ARTICLE

Syntax-Enhanced Entity Relation Extraction with Complex Knowledge

Mingwen Bi[1], Hefei Chen[2,*] and Zhenghong Yang[3,*]

[1]Business School, Beijing Wuzi University, Beijing, 101149, China

[2]National Engineering Research Center for Agri-product Quality Traceability, Beijing Technology and Business University, Beijing, 100048, China

[3]College of Information and Electrical Engineering, China Agricultural University, Beijing, 100083, China

*Corresponding Authors: Hefei Chen. Email: hff125125@163.com; Zhenghong Yang. Email: yangjohn@cau.edu.cn

ABSTRACT: Entity relation extraction, a fundamental and essential task in natural language processing (NLP), has garnered significant attention over an extended period., aiming to extract the core of semantic knowledge from unstructured text, i.e., entities and the relations between them. At present, the main dilemma of Chinese entity relation extraction research lies in nested entities, relation overlap, and lack of entity relation interaction. This dilemma is particularly prominent in complex knowledge extraction tasks with high-density knowledge, imprecise syntactic structure, and lack of semantic roles. To address these challenges, this paper presents an innovative "character-level" Chinese part-of-speech (CN-POS) tagging approach and incorporates part-of-speech (POS) information into the pre-trained model, aiming to improve its semantic understanding and syntactic information processing capabilities. Additionally, A relation reference filling mechanism (RF) is proposed to enhance the semantic interaction between relations and entities, utilize relations to guide entity modeling, improve the boundary prediction ability of entity models for nested entity phenomena, and increase the cascading accuracy of entity-relation triples. Meanwhile, the "Queue" sub-task connection strategy is adopted to alleviate triplet cascading errors caused by overlapping relations, and a Syntax-enhanced entity relation extraction model (SE-RE) is constructed. The model showed excellent performance on the self-constructed E-commerce Product Information dataset (EPI) in this article. The results demonstrate that integrating POS enhancement into the pre-trained encoding model significantly boosts the performance of entity relation extraction models compared to baseline methods. Specifically, the F1-score fluctuation in subtasks caused by error accumulation was reduced by 3.21%, while the F1-score for entity-relation triplet extraction improved by 1.91%.

KEYWORDS: Entity relation extraction; complex knowledge; syntax-enhanced; semantic interaction; pre-trained BERT

1 Introduction

In past years, the development of knowledge graphs tailored to specific vertical domains has demonstrated significant application potential and economic worth. Although companies can easily utilize tools to dump structured data directly into the knowledge base, most of the knowledge still exists in unstructured text. Improving the automatic extraction efficiency of entities and relations has emerged as the key to fixing the problem [1]. Entity relation extraction, the central goal in the domain of information extraction, tries to extract set types of entities and relations of unorganized text and construct a structured knowledge triplet (main entity, relation, and guest entity), which is of great significance for downstream applications such

as knowledge graph construction [2,3], intelligent question answering [4], and public opinion analysis [5]. Researchers typically categorize the problem into two main subtasks: named entity recognition [6–8] and relation extraction [9,10]. The two dominant research frameworks are the pipeline approach and the joint method.

As an early modeling method, the pipeline method trains two models to extract entities and classify relations, respectively. The entity model and the relation model can be trained using independent data sets, and it is not necessary to label them at the same time. According to the different execution orders of the model in the framework, it can be divided into the traditional entity relation model and the relation-entity model. However, because the two models are trained independently, and the post model is dependent upon the earlier model's prediction outcomes, this leads to the problems of error accumulation and lack of interaction in this method [11]. In addition, the traditional pipeline method also has some entity redundancy [12].

Considering the aforementioned issues with the pipeline method, the end-to-end joint method as a novel modeling method has received a lot of attention [13,14]. According to the number of framework stages, the joint method can be divided into multi-task learning (two-stage) and structured prediction (single-stage) [15].

Multi-task learning. Although this approach adopts an end-to-end training framework, it inherently involves cascading tasks, making it fundamentally similar to the pipeline method. Specifically, it still constructs separate models for entity recognition and relation extraction, but with a key distinction: both tasks share a unified encoder and leverage parameter sharing for joint optimization. For example, reference [16] employed Conditional Random Fields (CRF) to reframe entity recognition and relation extraction as a multi-head selection task, effectively addressing multi-relation challenges. In a related effort, reference [17] proposed the CASREL framework, which utilizes a cascading binary markup strategy to model relations as functions mapping subjects to objects within sentences, thereby resolving the issue of overlapping relation triples. Similarly, reference [18] introduced SPERT, a span-based joint entity and relation extraction model utilizing an attention mechanism. This model mitigates the problem of entity overlap by incorporating advanced entity recognition and filtering mechanisms. Further innovations include the work of [19], which approached joint entity and relation extraction as a direct set prediction problem. This strategy eliminates the complexity of predicting the sequence of multiple triples, streamlining the extraction process. To tackle entity overlap more effectively, reference [20] proposed an end-to-end learning model integrated with a replication mechanism, enabling the simultaneous extraction of related facts from sentences of varying complexity. Moreover, a novel reading comprehension-based paradigm has gained prominence, reframing entity and relation extraction as a multi-turn question-answering task. This method transforms the extraction process into one of identifying the answer's scope within the given context, offering a flexible and intuitive framework for handling intricate extraction scenarios [21]. However, similar to traditional pipeline methods, multi-task learning methods still have significant gaps in training and inference processes. This method has inherent leakage bias [22], and there is still an error accumulation problem in inference, thereby affecting the accuracy and reliability of the final results.

Structured prediction. Reference [23] introduced a relation-specific attention network (RSAN) that employs a relation-aware attention mechanism to generate tailored phrase representations for different relations. This model utilizes sequence labeling to identify and extract corresponding head and tail elements. Similarly, reference [24] proposed an innovative labeling scheme that reformulates the joint extraction task as a labeling problem, enabling the direct extraction of entities and their relations in a unified manner. Additionally, reference [22] presented a single-stage joint extraction model, TPLinker, which redefines the task as a token pair linking problem, effectively addressing the issue of exposure bias in the extraction process.

