Construction of a Question-Answer Dataset and Research on Named Entity Recognition in the Food Manufacturing Industry

Yuqi Li

School of Computer Science, Guangdong University of Technology, Guangzhou, Guangdong Province, 510000, China

2122105324@mail2.gdut.edu.cn

Abstract. In recent years, food safety concerns have gained considerable attention, leading to an increasing demand in the food manufacturing industry for real-time monitoring, quality control, and traceability. Precise entity recognition becomes an effective means of extracting key information in addressing food-related issues. However, the lack of publicly available classify datasets in this domain emphasizes the urgent need to construct datasets tailored for the food manufacturing industry. In this study, we built a question-answering dataset for named entity recognition tasks and employed various models for training and evaluation. Experimental tests were conducted on four identification models based on BERT as the baseline model, using both publicly available datasets and the constructed dataset. The results indicate that the constructed dataset is suitable for named entity recognition tasks in the food manufacturing industry. Furthermore, the BERT-BiLSTM-CRF model outperforms other models in precision, recall, and F1 score, demonstrating its effectiveness in this domain.

Keywords: Named entity recognition; Food manufacturing industry; BERT-BiLSMT-CRF.

1 Introduction

The food manufacturing industry encompasses multiple fields, including chemistry, biology, and engineering, among others.[1]Establishing a named entity recognition dataset for the food manufacturing industry contributes to a deeper understanding and analysis of relevant information, fostering the development of the industry. By accurately identifying key entities in the food manufacturing domain, such as food components, processes, equipment, and food safety, it is possible to deepen research on related issues, meeting the growing consumer concerns regarding food safety and ingredients. Constructing such a dataset aids in improving transparency, making product information more accessible to consumers, building trust, and promoting the healthy development of the food manufacturing industry.

Z. Zhan et al. (eds.), *Proceedings of the 2024 10th International Conference on Humanities and Social Science Research (ICHSSR 2024)*, Advances in Social Science, Education and Humanities Research 858,
https://doi.org/10.2991/978-2-38476-277-4_33

Named Entity Recognition (NER), a vital task in Natural Language Processing (NLP), aims to identify specific entities with particular meanings in text and finds widespread applications in information extraction, question-answering systems, and machine translation, among other fields.[2]Deep learning methods such as BERT pre-trained models have made significant strides in enhancing NER performance. Utilizing Recurrent Neural Networks (RNN) networks[3],effectively captures contextual information, improving recognition precision. Internationally, Damion et al.[4]introduced the FoodOn ontology concept, providing a standardized semantic representation for relevant knowledge in nutrition and food safety. There is a lack of publicly available entity recognition datasets in the food manufacturing industry. This paper's primary contributions include entity classification standards based on text properties in the food domain, a classification dataset constructed using supervised learning strategies, and an evaluation of named entity recognition models.

2 Dataset Construction

In the process of constructing the dataset for the food manufacturing industry, with the goal of ensuring high quality and extensive coverage, a comprehensive and diverse data collection is achieved by integrating different information sources.[5]This dataset comprises structured, semi-structured, and unstructured data, with national standard documents and data sheets as the primary sources of structured data, providing essential nutritional parameters and information on food additive components. Crawler to obtain specialized domain datasets, such as the Foodmate database, further enhances the diversity of comprehensive Chinese data. For unstructured data, materials such as textbooks, popular science literature, research papers, and online forum articles are processed, and data normalization is achieved through knowledge fusion techniques. Throughout the entire process, emphasis is placed on data cleaning, labeling, and dataset partitioning. This diverse data collection approach constructs a rich and comprehensive dataset for the food manufacturing industry, providing robust support for relevant research and applications.

