A Research on Event Extraction for illegal Cultivated land use Case Texts

Jun Zhang[1],Hao Wu[1],Ji She[1,2,4],Dongyang Hou[3]

[1] National Geomatics Center of China, Beijing, China, 100830 - junzhang@ngcc.cn, wuhao@ngcc.cn, 2100858168@qq.com,
[2] China University of Mining & Technology(Beijing), BeiJing, China, 100083-2100858168@qq.com
[3] School of Geoscience and Information Physics, Central South University, Changsha, China, 410083 - houdongyang1986@csu.edu.cn
[4] Corresponding author

Keywords: Illegal cultivated land use case, Event extraction, Argument recognition, BiLSTM-CRF model

Abstract

Cultivated land is fundamental to human survival and development. China attaches great importance to the utilization and protection of cultivated land. In the field of administrative law enforcement, the Ministry of Natural Resources, as the competent authority, has regularly released text of cases involving illegal occupation of cultivated land for a long time. Such unstructured data is of low value density and limited usability. In this paper, the structured information such as location, time, event type, offence and occupied area can be quickly and efficiently extracted from the administrative cases through event extraction technology, so as to promote the use of historical case information and the study of intelligent law enforcement and case handling for administrative agencies. Firstly, the real-case data is collected from the official website of the Ministry of Natural Resources using a thematic web crawler. Subsequently, it constructs an entity argument recognition framework, defining the types of illegal cultivated land use events and their argument compositions. A training dataset is built for illegal cultivated land use cases through manual annotation. Finally, two sets of models based on statistical-based machine learning and deep learning are employed to optimize the extraction results of event arguments. Experiments show that the F1 score for the event argument extraction reaches 90.81% with BiLSTM-CRF model. The research in this paper achieves effective extraction of case information, verifies the effectiveness of existing models in event extraction from administrative law enforcement case texts, and provides effective support for deeper research in this field in the future.

1. Introduction

Cultivated land refers to land used for growing crops (General Office of the State Council, 2020), primarily providing humanity with agricultural food supply and serving as the foundation for human survival and development since the advent of agricultural civilization. Influenced by human activities, the utilization of cultivated land often undergoes changes, including its conversion to non-agricultural uses ("non-agriculturalization") and non-food crop cultivation ("non-grainification"). To prevent the "non-agriculturalization" and the "non-grainification" of cultivated land, the Chinese government has formulated policies and regulations for its protection. As the competent authority, the Ministry of Natural Resources has been regularly reporting cases of illegal and irregular occupation of cultivated land, storing a large amount of historical cases recorded. However, the unstructured data has low value density and limited usability(Zhang, et al., 2022). If specialized processing is employed to swiftly and efficiently extract structured, valuable information from these case (Sahnoun et al., 2020), it would enable the construction of knowledge graphs, predictive models, and other tools using historical cases. (Wang and Xiang, 2023) This would assist administrative agencies in resource allocation and law enforcement decision-making, enhance the quality and efficiency of law enforcement, and accelerate the process of intelligent law enforcement.

Event extraction tasks typically encompass two parts: entity recognition and event argument recognition (Ahn, 2006). Entity recognition aims to identify named entities such as time, location, people, and organizations from text. while event argument recognition focuses on identifying key information such as event trigger, event type, and their associated participants and causes. Traditional event extraction methods are mainly divided into pattern-based methods and statistical model-based methods like Hidden Markov Models (HMM) and Conditional Random Fields (CRF). With the application of deep learning technology (Li, et al., 2022), especially the use of pre-trained language models (Wang and Zhou, 2024), has significantly improved the performance of event extraction. Currently, research on event extraction tasks primarily focuses on public datasets production, covering areas such as social media analysis, meteorology (Hu, et al., 2022), law, and transportation (Zhang et al., 2021). However, study of event extraction for illegal cultivated land use case texts is still relatively rare.

