Postprint: Integrating Syntactic Features and Syntactic Similarity for Online Public Opinion Emergency Event Identification

Authors: Chen Jianyao, Zhai Shanshan, Xia Lixin, Liu Deyin

Date: 2023-04-01T16:02:50+00:00

Abstract

[Purpose/Significance] To rapidly and accurately identify events from texts of sudden network public opinion. [Method/Process] This paper proposes a method for identifying sudden events in network public opinion that integrates syntactic features and syntactic similarity. An event-oriented syntactic feature extraction method is proposed based on syntactic features, and an event syntactic feature library is constructed using event semantic annotation and syntactic feature extraction methods, through which network public opinion sudden events are identified by calculating the syntactic similarity between the text to be tested and the syntactic library. [Results/Conclusion] Taking the COVID-19 pandemic as an example, the proposed method achieves an optimal similarity of 0.93 for this public opinion scenario, identifying 160 events and 30 non-events from a new text segment at this similarity threshold, achieving an F1-score of 0.848. Method evaluation demonstrates the effectiveness of the innovations in the proposed network public opinion sudden event identification method, particularly in utilizing syntactic similarity for event identification and merging identical adjacent parts of speech.

Full Text

Network Public Opinion Emergency Event Recognition Based on Syntactic Features and Syntactic Similarity

Chen Jianyao, Zhai Shanshan, Xia Lixin, Liu Deyin

School of Information Management, Central China Normal University, Wuhan 430079

Abstract: [Purpose/Significance] This study aims to identify events quickly and accurately from sudden network public opinion texts. [Method/Process]

We propose a method for identifying network public opinion emergency events that integrates syntactic features and syntactic similarity. Combining syntactic features, we propose an event-oriented syntactic feature extraction method. We construct an event syntactic feature database using event semantic annotation and syntactic feature extraction methods, and identify network public opinion emergency events by calculating the syntactic similarity between the text to be tested and the syntactic database. [Result/Conclusion] Taking the COVID-19 pandemic as an example, the optimal similarity of the proposed network public opinion emergency event identification method under this public opinion scenario is 0.93. At this similarity level, 160 events and 30 non-events were identified from a new text, with the F1 value reaching 0.848. Method evaluation demonstrates the effectiveness of innovations in using syntactic similarity to identify events and merging identical adjacent parts of speech.

Keywords: Network Public Opinion; Event Identification; Syntactic Features; Syntactic Similarity

Classification Number: G250.2
DOI: 10.13266/j.issn.0252-3116.2021.09.005

According to the 44th "Statistical Report on China's Internet Development" released by the China Internet Network Information Center (CNNIC) [**?**], as of June 2019, China's internet user population reached 854 million, with an internet penetration rate of 61.2%, an increase of 1.6 percentage points from the end of 2018. The popularization and democratization of the internet have made it more convenient for the public to pay attention to and respond to public opinion events, resulting in richer content in network public opinion. Against this background, quickly and accurately identifying events that reflect public attitudes and public opinion trends from sudden network public opinion texts, and providing targeted reference opinions for government guidance strategies, has become an important research topic in the field of network public opinion.

Network public opinion emergency events refer to events that reflect different public views on sudden social issues online. Events possess characteristics of abstraction, generality, and semantic completeness, represented as event triples $E = (S, P, O)$, where P is the trigger word, S is the agent, and O is the patient [**?**]. A complete event must contain a trigger word, which determines the event type. Agents and patients can be partially omitted. For example, "Typhoon makes landfall (S, P)," "Watch movie (P, O)," and "Guizhou Liangshan mountain fire breaks out (S, P, O)" can all be called events.

The task of network public opinion emergency event identification mainly studies the identification of structured event texts containing event elements from unstructured social media data. Since 2005, event extraction has been included in the ACE evaluation conference [**?**], and event identification is an important component of event extraction tasks. Event extraction can be divided into topic event extraction and meta-event extraction. Topic events refer to a group of events related to a certain theme, consisting of a core event and all directly re-

lated events or activities [**?**]. Meta-events mainly describe the main component structure of action events, usually using verbs and nouns to represent the occurrence of actions or changes in states. Chinese event identification technology has also made considerable progress under the research of scholars both domestically and internationally [**?**]. Compared with the clear sentence structure in English, the arrangement and combination of Chinese words are more complex and flexible, and words have many polysemous phenomena. The meaning of events also needs to be distinguished based on contextual semantics, which brings certain difficulties to Chinese event identification technology. How to reduce the dimensionality of Chinese text and the semantic association between words has become a major challenge in Chinese event identification.