3 Text Data Processing in the Food Manufacturing Industry

3.1 Text Classification, Feature Analysis, and Construction of Entity Labels

The resources in the food manufacturing industry domain encompass vital information related to food classification, physiological characteristics, agricultural and technological practices, and production experiences. This paper constructs a dataset based on the characteristics of the food manufacturing industry, with domain data primarily involving source entities in the food domain, including additives, nutritional enhancers, processing techniques, and more. Data representation covers the biochemical characteristics of food, pest and disease issues, pharmaceuticals for livestock feeding, processing methods, and other relevant aspects. Data conditions are often context-specific, with a wide diversity of food categories, significant importance placed on additives and

ingredients, and various processing methods. The food manufacturing industry focuses on safety and nutrition, involving concerns such as harmful substances, microbial contamination, and nutritional components. Based on these characteristics, knowledge in the food manufacturing industry domain is categorized into 11 entity types. The classification details are shown in Table 1 below.

Table 1. Entity Type Table.

Entity Types	Chinese meaning	Provide examples.
FN	Nutrient fortifiers	Vitamin A, Calcium, Iodine
FAA	Food additives	Sorbitol, Potassium Sorbate
PT	Food Packaging Technology	Vacuum Packaging, Modified Atmosphere Packaging
FIA	Instruments in the Food Domain	High-Performance Liquid Chromatography (HPLC)
PA	Pathogenic Bacteria	Salmonella
CP	Contaminants	Benzopyrene
PRT	Processing Techniques	Pasteurization, Supercritical Extraction
MYC	Fungal Toxins	Aflatoxin
PAM	Packaging Materials	Polystyrene
VM	Edible and Non-edible Livestock Medications	Furacilin
EM	Edible and Non-edible Crop Medications	Benomyl

3.2 Construction and Processing Methods of the Corpus

The data is annotated using the BIOES labeling strategy. Taking the entity category "Food Additive" as an example, the initial part is labeled as "B-FAA," the middle entity content is labeled as "I-FAA," the ending entity content is labeled as "E-FAA," a single entity is labeled as "S-FAA," and other components in the sentence are labeled as "O."

4 Named Entity Recognition Model for Food Manufacturing Sector

In this study, BERT is adopted as the baseline model. To enhance named entity recognition performance, we experimented with the following model combinations: BERT-CRF model, applying CRF for sequence labeling to capture contextual information and label dependencies of entities; BERT-BiLSTM model, using BiLSTM for sequence labeling to better capture contextual features of entities; BERT-BiLSTM-CRF model, combining the strengths of BiLSTM and CRF with BERT to achieve more accurate named entity recognition results. The core of the BERT model consists of a multi-layer Transformer encoder[6], incorporating self-attention mecha-

nisms to capture dependencies between words. The input structure diagram of the BERT model is illustrated in Figure 1.

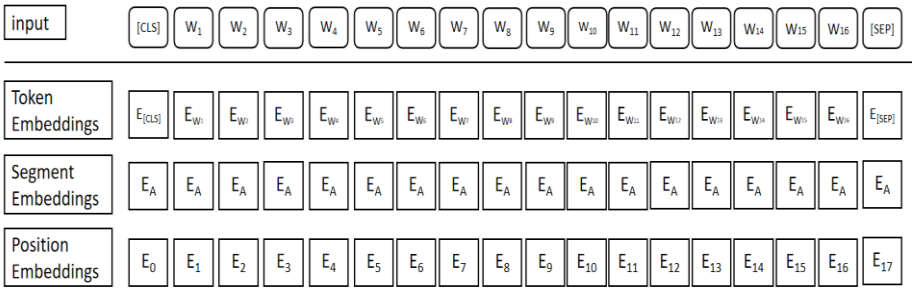

Fig. 1. Structure diagram of BERT model input

BiLSTM (Bidirectional Long Short-Term Memory) is a variant of RNN used for processing sequential data, capable of capturing semantic information from both past and future contexts simultaneously[7].In an input sentence T containing N characters, each character *ti (1≤i≤N)* undergoes word embedding processing to obtain a representation *xi=[wi]*, where *wi* represents the result obtained from a pre-trained word embedding model. The formula for the BiLSTM model is as follows[8]:

$$i_t = sigmoid(W_{xi}x_t + W_{hi}h_{(t-1)} + b_i) \tag{1}$$

$$f_t = sigmoid(W_{xf}x_t + W_{hf}h_{(t-1)} + b_f) \tag{2}$$

$$g_t = \tan h\,(W_{ig}x_t + W_{hg}h_{(t-1)} + b_g) \tag{3}$$

$$o_t = sigmoid(W_{xo}x_t + W_{ho}h_{(t-1)} + b_o) \tag{4}$$

$$C_t = f_t \otimes C_{(t-1)} + i_t \otimes g_t \tag{5}$$

$$h_t = o_t \otimes tanh(C_t) \tag{6}$$

Where W and b are the parameters of the BiLSTM, with W as the weight matrix and b as the bias vector., $sigmoid$ represents the activation function,⊗denotes the element-wise multiplication,i_t, f_t, o_t epresent the input gate, forget gate, and output gate at time step t, respectively,C_t, h_t, g_t represent the cell state, output state, and new state in the BiLSTM.

CRF (Conditional Random Field) is a probabilistic graphical model primarily used for modeling the joint probability distribution of sequential data. It finds extensive applications in tasks such as sequence labeling in fields like NLP and computational biology[9]. The basic idea involves describing the relationship between input and output sequences through a set of feature functions[10]. The conditional probability formula for CRF is as follows:

$$P(Y|X) = \left(\frac{1}{Z(X)} \times emp \sum_i \theta_j \times f_j(X, Y)\right) \tag{7}$$

Here, $Z(X)$ is the normalization factor used to ensure that the sum of conditional probabilities is equal to 1.θ_j represents the parameters of the model, and $f_j(X, Y)$ is the feature function defined on the input sequence X and the label sequence Y, used to capture the relationship between input and labels.

5 Experimental Design and Results Analysis

5.1 Experimental Environment and Parameter Settings

This experiment utilized Python version 3.8.12 and compared the NLPCC-ICCPOL, MSRA, RESUME public datasets, and an experimental dataset. All experiments were conducted on a host configured with an Intel(R) Xeon(R) CPU E5-2678 v3 @ 2.50GHz, 62GB of RAM, NVIDIA GeForce RTX 2080Ti 11GB graphics card, a 100GB hard drive, and running the Linux operating system. The model training parameters were set as follows: a learning rate (lr) of 0.005, 50 epochs, word vector dimensionality of 768, and BiLSTM hidden vector dimensionality of 128.

Firstly, the named entity recognition performance of the BERT-BiLSTM-CRF model was validated on the NLPCC-ICCPOL, MSRA, and RESUME datasets, employing deep learning models for comparison. The experimental results demonstrated the generalization ability and language understanding capability of the BERT model, combining the advantages of BiLSTM and the overall performance improvement provided by CRF enhanced named entity recognition. Subsequently, the model was applied to the experimental dataset to validate the usability of the constructed dataset in the food manufacturing industry domain.

5.2 Experimental Results and Analysis

This paper analyzed the named entity recognition performance of four models, namely BERT, BERT-CRF, BERT-BiLSTM, and BERT-BiLSTM-CRF, on the NLPCC-ICCPOL, MSRA, and RESUME public datasets. Simultaneously, ten-fold cross-validation experiments were conducted on the classified and labeled datasets, providing a comprehensive analysis of the recognition performance of the models on different datasets.

5.2.1 Comparative Experiments of Various Models on Public Datasets

Table 2 presents the experimental results of four deep learning models based on BERT on the NLPCC-ICCPOL, MSRA, and RESUME public datasets. The BERT-BiLSTM-CRF model surpasses other models in overall evaluation metrics, demonstrating superior recognition capabilities. This model will be further utilized to validate the rationality and effectiveness of the constructed textual dataset in the food manufacturing industry domain.