While the structured prediction method successfully mitigates issues such as exposure bias, entity overlap, and missing interactions, it does not necessarily imply that the joint method outperforms the pipeline method [15]. In fact, current methods for entity and relation extraction have some limitations, such as difficulty in handling nested entity and relation overlap, which pose challenges for models in extracting complex knowledge. In addition, when the encoding dimension of the data is very high, the model may encounter difficulties in training due to extremely sparse data. Consequently, the development of novel syntax-enhanced models is essential to effectively overcome these limitations and enhance the accuracy and robustness of entity relation extraction.

This paper adopts the pipeline extraction method of relation before entity, based on the Chinese pre-trained BERT [25] (i.e., Bidirectional Encoder Representation from Transformers) model called "BERT-wwm-ext" [26]. Two independently trained models are included in the pipeline extraction method, which are called entity model and relation model. The predicted value of the relation model is supplemented by the input of the entity model as an additional semantic feature and combined with the "Queue" input scheme proposed in this paper. The model effectively solves the problems of relation overlap and entity overlap that were neglected by previous research work. In addition, it benefits from the excellent performance of the pre-trained model BERT in the field of text classification (96%), we combine BERT with POS information to further enhance the performance of the relation model. On the basis of alleviating the exposure bias of the pipeline method as much as possible, it effectively avoids the limitation of the joint extraction model in processing complex samples.

In this paper, a comprehensive comparative experiment is carried out on the self-constructed data set EPI. The experimental results show that the performance fluctuation of subtasks caused by error accumulation decreased by 3.21%, while the entity relation triplet extraction performance increased by 1.91%, which proves the effectiveness of the model.

The main contributions of this paper are as follows:

(1) This paper proposes CN-POS and introduces the POS information represented by it in the pre-trained language model, which solves the problem that Chinese POS and character-level encoding cannot be aligned, and enhances the model's semantic understanding ability of syntactic information.

(2) This paper introduces the RF framework, which employs a "Queue" scheme to link the relation model and entity model. This approach directs the model's attention to specific parts of the input sentence corresponding to different relations, thereby strengthening the semantic interaction between relations and entities and mitigating the effects of knowledge overlap.

(3) This paper constructs an EPI dataset and proposes SE-RE, achieving optimal performance on the dataset.

The remainder of this paper is structured as follows: Section 2 presents the dataset and task description. Section 3 details the proposed model and algorithm. Section 4 outlines the experimental setup, followed by results and discussion in Section 5. Finally, Section 6 concludes the paper.

2 Dataset and Task

The starting point of this paper is to provide a high-performance model of automatic extraction of entity relation for the vertical domain knowledge graph based on the construction of the ontology layer [27]. However, most of the existing public datasets are oriented to the public domain and rarely involve professional fields. Therefore, they are mostly simple sentences with low semantic complexity, which are reflected in the following characteristics: (1) short sentences (2) each sentence contains only one or two knowledge triples (3) the semantic differences of the entities in sentences are obvious and easy to recognize

(e.g., person, location, organization, time, etc.), and knowledge extraction can be achieved through some excellent NLP tools.

A dataset called EPI describing e-commerce listings was constructed in this paper, which is a complex sample with high-density knowledge in the e-commerce field. The data for the EPI dataset is obtained from commonly used e-commerce platforms without additional processing or reduction, preserving the integrity of the original data. This characteristic ensures its applicability in a wide range of scenarios, accurately reflecting the features and complexities of real-world e-commerce data. We extracted information about 1653 products and obtained 5498 sample data through sentence segmentation. Each sample describes the information on related products. A total of 14 relation types and 12 entity categories are defined, which constitute 14 schemas. The goal is to extract structured knowledge triple RDF (Resource Description Framework). A specific example is shown in Table 1.

Table 1: A specific example in EPI

Sample	RDF
ZH: 华硕灵耀是笔记本电脑, 配备锐龙处理器, 适用于办公和娱乐。	**ZH:** (华硕灵耀, 类别, 笔记本电脑), (华硕灵耀, 品牌, 华硕), (华硕灵耀, 规格, 锐龙处理器), (华硕灵耀, 适用场景, 办公), (华硕灵耀, 适用场景, 娱乐)
EN: ASUS Zenbook is a laptop that is equipped with a Ryzen processor for both work and play	**EN:** (ASUS Zenbook, Category, laptop), (ASUS Zenbook, Brand, ASUS), (ASUS Zenbook, Specification, Ryzen processor), (ASUS Zenbook, Usage, work), (ASUS Zenbook, Usage, play)

It is worth noting that: (1) A product can have multiple aliases, be sold on multiple platforms, and be suitable for a variety of scenarios (2) Different styles and colors of a product will interfere with the price (3) Different models of a product have differences in function, material and price (4) The promotional price of the product in different time periods is different (5) Because Chinese places a strong focus on semantics and less on framework, there are a large number of reference omissions and irregular splicing vocabulary in the samples.

