Research on Entity Extraction for Food Safety Incidents Based on Multi-Feature Knowledge (Postprint)

Authors: Wang Dongbo, Wu Yi, Ye Wenhao, Liu Ruilun

Date: 2017-11-08T00:00:00+00:00

Abstract

Objective: To extract food safety incident entities from large-scale food safety incidents. Method: Based on occurred food safety incidents, combined with informatics approaches to data acquisition, annotation, and organization, and integrating various distributional characteristic knowledge of food safety incident entities, the Conditional Random Field model is employed to construct food safety incident corpora and extract corresponding entities. Limitation: The feature templates formulated during the food safety incident entity extraction process have certain limitations in domain adaptation. Results: On the scale of existing 15-million-word annotated food safety incident corpora, by statistically analyzing internal and external features of food safety incident entities, and based on the Conditional Random Field machine learning model, we constructed an extraction model for food safety entities, which achieved a maximum F-value of 91.94%. Conclusion: Through analysis of food safety incident entity extraction results, on domain-specific corpora in the food sector, entity extraction based on Conditional Random Fields is feasible.

Full Text

Preamble

Extracting Food Safety Event Entities with Multi-Feature Knowledge

Wang Dongbo[1],[2], Wu Yi[1], Ye Wenhao[1], Liu Ruilun[1]

[1](College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, China)

[2](Research Center for Domain Knowledge Correlation, Nanjing Agricultural University, Nanjing 210095, China)

Abstract

[**Objective**] To extract food safety event entities from large-scale food safety incidents. [**Methods**] Based on historical food safety events, we integrated multiple distribution characteristics of food safety event entities and constructed a food safety event corpus using information science methods for data acquisition, annotation, and organization. We then extracted corresponding entities through a Conditional Random Field (CRF) model. [**Limitations**] The feature templates developed during entity extraction have certain limitations in domain migration. [**Results**] Using a 15-million-word annotated food safety event corpus, we constructed an entity extraction model based on the CRF machine learning model by statistically analyzing internal and external features of food safety event entities. The model achieved a maximum F-score of 91.94%. [**Conclusion**] Analysis of extraction results demonstrates that CRF-based entity extraction is feasible on domain-specific corpora such as food safety.

Keywords: Feature Knowledge; Conditional Random Field Model; Entity; Food Safety Event
Classification Number: G350

1. Introduction

To address prominent food safety incidents such as the "Shuanghui Clenbuterol scandal," "Aged Yogurt controversy," "Jiugui Liquor plasticizer exceeding standards,""Carcinogenic Enoki Mushrooms,""Friso Milk Powder,""Sulfur-fumigated Goji Berries," and "Cadmium Rice," the Central Rural Work Conference held on December 23-24, 2013 explicitly proposed establishing a unified national traceability platform for agricultural products and food safety information. The foundation for building such a platform lies in identifying key entities within food safety incidents, particularly when processing food safety-related public opinion, where entity extraction becomes increasingly critical.

In response to this need, this study conducts entity extraction experiments on food safety events based on a constructed food safety event corpus and a CRF machine learning model, leveraging multi-feature knowledge of food safety event entities. This work provides fundamental knowledge anchors for constructing a food safety event knowledge base while laying groundwork for in-depth mining, analysis, and strategy formulation for addressing food safety incidents.

2. Related Work

Research on food safety events has primarily focused on case studies, policy analysis, and emergency response. Representative studies include: Wu Heng, a graduate student from Fudan University, collaborated with 34 online volunteers to create the "Throw It Out the Window" website [1], which collected food safety incidents and built a database. This database provided a substantial amount of text for our corpus construction and served as its foundation.

Most food safety research adopts a management perspective. Notable examples include: Zhang Mujie et al. [2] analyzed the hazards of information non-disclosure in emergency management through two typical cases and examined common reasons for non-disclosure. Their case selection methodology informed our approach to determining corpus texts. Ma Ying et al. [3] constructed an epidemic model for risk perception in food industry incidents, using Japan' s earthquake-induced "salt panic buying" as a case study for numerical analysis and model validation. This study provided valuable insights for annotating food safety event names.