Table 2. Comparison experiment results of different models on the public dataset

Dataset	Model	P	R	F1
NLPCC-ICCPOL	BERT	82.04	90.48	86.05
	BERT-BiLSTM	94.60	96.03	95.31
	BERT-CRF	95.68	94.90	95.29
	BERT-BiLSTM-CRF	96.37	96.53	96.45
MSRA	BERT	56.35	76.54	64.91
	BERT-BiLSTM	85.66	90.99	88.24
	BERT-CRF	88.80	88.74	88.77
	BERT-BiLSTM-CRF	92.13	91.86	91.99
RESUME	BERT	73.46	89.33	80.62
	BERT-BiLSTM	88.97	94.48	91.64
	BERT-CRF	93.41	93.87	93.64
	BERT-BiLSTM-CRF	93.32	95.09	94.20

Experimental data show that the BERT model has a basic effect on entity recognition tasks, but the effect is worse than that of the other three models, with F1 of 86.05%, 64.91% and 80.62%, respectively. This may be due to the fact that BERT is pre-trained with local context information, while NER tasks require global label information to ensure entity boundary consistency. When using BERT directly, it only focuses on the independent prediction of each marker. In contrast, CRF introduces a loss function on a global label sequence, which more comprehensively considers the label combination of the entire sequence. The BERT-BiLSTM model can better capture the context information through BiLSTM. The BERT-BiLSTM-CRF model is an end-to-end structure, which is trained end-to-end on the NER task, which is direct and synthetic, and is helpful for learning task-related features. In terms of various evaluation indicators, the BERT-BiLSTM-CRF model has better entity naming recognition performance in public datasets, with F1 values of 96.45%, 91.99% and 94.20%, respectively, which are better than other models.

5.2.2 Comparative experiments of various models on datasets in the field of food manufacturing

The author carried out a ten-fold cross-validation experiment in the field of food manufacturing, divided the dataset into 10 parts, selected 9 as the training set and 1 as the test set each time, and carried out comparative experiments on the four models. The comparison results are shown in Table 3, and the optimal results of precision and F1 value are presented in bold. The BERT-BiLSTM-CRF model has the best learning effect, and compare to BERT-BiLSTM the F1 value is increased by 7%~12%.On average, 4% higher than the BERT-CRF model. Through the comparison of the average F1 value of the ten-fold cross-validation experiment, the overall recognition effect of the BERT-BiLSTM-CRF model was better than that of other models, and the F1 value of each fold was better than that of other models, so the performance was stable.

Table 3. Tenfold cross-validation experiments under different models

Dataset	BERT			BERT-BiLSTM			BERT-CRF			BERT-BiLSTM-CRF		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	43.02	65.77	52.01	83.09	91.04	86.89	88.54	90.06	89.29	**92.77**	**94.48**	**93.62**
2	44.66	67.88	53.87	82.39	89.58	85.83	87.00	87.64	87.32	**93.12**	**91.88**	**92.50**
3	44.43	68.72	53.97	81.40	90.02	85.50	90.37	90.15	90.26	**94.31**	**93.84**	**94.07**
4	46.61	70.20	56.02	78.03	89.66	83.44	89.58	92.12	90.83	**94.33**	**94.21**	**94.27**
5	49.66	70.98	58.43	82.92	89.39	86.03	91.43	89.76	90.58	**96.02**	**94.15**	**95.07**
6	46.67	67.48	55.18	83.58	90.50	86.90	91.19	92.08	91.64	**94.82**	**93.67**	**94.24**
7	48.27	71.12	57.51	82.98	91.14	86.87	93.20	91.50	92.35	**94.87**	**94.17**	**94.52**
8	43.23	65.42	52.06	77.77	87.23	82.23	90.21	89.88	90.04	**90.44**	**90.00**	**90.22**
9	45.45	67.93	54.47	79.59	89.72	84.35	85.09	88.74	86.88	**93.62**	**93.39**	**93.50**
10	45.95	68.73	55.08	80.68	88.61	84.46	88.22	90.79	89.49	**94.64**	**94.18**	**94.41**