To explore an identification method suitable for the Chinese environment for network public opinion emergency events, we propose a network public opinion emergency event identification model that integrates syntactic features and syntactic similarity. Taking the COVID-19 pandemic as an example, we construct a network public opinion event syntactic feature database and use the syntactic similarity between sentences to identify new events in network public opinion.

2 Related Research

Different researchers in different fields have different definitions of events. Event logic graph researchers define events as abstract, generalized event triples with complete semantics [**?**]. In the field of linguistics, events are considered terms composed of predicate verbs and the time and circumstances of action occurrence [**?**]. The ACE evaluation conference believes that events are things or state changes that occur at a specific time point or period, within a specific geographical range, and consist of one or more actions participated in by one or more roles [**?**]. Among the above definitions, the event definition given by event logic graph researchers is more suitable for network public opinion emergency events in this study due to its structured characteristics, so this definition is adopted.

2.1 Related Research on Event Identification Methods

Event identification methods mainly fall into two categories: pattern matching-based methods and machine learning-based methods. Pattern matching-based methods identify and extract events under the guidance of certain patterns. Patterns are mainly used to specify the contextual constraints that constitute the target information, embodying the fusion of domain knowledge and linguistic knowledge [**?**]. These methods can be divided into rule-based extensions and relationship-based restrictions. The former tends to expand trigger word lists and improve knowledge base construction at the macro level, while the latter tends to integrate text information units, semantic consistency reasoning, and semantic constraints at the micro level. Currently, some scholars use pattern matching for war event extraction [**?**]. Machine learning-based methods use statistical models for event identification and extraction. This method

has been more mainstream in recent years. Common learning methods include Conditional Random Field models [**?**], Hidden Markov models [**?**], and Support Vector Machine models [**?**]. He Ruifang et al. [**?**] treated event extraction as a sequence labeling task and constructed a Chinese event extraction joint model based on CRF multi-task learning, expanding on the defects of models based only on CRF. Liu Zhongbao et al. [**?**] used the BERT model and LSTM-CRF model to extract historical events and their constituent elements.

It can be seen that research on event identification methods has made great progress and can accurately identify events from texts. However, most of these methods rely on large-scale, comprehensive training sets, requiring the construction of knowledge bases for event identification. When applied to the field of network public opinion emergency events, these methods face the problem of insufficient training set corpora in the early stage of network public opinion. Therefore, this study aims to propose an emergency event identification method that can be applied to the network public opinion field based on current event identification methods.

2.2 Research Status of Network Public Opinion Emergency Event Identification

Wei Yongqing et al. [**?**] studied event feature extraction methods and emotional feature burstiness on the basis of researching the propagation patterns of sudden events in network public opinion, using them to identify sudden events and provide data support for predicting event development. Wu Peng et al. [**?**] used game analysis to derive a probability model for sudden event information publishers being followed on Weibo, laying a foundation for identifying key nodes in network public opinion emergency information transmission. Liu Yashu et al. [**?**] used the LDA method to divide topics in network public opinion emergency comment data and construct an event evolution topic graph to dynamically track public opinion and understand the development direction of network public opinion emergencies. Lan Yuexin [**?**, **?**] established derivative public opinion monitoring and warning models and research on the propagation patterns of sudden event network public opinion information, providing references for government network public opinion management and early warning research. Zhang Yuliang [**?**] divided sudden event network public opinion into three stages: generation, diffusion, and decline/calm, providing effective theoretical support for government departments to assess the actual situation of sudden events and grasp the development trend of sudden event network public opinion. Chen Sijing et al. [**?**] used a dynamic identification method for key nodes considering user behavior characteristics, network global information, and influence decay mechanisms to identify key nodes and their evolution characteristics in different stages of sudden event information dissemination. Li Gang et al. [**?**] used the LDA topic model and Maximum Entropy model with Mutual Information (MaRxEnt-MI) to extract event summary keywords and generate event summaries. Xia Lixin et al. [**?**] constructed network public opin-

ion event features from multiple dimensions from a visualization perspective to form visual event summaries.

It can be seen that current scholars have conducted event identification and application research in the field of network public opinion, but most research focuses on event propagation and public opinion development, without proposing an emergency event identification method applicable to the network public opinion field. Therefore, based on current research on event identification and application in the network public opinion field, we propose a general emergency event identification method for network public opinion to provide references for subsequent researchers studying network public opinion based on events.