These studies offered both macro-level methodological guidance and specific criteria for identifying food safety event entities.

Recent research on entity extraction primarily employs machine learning methods for extracting entities from unstructured text. Representative approaches include: Neural network-based strategies, where Chen Yu et al. [4] attempted to extract entities and their relationships using Deep Belief Nets. This study guided our determination of feature quantities. Semantic knowledge-based extraction is also popular; Shao Fa et al. [5] addressed word sense disambiguation using HowNet and Bayesian classification resources to extract entities. While scientifically sound, this approach' s overall performance on large-scale corpora requires further validation.

For rapidly increasing electronic medical texts, Xu Hua et al. [6] used rule-based methods on segmented and POS-tagged medical corpora to extract entities, achieving over 80% performance. Although rule-based methods adapt well to specific feature corpora, insufficient exploration of rules embedded in vocabulary can lead to relatively poor coverage. This represents one primary reason for selecting the CRF model for food safety event entity extraction.

Current information extraction research related to food safety focuses on vocabulary-level knowledge extraction from food complaint texts, exemplified by Wei Xiuzhuo' s [7] work on sensitive word extraction and Gao Rui' s [8] ontology-based hazard information extraction from food complaint texts. Compared to entity extraction, vocabulary-level extraction is relatively simpler, primarily due to shorter word lengths and simpler internal composition.

CRF models are widely applied for extracting sequential entities and terms. Representative studies include: Li Lishuang et al. [9] extracted automotive terms using simple feature templates; Wang Wenlong et al. [10] extracted entities from project proposals using composite feature templates; Liu Kai et al. [11] built an entity extraction model for Traditional Chinese Medicine electronic records using characteristic knowledge of TCM vocabulary. These CRF-based studies utilized only simple feature knowledge of entities themselves without incorporating contextual information. This paper addresses this limitation by constructing complex feature templates for identifying food safety event entities.

3. Methodology

3.1 Food Safety Event Corpus Construction and Entity Definition

Based on collection, annotation, and organization of food safety incidents, we constructed a food safety event corpus covering 2005-2015. Sources included internet-based incidents (collected through vertical search engine technology targeting event themes from news portals, forums, and blogs) and print media cases (manually entered and proofread). Heterogeneous collected data underwent cleaning, transformation, and statistical storage in a database. Figure 1 [Figure 1: see original paper] shows a screenshot of the food safety event crawling software.

Figure 1. Screenshot of Food Safety Event Crawling Software

Annotation involved word segmentation and POS tagging, with longer food safety terms receiving higher-level POS tags. Unlike general corpora, food safety event vocabulary tends to be longer; such terms were treated as single words during segmentation and POS tagging. Organization involved category annotation based on China' s Food Safety Law.

After processing, the corpus reached 15 million characters and 6.87 million words, comprising 2,800 food safety incidents.

This paper defines entities as food names and specific factors causing food safety incidents. Food names include terms like "milk powder," "soy sauce," "rice," and "milk," while specific factors include "additives," "formaldehyde," "benzoyl peroxide," and "trans fatty acids." Our primary task involves building a machine learning model to automatically extract these entities. Sample CRF training and testing corpora appear as follows:

```
Enterprise/n or/c individual/n of/u "/w illegal/vn behavior/n "/w in/f ,/w including/v "/w p
```

3.2 Statistical Analysis of Internal and External Features

We manually annotated food names and specific factors across 2,800 food safety events and statistically analyzed their internal and external features.

(1) Internal Features

Word Length: Analyzing entity length helps assess extraction difficulty and determine the CRF tag set size. Table 1 shows the length distribution.