3 Proposed Model

Considering the limitations of existing entity relation extraction methods in effectively modeling the interactions between entities and relations, this paper proposes SE-RE, a novel approach designed to tackle the challenges of entity nesting and relation overlap in complex semantic texts. The overall structure of the model is illustrated in Fig. 1, while Fig. 2 provides a detailed case study. The model architecture comprises the following three main components: (1) To enhance the semantic modeling ability of each extraction model for Chinese syntactic information, this paper proposes CN-POS and introduces it into the Chinese pre-trained BERT model to compensate for its shortcomings in POS feature modeling. (2) To improve the interaction between entity recognition and relation extraction models, this paper proposes a "Queue" input scheme and RF to achieve syntactic and semantic enhancement based on different relations. (3) To alleviate the problems of relation overlap and entity redundancy, this paper employs a "Queue" input framework to feed the strengthened input sequences into the entity recognition model in order to obtain entity pairs cascaded by the current relation.
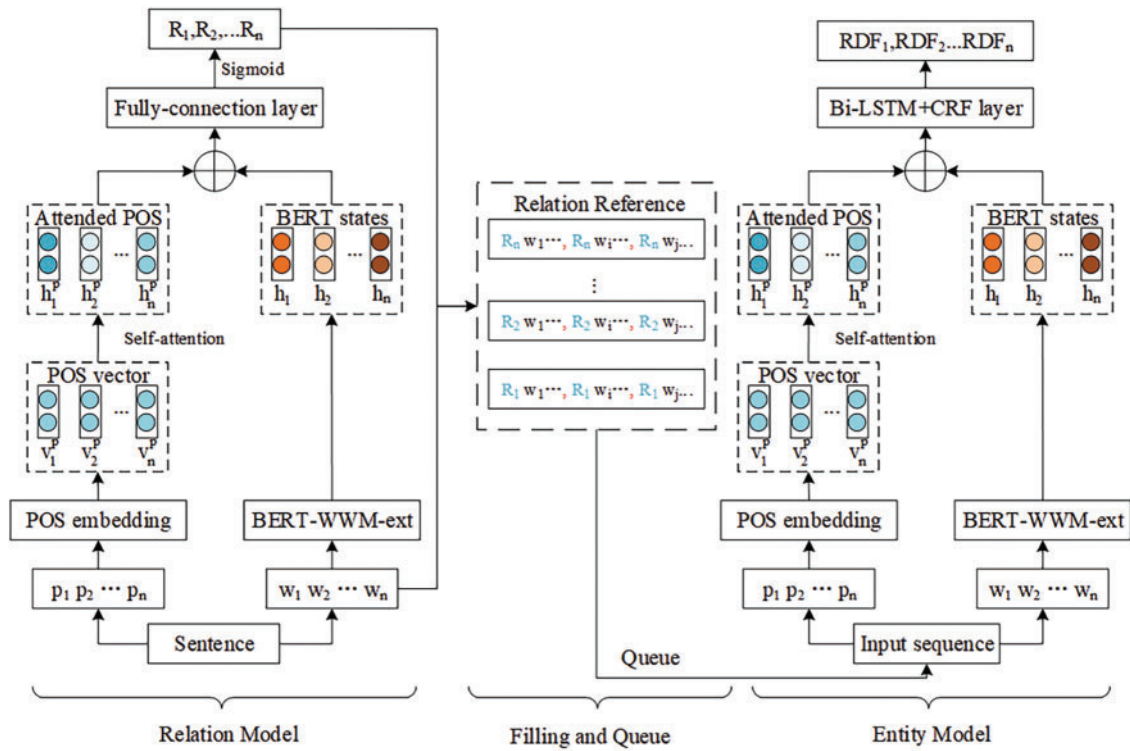
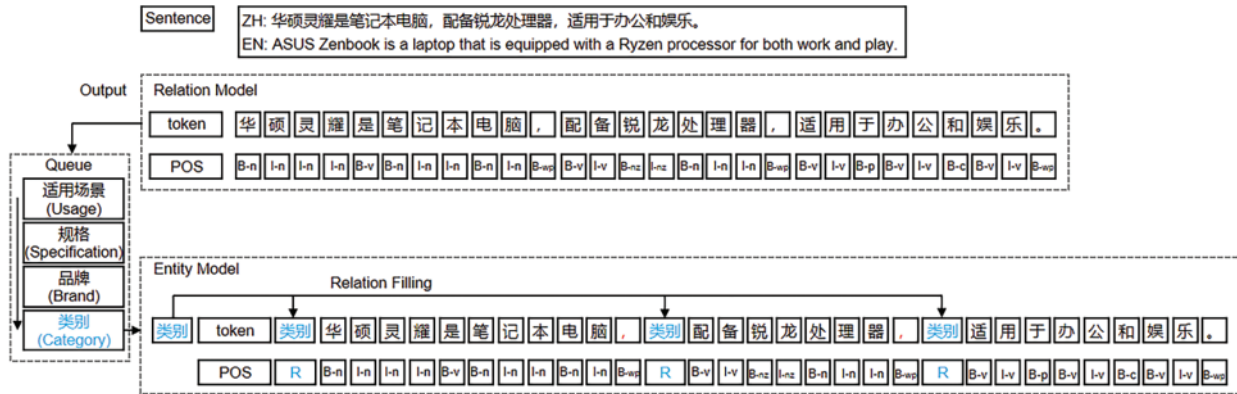**Figure 1:** The overall architecture of the SE-RE model

Figure 2: A specific case study

3.1 Problem Definition

Traditional entity relation extraction methods only extract entities and relations from the input sequence, making it difficult to achieve cascading matching between entities and relations. But in this paper, for a given input sequence $S = \{W_1, W_2, \ldots, W_n\}$, the SE-RE can extract all knowledge triples RDF that conform to the predefined schema, Equations such as Eq. (1).

$$RDF = \left\{ (subject, relation, object)_j \,\middle|\, j \in [1, m] \right\} (T_s, R, T_o) \in Schema, \tag{1}$$

where $(subject, relation, object)$ represents the knowledge triples extracted from the input sequence, m represents the total number of triples, $Schema$ is a predefined set of triplet categories, and T_s, R, and T_o respectively represents the categories of the main entity, relation, and guest entity set in the schema.

3.2 Input Representation

For the given text sequence $S = \{W_1, W_2, \ldots, W_n\}$, this paper uses the large-scale Chinese pre-trained BERT model to encode it into a vector S^{emb}. To meet the special input format of the BERT model, special identifiers $[CLS]$ and $[SEP]$ are first added at the beginning and end of S to form $\{[CLS], W_1, W_2, \ldots, W_n, [SEP]\}$. If there is context in the input sequence, that is, S covers two different contexts of text sequence $text_a + text_b$, we need to add the $[SEP]$ mark between the sentences to form $\{[CLS], text_a, [SEP], text_b, [SEP]\}$.

3.3 CN-POS Tagging Strategy

Due to BERT being a character-level encoding model, while Chinese POS information is a word-level feature. For instance, the Chinese "类别" corresponds to the English word "Category". As a token, the POS of "Category" is noun (n). Similarly, the POS of "类别" is also n. But as tokens, "类" and "别" obviously cannot be tagged in a traditional manner, because the characters do not express complete semantics. In addition, unlike the clear space boundaries between English words, Chinese needs additional word segmentation.

