Recognition of the Name of Quality Inspection Institutions Based on CRF and Rules

Xiuli Ning*, Xueyan Su and Bin Peng
Quality Management Branch, China National Institute of Standardization, Beijing 10001, China
*475554762@qq.com

Abstract: In recent years, the case of injuries, deaths and property losses caused by product quality and safety issues occur frequently. The analysis and processing of product quality and safety data can provide data support for government departments in decision-making, which is significant for improving product quality and safety in general. Institution name recognition is an important link in the analysis and processing of quality inspection data. A recognition method which combined CRF and rules for the name of quality inspection institution is put forward in this paper. Firstly, summarizing the rules and features of the name of quality inspection institution by the analysis of the composition characteristics, and then the entity recognition model based on statistics and rules is constructed in combination with CRF. The experimental results show that the method proposed this paper has high accuracy, recall rate and F value for the quality inspection institution names of different scales.

1. Introduction
Quality problem, which is involved in everyone’s basic survival interest, is the focus livelihood issue of the whole society. Party and state leaders have always attached great importance to quality development, and elevated “building a powerful nation with quality” as a national development strategy. Product quality is inseparable from consumers, which has attracted unprecedented attention from all walks of life. The analysis of product quality and safety data from a wide range of sources can be processed, which is conducive to improving the overall quality and safety issues in China. Named entity recognition is an important part of data processing, so the entity recognition technology for quality inspection institution name is studied in this paper.

In recent years, a large number of scholars at home and abroad have carried out research on the named entity recognition technology. Chen et al. compared the influence of three sampling strategies based on uncertainty, diversity and baseline on the results of named entity recognition [1] Derczynski et al. described a new Twitter entity ambiguity data set for named entity recognition [2]. Bhasaran et al. put forward a stack integration method combining fuzzy matching for the recognition of disease names [3]. Gridach et al. combined deep neural network, CRF, word embedding and word-level representation for named entity recognition [4]. Ghiasvand et al. put forward a new method that did not require manual annotation to train the clinical NER system for entity recognition [5]. Lee et al. proposed an integrated neural network model for simultaneous execution of MA and NER [6]. Xu et al. put forward a new neural network method for disease name recognition [7]. Chen et al. identified the entity names in adverse drug event reports through BiLSTM-CRF based on lexical characteristics and
three-level training [8]. Liu et al. combined Bi-LSTM and CRF, and made extraction with online data [9].

2. Analysis of quality inspection institution name

2.1. Semantic structure analysis of institution name

The name of quality inspection institution is composed of the regional modifier (F), central modifier (C) and attribute modifier (A), so the name of quality inspection system can be expressed as “F+C+A”. Among them, regional modifiers represent the geographical information involved in the institution name, including national level, provincial level, municipal level, county, district, etc.; central modifiers are the adjunct words special for the institution name of quality inspection system, including standardization, quality, inspection and quarantine, measurement, etc. Attribute modifiers are suffix words of quality inspection institution names, including institute, association, and etc. This research focuses on the analysis of quality inspection institution name from the construction mode of semantic level.

Firstly, the number of regional modifiers in the quality inspection system is limited and fixed, including the national, provincial, municipal, county and district names, for example, the State Administration for Market Regulation, Zhejiang Provincial Administration for Market Regulation, Hangzhou Municipal Administration for Market Regulation, and Administration for Market Regulation of Xihu District, Hangzhou.

Secondly, the central feature word is the key to distinguish the quality inspection institution name from that of other fields, and its form is relatively fixed, which can be obtained automatically from the institution names of large-scale quality inspection systems. But there are still a few names that don't have uniform rules or features and need to be manually revised and expanded to ensure full coverage.