In this paper, the structured information such as location, time, event type, offence and occupied area can be quickly and efficiently extracted from the administrative cases through event extraction technology, to promote the use of historical case information and the study of intelligent law enforcement and case handling for administrative agencies.

2. Methodology

This paper proposes an event extraction method for illegal cultivated land use case texts in natural resources, as illustrated in Figure 1. Firstly, a thematic web crawler is utilized to collect texts on illegal cultivated land use cases from official websites such as the Ministry of Natural Resources. Subsequently, the original texts are pre-processed to retain core information relevant to illegal cultivated land use events, thereby obtaining the event skeleton. Next, an entity argument recognition framework is constructed, defining the types of illegal cultivated land use events and their argument compositions. The BMES (Begin, Middle, End, Single) annotation method is then employed to annotate key information in the texts of illegal cultivated land use events, resulting in a training dataset. Finally, within the PyTorch environment, two sets of models based on statistical and deep learning approaches are used to select the optimal event argument extraction results. Evaluation metrics such as precision, recall, and F1-score are adopted to determine

the most suitable model for the task. Ultimately, based on the event type and event argument, a semantic relationship network for illegal cultivated land events is constructed.

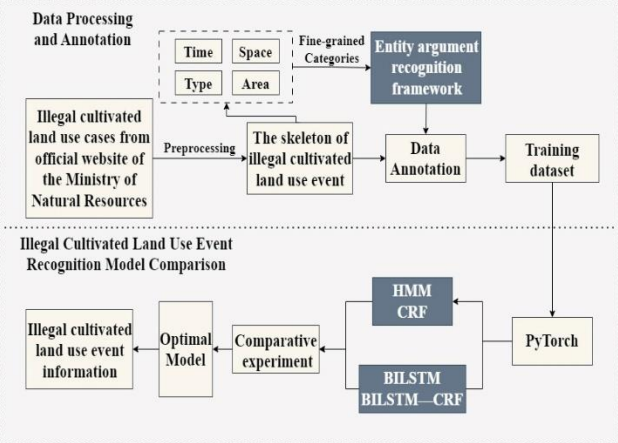

Figure 1. The flowchart of event extraction for illegal cultivated land use case text

2.1 Entity Argument Recognition Framework Construction

In the process of analysing illegal cultivated land use cases, systematically recording and analysing various key elements is crucial to ensure that farmland events are correctly interpreted and analysed. This paper constructs a multi-dimensional and systematic fine-grained recognition framework for entity arguments in illegal cultivated land use events. By defining specific event-related arguments, it provides a comprehensive framework for identifying illegal cultivated land use events, aiming to recognize, describe, and quantify specific illegal farmland use incidents to reflect the violations and their impacts.

Firstly, for any event, there is always information about its location and time. Therefore, the arguments Location (Hebei Province) and Time (2022) are defined to reflect the spatiotemporal information of illegal cultivated land use events. The event type "Type" of illegal cultivated land use cases is then defined as "non-agriculturalization" and "non-grainification", which helps to clarify the change in the nature of cultivated land use in such cases. By identifying farmland events as non-agricultural or non-grain purposes, the loss of farmland resources and its impact on agricultural production can be effectively recognized. The argument composition includes five elements. The argument "Offence" , which refers to specific illegal activities occurring on cultivated land, typically including illegal occupation, alteration, or destruction of cultivated land use, for example, illegal land occupation for construction or landscaping with lakes, and other actions. Identifying the argument “Offence” can specifically explain which actions have led to the loss of cultivated land resources, providing a basis for administrative penalties and actions by relevant law enforcement departments. The specified use of occupied cultivated land "SpecifiedUse" (e.g., facility agricultural land) and the actual use type after occupying the cultivated land "ActualUse" (e.g., landscape roads, landscape plazas, etc.) are defined to describe the actual land use before and after the illegal cultivated land occupation, and help to determine whether the land still has potential agricultural value or has completely deviated from agricultural use. These two arguments contribute to assessing the loss of land resources and provide a basis for land reclamation and resumption of farming. The total occupied area "TotalArea" (e.g., occupying 137.91 acres of land), and the cultivated area within the occupied land "CultivatedArea" (e.g., 51.6 acres of cultivated land) are defined, which help to evaluate the illegal impact on cultivated land from the perspective of the scope of the violation, identifying the percentage of cultivated land occupied and the degree of loss of this portion of cultivated land in the illegal activity. The recognition framework for the major categories of entity arguments in illegal cultivated land use events are shown in Table 1.

