding author）

Keywords:information extraction, named entity recognition, public health, hidden Markov model, web document

Abstract. Globalization and economic growth have brought more and more uncertain factors that would lead to the occurrence of public health emergencies, which greatly threaten people’s lives and properties. The occurrence of a public health emergency is often accompanied by the appearance of a huge amount of related documents on the Internet, and these documents carry a lot of important information. To extract such information, which are usually stored in the form of plain texts (unstructured documents) and cannot be reused directly, it is crucial to automate the extraction process. This work proposed a method for the recognition of named entities with H7N9 public health emergency-related web documents as the research subject, using Hidden Markov Models. The experimental results showed that the proposed method could effectively extract time, location and symptom information.

1 Introduction

In recent years, more and more uncertain factors would lead to the continuous occurrence of public health emergencies. As the increasing globalization causes the close connection among various fields in the society, the impact of public health emergencies is getting intensified. The public is more sensitive to public health emergencies compared with other industries. Inaccurate and incomplete information quickly spreading over the web could easily cause panic, even riot among residents. Therefore, quick and accurate information extraction from web documents is of great significance to the data analysis, including data mining and trend prediction, in the public health emergency field.

Generally, the basic informative elements of a document are the basis of correctly understanding its content. Named Entity Recognition (NER) refers to the task that identifies the NAME, LOCATION and TIME in an unstructured document (text)[1]. Recently, with the growing number of documents on the Internet, NER is becoming an essential component of information extraction, information retrieval, machine translation and problem solving, etc., which are all key issues in the natural language processing area.

Due to the huge amount of expert and skilled knowledge required in rule-based methods, machine learning techniques, as a substitution, have become the current popular research method, especially in medical science[2-4]. An early and well-known application of Hidden Markov Models (HMMs) in NER was carried out by Bikel et al. [5]. After that, more and more researchers have applied HMM in NER studies [6-8]. However, the NER in the public health emergency field has rarely been studied. In this study, H7N9 emergency-related web documents were collected, and using Hidden Markov Model, the named entities in the documents were extracted, including persons’ names, locations, times, and dates, even the symptoms, etc.

2 Methodology

Hidden Markov model is a powerful statistical machine learning algorithm widely employed in various natural language processing tasks. With its strong statistical foundation, HMM can tolerate more noise in the data comparing with other statistical algorithms, thus it can handle new data robustly.

2.1 HMM

In simpler Markov models, the state is directly visible to observers. That is to say, the state transition probabilities are the only parameters. However, in Hidden Markov model, the state is not directly visible; only the output, which depends on the state, is visible. Each state has a probability distribution over the possible output tokens.

The sequence of tokens generated by an HMM gives some information about the sequence of states. An HMM is composed of two sequences of random variables:

A sequence of states {Xt|t∈T} = {X1, X2, ... ,XT} , and

A sequence of observations {Yt |t∈T} ={Y1, Y2, ... ,YT}.

The parameters of the stochastic process are estimated, and a statistical description is given, in which the system being modeled is assumed to be a Markov process with unknown states. Each hidden state has a probability distribution over the possible output tokens. Therefore, the sequence of tokens (observable parameters) generated by an HMM implies the sequence of hidden states. The task of HMMs is to determine the hidden states based on the observable parameters, and the determined hidden states can be used to for further analysis.

NER tasks can be treated as classification problems, where every word is either part of name classes or not part of any name. When applying HMMs to an NER task, words (W) are the observable parameters and name-classes (N-C) are the hidden states of the associated Markov process. Therefore, given a model and all its parameters, the NER task is completed by determining the sequence of name-classes N-C = {Name1, ...,NameT} that is most likely to have generated a sequence of words W = {W1, ...,WT} in a sentence of T words.

2.2 Outline of the system

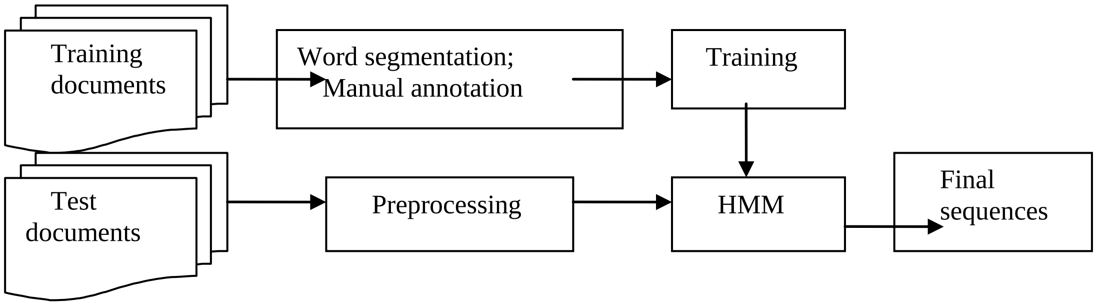

Fig. 1. Outline of the proposed system

As shown in Fig.1, the proposed approach includes two parts, training and testing. The training documents are used to train the HMM model. Then, the obtained HMM model is used to extract the sequence of target information. Some important components of the system, especially for Chinese web documents, are introduced as follows.