Finally, attribute modifiers are suffixes of quality inspection institution names, which are limited and fixed in number, including institute, association, bureau, and etc. Main modifiers of quality inspection institution name are shown in Table 1:

| Type of modifier | Modifiers | Examples |
|------------------|-----------|----------|
| Regional modifier | China, national, 34 provincial-level administrative regions, and cities, counties and districts under each provincial-level administrative region | Jilin Provincial Institute of Standards, Anhui Provincial Institute of Quality and Standardization, Beijing Institute of Standardization |
| Central modifier | Standardization, inspection and quarantine, certification and accreditation, measurement, testing, market supervision, quality, etc | China National Institute of Metrology, China National Institute of Standardization, Chinese Academy of Inspection and Quarantine |
| Attribute modifier | Center, association, research institute, company, society, bureau, etc | China Association for Standardization, China National Institute of Standardization |

2.2. Context feature analysis

Based on the analysis of the context information of the name of quality inspection organization, the boundary of institution name is defined as the prefix words and suffix words, and the functions of
coordinating relation words and mutually exclusive words in the context of institution name are also defined.

(1) Prefix words: A category of words or phrases that appear before the names of quality inspection organizations are called prefix words. For example, the word “in” in the sentence of “the 13th Standard Documentation Work Conference is held in Shanghai Institute of Quality and Standardization”; the word “at” in the sentence of “the standardization engineers from all over the country participate in the standard establishment training at China National Institute of Standardization”.

(2) Suffix words: A category of words or phrases that appear after the name of the immediate system institution are called suffix words. For example, the word, “on” in the sentence that “Beijing Product Quality Supervision and Inspection Institute announced the random checking results of laundry detergent products on March 15”; the phrase, “CCYL committee” in the sentence that “the CCYL committee of China National Institute of Metrology organized the educational activities to commemorate the 100th anniversary of the May 4th Movement.

(3) Coordinating relation words: the conjunctions between the names of two quality inspection organizations. For example, the word, “and” in the sentence of “the supervision and random inspection of refrigerator products shall be jointly undertaken by Nanjing Product Quality Supervision and Inspection Institute and Hunan Product Quality Supervision and Inspection Institute”. If one side of the conjunction is the name of the quality inspection organization, the other side may also be the name of other quality inspection organization.

(4) Mutually exclusive words. For example, “multiple” in “multiple standardization institutes jointly participate in the standard development work of sweeping robot”, “national” in “national market supervision departments actively crack down on the production and sale of counterfeit and shoddy products”, usually refers to the general name of some institutions, rather than a specific institution name.

3. Recognition algorithm

3.1. Recognition process

The core idea of conditional random field (CRF) is derived from the maximum entropy model, which can be regarded as a markov random field or an undirected graph model, namely, an undirected graph model to calculate the conditional probability of the output node under the condition of a given input node.

In terms of general linear CRF, according to the theory of conditional random field, the conditional probability of its state sequence is:

\[
p(y|x, \lambda) \propto \exp \left( \sum_j \lambda_j t_j(y_{i-1}, x, i) + \sum_k \mu_k s_k(y_i, x, i) \right)
\]

Where, \( x = \{x_1, x_2, \ldots, x_n\} \) represents the observation sequence, \( y = \{y_1, y_2, \ldots, y_n\} \) is the set of finite state sequences, \( t_j(y_{i-1}, x, i) \) represents the transfer characteristic function between the mark positions \( i - 1 \) and \( i \) of the observation sequence, and \( s_k(y_i, x, i) \) represents the state characteristic function of the observation sequence position \( i \).

After the two functions are unified into one function, \( f_j(y_{i-1}, x, i) \):

\[
p(y|x, \lambda) = \frac{1}{Z(x)} \exp \left( \sum_{i=1}^n \sum_j \lambda_j f_j(y_{i-1}, x, i) \right)
\]

Where, \( Z(x) \) is the normalization factor:
\[
z(x) = \sum_{j} \exp \left( \sum_{i=1}^{n} \lambda_{ij} f_{ij}(y_{c-1}, x, i) \right)
\]  

(3)

\(\lambda\) is the weight of the characteristic function, which is a parameter learnt from training data. The recognition method for the names of quality inspection institutions based on CRF and rules is studied in this paper, with the recognition process shown in figure 1:

![Flow chart of cognition of name of quality inspection system](image)

**Figure 1.** Flow chart of cognition of name of quality inspection system

The recognition process of the names of quality inspection organizations can be summarized into two parts:

1. **Analysis and summary of the structure of institution names**
   
   Through the analysis and summary of the structure of quality inspection institution names, namely the context, five dictionaries closely related to the institution names are sorted out, such as common feature modifiers, common central modifiers, common attribute modifiers, common high-frequency words, common prefix words, suffix words, juxtaposing relation words, and mutually exclusive relation words.

2. **Verification of CRF-based preliminary recognition level based on rules**

   Firstly, the randomly selected corpus containing the names of quality inspection organizations are processed; secondly, the word segmentation and part-of-speech tagging is carried out by using the word segmentation system and corresponding dictionary in combination with the analysis of the first part, and the CRF feature template with the characteristics of quality control system is established; thirdly, the preliminary recognition of institution names is made based on the CRF; finally, the rules and patterns established in the first part are used to screen the preliminary recognition results, thus outputting the final results.

### 3.2. Feature template selection

In the CRF model, the key to model performance is to select the appropriate feature set for a particular task. The characteristic attributes closely related to the name recognition of quality inspection system include the regional modifier, central modifier and attribute modifier.

Institution name is based on word segmentation and part-of-speech tagging. The word itself, the word position (previous position, current position, latter position) and part-of-speech contain rich information, which are of great significance to the recognition of institution name and indispensable features of institution name recognition.

The atomic feature template is constructed for the name recognition of quality inspection system, and the composite feature template composed of independent attributes is established for the complex context. Some feature templates are shown in Table 2:
Table 2. Some feature templates.

| No. | Type               | Feature                  | Meaning                                                                                      |
|-----|--------------------|--------------------------|----------------------------------------------------------------------------------------------|
| 1   | Atomic features    | (IsAreaWord)             | Area words of quality inspection institution names or not                                    |
| 2   |                    | (IsCenterWord)           | Center words of quality inspection institution names or not                                   |
| 3   |                    | (IsAttributesWord)       | Attribute words of quality inspection institution names or not                                |
| 4   |                    | (IsCommonWord)           | Common words of quality inspection institution names or not                                   |
| 5   | Composite features | (P_word, C_word)         | A combination of the previous word marker and the current word                               |
| 6   |                    | (C_word, A_word)         | A combination of the latter word marker and the current word                                |
| 7   |                    | (P_word, C_word, A_word) | A combination of the previous word, the current word, and the latter word                    |

3.3. Corpus processing
The news that contains the names of quality inspection organizations is marked randomly, and the following marks are set for all kinds of dictionaries:

AR: Area
CT: Central word
AB: Attribute
HF: High-frequency words
OR: Others

Therefore, the sentence to be recognized can be represented by a marker symbol. For example, “Shanghai Municipal Institute of Quality Supervision and Inspection announces the random inspection result of hand sanitizer” can be marked as: “Shanghai\AR quality CT\supervision\CT inspection\AB\OR hand sanitizer\OR random inspection\HF result\OR”, and the annotated training data is used to train the model.

3.4. CRF model training
CRF training consists of two steps: the establishment of characteristic function, and the calculation of the weight of the characteristic function. The training process is as follows:

1. The feature function set generated according to the feature template;
2. The model expectation and empirical expectation of characteristic function by dynamic programming method;
3. \[
\frac{\partial L(\lambda)}{\lambda} \text{ obtained by smoothing factor } \sigma, \text{ feature model expectation and experience expectation is substituted into the L-BFGS module for training, so as to obtain the modified parameter } \lambda. \text{ If the iteration termination condition is reached, the training will be finished; if the termination condition is not reached, continue training in (2). The termination condition converges when the parameter } \lambda \text{ in the L-BFGS algorithm is 0 or smaller. }
\]