Categories	Definition	Example
Location	Province, City, District, County, Township	Hebei Province
Time	Year, Month, Day	Year 2024
Type	illegal cultivated land use cases	Non--agriculturalization
Offence	specific illegal activities occurring on cultivated land	illegal land occupation for construction/landscaping with lakes
SpecifiedUse	the specified use of occupied cultivated land	facility agricultural land
ActualUse	the actual use type after occupying the cultivated land	landscape roads, landscape plazas, etc.
TotalArea	the total occupied area	occupying 137.91 acres of land
CultivatedArea	the cultivated area within the occupied land	51.6 acres of cultivated land

Table 1. The recognition framework for the major categories of entity arguments in illegal cultivated land use events.

2.2 Named Entity Recognition (NER) and Text Annotation

Named Entity Recognition (NER) is a significant component of an important task in natural language processing (NLP), with the objective of identifying named entities in text. For example, to identify and classify names of people, places, times, organizations, etc.; At the same time, NER is also a key step in the field of Information Extraction (IE), commonly used in in text analysis, question answering, machine translation, speech recognition and many other applications. NER facilitates the automated identification and extraction of key information from text, help improve the efficiency and accuracy of text analysis, support cross-domain and multi-language applications, reduce manual intervention, and conserving time and cost. With the advancement of technology, NER has been applied widely in various fields, and the commonly used methods have gradually changed from traditional rule-based methods to statistical principle-based machine learning model methods and deep learning-based methods, which provide strong support for many intelligent text systems. Therefore, this paper adopts these two types of methods to complete the NER task of identifying and analysing textual events in cultivated land.

A plethora of Named Entity Recognition (NER) text annotation methods, with the most common ones being the BIO and BMES annotation methods. The BIO annotation method is considered to be uncomplicated and suitable for the majority of named entity recognition tasks, especially when dealing with multiple entities, as it can accurately handle the boundaries between different entities. BMES is another widely used annotation method for named entity recognition. It is similar to BIO but with finer granularity, allowing for more precise handling of multi-word and single-word entities. A key feature of the BMES method is its explicit delineation of the beginning, middle, and end words of entities, while also avoiding some ambiguities inherent in the BIO method, particularly in the case of multi-word entities, making it more intuitive. Therefore, the BMES method is selected to annotate the text data and generate BMES files. In this standard, B (Begin) denotes the first word of the entity, M (Middle) represents the middle part of the entity, E (End) indicates the last word of the entity, S (Single) signifies a single-word entity, and O (Other) represents non-entity portions.

2.3 Event Extraction Methods Based on Machine Learning and Deep Learning

The event argument extraction task primarily utilizes two sets of models for comparison and optimal selection. The first set includes traditional machine learning models, specifically the HMM (Hidden Markov Model) and CRF (Conditional Random Field) models. The second set comprises deep learning models, namely the Bi-directional Long Short-Term Memory model(BiLSTM) and BiLSTM-CRF models. The main reason for comparative experiment is that traditional machine learning models and deep learning models have different characteristics and strengths when handling text extraction tasks. Comparing different models not only helps evaluate their respective advantages and limitations but also provides a basis for identifying the optimal model for extracting illegal farmland use events.

