Feature Selection and Model Optimization for Chinese Word Segmentation in Food Safety Emergencies (Postprint)

Authors: Zhang Yue, Wang Dongbo, Zhu Danhao

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

Abstract

Objective: In the domain of food safety, establishing relevant databases significantly facilitates the supervision and control of food safety. Automatic word segmentation plays a vital role in index construction, index utilization, and corpus building. This study applies a character-based tagging statistical learning method using Conditional Random Fields to automatic word segmentation of corpora concerning food safety emergency incidents.

Methods: By analyzing corpus characteristics such as word length distribution, we conducted various experiments on feature selection and feature templates involved in the automatic word segmentation process, to ascertain the effects of different feature selections and different feature templates on segmentation performance.

Results: Experimental results demonstrate that in feature selection, more features do not necessarily yield better segmentation performance, as feature interference may occur. In the food safety emergency incident corpus, where two- and three-character words constitute 46.62%, the feature templates representing the current character and its first preceding and succeeding characters exert a significant influence on segmentation performance.

Conclusion: Through experiments on different feature selections, feature templates, and their combinations, we identified the optimal features and feature templates for automatic word segmentation of the corpus studied. Using the 5Tag feature tagging scheme with corresponding feature templates, the F-score for segmenting the target corpus reached 92.88%.

Full Text

Preamble

ChinaXiv Collaborative Journal

Feature Selection and Model Optimization for Chinese Word Segmentation in Food Safety Emergencies

Zhang Yue[1], Wang Dongbo[1],[2], Zhu Danhao[3]
[1](College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, China)
[2](Research Center for Correlation of Domain Knowledge, Nanjing Agricultural University, Nanjing 210095, China)
[3](Library of Jiangsu Police Institute, Nanjing 210031, China)

Abstract

[**Objective**] In the food safety domain, establishing relevant databases significantly aids in monitoring and controlling food safety, where automatic word segmentation plays a crucial role in index construction, index utilization, and corpus building. This study applies a character-based statistical learning method using Conditional Random Fields (CRF) to automatic word segmentation of food safety emergency corpora. [**Methods**] We analyze the word length distribution characteristics of the corpus and conduct various experiments on feature selection and feature templates involved in the automatic segmentation process to determine how different features and templates affect segmentation performance. [**Results**] Experimental results demonstrate that more features do not necessarily yield better segmentation outcomes due to feature interference. In the food safety emergency corpus, where two- and three-character words account for 46.62% of the total, feature templates representing the current character and its immediate neighboring characters (both preceding and following) have a particularly significant impact on segmentation accuracy. [**Conclusions**] Through systematic experiments on different feature selections, feature templates, and their combinations, we identify the optimal features and templates for automatic segmentation of our target corpus, achieving an F-score of 92.88% using 5Tag features with corresponding templates.

Keywords: Chinese Word Segmentation; Food Safety; Conditional Random Field; Feature Template; Feature Selection
Classification Number: G351

1. Introduction

Food safety incidents have emerged with increasing frequency in recent years, with numerous severe cases profoundly impacting social production and public life. The rapid proliferation of information related to food safety emergencies has drawn widespread attention. Since food safety concerns public health and lives,

addressing these issues requires not only top-down administrative supervision and corporate self-regulation [1], but also bottom-up participation from social oversight forces [2]. In today' s fast-paced information environment, networks, newspapers, and books serve as primary carriers for the rapid dissemination of food safety emergency information and constitute a major channel for public access to such information.

With the vigorous development of natural language processing, research on Chinese word segmentation technology has achieved substantial progress in both accuracy and speed. This technology has become critical in various Chinese information processing tasks, including text classification, information retrieval, information filtering, automatic indexing, and automatic summarization [3]. However, applications and research on automatic word segmentation in food safety information processing remain limited and warrant further exploration.

In the food domain, establishing relevant databases significantly contributes to food safety supervision and control. Zhang Xinglian et al. [4] highlighted the importance of building food safety early warning database systems. Information asymmetry and inaccuracy represent fundamental causes of misconduct in the food sector, and establishing effective electronic food supervision systems—dynamic databases with timely, accurate, and transparent updates—can ensure safer food products and more stable industry order [5]. As the food industry rapidly expands in both scope and depth, Yu Qing et al. [6] analyzed the necessity of constructing processed food risk databases that provide inspection information and hazard risk coefficients, greatly facilitating research on processed food risk data. In practical implementation, Jia Kai et al. [7] established a traceability database for fresh agricultural products in Sanjie Town, Pengzhou City, building upon domestic and international traceability system research while addressing local food conditions to provide information management and application functions.

