Research on targeted service technology for early warning information of meteorological disaster based on NLP

Yuxing Feng[1], *Wei* Tang[1,*], *Muhua* Wang[1], *Zhiyu* Cao[1], *Liang* Qiao[2], and *Lei* Cui[1]

[1]CMA Public Meteorological Service Center, 100081 Beijing, China
[2]Meteorological Service Center in Heilongjiang Province, 150030 Heilongjiang Province, China

Abstract. Affected by global climate change and the superposition of urban construction, the climate of city meteorological guarantee service puts forward new requirements and new challenges. As the first line of defense of meteorological disaster prevention and reduction, it is very important to realize the refinement targeted service of early warning information. There are some problems in the traditional early warning information service, such as the difficulty of accurate service of early warning information, the lack of precision, and the insufficient mining of early warning text information. This paper mainly analyzes the text description characteristics of early warning information of meteorological disaster, constructs the early warning information knowledge extraction process, constructs the early warning information labeling system, and realizes the early warning effective time extraction method based on conditional random field model,Early warning affected areas extraction method based on bidirectional long-term and short-term memory neural network and early warning cautions extraction method based on bootstrapping weak supervised learning method. Finally, taking the early warning information targeting service of meteorological information decision support system as an example, this paper tests the early warning information extraction methods, and preliminarily realizes the early warning precision targeting service in the decision support service of meteorological disaster prevention and reduction.

1 Introdution

Xi Jinping, Chinese President in celebrating the 100th anniversary of the founding of the communist party of China emphasized when delivered an important speech at the conference: " To take history as a mirror and open up the future, we must unite and lead the Chinese people in constantly striving for a better life." Meteorological service is a scientific and technology-based and basic social public service. It serves the country's economic and social development and protects the safety and well-being of the people. This requires us to implement Xi's important instructions on "playing the role of meteorological as the first line

* Corresponding author: tangw8@cma.gov.cn

of defense in disaster prevention and reduction", vigorously promote the high-quality development of meteorological services, improve the people's ability to provide meteorological services for their work and life, and maintain social stability and security [1]. In the past 40 years, the frequency of major natural disasters in the world has been increasing, the number of disasters in the past 20 years has increased significantly compared with the previous 20 years. China is in the East Asian monsoon region. Affected by geographical location, landform and climate characteristics, China has more kinds of meteorological disasters, wide geographical distribution, high frequency of occurrence and heavy losses than most countries in the world. According to statistics, since the beginning of this century, China has suffered an average annual direct economic loss of 290 billion yuan due to meteorological disasters, which have seriously affected and threatened the safety of people's lives and property. The role of meteorological disaster prevention and reduction as the first line of defense is of special practical significance in ensuring the safety and well-being of the people [2-3].

At present, meteorological disaster warning information service only achieves "warning" in the traditional macro sense. It requires the emergency responsible person, the industry, and the public to have a certain knowledge of meteorological space-time expression to understand the content of early warning, which does not reach the efficient service of early warning. Moreover, the warning text contains more valuable fragmented information, which has not been fully utilized and mined. Therefore, extracting and digging from early warning text to break the traditional service mode has become an urgent problem to be solved in the accurate and targeted early warning service at present. This paper will analyze the text features of meteorological disaster warning and build a warning information extraction process, based on natural language processing technology, research on extraction methods for the effective time of early warning, the areas affected of early warning and warning cautions, improve the precision targeted service ability of meteorological early warning, and guard the "first line of defense" of meteorological disaster prevention and reduction.

2 Extraction process of early warning information

Meteorological disaster early warning information is issued uniformly by the national early warning information release system of the National Early warning Information Release Center. It is the four-level (national, provincial, city and county) early warning issuing units that are responsible for making corresponding national, provincial, city and county early warning contents in accordance with relevant standards and norms. The text of early warning information mainly includes three categories: effective time, affected areas and warning cautions, etc., with space-time characteristics [2]. In comparison, the content of national early warning is more complex, and the content of county early warning is simpler.

We will analyze the time, areas and cautions characteristics of warning cautions of four-level warning text, and build the early warning information extraction process,Through the research of natural language processing technology, the time, areas, and warning cautions in the warning content are accurately extracted and structured storage for precise targeted services [3]. The extraction process includes three parts: early warning information collection, early warning information extraction and early warning information application, as shown in Figure 1.

Early warning information collection: Data can be collected in two ways. One is to obtain real-time data through the early warning Webservice interface and store it in the database. The other is to obtain the early warning CAP file package for file storage.

Early warning information extraction: according to the text description rules and characteristics of early warning content, the effective time of early warning, the affected

areas of early warning, and the cautions of early warning are extracted, and the data is standardized [4-5].