Statistical-based machine learning models, such as HMM and CRF, have been widely used in early text processing tasks due to their simplicity and interpretability. The HMM model addresses sequence labeling problems by estimating state transition probabilities, emission probabilities, and initial state probabilities. Its computational principles are straightforward, making the model easy to understand and implement. HMM performs well on short-sequence tasks, particularly for texts with simple annotation structures. Even with relatively small datasets, HMM can still effectively annotate by estimating state transitions and emission probabilities. However, HMM fails to capture long-distance dependencies and performs less well than complex models like CRF in multi-label tasks. The ability to model contextual information is also limited. CRF, on the other hand, predicts labels by seeking a globally optimal solution, offering greater flexibility in feature selection and the ability to incorporate more global information. Additionally, CRF can improve annotation accuracy by modeling dependencies between labels, especially when the dependencies are complex. This makes CRF particularly advantageous in such scenarios. Both HMM and CRF have their strengths and weaknesses in text recognition tasks. As a result, these two statistical machine learning models are selected for comparative analysis to evaluate their effectiveness in identifying illegal land use incidents.

Deep learning-based models, such as BiLSTM and BiLSTM-CRF, leverage neural network architectures, particularly the introduction of LSTM networks, to deeply learn contextual information from text. The BiLSTM model exhibits enhanced long-term memory capabilities, enabling it to recognize implicit, deep-level semantic dependencies in text. By learning text sequences from both forward and backward directions, BiLSTM effectively processes information flow in long sequences, demonstrating superior semantic modeling abilities. On the other hand, BiLSTM-CRF combines the powerful context-capturing capability of BiLSTM with CRF's ability to model label dependencies, making it particularly effective in long-sequence labeling tasks. This hybrid model excels at identifying entities and relationships in text. During training, the CRF layer further improves labeling accuracy by optimizing the global label sequence for the most optimal path. However, both models require longer training times and may underperform in scenarios with limited sample sizes compared to traditional statistical-based machine learning models. Therefore, these two deep learning-based models are selected for comparative analysis to evaluate their effectiveness in identifying illegal land use incidents. Additionally, their performance will be compared with that of statistical machine learning models to provide a comprehensive assessment.

2.4 Evaluation Metrics

In order to comprehensively evaluate the performance of two sets of models in the cultivated land illegal event argument extraction, the following common evaluation metrics such as Precision, Recall, and F1-score were employed to measure the models' performance in various aspects. Precision, which is defined as the proportion of actual positive samples among those predicted as positive by the model, focuses on the accuracy of the model's positive predictions. The formula is:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

where TP=True Positives, which refer to correctly predicted positive samples.
FP=False Positives, which refer to incorrectly predicted positive samples.

Recall, which is defined as the proportion of actual positive samples among all actual positive samples, focuses on the model's ability to capture positive samples. The formula is:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

where FN = False Negatives, which refer to Incorrectly predicted negative samples.

F1-score, which is the harmonic mean of precision and recall, balancing the Precision and Recall, is suitable for imbalanced datasets. The formula is:

$$\mathrm{F1-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

3. Experiment and Results

3.1 Data Source and Preprocessing

The text of illegal cultivated land use cases used in this study is sourced from notifications issued by the Ministry of Natural Resources of China, as well as the natural resources departments of various provinces and local natural resources authorities, spanning from 2022 to 2024. These notifications compile over 400 texts of illegal cultivated land use cases from various regions

across China. The data from these cases are publicly available and authoritative, encompassing structured information such as the location, time, type of event, illegal activities, and occupied area of the cases. This provides crucial reference information for researching the spatial distribution, temporal changes, and primary types of illegal cultivated land use activities. An example of the illegal cultivated land use case is shown in Figure 2 as bellowing.