Current Chinese automatic word segmentation methods primarily fall into four categories: mechanical segmentation, statistical segmentation, character-based statistical learning, and deep neural network-based methods. Before 2002, automatic segmentation was essentially dictionary-based [8], further divisible into rule-based mechanical methods and statistical methods. These approaches rely entirely on dictionaries as their sole information source [9]. While they can achieve good domain-specific accuracy when supplemented with extensive disambiguation information, their complete dependence on dictionaries limits adaptability. Moreover, dictionary construction requires substantial time and manpower, and maintenance becomes increasingly difficult as more unknown words emerge.

With the launch of the SIGHAN international Chinese word segmentation evaluation (Bakeoff), treating Chinese word segmentation as a sequence labeling problem gradually became mainstream. Character-based statistical learning methods demonstrate superior performance in handling unknown words and disambiguation, clearly outperforming dictionary-based methods without requir-

ing lexicons. Meanwhile, deep neural network-based methods remain immature with limited applications in natural language processing, and thus are not discussed in this paper.

Chinese word segmentation using character-based statistical learning is essentially a sequence labeling process that abstracts text information into an observation sequence and tags each character [10]. The key lies in selecting an appropriate machine learning model, with Hidden Markov Models (HMM), Maximum Entropy (ME) models, and Conditional Random Fields (CRF) being the most commonly used.

HMM' s primary drawback is its output independence assumption, which prevents consideration of contextual features and limits feature selection. The ME model addresses this issue by allowing arbitrary feature selection, but its per-node normalization leads only to local optima and introduces label bias, where all unobserved cases in training data are ignored. The CRF model effectively resolves these problems by performing global normalization across all features rather than at each node, thereby obtaining global optimal solutions. Previous research has demonstrated that CRF-based segmentation systems outperform both ME and HMM approaches.

This paper proposes a feature selection and model optimization method for Chinese word segmentation specifically tailored to food safety emergency corpora. The research focuses on two aspects: applying chain-structured CRF-based Chinese automatic segmentation to food safety corpora, and analyzing the corpus to propose feature templates, feature selections, and feature tag schemes that align with its characteristics. Compared with other segmentation systems, this approach better addresses overlapping ambiguities and unknown word problems in dense food safety case texts, effectively improving segmentation precision and recall.

2. Conditional Random Fields Model

Conditional Random Fields (CRFs), proposed by Lafferty et al. in 2001 based on Maximum Entropy models and Hidden Markov Models, represent an undirected graphical learning model for labeling and segmenting sequential data [11].

Undirected graphical models, also known as Markov Random Fields or Markov networks, were introduced by Pearl [12]. An undirected graph consists of vertices/nodes representing random variables and edges/arcs representing conditional dependencies between them.

Although assigning conditional probabilities to each node is possible, the undirectionality prevents parameterizing joint probabilities using conditional probabilities. Instead, joint probabilities must be represented as products of local functions derived from conditional independence principles.

[Figure 1: see original paper] Example of maximal cliques in an undirected graph

Figure 1 illustrates a simple example where maximal cliques in the undirected graph are $\{X_1, X_2\}$, $\{X_1, X_3\}$, $\{X_3, X_4\}$, and $\{X_2, X_4, X_5\}$. The joint probability distribution for this undirected graphical model can be easily obtained as:

When each random variable in a Markov Random Field has observed values, we must determine the distribution of the field given the observation set—that is, the conditional distribution. This conditional Markov Random Field is called a Conditional Random Field, with a form similar to the Markov Random Field distribution but with an additional observation set X.

CRFs were proposed to solve sequence labeling problems for discrete data. Given a sequence $X = \{x_1, x_2, x_3, \cdots, x_{\ 1}, x\}$ and a finite state set $Y = \{y_1, y_2, y_3, \cdots, y_{\ 1}, y\}$, let $G = (V, E)$ be an undirected graph where Y (v V) represents a set of random variables indexed by nodes in G. If each random variable satisfies the Markov property—that is, $P(Y|X, Y, u \neq v) = P(Y|X, Y, u \ v)$, where u v indicates adjacent nodes—then (X, Y) constitutes a Conditional Random Field.