2.3 Related Research on Syntactic Analysis

Syntactic analysis research is generally divided into rule-based methods and statistical methods. The former is based on linguistic theory, while the latter describes grammatical rules and language forms in some way [**?**]. Yuan Lichi [**?**] established a statistical model for syntactic analysis based on dependency relations. Guo Xiyue et al. [**?**] integrated syntactic features, semantic features, dependency relations, core predicates, and semantic role labeling features for entity relation extraction, and experimental results showed the effectiveness of incorporating syntactic features. Xu Fei et al. [**?**] used the BiLSTM-CRF model for part-of-speech tagging of food events and achieved good results. Hu Baoshun et al. [**?**] proposed a new answer extraction algorithm based on syntactic structure feature analysis and classification technology, and experimental results proved that the method based on syntactic structure features outperformed current typical algorithms. Chen Yongbo et al. [**?**] proposed a dependency parsing algorithm combining simple edge priority and SVM, and experimental results proved that for complex noun phrase dependency parsing, the algorithm's accuracy was significantly improved compared to simple edge priority algorithms.

Research and application of syntactic analysis by scholars have proven the effectiveness of syntax in expressing text features. Based on this, we believe that Chinese texts expressing specific events have certain syntactic patterns. In the case of insufficient training set corpora in the early stage of network public opinion, syntactic features can replace text features for event identification. Using event syntax to identify events can effectively reduce the dimensionality of Chinese text, greatly reducing the workload and complexity of event identification. At the same time, this method can reduce dependence on specific public opinion domain dictionaries, making event identification methods more widely applicable.

3 Syntactic Similarity Measurement Based on Syntactic Feature Extraction

Syntactic similarity measurement based on syntactic feature extraction is mainly divided into two sub-modules: (1) sentence syntactic feature extraction for event recognition, where event syntax contains event semantic logic under the event framework; (2) event syntactic similarity calculation method based on syntactic features. The higher the similarity between two event syntactic structures, the more similar the two events are at the syntactic-semantic structure level.

3.1 Sentence Syntactic Feature Extraction for Event Recognition

We use a word segmentation tool to segment event texts and perform part-of-speech tagging. Taking the event sentence "Turkey fires again at a nuclear power" as an example, after word segmentation and part-of-speech tagging, we obtain the text vector:

$$E = [\text{"Turkey"} : n, \text{"again"} : d, \text{"at"} : p, \text{"—"} : m, \text{"nuclear power"} : n, \text{"fire"} : v]$$

After syntactic feature extraction of text vector E, we obtain the syntactic feature vector:

$$P = [n, d, p, m, n, v]$$

Through sentence syntactic feature extraction, the representation of events is converted from text vector E to syntactic feature vector P, which shifts the dimension of event identification from text features to syntactic features. However, this also presents a problem: the types of part-of-speech are far fewer than the types of words, and many different words have the same part-of-speech, causing redundancy in syntactic feature vectors. Due to different colloquial expressions, the same event may use multiple language expressions. To reduce this redundancy and the complexity of syntactic features, we adopt the method of "merging identical adjacent parts of speech" to reduce the types of syntax and vector dimensions for the same event. For example, for the syntax:

$$P = [n, n, d, p, m, m, n, v]$$

We simplify it to:

$$P = [n, d, p, m, n, v]$$

The purpose of merging identical parts of speech is to generalize syntactic types, assuming that adjacent words with the same part-of-speech express the same

semantic features. This allows the model to maximize its advantages even when lacking a large text training set. However, due to phenomena such as "polysemy" and "multiple words with the same meaning," some syntax annotated by word segmentation tools has semantic conflicts, leading to errors. To reduce such errors, we adopt manual inspection to identify these conflicting syntactic patterns. Through manual identification of erroneous syntax via event semantic relations, we form an erroneous syntactic pattern dictionary. This dictionary is placed in the subsequent network public opinion emergency event identification model. When the syntax to be identified exists in the erroneous syntactic pattern dictionary, it indicates that the syntax is erroneous and is directly judged as a non-event. The workload of manual inspection of erroneous syntactic patterns is the same as event annotation workload in the initial stage, but as the erroneous syntactic pattern dictionary grows, fewer syntax will have semantic conflicts, and the corresponding workload will decrease. Compared with other error reduction methods, manual inspection is more suitable for syntactic feature extraction due to its convenience and controllability. The specific process of event syntactic pattern extraction is shown in Algorithm 1.