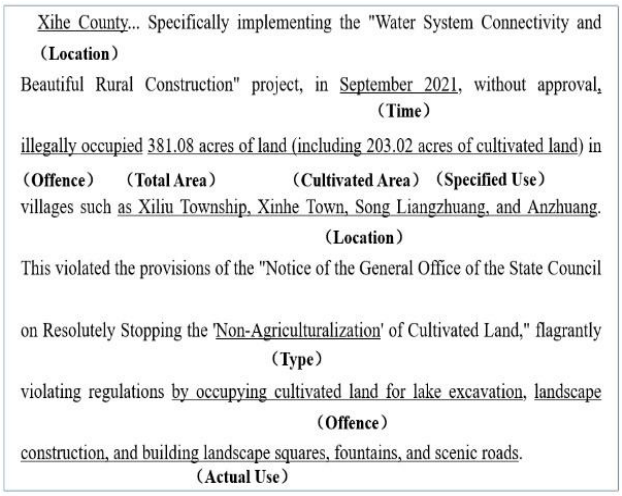
Xihe County... Specifically implementing the "Water System Connectivity and (Location)
Beautiful Rural Construction" project, in September 2021, without approval, (Time)
illegally occupied 381.08 acres of land (including 203.02 acres of cultivated land) in
(Offence) (Total Area) (Cultivated Area) (Specified Use)
villages such as Xiliu Township, Xinhe Town, Song Liangzhuang, and Anzhuang. (Location)
This violated the provisions of the "Notice of the General Office of the State Council
on Resolutely Stopping the 'Non-Agriculturalization' of Cultivated Land," flagrantly (Type)
violating regulations by occupying cultivated land for lake excavation, landscape (Offence)
construction, and building landscape squares, fountains, and scenic roads. (Actual Use)

Figure 2. An example of the illegal cultivated land use case (https://www.mnr.gov.cn/dt/ywbb/202304/t20230414_2781713.html)

In order to effectively remove redundant and irrelevant information and ensure that the text data retained in the constructed dataset contains the key information of the event, this paper conducted preprocessing on the original text. Specifically, irrelevant content such as serial numbers in the titles of each event, redundant spaces, and special characters were removed, while the core information related to illegal cultivated land use events was retained. Additionally, all events were arranged in chronological order based on their occurrence time. This processing not only enhanced the accuracy of the text data but also improved its practicality for subsequent research and analysis. Ultimately, the processed texts of illegal cultivated land use cases contained a total of over 53,000 characters of information.

3.2 Experimental Results and Analysis

3.2.1 Experimental Parameter Settings: When training models using HMM and CRF, in order to ensure fairness and comparability among the models, this paper adopted the default feature template configuration and made appropriate parameter adjustments to achieve optimal classification performance. When using the BiLSTM and BiLSTM-CRF models, a PyTorch environment was set up for model training, validation, and testing, with the basic parameter configurations as shown in Table 2. Additionally, the Adam optimizer was used in the training process of all models. By adjusting the learning rate for each parameter, Adam allows for smoother and more efficient updates of the parameters during training. At the same time, GPU acceleration was utilized to speed up the model training process, enabling efficient handling of large-scale text data.

Deep Learning Parameters	Value
Learning rate	0.001
batch_size	64
stride	5
epoch	200
emb_size	128
hidden_size	128
Optimizer	Adam

Table 2. Parameter Settings for Deep Learning Models

3.2.2 Results and Discussion: The annotated information from over 400 case texts was converted into BEMS format files, which can be processed by machine learning and deep learning models. The experiments were conducted with a 7:2:1 ratio for the training set, validation set, and test set. Ultimately, the argument information of illegal cultivated land use events in natural resource cases was extracted. Figure 3 provides examples of the extracted argument information for cultivated land non-agriculturalization events and cultivated land non-grainification events. In the figure, both types of events occurred in 2021. The left part of the figure shows an example of a non-agriculturalization event, where the Agriculture and Rural Affairs Bureau of Xinhe County dug lakes and created landscapes in villages such as Xiliu Township, illegally constructing scenic roads on cultivated land, occupying a total area of 381.08 acres, including 203.02 acres of cultivated land. The right part of the figure shows an example of a non-grainization event, where a company in Chongqing illegally occupied cultivated land in Zhongfeng Town to construct houses, occupying a total area of 106.5 acres, including 97.5 acres of cultivated land.