This paper proposes CN-POS to achieve alignment between Chinese POS information and token sequences. The specific process is shown in Fig. 3. For the input sequence $S = \{W_1, W_2, \ldots, W_n\}$, CN-POS first uses a hybrid NLP tool based on "Jieba Segmentor, custom dictionary and LTP(Language Technology Plantform) POS-tagger" to perform word segmentation and POS tagging, resulting in a sequence A of words and their corresponding POS tags $CP = \{C_i : P_i\}_{i=1}^m$. For each character $token_{C_i}^j$ in any word C_i, CN-POS assigns a POS label $P, P \in \{B - P_{C_i}, I - P_{C_i}\}$, which indicates the beginning and inside of the current POS, respectively. For example, for the noun "Category" with a part of speech of n, the alignment of its token sequence is $\{B - n, I - n\}$.
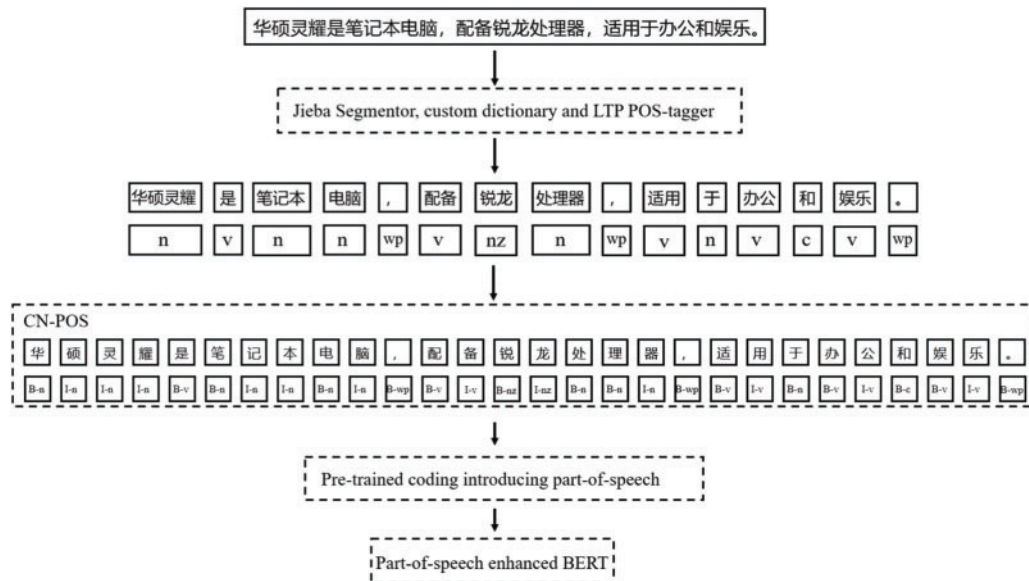

Figure 3: A BERT pre-training flowchart of CN-POS annotation and POS enhancement

3.4 POS Embedding

This paper references the POS embedding method in [28]. Specifically: we obtained the input POS tagging sequence $P = \{p_1, p_2, \ldots, p_n\}$. The POS embedding layer leverages the sparse vector representation P to generate a corresponding dense vector representation, enabling more efficient encoding of part-of-speech information V^P, equations such as Eq. (2).

$$V^P = \{V_i^p | i \in [1, n]\}, \tag{2}$$

where $V_i^p \in R^{h\_pos\_emb}$, and h_pos_emb denotes the size of the hidden layer in the POS embedding. Subsequently, the self-attention layer is employed to capture and analyze the entire sequence of POS tags, facilitating a comprehensive understanding of their contextual relationships and extract the grammatical dependencies in the input sentence h^P.

3.5 Relation Model

The relation extraction model aims to predict the relation categories contained in the input sequence. As shown in the "Relation Model" section of Fig. 1, the common "BERT+Bi-LSTM+FC" (Bi-directional Long Short-Term Memory, Fully Connected) network architecture is used for the relation extraction task. For each input sequence S, the input feature encoder is employed to obtain the character encoding S^{emb} and the POS encoding P^{emb}. Unlike the simple vector concatenation method used in previous research, this paper further maps the POS encoding P^{emb} to the $input\_ids$ sequence of the BERT model to alleviate its interference with the character encoding S^{emb}, and utilizes the fine-tuning of the BERT model during the training process to optimize it. Then, the concatenated vector of the two is further passed through a layer of Bi-LSTM to obtain the final hidden layer state vector representation $H = \{h_1, h_2, \ldots, h_n\}$ of the input sequence. Afterwards, it is mapped onto the target relation category set through a FC network, and its probability distribution logit is calculated using the sigmoid function. Finally, the predicted set of relation categories $\{R_1, R_2, \ldots, R_n\}$ is output by comparing it to a pre-set threshold.

In the training stage of the model, $BCEWithLogitsLoss$ is used as the LOSS function of the model training, which combines the $Sigmoid$ layer and $BCELoss$, It is widely used in classification tasks, and has excellent performance in processing imbalanced samples, equations such as Eqs. (3) and (4).

$$Loss_c = \{Loss_{1,c}, Loss_{2,c}, \ldots, Loss_{n,c}\}^T, \tag{3}$$

$$Loss_{n,c} = -W_{n,c}[P_c y_{n,c} \cdot \log(logit_{n,c}) + (1 - y_{n,c}) \cdot \log(1 - logit_{n,c})] \tag{4}$$

where c is the class number, n is the number of the sample in the batch and p_c is the weight of the positive answer for the class c. $y_{n,c}$ is the true value, $logit_{n,c}$ is the predicted value, and $W_{n,c}$ is the loss weight.