Table 1. Entity Length Distribution in Food Safety Events

As Table 1 shows, entity lengths primarily range from 1-5 characters, accounting for 99.25% of all entities. Lengths of 2-3 characters constitute 82.18% (55.19% at length 2, 27.00% at length 3). Consequently, 2-3 character entities are the primary extraction focus (e.g., "milk powder," "milk," "pork," "additives," "gutter oil"). Entities longer than 8 characters are mostly complex proper nouns or adjective-noun combinations (e.g., "sodium cyclamate").

Specific Entity Distribution: Table 2 shows the distribution of specific food safety entities.

Table 2. Distribution of Specific Food Safety Event Entities

This analysis covers 87,042 entities. The top 10 account for 13.16%, the top 5 for 8.23%, with "milk powder" ranking second at 1.91% and "additives" first at 2.58%. Statistical analysis of specific entities aids in understanding entity content and identifying boundary features.

(2) External Features

The left and right boundaries of "food names" and "specific factors" vary significantly across different food safety event corpora. We statistically analyzed these boundaries, which proved valuable for subsequent model construction. Boundary scope was limited to clauses ending with "。!?" punctuation. Left boundaries never crossed the entity's first marker, so analysis was confined to the range from sentence start to first marker (denoted as β). Right boundaries never crossed the last marker, so analysis covered the range from last marker to sentence end (denoted as α).

The formula for selecting left boundary words appears in Equation (1) [12]:

$$P = \frac{f(W_{\text{left\_outside}})}{f(W_{\text{left}})}$$

where $f(W_{\text{left\_outside}})$ represents the frequency of word W in range β, and $f(W_{\text{left}})$ represents its frequency in β plus within entities. With an empirical threshold of $P \geq 0.8$, we identified seven left boundary words: "de (的), yong (用), he (和), shi (是), shipin (食品), chaobiao (超标), zhong (中)."

Similarly, Equation (2) [12] selects right boundary words:

$$P = \frac{f(W_{\text{right\_outside}})}{f(W_{\text{right}})}$$

where $f(W_{\text{right\_outside}})$ represents frequency in range α, and $f(W_{\text{right}})$ represents frequency in α plus within entities. With the same threshold, we identified ten right boundary words: "de (的), yong (用), pin (品), you (有), zhong (种), he (和), shi (是), chao (超), zhong (中), chan (产)."

3.3 Machine Learning Model

The Conditional Random Field, proposed by Lafferty et al. [13], is an optimal model for sequence labeling tasks. It is an undirected graphical model calculating the joint conditional probability distribution of state labels for an entire observation sequence. Given input node values, it computes conditional probabilities of output nodes, with training maximizing these probabilities. The most

common CRF model is the first-order chain structure (linear chain), shown in Figure 2 [Figure 2: see original paper].

Figure 2. Topology of Linear Chain CRF Model

Let $x = \{x_1, x_2, \ldots, x_{n-1}, x_n\}$ represent the observed input sequence (e.g., segmented words), and $y = \{y_1, y_2, \ldots, y_{n-1}, y_n\}$ represent the finite state set where each state corresponds to a label. Given input sequence x, the conditional probability of state sequence y for linear chain CRFs with parameters $\lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_{n-1}, \lambda_n\}$ is given by Equations (3) and (4) [13]:

$$p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{i,k} \lambda_k f_k(y_{i-1}, y_i, x, i)\right)$$

$$Z(x) = \sum_{y} \exp\left(\sum_{i,k} \lambda_k f_k(y_{i-1}, y_i, x, i)\right)$$

where $Z(x)$ is the normalization factor ensuring all conditional probabilities sum to 1, and $f_k(y_{i-1}, y_i, x, i)$ are unified feature functions (typically binary indicator functions). Weights λ_k are learned from training data.

The Maximum Entropy model, based on McCallum et al.' s [14] principle, states that when probability distribution information is uncertain, the least biased approach treats all distributions equally without subjective assumptions. The distribution maximizing entropy under training data constraints is optimal. While widely used in AI and NLP, ME models suffer from label bias issues, leading to more errors and unrecognized cases, making them less effective than CRF in certain scenarios.