[Figure 2: see original paper] Linear-chain CRF graphical structure

As shown in Figure 2, CRFs employ undirected graphical models to describe states of given sequences, where each element in X corresponds to a node in the graph and each edge represents a node' s state. A CRF is essentially an undirected graphical model given an observation set [13].

According to CRF theory: $P(y|x) \ \exp(\ \lambda f(y_{\ 1}, y, x, i))$, where f represents feature functions and λ represents corresponding weights.

3. Experimental Design

3.1 Food Safety Corpus Description

Based on collecting, annotating, and organizing food safety emergency events, this study constructed a corpus covering 2005-2015. The final experimental data was created through coarse segmentation and manual correction, as illustrated in Figure 3 [Figure 3: see original paper].

[Figure 3: see original paper] Corpus construction process for experiments

The corpus construction process involved three main steps:

(1) **Data Collection**: Targeting both online and print sources of food safety emergencies. Online data was automatically collected using a self-developed program with a vertical search engine technology for emergency topics, covering news portals, forums, and blogs. Heterogeneous data underwent cleaning and transformation before database storage. Print cases were manually entered and proofread, yielding approximately 5,000 collected events and about 4,500 entries after cleaning (approximately 20MB total).

(2) **Coarse Segmentation**: The NLPIR segmentation software from the Institute of Computing Technology, Chinese Academy of Sciences, was used for initial annotation. However, results showed many inaccurate unknown word identifications. Food safety case libraries constitute dense texts with numerous Chinese unknown words and ambiguous words. Among various food descriptions, geographical descriptions, and chemical compound descriptions, food safety descriptive texts are highly representative, involving numerous food and chemical names that machines often mis-annotate. Therefore, NLPIR was only used for coarse segmentation to reduce manual workload.

(3) **Manual Annotation**: The coarsely segmented corpus underwent manual proofreading to identify and correct segmentation errors, ensuring maximum accuracy of the training corpus.

3.2 Implementation Method

CRF++ [14] is a customizable, open-source CRF toolkit for continuous sequence labeling, widely recognized as the most user-friendly, accurate, and stable CRF tool available. Designed for general purposes, it has been applied to various natural language processing tasks including named entity recognition, information extraction, and semantic analysis. This study employed CRF++ version 0.58 for Chinese text segmentation under Linux.

As shown in Figure 4 [Figure 4: see original paper], the experimental process comprised four stages: training, testing, evaluation, and optimization. The training stage involved feature extraction to identify suitable features for food safety corpora, assigning different feature tags and converting them into CRF++-readable format. Different features and combinations were selected to construct feature templates. The training data and feature template files were then used to train segmentation models. The testing stage applied these models to similarly formatted test data to produce final segmentation results, as exemplified in Table 1 . The evaluation stage assessed output results, comparing them with other experiments and iteratively adjusting feature selection, tagging schemes, and templates to achieve optimal performance.

[Figure 4: see original paper] Experimental workflow

Given the large corpus size and lack of initial annotations, all training and testing data required segmentation. We employed a hybrid approach combining automatic segmentation with manual correction. First, the NLPIR system performed automatic segmentation [15]. Due to the strong domain specificity of food safety events, researchers in food science supervised systematic and comprehensive manual proofreading of segmentation results, creating a high-quality segmented corpus [16]. All experiments were conducted based on this manually corrected data.

Chinese word segmentation for food safety emergencies can be abstracted as a

sequence labeling task, where input text represents the observation sequence. Given this sequence, the CRF model computes the maximum joint probability distribution across the entire sequence. Effective feature tag selection significantly impacts final segmentation quality. Therefore, we first screened and determined appropriate feature selections and corresponding tags, then constructed feature templates for model training. With training data and templates established, the CRF model was trained to produce final segmentation results.

The model optimization stage represents this paper' s core contribution. Through continuous experimentation with new feature tags, different feature combinations, and templates better suited to text characteristics, we achieved improved evaluation metrics: higher precision (P), recall (R), and F-score. The harmonic mean formula is:

F = 2PR / (P + R)

In practice, P, R, and F values were calculated for each positional feature tag, then weighted by their proportional frequency in the test corpus to obtain final weighted averages.

(1) Feature and Feature Tag Selection In CRF-based Chinese word segmentation, the training stage requires correctly segmented corpora (after machine segmentation and manual correction). Different feature selections and tagging schemes yield varying segmentation performance.