(1) Data preprocessing

All the training data in the corpus were in the form of plain text. The proposed model takes the words as observation values, and each sentence as a sequence of observation values. Therefore, in data preprocessing step, all the texts are read from the corpus, split into sentences according to the punctuations or line breaks, and segmented into words. This process returns sequences of the observation values, which are the input data of the HMM model.

(2) Manual annotation

After data preprocessing, some typical training documents are selected. According to the defined model state, the observation value of each word is annotated manually. All the values are to be used in the calculation of the Maximum-Likelihood (ML) algorithm.

(3) ML algorithm

This algorithm is used to calculate the model parameters, including initial state probability, transition probabilities and emission probabilities.

(4) Viterbi algorithm

The Viterbi algorithm is a dynamic algorithm that computes the most likely state transition path given an observed sequence of symbols.

3 Experiment

3.1 Web document analysis

3.1.1DOM Tree

To extract the body plain text from each web document, Document Object Model (DOM) was utilized. DOM is an application programming interface (API) for valid HTML and well-formed XML documents. It defines the logical structure of a document and the way the document is accessed and manipulated. XML is extensively used to represent many different kinds of information, and the DOM may be used to manage such data. Through DOM, programmers can build documents, navigate document structures, and add, modify, or delete elements and content of documents.

3.1.2 Chinese word segmentation

Written Chinese lacks of explicit word delimiter (equivalent to the blank space in written English). In addition, English characters and words, numbers, and symbols are often included in many documents. In order to convert sentences from continuous characters strings into sequences of words, a Chinese language analyzer LTP2.0 was used to perform segmentation. This software program is developed by Harbin Institute of Technology Information Retrieval Lab. The newest version of LTP is available at http://www.ltp-cloud.com/.

3.1.3 Sentence splitting

Since in the HMM model for Chinese language, each word is an observation value and each sentence is an observation sequence, every sentence in the plain text needs to be split into a single string. The symbols used to split plain texts include comma, full stop, semicolon, question mark, exclamatory mark and blank. Some symbols need to be processed specially, for example the brackets. If a sentence in brackets contains a symbol, the sentence is split by the symbol. If the sentence does not contain any symbol, and the number of Chinese characters in the sentence is greater than 10, the sentence is treated as an absolute one. If the sentence does not contain any symbol while containing less than ten Chinese characters, the brackets are ignored, and the characters inside and outside the brackets are treated as a whole sentence.

3.2 Manual annotation

After being analyzed, the web documents were annotated. The target information included location, time, symptom, etc. Only the sentences containing the target information were annotated, others were simply ignored.

The annotation of symptoms is taken as an example here, as shown in Table 1.

Therefore, for the following original sentence,

该患者合并诊断有病毒性肺炎，急性呼吸窘迫综合征，感染中毒性休克，急性肾功能衰竭，弥散性血管内凝血。

after segmentation and manually annotation, it is converted into

该|s_w患者|s_w合并|s_w诊断|有|s_a病毒性肺炎|s_d，|s_e急性呼吸窘迫综合征|s_d，|s_e感染中毒性休克|s_d，

|s_e急性肾功能衰竭|s_d，|s_e弥散性血管内凝血|s_d。

Tables 2 and Table 3 illustrate the manual annotation of location and time.

Table1Manual annotation of symptom information

State	Meaning	Examples
S_W	Background	
S_A	Above	appear | multiple organ failure |
S_B	Organ + Symptoms	left lung | Pneumonia |
S_C	Symptoms + Quantifier	fever | 5 | days
S_D	Symptoms	The patients with | Acute Respiratory Distress Syndrome |
S_E	Symptoms + punctuation / conjunctions + symptoms	Fever |, | cough |, / and | dyspnea |
S_F	Below	BP | is | 0 |
S_G	Below2	BP |is | 0 |

