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

Authors: Zhang Yue, Wang Dongbo, Zhu Danhao

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

Abstract

【Objective】 In the field of food safety, establishing relevant databases greatly facilitates the supervision and control of food safety, and automatic word segmentation plays a crucial role in index construction, index usage, and corpus construction. This study applies a character-based tagging statistical learning method based on Conditional Random Fields (CRF) to automatic word segmentation of food safety emergency event corpora. 【Method】By analyzing characteristics such as word length distribution in the corpus, various experiments were conducted on feature selection and feature templates involved in the automatic word segmentation process of this method, to determine the impact of different feature selections and different feature templates on segmentation results. 【Results】 Experimental results demonstrate that more features do not necessarily yield better segmentation performance during feature selection, as feature interference may occur. In the food safety emergency event corpus where two- and three-character words account for 46.62%, the feature templates representing the current character and the first preceding and following characters have a significant impact on segmentation performance. 【Conclusion】 Through experiments on different feature selections, feature templates, and their combinations, the optimal features and feature templates for automatic word segmentation in the corpus studied in this paper were selected. Under the 5Tag labeling scheme with corresponding feature templates, the F-measure for segmenting the target corpus reached 92.88%.

Full Text

Research on Feature Selection and Model Optimization for Chinese Word Segmentation of 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 provides substantial support for monitoring and controlling food safety, where automatic word segmentation plays a critical role in index construction, index utilization, and corpus development. This study applies a character-labeling statistical learning method based on 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 feature selections and templates affect segmentation performance. [**Results**] Experimental results demonstrate that more features do not necessarily yield better segmentation outcomes, as feature interference may occur. 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 significantly impact segmentation effectiveness. [**Conclusions**] Through systematic experiments on different feature combinations and templates, we identify the optimal features and templates for automatic segmentation of our target corpus, achieving an F-score of 92.88% using 5Tag feature labeling with corresponding feature templates.

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

Introduction

Food safety incidents have occurred with increasing frequency in recent years, with numerous severe cases profoundly impacting social production and public life. The rapid proliferation of information regarding food safety emergencies has attracted widespread attention. As food safety concerns public health and lives, addressing these issues requires not only top-down administrative supervision and corporate self-regulation but also bottom-up participation from social oversight forces. In today's fast-paced information environment, networks, newspapers, and books serve as primary channels for disseminating food safety emer-

gency information and constitute a major source for public awareness. With the vigorous development of natural language processing, research on Chinese text automatic segmentation has achieved significant progress in both accuracy and speed, playing a pivotal role in various Chinese information processing tasks such as text classification, information retrieval, information filtering, automatic indexing, and automatic summarization. However, applications and research on automatic segmentation in food safety information processing remain limited and warrant further exploration.

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

Current Chinese automatic segmentation methods primarily fall into four categories: mechanical segmentation, statistical segmentation, character-labeling statistical learning, and deep neural network-based methods. Before 2002, automatic segmentation methods were essentially dictionary-based, further divided into rule-based mechanical methods and statistical methods. These approaches rely entirely on dictionaries as their sole information source. 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 with the emergence of out-of-vocabulary words.

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-labeling statistical learning methods demonstrate superior performance in handling out-of-vocabulary words and disambiguation, clearly outperforming dictionary-based methods without requiring lexical resources. Deep learning-based methods remain immature with limited applications in natural language processing and are not discussed in this paper.

Chinese word segmentation using character-labeling statistical learning is essentially a sequence labeling process that abstracts text information into an observation sequence and labels each character. 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 commonly used. HMM' s primary limitation is its output independence assumption, which prevents consideration of contextual features and restricts feature selection. ME models address this issue by allowing arbitrary feature selection but require normalization at each node, leading to only local optima and the label bias problem, where all unseen cases in training data are ignored. CRF models resolve these issues by performing global normalization across all features rather than at each node, thereby obtaining globally 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 of food safety emergencies based on corpus characteristics. 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 selection, and labeling schemes tailored to its characteristics. Compared with other segmentation systems, this approach better resolves overlapping ambiguities and out-of-vocabulary word problems in dense texts like food safety case databases, effectively improving segmentation accuracy and recall rates.

2. Conditional Random Fields Model Introduction

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 and a conditional probability model for labeling and segmenting sequential data.

Undirected graphical models, also known as Markov random fields or Markov networks, were introduced by Pearl. An undirected graph consists of nodes representing random variables and edges representing conditional dependencies between them. Due to the undirectionality of edges, conditional probability parameterization cannot represent joint probability directly. Instead, joint probability must be expressed as a product of local functions derived from conditional independence principles.