It is necessary to provide the first derivative of likelihood function for CRF training by using L-BFGS algorithm, and it is assumed that the annotation of the Jth training instance makes its state sequence not ambiguous, and represents that path. The first derivative of Log likelihood for the training data set is:
\[
\frac{\partial L}{\partial f_k} = \left( \sum_{j=1}^{N} C_i \left( s^{(j)}, o^{(j)} \right) \right) - \left( \sum_{j=1}^{N} \sum_{k=1}^{K} P_{s} \left( s | o^{(j)} \right) C_i \left( s | o^{(j)} \right) \right)
\]

(4)

Where, \(C_i \left( s | o \right)\) represents the sum of feature \(f_k\) at each position \(t\) in string \(s\).

3.5. Verification of rules
The names of quality inspection organizations recognized by CRF statistical model cannot cover all situations, so the rules are verified by using linguistic features, with the verification rules shown as follows:

(1) Any mutually exclusive word in the recognition results is removed. For example, “many regions” in “market supervision departments in many regions” is a mutually exclusive term, so this information is excluded.

(2) Any cardinal number and quantifier contained in the recognition results will be omitted. For example, “three standardization institutes undertake risk monitoring tasks at the same time” involves cardinal number and quantifier, so this information is excluded.

4. Experimental verification
In order to verify the effectiveness of the algorithm in this paper, the data of product quality supervision and random checking collected manually in various provinces and cities were selected for the experiment, and the method based on CRF and rules was compared with the CRF method. The accuracy, recall rate and F value were used to evaluate the recognition results of quality inspection system names, with the experimental results shown in Table 3 and Table 4.

1. Accuracy (P) = (number of institution names correctly identified by the system / number of institution names identified by the system) × 100%

2. Recall rate (R) = (number of institution names correctly identified by the system / number of institution names in the document) × 100%

3. Value F = [2 × P × R / (P + R)] × 100%

Table 3. Recognition method based on CRF.

| The number of training articles | The number of training sentences | Actual number of institution names | The number of institution names identified | The number of institution names correctly identified | Accuracy rate | Recall rate | Value of F |
|--------------------------------|----------------------------------|-----------------------------------|--------------------------------------------|--------------------------------------------------|--------------|------------|------------|
| 100                            | 1254                             | 1116                              | 1065                                       | 918                                              | 86.16%       | 82.22%     | 84.14%     |
| 200                            | 2096                             | 1893                              | 1789                                       | 1526                                             | 85.30%       | 80.61%     | 82.89%     |
| 300                            | 3297                             | 2934                              | 2698                                       | 2278                                             | 84.44%       | 77.65%     | 80.90%     |
| 400                            | 4394                             | 4003                              | 3701                                       | 3137                                             | 84.75%       | 78.36%     | 81.43%     |
| 500                            | 5546                             | 5291                              | 4749                                       | 3980                                             | 83.80%       | 75.22%     | 79.28%     |
| 600                            | 6358                             | 6049                              | 5564                                       | 4518                                             | 81.20%       | 74.69%     | 77.81%     |
| 700                            | 7339                             | 7118                              | 6358                                       | 5102                                             | 80.25%       | 71.68%     | 75.72%     |
| 800                            | 8006                             | 7895                              | 7042                                       | 5762                                             | 81.82%       | 72.98%     | 77.15%     |
| 900                            | 8821                             | 8654                              | 7544                                       | 6141                                             | 81.40%       | 70.96%     | 75.82%     |
| 1000                           | 9645                             | 9373                              | 8351                                       | 6603                                             | 79.07%       | 70.45%     | 74.51%     |
Table 4. Recognition method based on CRF and rules.