This study tested three positional tagging schemes based on character positions within words, as shown in Table 2 . Positional features were placed in the final column of training data as CRF++ output. After segmentation, these tags were used to compose words from characters.

[Table 2 content about tagging schemes would be preserved here with proper formatting]

When calculating P, R, and F values, each tag required a weight. As shown in Tables 3 through 5 , we counted each feature tag' s frequency in the test corpus and used its percentage of total tags as its weight.

[Tables 3-5 content about tag distributions would be preserved here]

Most existing CRF-based segmentation research remains at the single-feature-tag stage. While some studies explore combined features through simple addition, our experiments revealed that different feature combinations produce varying effects—more features do not guarantee better performance. Feature interference and redundancy from excessive features can actually degrade segmentation quality [17].

Beyond positional features, we conducted individual and combined experiments on other common features for modern Chinese texts, such as phonetic features and word length features. Training corpora with selected feature combinations

were processed into CRF++-readable format, as shown in Table 6 , with positional features in the final column since segmentation was our primary task.

[Table 6 content about multi-feature training data would be preserved here]

During training and testing, the corpus was divided into 10 parts using 7:3 training-to-testing ratio with 10-fold cross-validation for stability assessment.

Experiments on different features and combinations yielded the results shown in Table 7 . The data indicate that among 4Tag, 5Tag, and 6Tag schemes, 4Tag and 5Tag generally achieved higher P, R, and F values than 6Tag. Adding other features caused performance degradation across all tag types.

[Table 7 content about segmentation evaluation results would be preserved here]

(2) Feature Template Construction and Optimization CRF++ feature templates define how features are extracted from training data. Extracted feature strings serve as binary functions in CRF++ to determine whether specific labels should be output.

Template files primarily use Unigram Templates (starting with "U"), where each line (e.g., U01:%x[-2,0]) represents a feature. The macro "%x[row offset, column position]" indicates relative row and absolute column positions from the current token, as shown in Table 8 . Each "%x[row, column]" generates a state function f(s, o), where s is the tag at time t and o is the context.

[Table 8 content about basic feature templates would be preserved here]

Using the corpus from Table 6 with these templates, features like "current character,""preceding first character,"and "following first character"are extracted as state functions.

Different feature templates and their combinations with various features significantly impact segmentation performance [18]. Using the effective 5Tag scheme, we experimented with different templates, with results shown in Table 9 . Removing binary features caused significant F-score drops, while unigram feature changes had minimal impact. Adding binary features without the current character (%x[0,0]) also showed little effect.

[Table 9 content about template comparison results would be preserved here]

4. Experimental Results Analysis

Comparative analysis of feature selection experiments, visualized in Figure 5 [Figure 5: see original paper], grouped data into three sets (4Tag, 5Tag, 6Tag). Results show that original positional features alone yielded the best F-scores, with performance declining after adding other features.

[Figure 5 reference and Table 10 content about word length distribution would be preserved here]

Our food safety emergency corpus segmentation performance is slightly lower than results on standard test sets [19]. Primary reasons include: (1) manual correction errors during post-processing, where many machine segmentation errors went uncorrected or correct segmentations were mistakenly altered, affecting both training and evaluation; and (2) the specialized nature of the domain.

CRF' s greatest advantage lies in incorporating not only current character features but also left and right contextual features to form effective templates [18]. Experiments showed that removing features involving the current character with its immediate neighbors caused significant F-score reductions. Table 10 shows the corpus word length distribution, where two- and three-character words account for 46.62%, explaining why these neighboring character features are indispensable. With words of three or more characters comprising only about 2.28%, features involving second or third neighboring characters had minimal impact.

Applying the CRF model to food safety emergency corpora using the mature, stable CRF++ toolkit, we conducted comprehensive experiments considering multiple features and their combinations. Results demonstrate that feature tag selection and combination choices affect segmentation performance, with 4Tag and 5Tag positional features achieving optimal F-scores of 92.87% and 92.88%, respectively. Additional features caused performance degradation. Through comparative analysis of different templates, we identified optimal templates matching our feature selections and research objectives.

Future research will further explore text features to incorporate contextual semantics and structural information, aiming for improved automatic segmentation performance.

References

[1] Li Hongfeng. Analysis of Realistic Plights and Countermeasures in Social Co-governance on Food Safety in China[J]. Food & Machinery, 2016, 32(4): 234-236.