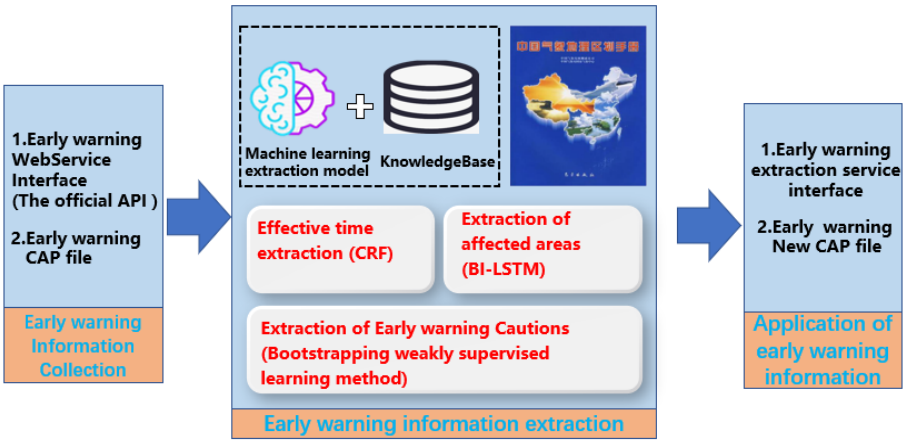

Fig. 1. Extraction process of early warning information.

Application of early warning information: For different service methods, the results of early warning standardized processing can be applied in two ways, including early warning extraction result service interface and new CAP file of early warning.

3 Research on extraction technology of early warning information

3.1 Early warning information labeling system

The early warning information organizes and expresses the information of various elements (time, areas, cautions, etc.) in order according to the unified space-time standard,by analyzing the difference of the expression mechanism of early warning information in natural language and meteorological disaster service system, the paper summarizes the linguistic description characteristics of early warning information. From the three levels of vocabulary, syntax and semantic structure, the warning information annotation system for Chinese text is constructed, and specific annotation modes and norms are formulated, and large-scale experimental data of Chinese text annotation are established.

The goal of establishing the warning information labeling system is to analyze the warning information described in natural language, find the language structure and semantic expression rules of the relevant elements information in the text, and establish the metadata to describe them. Based on the analysis of the typical example of the description of the national level 4 warning information, to develop early warning information system based on XML schema and specific labeling operation specification,In addition, ICTCLAS developed by the Institute of Computing Technology of the Chinese Academy of Sciences was used as the natural language processing platform, and GATE (General Architecture for Text Engineering) was used as the annotation platform to annotate large-scale data and provide standardized experimental data for early warning information extraction [6]. GATE can accept SCHEMA files in XSD format and provide annotation data management solutions. The processed corpus can be uniformly stored in XML format, facilitating data invocation and shared integration.

3.2 Extraction method of early warning effective time

The effective time of warning information includes "daytime", "tomorrow", "next 8 hours" and other abstract time words, which have their own language expression characteristics. According to natural language processing technology, the time information extraction problem can be transformed into a sequence labeling problem. Based on the Condition Random Fields (CRF) model, which is the mainstream in the field of named entity recognition, this paper adds the warning effective time description feature to realize the extraction of warning effective time [7]. The extraction process is shown in Figure 2.

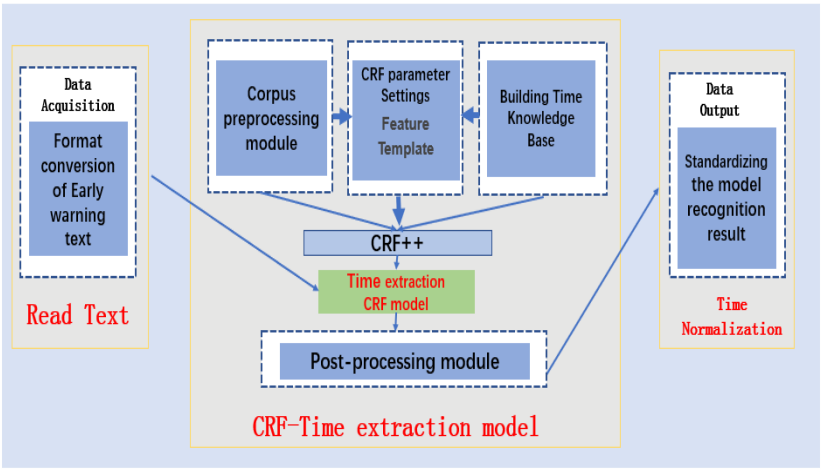

Fig. 2. Extraction process of early warning effective time.