3.4 Corpus Selection and Processing

Entities were annotated with "【】" markers, e.g., "【/wky milk/n】/wky" . Based on feature statistics and the CRF model definition, we determined the tag set size using Equation (5) [14]:

$$L = \sum_{i=j}^{k} \frac{N_i \times i}{N}$$

where L represents the average weighted length of entities when $i \leq k$, N_i is the count of entities of length i, k and j are the maximum and minimum entity lengths, and N is the total entity count.

Based on Equation (5) and experimental results, we adopted a 5-tag set $R = \{B, C, E, S, A\}$: B marks entity beginnings, C marks middle words, E marks endings, S marks non-entity words, and A marks single-word entities. For entities longer than 3 characters, C denotes extended words.

We developed Java programs to automatically annotate all corpora using the "[]" markers and feature templates. Table 3 shows annotation examples.

Table 3. Training and Testing Corpus for "Food Names" and "Specific Factors"

Word	POS	Word Length	Right Boundary	Tag
Zhejiang Province	ns	···	···	···
Jinhua City	ns	···	···	···

3.5 Feature Selection and Template Design

Feature selection critically impacts CRF model performance. Features comprise atomic and composite features. Our atomic features include: word itself, POS, word length, entity indicator, left boundary indicator, and right boundary indicator (6 total). Composite features combine atomic features to represent complex linguistic characteristics. Feature window sizes are 7, 3, 5, 5, 5, and 5 respectively: 7-window range is $\{-3, -2, -1, 0, 1, 2, 3\}$; 5-window range is $\{-2, -1, 0, 1, 2\}$; 3-window range is $\{-1, 0, 1\}$. For extraction performance, POS and word itself are most important, followed by boundary indicators and entity indicators, with length being least important.

4. Entity Extraction Experiments

We evaluated model performance using Precision, Recall, and F-measure. Both CRF and Maximum Entropy models were tested on annotated corpora using cross-validation, splitting 2,800 documents into 9:1 training-to-testing ratios. Results appear in Tables 4 and 5 , with training/testing time comparisons in Table 6 .

Table 4. Performance Comparison of CRF Model for "Food Names" and "Specific Factors" Extraction

Metric	Value
Precision	89.95% - 91.94%
Recall	88.35% - 92.12%
F-score	90.06% - 91.94% (Average: 90.88%)

Table 5. Performance Comparison of Maximum Entropy Model for "Food Names" and "Specific Factors" Extraction

Metric	Value
Precision	59.97% - 81.90%

Metric	Value
Recall	61.89% - 86.52%
F-score	62.49% - 73.38% (Average: 70.48%)

Table 6. Training and Testing Time Comparison Between CRF and Maximum Entropy Models

Model	Time
CRF	~50,000 seconds
Maximum Entropy	~100 seconds

CRF significantly outperformed Maximum Entropy, with F-scores ranging 90.06%-91.94% (average 90.88%) versus 62.49%-73.38% (average 70.48%). However, Maximum Entropy was far more efficient (~100s vs. ~50,000s). Since our focus is extraction performance rather than speed, we selected CRF.

Error analysis revealed that misidentified entities were primarily lengthy ones with complex compositions, such as "副溶血弧菌细菌" (Vibrio parahaemolyticus bacteria), "乔家栅高庄馒头" (Qiaojia Gaozhuang steamed buns), "兽用加硒腐殖酸钠" (veterinary sodium selenite humate), "受蜡样芽孢杆菌污染" (Bacillus cereus contamination), and "汪氏蜂胶软胶囊" (Wang' s propolis soft capsules). These contain challenging elements like multiple place names, adjectives, or name-noun combinations that reduce accuracy and recall.

We developed software to automatically crawl and clean food safety news reports from CNKI (China National Knowledge Infrastructure) for 2005, then applied our extraction model. Figures 3 [Figure 3: see original paper] and 4 [Figure 4: see original paper] show screenshots of the crawling and extraction functions.