3.6 "Queue" Scheme and RF Mechanism

RF mechanism aims to alleviate the phenomenon of entity nesting and relation overlap. The essence of the entity relation extraction task is to strengthen the interaction between entities and relations by leveraging the strong correlations within triplets. This approach aims to enhance the performance of subtask models and improve the accuracy of cascading processes. Inspired by the semantic core of predicates in Chinese [29], this paper proposes the RF using a "Queue" input to model the interaction between entity and relation. Specifically, as shown in the "Filling and Queue" section of Fig. 1, for the input sequence S and the relation category set $R = \{R_1, R_2, \ldots, R_n\}$ output by the relation extraction model, each relation R_i is filled into the beginning of each clause in S (clauses are sub clauses separated by commas, semicolons, and periods), forming n sequences $SR = \{SR_1, SR_2, \ldots, SR_n\}$ with enhanced relation semantics, where $SR_i =$

$\{R_i W_1, \ldots, R_i W_j, \ldots, R_i W_k, \ldots\}$, which aims to use relations to guide the entity recognition model to focus on the corresponding entity categories. Therefore, based on relational semantics and entity category semantics, the impact of nested entities on model performance can be mitigated. Formally, the sequence representation formula for relational semantic enhancement is in Eq. (5).

$$SR = \{RF(S, R_i) | R_i \in R\}, \tag{5}$$

In addition, to address the issue of relation overlap in entity relation extraction tasks, this paper inputs the n semantically enhanced sequences constructed above into the subsequent entity recognition model one by one. Since each input sequence SR_i highlights the semantics of only a single relation, the entity recognition model can generate an independent semantic understanding of the same initial sequence S, thereby avoiding the impact of relation overlap on model performance while maintaining flexible execution of subtasks.

3.7 Entity Model

The entity model aims to identify entities related to specific relations by analyzing the contextual connections between various vocabulary and phrases. For the aforementioned relational semantic-enhanced text sequence SR_i with "Queue" input, the goal of the entity recognition model is to extract the entity pairs that have a semantic cascade relation with the relational R_i. As shown in the "Entity Model" section of Fig. 1, the network architecture of "BERT+Bi-LSTM+CRF" [30,31] is used to perform the entity recognition task, and the specific execution process of the model is as follows:

Input layer. For each input sequence $SR_i = \{R_i, W_1, \ldots, R_i, \ldots, W_n\}$, the complete annotation sequence is $y = \{y_1, y_2, \ldots, y_n\}$, $y_i \in \{B - type, I - type, O, X\}$, where $\{B - type, I - type, O\}$ represents the entity beginning, the entity interior and the entity exterior of the current W_j. In order to eliminate the influence of relation words on label sequence, X is assigned to R_i as a special entity label. Similarly, to avoid the influence of R_i on the POS tagging sequence and the inherent $input\_ids$ sequence in BERT, this paper performs the following relational word mapping operations.

Input_ids sequence: $input\_ids$ is the encoding mapping strategy used in the BERT model to convert tokens into corresponding ids in the vocabulary verbs. Its value range is $input\_ids \in [100, len(verbs)]$, where $len(verbs)$ represents the total number of words in the BERT vocabulary. To avoid inserting the relational word R_i in verbs or affecting the encoding mapping of the original input sequence, This paper does not perform character-level tokenization operations on R_i. Instead, it is treated as a complete word and assigned a special $input\_ids$ encoding mapping, with a value range of $input\_ids \in [1, n]$, and n being the total number of relation categories in the pre-set schema. It is worth noting that the above method is only applicable when $n < 100$.

POS tagging: For the token W_j in the initial text sequence S, its POS label adheres to the CN-POS proposed in this paper. While for the relational word R_i, it is given a special POS label in this paper.

Context embedding layer. Through operations performed at the input layer, the model eliminates the influence of the additional inserted relational words on each feature sequence. At the context encoding layer, the entity recognition model obtains the word vector encoding SR^{emb} and the POS encoding POS^{emb} corresponding to the input sequences through the input feature encoder. The two are concatenated to obtain the hidden layer vector representation, which is then used to derive the final hidden layer state $H = \{h_1, h_2, \ldots, h_n\}$ is obtained by Bi-LSTM.

Output layer. The entity determination for the named entity recognition task is based on the prediction results of the entity label sequence corresponding to the token sequence. To enhance the temporal dependency of entity label sequences and avoid model outputs with illogical label orders (e.g., $\{\ldots OI \ldots\}$), this

paper uses CRF to decode the final hidden layer state vector *H* output by the Bi-LSTM network. The training loss function is calculated as in Eq. (6).

$$LOSS = \sum_{i=1}^{n} l(CRF(h_i), y_i), \tag{6}$$

where $l(CRF(h_i), y_i)$ is the negative $log\_likelihood$ of the comparison between the predicted value $CRF(h_i)$ of the entity model and the true label y_i of the entity label.

4 Experiment

Dataset. This paper verifies the proposed SE-RE model on the self-constructed EPI data set, which is oriented to complex samples with high-density knowledge in the e-commerce domain. The dataset is divided into training, development, and testing sets using random sampling while ensuring that the sample distributions accurately represent the characteristics of the original dataset. This approach maintains consistency in data patterns across all subsets, enabling robust model training and evaluation (See Section 2 for a detailed introduction of the dataset). Table 2 presents the sample distribution for the relation model, illustrating the frequency and proportions of various relation types within the dataset. Similarly, Table 3 details the sample distribution for the entity model (i.e., the single relation sample obtained by the RF mechanism), showcasing the distribution of different entity categories. These tables provide a comprehensive overview of the dataset's composition for both models, ensuring transparency in the data preparation process. The overall sample distribution of each relation can be observed in Table 4.

Table 2: The sample distribution of the relation model and the average length of each sample

	Train	Dev	Test
Total number	3668	814	1016
Average length	29.47	29.10	29.10

Table 3: The sample distribution of the entity model and the total amount of knowledge triples

	Train	Dev	Test
Total number	9017	1914	2385
RDF num	12,277	2639	3273

Table 4: Sample distribution of each relation, corresponding entity label, and F1-score of relation classification

Relation	Total number	Entity_label	F1-score (%)
类别 (Category)	973	CAT	98.53
别名 (Nickname)	1056	PNN	99.73
功能 (Application)	1204	APP	97.87
发货地点 (Location)	778	LOC	99.63
品牌 (Brand)	1541	BRD	99.45
价格 (Price)	1475	PRC	99.63
售后 (After-sales Service)	60	AFS	93.31
销售平台 (Platform)	313	PLT	96.36

(Continued)

Table 4 (continued)

Relation	Total number	Entity_label	F1-score (%)
适用场景 (Usage)	174	APP	98.41
材料 (Material)	836	MAT	100
配件 (Accessory)	1108	ACC	98.58
规格 (Specification)	1553	SPEC	100
评价 (Review)	1155	REV	98.63
型号 (Model)	1097	MDL	99.75

Parameter setting. During the training phase, the model employs an Adam optimizer with a linear scheduler for parameter adjustment, incorporating a warmup ratio of 0.1. The training process spans 10 epochs with a dropout rate of 0.1 (keep ratio of 0.9) and a batch size of 32. The learning rate is set at $5e-5$, the maximum sequence length is limited to 128, and the hidden layer dimension of the Bi-LSTM network is 256. Notably, the output of the relation extraction model plays a pivotal role in determining the overall performance of the SE-RE model. Therefore, in order to improve the performance of the relation extraction model as much as possible, this paper incorporates an F1-score threshold into the training of each epoch. Finally, when the threshold is 0.45, the F1-score reaches the optimal value.