According to the Table 4, it is evident that BERT-BiLSTM-CRF outperforms BERT-BiLSTM in terms of precision, recall, and F1 score. This is attributed to the CRF layer's ability to globally adjust sequence labeling tasks, considering constraints on the overall label sequence to enhance model performance. Although both BERT-CRF and BERT-BiLSTM-CRF utilize the CRF layer, the BiLSTM layer in BERT-BiLSTM-CRF may provide more contextual information to CRF. BERT-BiLSTM-CRF combines the advantages of the BiLSTM and CRF model and improves the text vectorization process on this basis, fully leveraging the advantages of the BERT pretrained language model. It automatically extracts rich semantic, word-level, and syntactic structural features from the text sequence. Therefore, in entity recognition tasks, BERT-BiLSTM-CRF is significantly superior to the other two models, achieving precision, recall, and F1 scores of 93.90%, 93.40%, and 93.64%, respectively.

Table 4. Average experimental results of different models on the experimental dataset

Model	P	R	F1
BERT	45.80	68.42	54.86
BERT-BiLSTM	81.24	89.69	85.25
BERT-CRF	89.48	90.27	89.87
BERT-BiLSTM-CRF	**93.90**	**93.40**	**93.64**

6 Conclusion

This paper provides a quick solution for named entity recognition in the field of food manufacturing, alleviating the pressure on professionals, researchers, and workers in the food manufacturing industry. Additionally, it lays the foundation for building domain knowledge graphs and developing intelligent question-answering systems for the food manufacturing industry, improving query efficiency and accuracy. The paper establishes entity classification standards for the food manufacturing industry, constructs a dataset covering 11 entity categories, and validates the feasibility of the dataset in named entity recognition tasks. The effectiveness of four models, including

BERT-BiLSTM-CRF, is verified, with BERT-BiLSTM-CRF outperforming other models in recognition tasks, demonstrating high precision, recall, and F1 scores. Future research will focus on expanding domain datasets and optimizing NLP models to offer better research tools for the food manufacturing industry.

References

1. Zhou J, Jin Y, Liang Q, et al. Effects of regulatory policy mixes on traceability adoption in wholesale markets: Food safety inspection and information disclosure[J].Food Policy, 2022, 107:102218-.DOI:10.1016/j.foodpol.2022.102218.
2. Wu L, Xie P, Zhou J, et al. Robust Self-Augmentation for Named Entity Recognition with Meta Reweighting[J]. 2022.DOI:10.48550/arXiv.2204.11406.
3. Soltau H, Shafran I, Wang M, et al. RNN Transducers for Nested Named Entity Recognition with constraints on alignment for long sequences[J]. 2022. DOI: 10. 48550/ arXiv. 2203.03543.
4. Dooley D M, Griffiths E J, Gosal G S, et al. FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration[J]. Npj Science of Food, 2018, 2(1).DOI:10.1038/s41538-018-0032-6.
5. Qian L, Cui X. Research on the construction method of domain knowledge graph based on data augmentation [J]. Modern Intelligence, 2022, 42 (3): 9. DOI: 10. 3969/ j. issn. 1008-0821.2022.03.004.
6. Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.DOI:10.48550/arXiv.1810.04805.
7. Ding S, Fang Z, Wang N. Named Entity Recognition in the Business Domain Based on Bi-LSTM-CRF[J]. Modern Information, 2020, 40(3): 8. DOI: CNKI: SUN: XDQB. 0. 2020-03-010.
8. Jing X. Research on the sub-module NER of the nuclear safety question answering system based on deep learning [D]. North China University of Technology, 2021. DOI: 10. 26926/d.cnki.gbfgu.2021.000401.
9. Sharma R, Morwal S, Agarwal B. Named entity recognition using neural language model and CRF for Hindi language[J].Computer Speech & Language, 2022, 74: 101356-. DOI: 10.1016/j.csl.2022.101356.
10. An Y, Xia X, Chen X, et al. Chinese clinical named entity recognition via multi-head self-attention based BiLSTM-CRF[J].Artificial intelligence in medicine, 2022 (May): 127.DOI:10.1016/j.artmed.2022.102282.