[Figure 1: see original paper] illustrates a simple example where the maximal cliques in the undirected graph are $\{X_1, X_2\}$, $\{X_1, X_3\}$, $\{X_3, X_4\}$, and $\{X_2, X_4, X_5\}$, yielding the joint probability distribution:

$$P(X) = \frac{1}{Z} \prod_{c \in C} \psi_c(X_c)$$

where Z is the partition function and ψ_c are potential functions over cliques.

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 including an additional observation set X.

CRFs aim to solve sequence labeling problems for discrete data. Given a sequence $X = \{x_1, x_2, x_3, \dots, x_{n-1}, x_n\}$ and a finite state set $Y = \{y_1, y_2, y_3, \dots, y_{n-1}, y_n\}$, let $G = (V, E)$ be an undirected graph where $Y_v \in V$ represents random variables indexed by graph nodes. If each random variable satisfies the Markov property:

$$p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v)$$

where $w \sim v$ indicates adjacent nodes, a conditional random field is formed.

As shown in [Figure 2: see original paper], CRF employs an undirected graphical model to describe states given a sequence. In a conditional random field, each element corresponds to a graph node, and each edge represents node states. CRF is essentially an undirected graphical model conditioned on observation sets.

According to CRF theory:

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

3. Experimental Setup

3.1 Food Safety Corpus Description Based on the collection, annotation, and organization of food safety emergencies, we constructed a corpus covering 2005–2015. The final experimental data were produced through coarse segmentation and manual correction, as illustrated in [Figure 3: see original paper].

The corpus construction process involved three main steps. First, we collected food safety emergencies from both online sources and printed newspapers/books. Online data were automatically gathered using a custom program with a vertical search engine for emergency topics, covering news portals, forums, and blogs. Heterogeneous collected data were cleaned, transformed, and stored in a database. Approximately 5,000 printed cases were manually entered and proofread, resulting in about 4,500 entries after cleaning, totaling approximately 20MB.

Second, we used the NLPIR segmentation software from the Institute of Computing Technology, Chinese Academy of Sciences, for initial annotation. However, NLPIR exhibited poor recognition of out-of-vocabulary words. Food safety case databases constitute dense texts with numerous Chinese out-of-vocabulary words and ambiguous terms. Texts describing foods, geographical locations, and chemical compounds are particularly representative, involving extensive

food and chemical names that machines often mislabel. Therefore, NLPIR was only used for coarse segmentation to reduce manual workload.

Third, we manually annotated the coarsely segmented corpus. Due to numerous unrecognized out-of-vocabulary words and misidentified ambiguous terms, we systematically reviewed each word, corrected segmentation errors, and ensured maximum accuracy in the training corpus.

3.2 Implementation Method CRF++ 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 employs CRF++ version 0.58 for Chinese text segmentation under Linux.

The experimental process, shown in [Figure 4: see original paper], comprises four stages: training, testing, evaluation, and optimization. The training stage involves feature extraction, selecting appropriate features for food safety corpora, assigning feature labels, and formatting data for CRF++ processing. Different features and combinations are selected to construct feature templates, which together with training data generate segmentation models. The testing stage applies these models to similarly formatted test data to produce segmentation results, as exemplified in . The evaluation stage assesses output results, compares them across experiments, and iteratively refines feature selection, labeling schemes, and templates to achieve optimal performance.

Given the corpus size and lack of initial annotations, we employed a hybrid approach combining automatic segmentation with manual correction. The NLPIR system first performed automatic segmentation, after which domain experts guided systematic manual proofreading to correct errors and produce high-quality segmented corpora for all experiments.

We conceptualize Chinese word segmentation of food safety emergencies as a sequence labeling task where input text represents the observation sequence. The CRF model computes the maximum joint probability distribution for the entire sequence given these observations. Effective feature labeling critically impacts segmentation performance, requiring careful selection of features and corresponding labels before determining training templates. Once training data and templates are established, the CRF model can be trained to produce final segmentation results.

The optimization stage represents our core research contribution. Through continuous experimentation with new feature labels, different feature combinations, and templates better suited to textual characteristics, we improved model evaluation metrics including precision (P), recall (R), and F-score. The formulas are:

$$P = \frac{\text{Number of correctly segmented words}}{\text{Total number of words segmented by the system}}$$

$$R = \frac{\text{Number of correctly segmented words}}{\text{Total number of words in the test set}}$$

$$F = \frac{2 \times P \times R}{P + R}$$

In practice, we calculate P, R, and F values for each position label, then compute weighted averages based on each label' s proportion in the test corpus.