Baselines. This article conducts comparative experiments between the SE-RE model proposed for the entity relation extraction task and the most representative state-of-the-art models on the EPI dataset. The detailed introduction of the baseline models is as follows:

Multi-head [16]: As a standard method for joint entity relation extraction, Multi-head adopts a span-based extraction framework that can simultaneously perform entity recognition and relation extraction without any manual input or additional tools, effectively addressing the situation in which the same entity connects to multiple relations.

Pipeline [15]: The entity recognition stage is carried out before the relation extraction step in this traditional pipelined approach to entity-relation extraction, which splits the operation into two separate subtasks. Special entity boundary labels are then constructed from the output of the entity model to model the dependency relation between entity pairs.

SpERT.PL [32]: This method further enriches the context embedding of pre-trained converters by embedding POS tags. Specifically, The method assigns the POS of the entity to each token that constitutes the entity and incorporates the POS into the Transformer-based pre-trained language model. Additionally, the method introduces a large number of additional features, such as the span range feature within the entity, the overall span feature, the correlation feature between entities, and the correlation feature between entity categories and relations.

SE-RE(PL): To further demonstrate the effectiveness of the proposed CN-POS, this paper replaces the POS feature encoding strategy in the SE-RE model with the direct splicing method used in the SpERT.PL model.

SE-RE(MRC): To further demonstrate the effectiveness of the RF mechanism proposed in this paper, the relation syntax-enhanced strategy in the SE-RE model is replaced with the contextual sentence format used in the BERT-MRC model, namely "text_b: original text sequence".

SE-RE(-POS): To further validate the effectiveness of incorporating POS information into the entity relation extraction task and highlight the advancements achieved by the CN-POS enhanced pre-trained

encoding model (i.e., the improved version of the traditional BERT-wwm-ext model) proposed in this paper, detailed experiments and evaluations were conducted. We remove all POS information from the SE-RE model for comparative analysis.

Evaluation. In this paper, we use the standard evaluation metrics for the entity relation extraction task: Precision (P), Recall (R), and F1-score (F1) to evaluate the relation extraction model, the entity recognition model, and the final ternary cascade results. It is worth noting that some baseline models adopt a span-based modeling strategy. Therefore, in the entity recognition task, this paper adopts a strict evaluation principle, which means that the output entity span results need to be completely consistent with the true values in order to be considered accurate prediction results.

5 Results and Discussion

In this section, we analyze the experimental results from three key perspectives: the entity-relation extraction framework, the entity-relation model feature interaction strategy, and the ablation experiment.

5.1 Entity Relation Extraction Framework

Table 5 shows the experimental results of the SE-RE model proposed in this paper, compared with other baseline models on the EPI dataset. Among them, Pipeline, SE-RE, and its variants use independent encoders to execute entity and relational models, while Multi-head and SpERT.PL uses a unified encoder with shared parameters. In addition, Multi-head, Pipeline, and SpERT.PL all adopt the span-based entity before the relation extraction framework, while the SE-RE series model adopts the execution order of relation before the entity.

Table 5: Performance comparison of entity-relation extraction models on EPI

Model	Relation model			Entity model			RDF
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	F1 (%)
Multi-head [16]	79.16	83.88	81.45	85.47	90.05	87.70	82.15
Pipeline [15]	82.04	87.39	84.63	89.94	91.70	90.81	86.84
SpERT.PL [32]	84.88	88.68	86.74	91.39	92.76	92.07	88.41
SE-RE	88.49	92.73	90.56	91.79	93.59	92.68	90.32
SE-RE(PL)	87.97	92.27	90.07	90.54	93.36	91.93	89.33
SE-RE(MRC)	88.49	92.73	90.56	91.44	92.97	92.20	89.76
SE-RE(-POS)	86.42	91.80	89.03	90.26	92.47	91.35	88.70

From the model performance shown in Table 5, it can be observed that: (1) In all subtasks and the triplet extraction task as a whole, the SE-RE model put forward in this study performs the best. The F1-score of the most representative RDF improves by 1.91% compared to that of the SpERT.PL model (The F1-score is 90.32% and 88.41%, respectively), which proves the effectiveness of the proposed method. (2) Models based on the entity before the relation extraction framework (Multi-head, Pipeline, SpERT.PL) rely on the output results of the entity model, and incorrect entity predictions and redundant candidate entity pairs can cause serious error accumulation problems. The performance of the relation models in these frameworks is generally poor (The F1-scores of the relation models are 81.45%, 84.63%, and 86.74%, respectively), which in turn affects the final triplet cascade matching results. (3) The SE-RE proposed in this paper (90.56% for the relation model

F1-score) does not affect the label prediction performance of the entity model itself (the entity model F1-score reaches an optimal 92.68%), although the output results of the relation model are still used as part of the entity model's inputs, and it only affects the final triplet cascade results, which further proves that the relation-based semantic core of SE-RE model can effectively alleviate the relation overlap problem through the pipeline-based relation input method.

To further validate the stability and generalization ability of the proposed SE-RE model across different datasets and complex scenarios, we conducted additional comparative experiments on the publicly available FS-QA [33] dataset. FS-QA focuses on the Chinese food safety domain and serves as an authoritative benchmark for information extraction and coreference resolution tasks. As shown in Table 6, the experimental results demonstrate that the SE-RE model significantly outperforms existing baseline methods, achieving the best performance. Notably, the F1-score for the most representative RDF task improves by 1.12% compared to the SpERT.PL model. These findings further confirm the superiority of the SE-RE model in cross-domain applications, highlighting its robustness and strong generalization capability.