Construct the effective time of early warning information corpus. In this paper, 10,000 warning data are randomly selected from the warning history database as a data set, and the text content of the data set is cleaned with special characters. Then, the warning information labeling system is used to libel the corpus.

Construct time knowledge base and define feature. In this paper, time knowledge base is constructed by sorting out common time noun dictionary and time information linking word list. Then, the characteristics of early warning time information are analyzed and the feature template is established. Features mainly involve part-of-speech features, including words, part of speech, combination of words and their part of speech, and contextual features of words [8].

Model training. The corpus of early warning effective time was randomly divided into training set, test set and verification set in a ratio of 7:2:1. The CRF++ tool was used for training, and the early warning effective time extraction model was generated [9].

Standardization of results. The result of model recognition is processed by merging adjacent time units, matching, and merging time continuous words, and the result of model recognition is generated. Finally, the model recognition results are normalized and converted according to the time standard format.

The model uses the warning effective time labeling corpus to conduct closed test and open test. Through fuzzy qualitative judgment, the experiment shows that CRF and rule hybrid model has a good effect on the accuracy of time information recognition, especially the conditional random field model can use the characteristics of semantic information in the text to determine whether the ambiguous words are time information, so it has a high accuracy. The accuracy and recall rate of time information extraction by using this model are above 90%.

3.3 Extraction method of early warning affected area

The text description of the affected regions of early warning information is characterized by the combination of Chinese administrative regionalization words and meteorological geographical regionalization words, such as "northeast Jiangxi", "east Guizhou", "south Guizhou", etc. According to the different level of warning issuing unit, the description of affected areas have its own characteristics. Although the sequence information of context is considered in the existing geographical name resolution models based on deep learning, the output results are still mutually independent in nature [10]., Therefore, based on CRF, Bidirectional Long short-term Memory (BI-LSTM) neural network is used to extract the affected areas of early warning. The extraction process is shown in Figure 3.

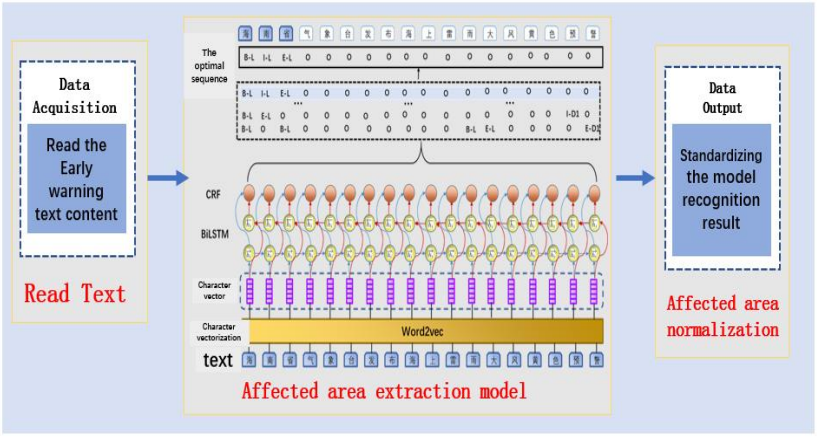

Fig. 3. Extraction process of early warning affected area.

3.3.1 Build the extraction model of early warning affected area

The bottom-up Embedding layer, two-way LSTM layer and CRF layer respectively.

Firstly, vector representation of words in sentences is carried out by Embedding layer, and each word is mapped into k-dimensional real number vector (K is the hyperparameter in the model) by word2vec word vector generation model. Semantic similarity between words is judged by the distance between words and a representation of words in vector space is obtained. Vector numerical representation as computer input, as input to bidirectional LSTM.

Then the text context information is obtained through the bidirectional LSTM layer, and the text information of the affected area is extracted. Through a forward LSTM and a reverse LSTM, the bidirectional LSTM layer calculates the corresponding vectors of each word considering the words on the left and right, and then connects the two vectors of each word to form the vector output of the word [11-12].

Finally, the CRF layer is used for identification. The CRF layer takes the context feature vector output by bidirectional LSTM as input, calculates the category probability value of each character in the recognition result of the affected area, and marks the affected area with the maximum probability index, to complete the identification of the affected area. As shown in Figure 3, input the text "Hainan Meteorological Observatory issued yellow warning of thunderstorm gale at sea" and identify the area name "Hainan province".

This method can effectively solve the traditional method of manual feature design, and the character vector representation replaces the traditional sparse representation. Using this

model, closed test and open test were carried out on the annotated corpus in the early warning affected area, and the accuracy and recall rate reached over 85%.