Algorithm 1: Sentence Syntactic Feature Extraction

Input: sentences[0..n-1]: an array containing n sentences to be processed; f1(sentence): a function for word segmentation; f2(word): a function for part-of-speech tagging; f3(pattern): a function for merging identical adjacent parts of speech in extracted syntax

Output: patterns: event syntactic set

1. function Pattern(sentences[0..n-1]: array of sentence; f1: function; f2: function; f3: function): patterns;
2. var
3. words[0..m-1]: an array containing m words;
4. nominal: part-of-speech tagging sequence;
5. begin
6. for i ← 0 to n-1 do
7. pattern ← null
8. words[0..m-1] ← f1(sentences[i])
9. for i ← 0 to m-1 do
10. nominal ← f2(words[i])
11. pattern ← pattern + nominal
12. pattern ← f3(pattern)
13. if pattern not in patterns then
14. patterns ← patterns + pattern
15. endif
16. return patterns
17. end

3.2 Event Syntactic Similarity Calculation Method Based on Syntactic Features

Text cosine similarity is a commonly used text similarity measurement standard. Traditional text vector cosine similarity can express the similarity between two text documents, judging the relationship between two documents through word vector distance. We modify text cosine similarity and apply it to event syntactic similarity calculation. Syntactic feature vector similarity can express the similarity between two events at the semantic logic level of syntactic structure. We calculate the similarity between the syntax of the event to be identified $P_1 = [x_1, x_2, ..., x_i]$ and the syntax in the event syntax library $P_2 = [y_1, y_2, ..., y_i]$, taking the maximum similarity as the final similarity. If the final similarity is 100%, it means that the syntax already exists in the event syntax library, and the text is an event text. In addition, event texts must contain trigger words σ. Texts without trigger words are directly judged as non-event texts. Considering all aspects, the final similarity calculation method of this model is shown in Formula (1):

$$\cos(\theta) = \frac{\sum_{i=1}^{n}(x_i \cdot y_i)}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{1}$$

The event syntactic similarity calculation algorithm is shown in Algorithm 2:

Algorithm 2: Event Syntactic Similarity Calculation

Input: patterns[0..n-1]: an array containing n event syntactic features (pattern); sentence: the syntax of the text to be tested; f1(sentence): returns 1 if the input syntax contains trigger word σ, otherwise returns 0; f2(pattern, sentence): calculates the cosine similarity between two event syntactic structures

Output: cos: event syntactic similarity

1. function Cos(patterns[0..n-1]: array of pattern; sentences: text to be tested; f1: function; f2: function): cos;
2. var
3. σ: trigger word identification variable;
4. cos: syntactic similarity;
5. temp: temporary variable;
6. begin
7. for i ← 0 to n-1 do
8. pattern ← patterns[i]
9. σ ← f1(sentence)
10. temp ← f2(pattern, sentence)
11. temp ← temp $\cdot \sigma$
12. if temp > cos then
13. cos ← temp
14. endif

15. return cos
16. end

4 Network Public Opinion Emergency Event Identification Method Integrating Syntactic Features and Syntactic Similarity

4.1 Advantages of Syntactic Features and Syntactic Similarity in Event Identification

Research by domestic and foreign scholars has proven that incorporating syntactic analysis or syntactic features into syntactic analysis statistical models [**?**], Chinese entity relation extraction [**?**], answer extraction algorithms [**?**], and Chinese complex noun phrase analysis [**?**] have all achieved good experimental results. The syntactic features of sentences describe the grammatical rules and language forms of texts from the semantic level [**?**], and syntactic features can express the semantic features of sentences. Different from traditional text features, syntactic features describe the dependency structure and phrase structure in sentences, making the event identification model pay more attention to the semantic logic and dependency relationships between event words during event identification, which greatly helps improve the recall and precision rates of event identification models. Events themselves have certain semantic logic and syntactic structures, and using syntactic features to express events has inherent advantages. Using syntactic features can shift the focus of event identification from text content to semantic logic, thereby avoiding the scale of Chinese text types and quantities to achieve event identification.

4.2 Overall Design Approach for Network Public Opinion Emergency Event Identification Method

We construct a network public opinion emergency event identification model as shown in Figure 1 [Figure 1: see original paper]. The model consists of two parts: (1) Event syntactic feature database construction. First, we obtain relevant public opinion corpora training sets from social media through web crawlers, manually annotate the event set $E = \{E_1, E_2, ..., E_j | E_j \in TD_i\}$ from document $TD = (TD_1, TD_2, ..., TD_i)$, then obtain the syntax $P_m = \{< E_1 : P_1 >, < E_2 : P_2 >, ..., < E_j : P_j >\}$ corresponding to event set E through syntactic feature extraction methods. We obtain all event syntax from the public opinion corpus training set documents using this method, and after deduplication and manual correction, form the event syntactic feature database of the public opinion corpus. (2) Text to be tested event identification. First, we perform sentence segmentation on the document to be identified $D = (D_1, D_2, ..., D_i)$, cutting document D_i into a sentence set $S = \{S_1, S_2, ..., S_j | S_j \in D_i\}$ contained by sentences. Then we obtain the syntactic feature set $P_n = \{< S_1 : P_1 > , < S_2 : P_2 >, ..., < S_j : P_j >\}$ of document D_i through syntactic feature extraction. P_n enters the model as the syntax of the text to be identified and

calculates similarity with existing event syntax in the public opinion event syntactic database. The text to be tested with syntactic similarity greater than or equal to the model similarity threshold α is identified as an event text.