(1) Feature and Feature Label Selection

In CRF-based Chinese word segmentation, the training stage requires correctly segmented corpora (after machine segmentation and manual correction). Different feature selections and labeling schemes significantly impact segmentation performance.

We tested three position labeling schemes based on character positions within words, as shown in : 4Tag {B, M, E, S}, 5Tag {B, I, M, E, S}, and 6Tag {B, I, J, M, E, S}, where B marks word-beginning characters, M marks middle characters, E marks end characters, and S marks single-character words. Position features occupy the final column of training data for CRF++ output, enabling word reconstruction from labeled characters.

When calculating P, R, and F values, we assign weights to each label based on its frequency in the test corpus. through show label distributions, with percentages serving as weights. Most CRF-based segmentation research remains at the single-feature stage. Our experiments reveal that different feature combinations produce varying effects—more features do not guarantee better performance, as feature interference and redundancy may degrade results.

Beyond position features, we conducted separate and combined experiments on common Chinese text features such as pronunciation and word length features. Training corpora with multiple features were formatted for CRF++ processing, as illustrated in . All experiments used the feature template shown in to ensure distinguishability, with multi-feature templates adding additional lines as needed.

Using 10-fold cross-validation with a 7:3 training-testing split, we evaluated different feature combinations. Results in show that among 4Tag, 5Tag, and 6Tag schemes, 4Tag and 5Tag generally achieve higher P, R, and F values than 6Tag. Adding pronunciation or length features to any tag scheme decreases performance, with combined feature additions showing further degradation.

(2) Feature Template Construction and Optimization

CRF++ feature templates define feature extraction methods from training data. Extracted feature strings function as binary functions determining whether specific labels should be output. The template primarily uses Unigram Templates (prefixed with "U"), where each line (e.g., U01:%x[-2,0]) represents a feature. The macro "%x[row_{offset}, column_{absolute}]" specifies token positions relative to the current token, as detailed in . Each template line generates a state function f(s, o), where s represents the label at time t and o represents the context.

Template construction significantly impacts segmentation performance. Using the effective 5Tag scheme, we tested various templates with results shown in . Removing unigram features caused minimal performance change, while removing bigram features substantially decreased F-scores. Adding unigram features showed negligible impact, and added bigram features without current character involvement (%x[0,0]) had little effect.

4. Experimental Results Analysis

Comparative analysis of feature selection experiments, visualized in [Figure 5: see original paper] from data, reveals consistent patterns across 4Tag, 5Tag, and 6Tag groups. Original position features alone achieve the best F-scores, with performance declining when adding other features.

presents word length distribution in our corpus, showing two-character words at 41.39% and three-character words at 5.23%, together accounting for 46.62% of the total. This explains why features involving the current character and its immediate neighbors are essential, while features involving second-order neighbors have minimal impact given that words longer than three characters represent only about 2.28% of the corpus.

Our segmentation performance on food safety emergency corpora is slightly lower than results on standard test sets. Error analysis reveals that during manual correction of machine-segmented results, many errors went uncorrected while some correctly segmented words were mistakenly altered, affecting both training and evaluation stages.

CRF' s primary advantage lies in incorporating not only current character features but also left and right contextual knowledge to form effective feature templates. Experiments demonstrate that removing features connecting the current character with its immediate neighbors significantly degrades performance. The word length distribution data confirm the importance of these neighboring character features.

Applying the CRF model to food safety emergency corpora using the robust CRF++ toolkit, we systematically explored multi-feature combinations. Results indicate that feature labeling schemes and combinations substantially impact performance, with 4Tag and 5Tag position features achieving optimal F-scores of 92.87% and 92.88%, respectively. Additional features consistently de-

creased performance. Feature template analysis identified configurations best suited to our corpus characteristics.

Future research will further explore textual features that incorporate contextual semantics and structural information to achieve improved 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): 115-125.

[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 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: Conceptualized the research and designed the study methodology.
Wang Dongbo, Zhang Yue, Zhu Danhao: Collected, cleaned, and analyzed the data.
Zhang Yue: Conducted experiments and drafted the manuscript.
Wang Dongbo, Zhang Yue: Revised the final version of the paper.

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

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

[1] Zhang Yue, Wang Dongbo. data.txt. Chinese word segmentation training and test data for 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. Chinese word segmentation results

for 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.