3.3.2 Standardization of extraction results

First, the meteorological geographical regionalization data set is made. Based on the picture data of meteorological geographical regionalization in Manual of Meteorological Geographical Regionalization in China, this paper makes structured data of meteorological geographical regionalization.

Then the extraction results are standardized. For example, the model extraction result is "northeast Chongqing", it will be associated with the code table of the administrative division and the geographical division data set,The final extraction results are shown in Table 1

Table 1. Example of standardization of extraction results for early warning affected areas.

Model extraction results	**Final extraction result**	
Northeast chongqing	chongqing (500000000000)	Wanzhou District (500101000000)
		Chengkou County (500229000000)
		Kaizhou District (500234000000)
		Yunyang County (500235000000)
		Fengjie County (500236000000)
		Wushan County (500237000000)
		Wuxi County (500238000000)

3.4 Extraction method of early warning cautions

The warning cautions in the text of early warning information are closely related to the types of early warning disasters, usually including the intensity of disasters and the related disasters that can be triggered”. Warning cautions need to be extracted by relational extraction method. This paper intends to use Bootstrapping weakly supervised learning method to extract warning cautions, the extraction process is shown in Figure 4.

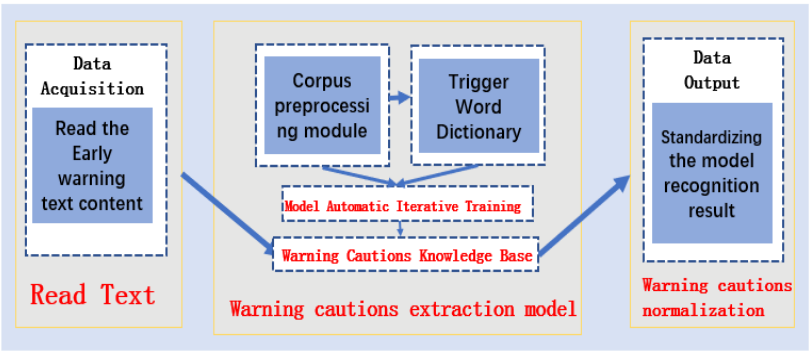

Fig. 4. Extraction process of early warning cautions.

Firstly, build a corpus of warning cautions. In this paper, 10,000 warning data in are randomly selected from the warning history database as a data set, and the characteristics of attribute words of various warning disaster warning cautions are learned from many samples. Then, the warning information labeling system is used to mark the corpus, and the trigger vocabulary dictionary of disaster warning items is constructed [13].

Then, bootstrapping weakly supervised learning method is used to obtain the information expression pattern automatically and iteratively, and the corresponding knowledge base of warning cautions is constructed. Bootstrapping manually obtains the seed of feature words (disaster warning trigger vocabulary dictionary) and then incrementally iterates the automatic training until convergence. It relieves the burden of manual annotation corpus [14-15].

Finally, a pattern matching algorithm is designed by using the knowledge base of warning information to extract warning cautions from warning text.

4 Application example

Taking the blue rainstorm warning issued at 22:02 on October 6, 2021 in Shaanxi Province as an example, the extraction method of the effective time, the extraction method of the affected areas and the extraction method of the warning cautions in this paper is used for extraction,The extraction results are shown in Table 2.

Table 2. Extraction results of application example.

Early Warning Content	Data processing	Effective time	Affected area	Warning caution
Shaanxi Meteorological Observatory will continue to issue blue rainstorm warning signal at 22:02 on October 06, 2021: It is expected that the rainfall in the following areas will reach 50 mm or more in the next 12 hours: Zhenba County of Hanzhong City, Hanbin District of Ankang City, Hanyin County, Shiquan County, Ziyang County, Langao County and Pingli County, please take precautions.	Model to extract	12 hours	Zhenba County of Hanzhong City, Hanbin District of Ankang City, Hanyin County, Shiquan County, Ziyang County, Langao County and Pingli County	More than 50 millimetres of rain will fall
	Normalized processing	October 07, 2021, 10:02 am	Zhenba County（610728000000）; Hanbin District（610902000000）; Hanyin County（610921000000）; Shiquan County（610922000000）; Ziyang County（610924000000）; Langao County（610925000000）; Pingli County（610926000000）;	Rainfall of 50 mm

The extraction results of the early warning were applied to the meteorological information decision support system serving the Ministry of Emergency Management, and the affected areas and levels of the early warning were colored on the map. Compared with the original full-text display of early warning, the important contents of early warning are more intuitive, precise, and efficient, which is conducive to the rapid perception and emergency decision-making of the decision-making department of emergency management department of disaster early warning, and the accurate targeted service of early warning content is preliminarily realized.