4.3 Implementation of Network Public Opinion Emergency Event Identification Method

4.3.1 Semantic Annotation of Network Public Opinion Emergency Events The purpose of semantic annotation of network public opinion emergency events is to annotate public opinion texts to obtain a certain scale of known events, laying the foundation for subsequent construction of the event syntactic feature database. When performing event semantic annotation, we retain not only the subject-predicate-object related entities contained in the event triple but also other entity information such as location entities, time entities, and event entities, making the obtained event syntactic patterns more complete and improving the accuracy of subsequently identified new events using syntactic patterns.

For event semantic annotation, we define the following annotation principles:

Principle 1: The annotated event text must be able to derive the occurrence of an event.

Principle 2: Under the condition of satisfying Principle 1, the derived event must be an event that has actually occurred or is occurring. For example, if the event text contains negative words, future occurrence words, possible occurrence words, or individual subjective speculation words, it is not counted as an event.

Principle 3: Under the condition of satisfying Principle 1, time entities and location entities in the event text belong to part of the event and should be annotated.

Principle 4: Under the condition of satisfying Principle 1, one event can serve as the agent or patient of another event, meaning that the event itself can also be an entity.

4.3.2 Event Identification Based on Network Public Opinion Emergency Event Identification Model Using the network public opinion emergency event identification model proposed in Section 4.2 as the core, we identify events in specific domain network public opinion emergencies. The event syntactic feature database serves as the event syntactic training set in the event identification method, so constructing a complete event syntactic feature database is the primary task. After completing the construction of the event syntactic feature database, we segment the text to be tested and perform event identification.

(1) Event Syntactic Feature Database Construction. The construction of the event syntactic feature database is divided into two sub-modules: (1)

Network public opinion corpus collection and semantic annotation. We obtain relevant domain network public opinion emergency event corpora through self-written Python crawlers. After cleaning the data to a certain extent, we annotate the collected corpus information using the network public opinion emergency event semantic annotation proposed in Section 4.3.1 to form a public opinion event corpus. (2) Sentence part-of-speech tagging and syntactic feature extraction. We use the jieba word segmentation tool to perform word segmentation and part-of-speech tagging on event texts in the event corpus in sequence, then extract syntactic features from events through the sentence syntactic feature extraction method for event recognition proposed in Section 3.1. The obtained syntax enters the event syntactic feature database. To avoid generating duplicate syntactic features, we need to perform duplicate judgment on newly entered syntactic features. If the syntax is duplicated, it will not be entered into the database. At the same time, we adopt an error pattern checking and feedback mechanism to reduce the generation of erroneous syntax through manual methods. The specific process of event syntactic feature database construction is shown in Figure 2 [Figure 2: see original paper]. To ensure the accuracy of word segmentation and part-of-speech tagging, in addition to using the jieba word segmentation dictionary, we will also combine custom dictionaries for specific network public opinion emergencies. For example, in the "Typhoon Lekima" event, we define "Lekima" as a word with part-of-speech "noun (n)" through the dictionary, so that this word will not be split during word segmentation and its part-of-speech will be correctly identified.

(2) Text to be Tested Segmentation and Event Identification. The main difficulty in segmenting text to be tested for events is not knowing the location of event texts in the public opinion corpus to be tested. Since the event structure in the text is not known in advance, the location of events may exist in half a sentence or a sentence, and the event itself can also become an element of another event. For this, we adopt the method of repeated segmentation of a piece of text to be tested. For example, for the text to be tested "The United States investigates ZTE again, causing its stock price to plummet," we can segment it into three texts: "The United States investigates ZTE again," "causing its stock price to plummet," and "The United States investigates ZTE again, causing its stock price to plummet." Repeated segmentation of a piece of text can identify all events contained in it and avoid omissions.

Document D_i contains a large number of network public opinion emergency events. First, we perform sentence segmentation on it. We extract syntactic features from the text to be tested through the sentence syntactic feature extraction method for event recognition proposed in Section 3.1, obtaining the syntactic feature set $P_n = \{< S_1 : P_1 >, < S_2 : P_2 >, ..., < S_j : P_j >\}$ of document D_i. P_n enters the model and calculates similarity with syntax in the event syntactic feature database according to Algorithm 2. Due to different categories of network public opinion, each specific category of network public opinion event identification model corresponds to a specific similarity threshold α. The text to be tested with final syntactic similarity greater than or equal

to similarity threshold α is identified as an event text. The syntactic similarity that makes the model's $F1$ value optimal is the similarity threshold α. The $F1$ value calculation method is shown in Formula (2):

$$F1 = \frac{2 \times P \times R}{P + R} \tag{2}$$

5 Empirical Analysis of Network Public Opinion Emergency Event Identification: The Case of COVID-19

To verify the effectiveness of the proposed network public opinion emergency event identification method integrating syntactic features and syntactic similarity, we take the COVID-19 pandemic as an example.