Table 6: Performance comparison of entity-relation extraction models on FS-QA

Model	Relation model			Entity model			RDF
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	F1 (%)
Multi-head [16]	81.42	86.26	83.77	85.98	90.41	88.14	82.58
Pipeline [15]	83.94	88.65	86.23	90.56	91.65	91.10	87.24
SpERT.PL [32]	86.43	90.86	88.59	92.07	92.85	92.46	90.02
SE-RE	89.23	93.62	91.37	92.25	93.93	93.08	91.14

In addition, the results in Tables 5 and 6 indicate that the SE-RE model achieves a significant improvement over the baseline methods. Statistical tests confirm the robustness of these results, with p-values < 0.01 in paired t-tests, and 95% confidence intervals consistently placing the SE-RE model's performance above the baseline.

5.2 Entity-Relation Model Feature Interaction Strategy

By comparing and analyzing the performance of each model in Table 5, it can be concluded that: (1) In the span-based entity before relation extraction framework, the SpERT.PL and Pipeline models outperform the Multi-head (F1-scores of RDF are 88.41%, 86.84%, and 82.15%, respectively), which is mainly due to the fact that the first two models consider the interactive dependencies between entities and model the internal features of entity spans, such as span width. Another reason for the better performance achieved by SpERT.PL is the further incorporation of the POS feature and the overall feature coding of the entity span. (2) The SE-RE model utilizes the CN-POS strategy to also take into account the POS features of the input sequences (F1-score of RDF is 90.32%) and further models the effects of relations on entities ignored by SpERT.PL through the RF mechanism, and utilizes the relations to guide the entity model to output the entity types oriented to a specific relation, which alleviates the entity redundancy problem facing the SpERT.PL model and to a certain extent avoids the interference of nested entities on the entity model. Therefore, in the entity recognition task, the SE-RE model (precision and recall of entity model are 91.79% and 93.59%, respectively) has higher recall while maintaining comparable precision compared to the SpERT.PL model (precision and recall of entity model are 91.39% and 92.76%, respectively). In addition, under the influence of the RF mechanism, the SE-RE model can greatly alleviate the problem of entity-relation cascade errors.

5.3 Ablation Experiment

The comparative analysis of the performance between the SE-RE model and its variants in Table 5 shows that: (1) The performance of the SE-RE model is higher than that of SE-RE(PL) (F1-scores of RDF is 90.32% and 89.33%, respectively), which indicates that the CN-POS strategy proposed in this paper is more effective than the direct splicing of POS features used in the SpERT.PL model. (2) Since SE-RE(-POS) does not consider POS features, it has the lowest performance (F1-score of RDF is 88.70%), which suggests that POS features can effectively contribute to the performance of entity relation extraction models. (3) Since SE-RE (MRC) only changes the form of relation-to-entity interactions, its relational modeling performance remains consistent with that of the SE-RE model (90.56% for both relational model F1-scores). However, with the entity-relation interaction model based on the MRC framework, additional relation features introduce noisy data. Moreover, the interaction distance between post-sequential entities and relations is too long, and cross-sentence dependency modeling reduces the ability of relations to guide the entity model, causing the model's performance to decline (F1-scores of the entity model are 92.20% and 92.68%, respectively, a decrease of 0.48%). (4) The SE-RE(MRC) outperforms SE-RE(PL) (F1-scores of RDF are 89.76% and 89.33%, respectively), which suggests that the CN-POS strategy is able to mitigate the effect of noise due to relation features to some extent.

5.4 Error Analysis

We also provide a failure case (as shown in Fig. 4). In this example, "Apple iPhone 13 Pro Max 512 GB 银色版本" is a complete and specific entity, while "iPhone 13 系列" is a generalized entity that encompasses multiple models. Although the SE-RE model, under the RF mechanism, can leverage the "类别" relation to model long-distance cross-sentence entity dependencies, the model may prioritize "iPhone 13系列" as the entity matching "旗舰手机." This prioritization overlooks the more precise boundary of "Apple iPhone 13 Pro Max 512 GB 银色版本," leading to errors in triplet extraction. This issue highlights the current limitations of the model in handling complex nested entities.

Figure 4: A failure case

To address this problem, future improvements should focus on enhancing the capacity of the model to discern entity boundaries at a finer granularity. Approaches such as multi-task learning or hierarchical representation methods could be adopted to balance relation guidance with precise entity semantic matching. Additionally, exploring the incorporation of an entity granularity control mechanism to prioritize entities at different granularities from a semantic perspective could further improve recognition accuracy.

6 Conclusion

This paper proposes SE-RE to address entity nesting and relation overlap in complex semantic texts. The aim is to extract structured entity relation triplets from text descriptions in the unstructured e-commerce product information domain. Taking inspiration from previous research that used POS information to enhance model performance, this paper constructs CN-POS and introduces it into the Chinese pre-trained BERT model, the problem of POS noise and insufficient feature training introduced by the traditional feature splicing approach is effectively alleviated. Inspired by the semantic core of predicates in Chinese, this paper adopts a model framework of relation before entity, effectively alleviating the problem of error accumulation. On this basis, the use of the "Queue" scheme and RF mechanism enhances the interaction between relation and entity, greatly improving the cascading errors caused by relation overlap. Additionally, this strategy can guide the entity model to obtain more accurate boundaries of nested entities.