Furthermore, by analysing the precision, recall, and F1-score of the extraction results from the four models, as shown in Table 3, the experiments demonstrated that combining the CRF model with BiLSTM can more effectively extract argument information for illegal cultivated land use events, achieving precision, recall, and F1-score values of 0.9128, 0.9147, and 90.81%, respectively. These values indicated that the BiLSTM-CRF model had a high degree of match between the extracted information and the actual information, demonstrating its effectiveness and reliability in this task. It is worth noting that the traditional CRF model based on statistics ranks second only to the BiLSTM-CRF model in terms of extraction results, with an F1-score value of 0.9013%.

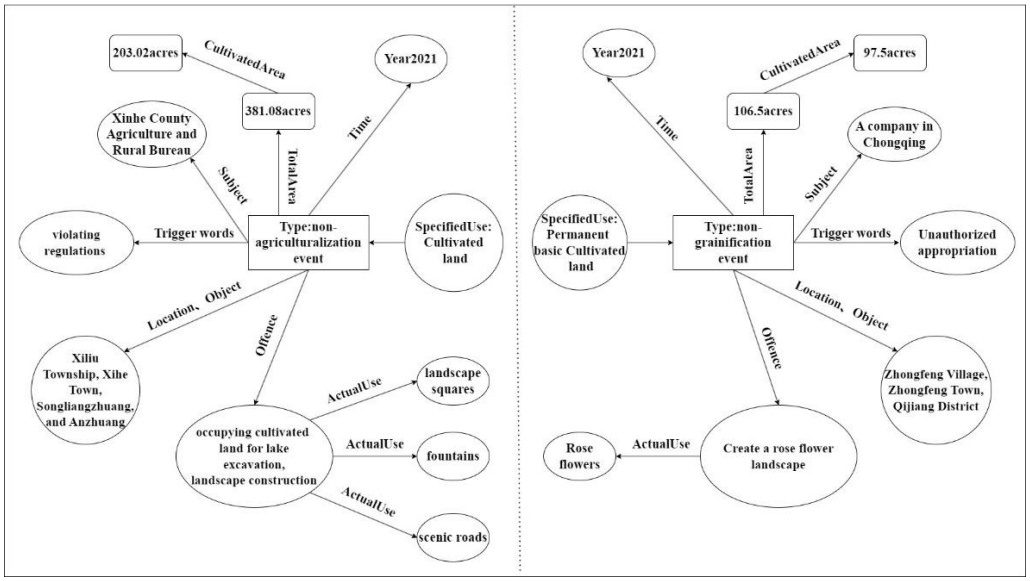

Figure 3. Example of the extracted argument information for cultivated land non-agriculturalization events and non-grainification events

Model name	Precision	Recall	F1_score
HMM	0.8882	0.8783	0.8793
CRF	0.9088	0.9095	0.9013
BILSTM	0.8938	0.8981	0.8913
BILSTM–CRF	0.9128	0.9147	0.9081

Table 3. Experimental results of illegal cultivated land use event arguments extraction

On the other hand, we observed that the training time for deep learning models such as BiLSTM and BiLSTM-CRF is relatively long (taking a total of 260 minutes in this experiment) and requires more computational resources. In contrast, the CRF model exhibits faster training and inference speeds (taking a total of 160 minutes in this experiment). Given the rapid nature of the CRF model, one potential future direction could be to leverage the preliminary processing results of the CRF model as prior inputs to enhance the processing speed and accuracy of the BiLSTM-CRF model. By incorporating the CRF model's rapid processing capability, we could potentially reduce the overall computational burden and improve efficiency while maintaining or even enhancing the accuracy of the argument information extraction task for cultivated land events. This approach would combine the strengths of both models, utilizing the speed of the CRF model for initial processing and the accuracy of the BiLSTM-CRF model for final output.

In summary, deep learning network models based on LSTM exhibit a strong capability in processing texts related to illegal farmland events by fully utilizing contextual information for semantic understanding. When faced with long texts, complex contexts, and multi-label entity recognition tasks, these models demonstrate more stable performance compared to traditional machine learning models. This underscores the powerful ability of LSTM-based models in text extraction related to illegal farmland cases, highlighting their potential and effectiveness in this specific domain.