5.1 Corpus Collection and Event Annotation

We use self-written Python crawlers to crawl relevant data from the Weibo platform using the keyword "COVID-19 pandemic," and perform cleaning and preprocessing on the data to form three different network public opinion corpus documents D_1, D_2, and D_3. Among them, document D_1 is used to construct the event syntactic feature database, document D_2 is used to calculate the similarity threshold α of the event identification model in the case of the COVID-19 pandemic, and document D_3 is used to test the ability of the event identification model to identify new events from unknown texts. According to the different functions of the documents, we perform event semantic annotation on documents D_1 and D_2, and perform repeated sentence segmentation on document D_3. After event semantic annotation, document D_1 obtained 1,353 events. Through syntactic feature extraction, we constructed an event syntactic feature database containing 1,328 effective syntactic patterns, with each syntactic pattern representing the syntactic logic-level features of an event.

The events obtained from document D_1 after event semantic annotation and the corresponding event syntactic feature database are shown in Figure 3 [Figure 3: see original paper].

Document D_2 obtained 65 events and 54 non-events after event semantic annotation. Through syntactic feature extraction, we obtained 119 syntactic patterns, which are used to determine the model similarity threshold α. The events and non-events obtained from document D_2 after event semantic annotation and the corresponding sentence syntactic features are shown in Figure 4 [Figure 4: see original paper].

Document D_3 formed a text to be tested expressed as a sentence set after repeated sentence segmentation. We perform corresponding syntactic feature extraction on the sentence set to form the test set. The sentence set formed after sentence segmentation of document D_3 and the corresponding syntactic features are shown in Figure 5 [Figure 5: see original paper].

5.2 Presentation and Analysis of COVID-19 Event Identification Results

To determine the optimal similarity threshold α, the network public opinion emergency event identification model sequentially calculates the syntactic similarity between the annotated events and non-events in document D_2 and the events in the event syntactic feature database, obtaining the syntactic similarity of all events and non-events in document D_2. Then, we take the similarity threshold α values in the interval $[0, 1]$ with a step size of 0.01 until we obtain the α value that makes the model's $F1$ value optimal. The final experimental results are shown in Table 1 . The similarity threshold α in the range of $[0.89, 1]$ can achieve the optimal result of the model. By comparing the trend charts of P value, R value, and $F1$ value under different α values (shown in Figure 6 [Figure 6: see original paper]), we can determine that the optimal similarity value of the event identification model in COVID-19 public opinion is 0.93. At this time, the $F1$ value corresponding to the experimental results of document D_2 is 0.786, the P value reaches 0.713, and the R value reaches 0.877.

With the optimal similarity threshold α determined as 0.93 for the COVID-19 pandemic, the model identifies events from document D_3 with a similarity threshold of 0.93. Finally, 160 events and 30 non-events are identified from document D_3. Partial event identification results are shown in Table 2 . The $F1$ value of the event identification model on document D_3 reaches 0.848, the P value reaches 0.769, and the R value reaches 0.946, as shown in Table 3 .

5.3 Method Evaluation

The main innovation of the proposed network public opinion emergency event identification model lies in using syntactic features to replace text features to solve the problem of insufficient training sets due to short text corpora in the early stage of network public opinion emergencies, effectively reducing the dimensionality of event identification. In the model, we further reduce syntactic redundancy by "merging identical adjacent parts of speech." To verify the effectiveness of these innovations, we select the network public opinion emergency event identification model without merging identical adjacent parts of speech and the text similarity-based event identification method as comparisons.

The text similarity-based event identification method constructs Chinese text vectors after text word segmentation for similarity calculation. Except for not using syntactic features to represent event features, the calculation steps of this control group are consistent with the steps of the proposed network public opinion emergency event identification model. All three methods use document D_2 as the test set for comparison.

Under the same training set and test set, the experimental results of the three event identification methods are shown in Table 4 . The proposed network public opinion emergency event identification model performs best, with $F1$ reaching 0.786, proving the rationality of using syntax to represent event features

and the effectiveness of adopting "identical adjacent parts of speech merging." At the same time, the network public opinion emergency event identification model performs better than the model without merging identical adjacent parts of speech. The main reason for this phenomenon is that after merging identical parts of speech, the network public opinion emergency event identification model reduces the dimensionality of syntactic vectors and reduces unnecessary calculations caused by part-of-speech redundancy, thereby improving the model's $F1$ value.