Table 2Manual annotation of time information

State	Meaning	Examples
T_W	Background	
T_M	Numerals	2010|/|05|/|12|
T_Q	After the numeral quantifier (year, day)	5|days
T_T	Temporal words (month, day)	March|5th|
T_V	First punctuation after a numeral (/, -, :)	08|:|34|
T_X	Second punctuation after a numeral (/, -, :)	2010|−|05|−|12|
T_L	Below	March|5th|report
T_U	Above	CNS|March|5th|report

Tables 3 Manual annotation of location information

State	Meaning	Examples
L_W	Background	
L_B	1 Above 1	Home / live / Shenzhen
L_C	2 Above 2	Where / Shenzhen / one / case
L_D	Location	Guangdong Province / people / infected / H7N9 / bird flu / diagnosis / 6 / cases
L_E	Prefixes of Locations	ShenzhenLonggang District
L_F	Below	Guangdong Province / person / infected / H7N9 / bird flu / diagnosis / 6 / cases

3.3 Information extraction

Viterbi algorithm was used to select the state sequence of the highest probability according to the observation sequence in order to extract target information. If the state sequence contains the target state, the target information is extracted with pattern matching algorithm. Finally, some noise data were removed using rule-based method.

4 Experimental results Experimental results

Of the 1267 H7N9 public health emergency-related documents collected in this study, 120 documents were used to train the HMM model, and 1067 documents were used to perform the test.

To assess the performance of the proposed method, the P (precision) and R (recall) values were used.

P = the number of correctly labeled entities/ the number of all extracted entities

R = the number of correctly labeled entities/ the number of all entities

Table 4Experimental results

Target	P (precision)	R (recall)
Time	0.7136	0.7038
Location	0.8167	0.7834
Symptom	0.6828	0.6678

5 Conclusion and Discussion

The P and R values shown in Table 4 validated the effectiveness of the presented method in extracting the target entities in the H7N9 public health emergency-related web documents. However, the final results are not satisfying enough since neither the precision nor the recall reached 0.9, which can be attributed to the following reasons,

The H7N9 public health emergency-related web documents contain medical vocabulary, which affected the segmentation accuracy as well as the extraction efficiency.

For the recognition of time entities, absolute temporal words can be found correctly. However, relative temporal words are difficult to be found. For example, in the phrase “几年前 (a few years ago)”, the character “几 (a few)” is a measure word instead of a numeral, thus the entity cannot be detected.

For the recognition of location entities, at present, the system can only recognize the names of provinces and cities. To improve the performance, names of districts and streets are to be added into the system.

For the recognition of symptom entities, some long descriptions of symptoms cannot be annotated according to the annotation rules (Table 1), for example, “整个肺部弥漫着大片白泡状的感染阴影”, “最高体温达到了 40.0℃”, “心率下降”, “血压无法维持”, etc. More special rules are needed to improve the coverage, and medical symptom dictionaries are to be added into the system.

Acknowledgement

This work is supported by the National Science Foundation of China (No.81271668) and Jiangsu Natural Science Foundation of College and University (14KJB310014，14KJB510031) and the Project of Six Talent Peak Program(2013-WSW-011).

References

[1] M. Bikel, Scott Miller, Richard M. Schwartz, and Ralph M. Weischedel. Nymble: a high-performance learning name-finder. In ANLP, 1997, pp.194–201.

[2] V.N. Vanpik. The nature of statistical learning theory(Second edition).New Yok,Spring,2000

[3] Erhu Zhang, Fan Wang, Yongchao Li, Xiaonan Bai, Automatic detection of microcalcifications using mathematical morphology and a support vector machine, Bio-Medical Materials and Engineering, 24(2014), 53-59

[4] Yifei Chen, Ping Hou, Bernard Manderick, An ensemble self-training protein interaction article classifier, Bio-Medical Materials and Engineering, 24(2014), 1323-1332

[5] M. Bikel, Richard L. Schwartz, and Ralph M.Weischedel. An algorithm that learns what’s in a name. Machine Learning, 34(1999),211–231.

[6] D. Freitag, A. McCallum and F. Pereira. Maximum Entropy Markov Models for Information Extraction and Segmentation.In:ProeeedingsofICML’2000.CA,USA:Morgan Kaufmann,2000,pp.591-598

[7] Souyma Ray and Mark Crven. Representing sentences structure in hidden Markov Models for information extraction. In: Proeeedings of the Seventeenth International Joint Conference on Artificial Intelligence. Seattle. WA:Morgan Kaufmann, 2001, pp. 1273-1279

[8] T,Seheffer, C.Deeomain and S.Wrobel. Active Hidden Markov Models for Information Extraction. In: Proceedings of the International SymPosiumon Intelligent Data Analysis.Lisbon,Portugal,2001,pp.309-318