4. Conclusion

This paper begins with the research needs for the efficient utilization of historical cases and intelligent law enforcement in the field of administrative law enforcement. It employs event extraction technology to transform unstructured text information from cases involving illegal occupation of cultivated land in natural resources into structured data for storage, facilitating subsequent research in this field, such as knowledge graph construction and predictive modeling. Firstly, the paper uses a thematic web crawler to collect real case text data on changes in cultivated land use from the official website of the Ministry of Natural Resources. Subsequently, it constructs an entity argument recognition framework, defines the types of cultivated land violation events and their argument structures, and builds a training dataset for cultivated land violation cases through manual annotation. Finally, the paper employs statistical models such as HMM and CRF, as well as deep learning models such as BiLSTM and BiLSTM-CRF, to optimize the extraction of event arguments. The experimental results demonstrate that the BiLSTM-CRF model achieves excellent event extraction performance, validating the effectiveness of existing deep learning models on text data related to illegal occupation of cultivated land in natural resources. This research holds both theoretical and practical significance.

Due to the cost of data annotation, this paper only extracts over 400 cases for experimentation. Future work will involve extracting more cases for further validation. Additionally, this study primarily focuses on the application of existing models. Subsequent research should explore how to innovate models by incorporating domain-specific text characteristics, and how to investigate methods for optimizing models under low-resource conditions.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (42201514).

References

Ahn, D.,2006: The stages of event extraction. In Proceedings of the Workshop on Annotating and Reasoning about Time and Ev ents., 1-8.

General Office of the State Council of the People's Republic of China., 2020. Notice of the General Office of the State Council on Resolutely Stopping the “Non-Agriculturalization” of Cultiv ated Land, https://www.gov.cn/zhengce/content/2020-09/15/con tent_5543645.htm. (15 September 2020).

Hu, D.M., Yuan, W., Niu, F.Q., et al. 2022: Multi-model Fusion Extraction Method for Chinese Text lmplicative Meteorologica l Disasters Event Information. *Joumal of Geo-information Scien ce*., 24(12):2342-2355. doi: 10.12082/dqxxkx.2022.220088.

Li, Q., Li, J., Sheng, J., et al. 2022: A survey on deep learning e vent extraction: Approaches and applications. *IEEE Transaction s on Neural Networks and Learning Systems*., vol. 35, no. 5, pp. 6301-6321, May 2024, doi: 10.1109/TNNLS.2022.3213168.

Sahnoun, S., Elloumi, S., Ben, Y.S., 2020: Event detection base d on open information extraction and ontology. *Journal of Infor mation and Telecommunication*., 4(3): 383-403. doi: 10.1080/24 751839.2020.1763007.

Wang, R.Y., Xiang, W., 2023: A Survey of Document level Event Extraction. *Journal of Chinese Information Processing*,. 37(06): 1-14. doi:10.3969/j.issn.1003-0077.2023.06.001

Wang, Y.Q., Zhou, Q.S., 2024: A Research on Internet Open So urce Information Extraction Based on Pre-trained Language Mo del and Intelligence Analysis Application: Take “Academic, Le cture, Forum” and Other Conference Activities as an Example. *I nformation Studies:Theory & Application*., 47(1): 154-163. doi. org/10.16353/i.cnki. 1000-7490.2024.01.019.

Zhang, M., Chen, J.H., Sun, R.R., 2021: Rule based information extraction of urban rail transit safety cases and its common kno wledge meta-model representation. *Science Technology and En gineering*., 21(15): 6435-6440. doi: 10.3969/j.issn.1671-1815.2 021.15.046.

Zhang, C.J., Zhang, L., Chen, Y.B., et al. 2022: Construction M ethod of Interactive Geological Entity Annotation Corpus Based on BERT. *Joumal of Geo-information Science*., 38(04): 7-12. d oi: 10.3969/i.issn.1672-0504.2022.04.002.