According to the experimental results, we draw a comparison of the results of the three different event identification methods under different similarity threshold α values, as shown in Figure 7 [Figure 7: see original paper]. It can be seen that among these three methods, the text similarity-based event identification method has the worst experimental results, with the optimal $F1$ value only reaching 0.657. The main reason for this poor experimental result is the small scale of the training set. The training set used contains only 1,353 events, which is small compared to traditional Chinese text training sets. However, this also proves from another aspect that under the condition of a small training set scale, using syntactic similarity for event identification is superior to using text similarity, and also proves the rationality of our strategy of merging identical adjacent parts of speech, providing a new idea for network public opinion emergency event identification.

6 Conclusion and Discussion

Facing network public opinion emergencies, quickly and accurately identifying events that can reflect network public opinion attitudes from social media is of great significance for government public opinion management and decision-making by relevant departments. From the perspective of text syntactic features, we believe that event syntactic features can replace text features to represent events, using this as a breakthrough for identifying network public opinion emergency events. On this basis, we propose a network public opinion emergency event identification method integrating syntactic features and syntactic similarity.

Compared with text features, syntactic features can effectively reduce the dimensionality of Chinese text, reducing sentences composed of tens of thousands of Chinese characters to syntax composed of dozens of parts of speech, greatly reducing vector dimensionality. On this basis, we merge identical adjacent parts of speech in syntax, further reducing the types of syntax. Therefore, even with a small-scale training set corpus, the model can still achieve good event identification results.

Taking the COVID-19 pandemic as an example, the proposed network public opinion emergency event identification model, under the optimal similarity threshold of 0.93, identified events and non-events from a text to be tested, with the $F1$ value reaching 0.848. Using the same training set and test set,

the proposed method is superior to the network public opinion emergency event identification model without merging identical adjacent parts of speech and the text similarity-based event identification method, proving the effectiveness of using syntactic similarity to identify events and merging identical adjacent parts of speech.

Identifying network public opinion emergency events from social media texts is of great significance for network public opinion feature analysis and evolution analysis. In future research, we will use the network public opinion emergency event identification method proposed in this study to identify unknown events contained in network public opinion and conduct further analysis and research on network public opinion based on the identified events.

References

[1] China Internet Network Information Center. Statistical Report on China's Internet Development [R]. Beijing: China Internet Network Information Center, 2019.

[2] DING X, LI Z, LIU T, et al. ELG: an event logic graph [J/OL]. arXiv preprint arXiv:1907.08015 [2021-04-11]. https://arxiv.org/abs/1907.08015.

[3] AGUILAR J, BELLER C, MCNAMAE P, et al. A comparison of the events and relations across ace, ere, tac-kbp, and framenet annotation standards [C]//Proceedings of the second workshop on events: definition, detection, coreference, and representation. USA: Association for Computational Linguistics, 2014: 45-53.

[4] Wu G. Research and Application of Chinese Event Extraction Technology Based on Topics [D]. Suzhou: Suzhou University, 2009.

[5] Xiang W, Wang B. A Survey of Chinese Event Extraction Research [J]. Computer Technology and Development, 2020(1): 1-9.

[6] CHUNG S, TIMBERLAKE A. Tense, aspect and mood [M]//Language typology and syntactic description. Cambridge: Cambridge University Press, 1985: 202-258.

[7] DODDINGTON G R, MITCHELL A, PRZYBOCKI M A, et al. The automatic content extraction (ACE) program tasks, data, and evaluation [C]//Proceedings of the international conference on language resources and evaluation. Portugal: European Language Resources Association, 2004: 837-840.

[8] Gao Q, You H. A Survey of Event Extraction Technology Research [J]. Information Studies: Theory & Application, 2013, 36(4): 114-117, 128.

[9] Li Z, Li Z, He L. Research on War Event Extraction Technology in "Zuo Zhuan" [J]. Library and Information Service, 2020, 64(7): 20-29.

[10] He R, Duan S. Chinese Event Extraction Joint Model Based on Multi-Task Learning [J]. Journal of Software, 2019, 30(4): 1015-1030.

[11] Yu Y. Recruitment Network Information Extraction Based on Hidden Markov Model [J]. Journal of Beijing Electronic Science and Technology Institute, 2008, 16(4): 93-98.

[12] Li X, Yang X, Wei Y, et al. News Event Type Recognition Based on Support Vector Machine [J]. Geomatics World, 2019, 26(2): 73-78.