| The number of training | The number of training sentences | Actual number of institution names | The number of institution names identified | The number of institution names correctly identified | Accuracy rate | Recall rate | Value of F |
|------------------------|----------------------------------|-----------------------------------|--------------------------------------------|-------------------------------------------------|-------------|------------|-----------|
| 100                    | 1254                             | 1145                              | 1052                                       | 973                                             | 92.52%      | 85.00%     | 88.60%    |
| 200                    | 2096                             | 1934                              | 1798                                       | 1635                                           | 90.92%      | 84.53%     | 87.61%    |
| 300                    | 3297                             | 2978                              | 2796                                       | 2493                                           | 89.17%      | 83.72%     | 86.36%    |
| 400                    | 4394                             | 4121                              | 3854                                       | 3421                                           | 88.75%      | 83.00%     | 85.78%    |
| 500                    | 5546                             | 5378                              | 4911                                       | 4412                                           | 89.84%      | 82.04%     | 85.76%    |
| 600                    | 6358                             | 5935                              | 5618                                       | 4935                                           | 87.84%      | 83.15%     | 85.43%    |
| 700                    | 7339                             | 7102                              | 6775                                       | 5834                                           | 86.11%      | 82.14%     | 84.08%    |
| 800                    | 8006                             | 7758                              | 7318                                       | 6309                                           | 86.21%      | 81.32%     | 83.69%    |
| 900                    | 8821                             | 8398                              | 7896                                       | 6789                                           | 85.98%      | 80.84%     | 83.33%    |
| 1000                   | 9645                             | 9123                              | 8664                                       | 7247                                           | 83.65%      | 79.44%     | 81.49%    |

It can be seen from Table 3 and Table 4 that the method based on CRF and rules has a good recognition effect on the corpus of different sizes.

5. Conclusion

Through the analysis of the composing mode of quality inspection system name, some rules of its structure is summarized, and CRF is combined with the constructed rules to identify the entity recognition of quality inspection system name; the simulation results verify the effectiveness of the method proposed. The next step is to investigate the cases with larger text content and excavate new rules and knowledge from the texts, and then add them to the CRF model for institution name recognition.

Acknowledgment

We would like to acknowledge that this research is supported and funded by the Market Supervision Technology Assurance Project "Research on Evolution and Response Strategy of Product Quality and Safety Emergencies Based on Domain Ontology", the the Basic Scientific Research Business Projects 552018Y-5927-2018, and National Science Foundation of China under Grant No. 91646122, 91746202.

References

[1] Yukun Chen, Thomas A. Lasko, Qiaozhu Mei et al. A study of active learning methods for named entity recognition in clinical text[J]. Journal of Biomedical Informatics, 2015, 58:11-18.
[2] Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp et al. Analysis of named entity recognition and linking for tweets[J]. Information Processing & Management, 2015, 51: 32-49.
[3] Balu Bhasuran, Gurusamy Murugesan, Sabenabanu Abdulkadhar, Jayakumar Natarajan. Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases[J]. Journal of Biomedical Informatics, 2016, 64:1-9.
[4] Mourad Gridach. Character-level neural network for biomedical named entity recognition[J]. Journal of Biomedical Informatics, 2017, 70: 85-91.
[5] Omid Ghiasvand, Rohit J. Kate. Learning for clinical named entity recognition without manual annotations[J]. Informatics in Medicine Unlocked, 2018, 13:122-127.

[6] Hyeon-gu Lee, Geonwoo Park, Harksoo Kim. Effective integration of morphological analysis and named entity recognition based on a recurrent neural network[J]. Pattern Recognition Letters, 2018, 1121:361-365.

[7] Kai Xu, Zhenguo Yang, Peipei Kang, Qi Wang, Wenyin Liu. Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition[J]. Computers in Biology and Medicine, 2019, 108: 122-132.

[8] Yao Chen, Changjiang Zhou, Tianxin Li, Hong Wu et al. Named entity recognition from Chinese adverse drug event reports with lexical feature based BiLSTM-CRF and tri-training[J]. Journal of Biomedical Informatics, 2019, 96:103252.

[9] Xin Liu, Yanju Zhou, Zongrun Wang. Recognition and extraction of named entities in online medical diagnosis data based on a deep neural network[J]. Journal of Visual Communication and Image Representation, 2019, 60: 1-15.