5 Conclusion

This paper mainly analyzed the text description characteristics of meteorological disaster warning information issued by the national warning issuing system, built the knowledge extraction process of warning information, and built the warning information labeling system. The extraction method of warning effective time based on conditional Random field model (CRF), extraction method of warning affected areas based on BI-LSTM deep learning model and extraction method of warning cautions based on Bootstrapping weak supervised learning method are implemented. Finally, this paper takes the targeted early warning information service of meteorological information decision support system as an example, and preliminarily realizes the accurate targeted early warning service in the decision-making guarantee service of meteorological disaster prevention and reduction. However, the following problems still exist: At present, the extraction model of early warning affected areas proposed in this paper only realizes the extraction of China's land area, but is not suitable for the extraction of China's sea area and river basin. The extraction model of warning cautions only realizes partial extraction of meteorological disaster types；Further studies will be conducted in the future to improve the extraction model. In addition, the research results of this paper have not been applied in practical business. In the future, accurate extraction results of early warning will be applied to multiple business systems, to truly provide accurate early warning service to every public, and do a good job as the first line of defense for disaster prevention and reduction to protect national security.

National Key R&D Program of China, 2018YFF0300105.

References

1. ZHUANG Guotai, To strengthen the first line of defense of meteorological disaster prevention and mitigation [J]. Qiushi,2021(14):72-77
2. ZHIYU CAO, YUXING FENG, XIAO LI. A Study on the Calculation Method for the Coverage Rate of Early Warning Release[J]. IOP Conference Series: Earth and EnvironmentalScience,2019,233(5):052034(9pp).DOI:10.1088/1755-1315/233/5/052034
3. WANG Muhua. Research on Meteorological Disaster Prevention and Mitigation Monitoring and Management Based on WSR and Hall 3 D Model [J]. Journal of Catastrophology,2020,35(4):103-107. DOI:10.3969/j.issn.1000-811X.2020.04.020
4. Zhang Xueying, Zhang Chunju, Wu Mingguang, et al. Spatiotemporal features based geographical knowledge graph construction [J]. Science in China (Information Sciences),2020,50(7):1019-1032

5. Wang Haofen, Qi Guilin, Chen Huajun. 《Knowledge Mapping: Methods, Practices and Application》[J]. Automation Panorama,2020,37(1):7. DOI:10.3969/j.issn.1003-0492.2020.01.011
6. Wang Feng-e. Recognition of Temporal Relation in Chinese Text [D]. Shanxi: Shanxi University,2012
7. Liu Wencong, Zhang Chunju, Wang Chen, et al. Geological Time Information Extraction from Chinese Text Based on BiLSTM-CRF [J]. Advances in Earth Science,2021,36(2):211-220
8. YAN Zi-fei,JI Dong-hong. Exploration of Chinese temporal information extraction based on CRF and semi-supervised learning [J]. Computer Engineering and Design,2015(6):1642-1646. DOI:10.16208/j.issn1000-7024.2015.06.044
9. Song Guomin, Zhang Sanqiang, Jia Fenli, et al. Temporal Information Extraction and Normalization Method in Chinese Texts [J]. Journal of Geomatics Science and Technology,2019,36(5):538-544. DOI:10.3969/j.issn.1673-6338.2019.05.017
10. Wang Muhua, Wang Tianyue, Li Yanpeng, et al. Research on Knowledge Graph Model about Rainstorm Disaster Based on Simple Event Model [J]. Journal of Catastrophology,2021,36(4):74-78. DOI:10.3969/j.issn.1000-811X.2021.04.013
11. Ma Jian-xia, YUAN Hui, JIANG Xiang. Extracting Name Entities from Ecological Restoration Literature with Bi-LSTM+CRF [J]. Data Analysis and Knowledge Discovery,2020,4(2):78-88. DOI:10.11925/infotech.2096-3467.2019.0034
12. Liu Jiaqi, Luo Yonglian. Study on the Algorithm of Extracting Toponym from Chinese Emergency News [J]. Information & Computer,2019(15):53-54,57
13. Gao Qi. Research on ontology annotation method based on Bootstrapping [D]. Chongqing: Chongqing University,2010
14. Yin Jihao,Fan Xiaozhong,Liu Shining,etc. Training Corpus Construction Based on Bootstrapping [J]. JOURNAL OF COMPUTER RESEARCH AND DEVELOPMENT,2007,44(z2):394-397
15. YU Li,LU Feng,LIU Xiliang. A Bootstrapping Based Approach for Open Geo-entity Relation Extraction [J]. Acta Geodaetica et Cartographica Sinica,2016,45(5):616-622