[13] Liu Z, Dang J, Zhang Z. Automatic Extraction of Historical Events from "Records of the Grand Historian" and Event Logic Graph Construction [J]. Journal of Intelligence, 2015, 38(2): 92-96.

[14] Wei Y, Yang Y, Fei S, et al. Online Sudden Event Identification Integrating User Emotions [J]. Journal of Intelligence, 2013, 32(5): 16-19.

[15] Wu P, Wang H, Liu Q, et al. Research on Threshold Model of Sudden Event Information Publishers Being "Followed" on Weibo [J]. Journal of Intelligence, 2012, 31(11): 11-13, 34.

[16] Liu Y, Zhang H, Xu H, et al. Research on Evolution Topic Graph of Network Public Opinion Emergencies Based on Multi-Dimensional Feature Fusion [J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(8): 798-806.

[17] Lan Y. Research on Derivative Public Opinion Monitoring Model for Sudden Events [J]. New Technology of Library and Information Service, 2013(3): 51-57.

[18] Lan Y, Zeng R. Research on Propagation Patterns and Early Warning Stages of Sudden Event Network Public Opinion Information [J]. Journal of Intelligence, 2012, 31(11): 11-13, 34.

[19] Zhang Y. Risk Evaluation Index System for Sudden Event Network Public Opinion Based on Occurrence Cycle [J]. Information Science, 2012, 30(7): 1034-1037, 1043.

[20] Chen S, Li G, Mao J, et al. Dynamic Identification of Key Nodes in Sudden Event Information Dissemination Network [J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(2): 178-190.

[21] Li G, Xu W, Wang X. Microblog Hot Event Summary Extraction Based on Combined Model of Event Elements [J]. Library and Information Service, 2018, 62(1): 96-105.

[22] Xia L, Chen J, Yu H. Research on Multi-Dimensional Feature Network Public Opinion Event Visual Summary Generation Based on Event Logic Graph [J]. Information Studies: Theory & Application, 2020, 43(10): 157-164.

[23] Zhang N, Zhu L. A Survey of Question Analysis Research in Chinese Question Answering Systems [J]. Technology Intelligence Engineering, 2016, 2(1): 32-42.

[24] Yuan L. Statistical Model for Syntactic Analysis Based on Dependency Relations [J]. Journal of Central South University (Science and Technology), 2009, 40(6): 1630-1635.

[25] Guo X, He T, Hu X, et al. Chinese Entity Relation Extraction Based on Syntactic and Semantic Features [J]. Journal of Chinese Information Processing, 2014, 28(6): 183-189.

[26] Xu F, Ye W, Song Y. Research on Automatic Part-of-Speech Tagging for Food Safety Events Based on BiLSTM-CRF Model [J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(12): 1204-1211.

[27] Hu B, Wang D, Yu G, et al. Answer Extraction Algorithm Based on Syntactic Structure Feature Analysis and Classification Technology [J]. Chinese Journal of Computers, 2008(4): 662-676.

[28] Chen Y, Tang A, Ji D. Chinese Complex Noun Phrase Dependency Parsing [J]. Application Research of Computers, 2015, 32(6): 1617-1620.

Author Contributions:

Chen Jianyao: Proposed the research idea, wrote the paper, and completed the experiments;
Zhai Shanshan: Revised the research framework;
Xia Lixin: Revised and guided the paper;
Liu Deyin: Data crawling and data organization.

Research on Network Public Opinion Emergency Recognition Method Based on Syntactic Features and Syntactic Similarity

Chen Jianyao, Zhai Shanshan, Xia Lixin, Liu Deyin

School of Information Management, Central China Normal University, Wuhan 430079

Abstract: [Purpose/significance] This study aims to identify events quickly and accurately from sudden network public opinion texts. [Method/process] This paper proposes a method to identify network public opinion emergencies by integrating syntactic features and syntactic similarity. An event-oriented syntactic feature extraction method is proposed based on syntactic features. The event syntactic feature database is constructed using event semantic annotation and syntactic feature extraction methods. Network public opinion emergencies are identified by calculating the syntactic similarity between the text to be tested and the syntax database. [Result/conclusion] Taking the COVID-19 pandemic as an example, the optimal similarity of the proposed network public opinion emergency identification method in this public opinion is 0.93. Under this similarity, 160 events and 30 non-events are identified from a new text, and the F1 value reaches 0.848. Through method evaluation, it is proved that the proposed method is effective in using syntactic similarity to identify events and merging the same adjacent parts of speech.

Keywords: internet public opinion; event identification; syntax features; syntactic similarity

Note: Figure translations are in progress. See original paper for figures.