[2] Wang Huixia. Public Participation in Food Safety Management of the Rule of Law[J]. Commercial Research, 2012(4): 170-177.

[3] Feng Guohe, Zheng Wei. Review of Chinese Automatic Word Segmentation[J]. Library and Information Service, 2011, 55(2): 41-45.

[4] Zhang Xinglian, Tang Xiaochun. Establishment on Database System of Food Safety Early-warning in China[J]. Food Science and Technology, 2008, 33(12): 250-254.

[5] Wu Yunhong, Zhu Liang, Chu Wei, et al. Key of Food Supervision and Administration Reform-dynamic and Third Party Database Based on Internet[J]. Science and Technology of Food Industry, 2009(9): 272-274.

[6] Yu Qing, Hong Yuan. Construction Idea for Risk Database of Processed Food[J]. Value Engineering, 2013(30): 174-175.

[7] Jia Kai, Peng Peihao, Ruan Weiling. Study on the Investigation of Farmer Cooperatives in Sanjie Town, Pengzhou City, Sichuan Province[J]. Beijing Agriculture, 2014(3): 247-248.

[8] Huang Changning, Zhao Hai. Chinese Word Segmentation: A Decade Review[J]. Journal of Chinese Information Processing, 2007, 21(3): 8-19.

[9] Zeng D, Wei D, Chau M, et al. Domain-specific Chinese Word Segmentation Using Suffix Tree and Mutual Information[J]. Information Systems Frontiers, 2011, 13(1): 103-114.

[10] Liu Zewen, Ding Dong, Li Chunwen. Chinese Word Segmentation Method for Short Chinese Text Based on Conditional Random Fields[J]. Journal of Tsinghua University: Science and Technology, 2015, 55(8): 16-20.

[11] Lafferty J D, 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.

[12] Pearl J. Bayes and Markov Networks: A Comparison of Two Graphical Representations of Probabilistic Knowledge[R]. Los Angeles, California, USA: University of California, 1986.

[13] Wallach H M. Conditional Random Fields: An Introduction[EB/OL]. (2004-02-24). http://www.inference.phy.cam.ac.uk/hmw26/papers/crf_{intro}.pdf.

[14] CRF++: Yet Another CRF Toolkit[EB/OL]. [2014-08-04]. http://crfpp.sourceforge.net/.

[15] Institute of Computing Technology of the Chinese Academy of Sciences. ICTCLAS Chinese Word Segmentation System[CP/OL]. (2016-02-17). [2016-06-30]. http://ictclas.nlpir.org/.

[16] Yue Jinyuan, Xu Jin' an, Zhang Yujie. Chinese Word Segmentation for Patent Documents[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2013, 49(1): 159-164.

[17] Chen L, Li M, Zhang J, et al. A Double-Layer Word Segmentation Combined with Local Ambiguity Word Grid and CRF[J]. Transactions on Computer Science & Technology, 2013, 2(1): 1-8.

[18] Huang Shuiqing, Wang Dongbo, He Lin. Exploring of Word Segmentation for Fore-Qin Literature Based on the Domain Glossary of Sinological Index Series[J]. Library and Information Service, 2015, 59(11): 127-133.

[19] Zhao H, Huang C N, Li M, et al. An Improved Chinese Word Segmentation System with Conditional Random Field[C]//Proceedings of the 5th SIGHAN Workshop on Chinese Language Processing. 2006: 162-165.

Author Contributions

Wang Dongbo: Conceived research ideas and designed the study; Wang Dongbo, Zhang Yue, Zhu Danhao: Collected, cleaned, and analyzed data; Zhang Yue: Conducted experiments and drafted the manuscript; Wang Dongbo, Zhang Yue: Revised the final manuscript.

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

Supporting data [1-3] are self-archived by authors, E-mail: db.wang@njau.edu.cn; Supporting data [4] is available in the journal' s online version at http://www.infotech.ac.cn.

[1] Zhang Yue, Wang Dongbo. data.txt. Training and testing data for Chinese word segmentation of food safety emergencies.
[2] Zhang Yue, Wang Dongbo. Template. Feature templates for Chinese word segmentation of food safety emergencies.
[3] Zhang Yue, Wang Dongbo. result.txt. Segmentation results for Chinese word segmentation of food safety emergencies.
[4] Zhang Yue, Wang Dongbo. Java project files for wordseg.

Received: September 22, 2016
Revised: October 31, 2016

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