Figure 3. Screenshot of CNKI Data Crawling Function
Figure 4. Screenshot of Entity Extraction Function

5. Conclusion

Automatic annotation of "food names"and "specific factors"provides foundational resources for building food safety event knowledge bases and mining response strategies. By statistically analyzing internal and external entity features on annotated corpora, we constructed a machine learning extraction model whose open-test results demonstrate strong performance, achieving practical applicability. Future research will apply this model to 1995-2004 corpora and improve it by incorporating new features based on overall performance.

References

[1] Throw It Out the Window [EB/OL]. [2014-02-18]. http://www.zccw.info/. (Original Chinese title: Zhi Chu Chuang Wai)

[2] Zhang Mujie, Shen Jianhua. About the Disposal of the Food and Drug Safety Incident Information to the Public Thinking about the Disposal of the Food and Drug Safety Incident Information[J]. Shanghai Food and Drug Information Research, 2012(2): 45-49.

[3] Ma Ying, Zhang Yuanyuan, Song Wenguang. Research on Epidemic Model of Emergency Events Risk Perception in Food Industry[J]. Science Research Management, 2013, 34(9): 123-130.

[4] Chen Yu, Zheng Dequan, Zhao Tiejun. Chinese Relation Extraction Based on Deep Belief Nets[J]. Journal of Software, 2012, 23(10): 2572-2585.

[5] Shao Fa, Huang Yin' ge, Zhou Lanjiang, et al. Chinese Entity Relation Extraction Based on Entity Disambiguation[J]. Journal of Shandong University: Engineering Science, 2014, 44(6): 32-37.

[6] Xu Hua, Liu Maofu, Jiang Li, et al. Disease and Bacteria Entity Extraction Based on Linguistic Rule[J]. Journal of Wuhan University: Natural Science Edition, 2015, 61(2): 51-55.

[7] Wei Xiuzhuo. Food Complaint Text Sensitive Words Extraction Research[D]. Changchun: Northeast Normal University, 2015.

[8] Gao Rui. Ontology-based Hazard Information Extraction from Chinese Food Complaint Documents[D]. Changchun: Northeast Normal University, 2011.

[9] Li Lishuang, Dang Yanzhong, Zhang Jing, et al. Automotive Term Extraction Based on Conditional Random Fields[J]. Journal of Dalian University of Technology, 2013, 53(2): 267-272.

[10] Wang Longwen, Wang Dongbo. Project Application-oriented Named Entity Extraction Model Construction[J]. Information Documentation Services, 2015(1): 30-34.

[11] Liu Kai, Zhou Xuezhong, Yu Jian, et al. Named Entity Extraction of Traditional Chinese Medicine Medical Records Based on Conditional Random Field[J]. Computer Engineering, 2014, 40(9): 312-316.

[12] Wu Yunfang. Researches of Modern Chinese Coordinate Construction for Language Information Processing[M]. Beijing: Beijing Normal University Press, 2004.

[13] Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]// Proceedings of the 18th International Conference on Machine Learning. 2001: 282-289.

[14] McCallum A, Freitag D, Pereira F. Maximum Entropy Markov Models for Information Extraction and Segmentation[C]//Proceedings of the 17th International Conference on Machine Learning. 2000: 591-598.

Author Contributions

Wang Dongbo: Framework design, manuscript writing and revision; Wu Yi: Model training and manuscript writing; Ye Wenhao: Data annotation; Liu Ruilun: Corpus preprocessing.

Conflict of Interest

All authors declare no conflict of interest.

Supporting Data

Supporting data is self-archived by the authors, E-mail: db.wang@njau.edu.cn.

[1] Wang Dongbo, Wu Yi, Ye Wenhao, Liu Ruilun. Event statistics programming. Entity Statistics Program Based on Food Safety Event Corpus.
[2] Wang Dongbo, Wu Yi, Ye Wenhao, Liu Ruilun. Event extracting platform. Entity Extraction Platform Based on Conditional Random Field Model.

Received: 2016-08-03
Revised: 2016-12-07

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