Author Contributions: The authors confirm their contributions to the paper as follows: Study conception and design: Mingwen Bi, Zhenghong Yang; Data collection: Hefei Chen; Analysis and interpretation of results: Mingwen Bi, Hefei Chen; Draft manuscript preparation: Mingwen Bi, Hefei Chen. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, Hefei Chen, upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Takanobu R, Zhang T, Liu J, Huang M. A hierarchical framework for relation extraction with reinforcement learning. AAAI Conf Artif Intell. 2019;33(1):7072–9. doi:10.1609/aaai.v33i01.33017072.
2. Mao Z, Chuan T, Zhou J, Su WW, He YJ, Mei DH, et al. Knowledge graph-based approach to unstructured multi-source data fusion. In: Advances in artificial intelligence big data and algorithms. IOS Press; 2023. p. 663–9.
3. Chen X, Li L, Qiao S, Zhang N, Tan C, Jiang Y, et al. One model for all domains: collaborative domain-prefix tuning for cross-domain NER. arXiv:2301.10410. 2023.
4. Ye R, Ge R, Fengting Y, Chai J, Wang Y, Chen S. Leveraging unstructured text data for federated instruction tuning of large language models. arXiv:2409.07136. 2024.
5. Nemes L, Kiss A. Information extraction and named entity recognition supported social media sentiment analysis during the COVID-19 pandemic. Appl Sci. 2021;11(22):11017. doi:10.3390/app112211017.
6. Cui X, Song C, Li D, Qu X, Long J, Yang Y, et al. RoBGP: a Chinese nested biomedical named entity recognition model based on roberta and global pointer. Comput Mater Contin. 2024;78(3):3603–18. doi:10.32604/cmc.2024.047321.
7. Dash A, Darshana S, Yadav DK, Gupta V. A clinical named entity recognition model using pretrained word embedding and deep neural networks. Decis Anal J. 2024;10:100426. doi:10.1016/j.dajour.2024.100426.
8. Xiang Y, Zhao X, Guo J, Shi Z, Chen E, Zhang X. A U-shaped network-based grid tagging model for Chinese named entity recognition. Comput Mater Contin. 2024;79(3):4149–67. doi:10.32604/cmc.2024.050229.

9.   Wei C, Li J, Wang Z, Wan S, Guo M. Graph convolutional networks embedding textual structure information for relation extraction. Comput Mater Contin. 2024;79(2):3299–314. doi:10.32604/cmc.2024.047811.

10.   Wang X, El-Gohary N. Deep learning-based relation extraction and knowledge graph-based representation of construction safety requirements. Autom Constr. 2023;147:104696. doi:10.1016/j.autcon.2022.104696.

11.   Li Q, Ji H. Incremental joint extraction of entity mentions and relations. Proc ACL. 2014;1:402–12. doi:10.3115/v1/P14-1.

12.   Chan YS, Roth D. Exploiting syntactico-semantic structures for relation extraction. Proc ACL. 2011;1:551–60. doi:10.5555/2002472.2002542.

13.   Dai D, Xiao X, Lyu Y, Dou S, She Q, Wang H. Joint extraction of entities and overlapping relations using position-attentive sequence labeling. AAAI Conf Artif Intell. 2019;33:6300–8. doi:10.1609/aaai.v33i01.3301630.

14.   Tan Z, Zhao X, Wang W, Xiao W. Jointly extracting multiple triplets with multilayer translation constraints. AAAI Conf Artif Intell. 2019;33(01):7080–7. doi:10.1609/aaai.v33i01.33017080.

15.   Zhong Z, Chen D. A frustratingly easy approach for joint entity and relation extraction. arXiv:2010.12812. 2020.

16.   Bekoulis G, Deleu J, Demeester T, Develder C. Joint entity recognition and relation extraction as a multi-head selection problem. Expert Syst Appl. 2018;114:34–45. doi:10.1016/j.eswa.2018.07.032.

17.   Wei Z, Su J, Wang Y, Tian Y, Chang Y. A novel cascade binary tagging framework for relational triple extraction. arXiv:1909.03227. 2019.

18.   Eberts M, Ulges A. Span-based joint entity and relation extraction with transformer pre-training. In: ECAI 2020. IOS Press; 2020. p. 2006–13.

19.   Sui D, Zeng X, Chen Y, Liu K, Zhao J. Joint entity and relation extraction with set prediction networks. IEEE Transacti Neural Networks Learn Systems. 2024;35(9):12784–95. doi:10.1109/TNNLS.2023.3264735.

20.   Zeng X, Zeng D, He S, Liu K, Zhao J. Extracting relational facts by an end-to-end neural model with copy mechanism. In: Proceedings of the Anterior Cruciate Ligament; 2018; Melbourne, VIC, Australia. p. 506–14. doi:10.18653/v1/P18-1047.

21.   Li X, Yin F, Sun Z, Li X, Yuan A, Chai D Entity relation extraction as multi-turn question answering. arXiv:1905.05529. 2019.

22.   Wang Y, Yu B, Zhang Y, Liu T, Zhu H, Sun L. TPLinker: single-stage joint extraction of entities and relations through token pair linking. arXiv:2010.13415. 2020.

23.   Yuan Y, Zhou X, Pan S, Zhu Q, Song Z, Guo L. A relation-specific attention network for joint entity and relation extraction. In: International Joint Conference on Artificial Intelligence; 2021.

24.   Zheng S, Wang F, Bao H, Hao Y, Zhou P, Xu B. Joint extraction of entities and relations based on a novel tagging scheme. arXiv:1706.05075. 2017.

25.   Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. 2018.

26.   Cui Y, Che W, Liu T, Qin B, Yang Z. Pre-training with whole word masking for chinese BERT. IEEE/ACM Trans Audio Speech Lang Process. 2021;29:3504–14. doi:10.1109/TASLP.2021.3124365.

27.   Tang X, Feng Z, Xiao Y, Wang M, Ye T, Zhou Y, et al. Construction and application of an ontology-based domain-specific knowledge graph for petroleum exploration and development. Geosci Front. 2023;14(5):101426. doi:10.1016/j.gsf.2022.101426.

28.   Phan MH, Ogunbona PO. Modelling context and syntactical features for aspect-based sentiment analysis. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020. p. 3211–20. doi:10.18653/v1/2020.acl-main.293.

29.   Wang Z, Jiang T, Chang B, Sui Z. Chinese semantic role labeling with bidirectional recurrent neural networks. In: Proceedings of the Conference Empirical Methods in Natural Language Processing (EMNLP); 2015; Lisbon, Portugal. p. 1626–31. doi:10.18653/v1/D15-1186.

30.   Catelli R, Casola V, Pietro GDe, Fujita H, Esposito M. Combining contextualized word representation and sub-document level analysis through Bi-LSTM+CRF architecture for clinical de-identification. Knowl-Based Syst. 2021;213:106649. doi:10.1016/j.knosys.2020.106649.

31. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. arXiv:1603.01360. 2016.

32. Santosh TYSS, Chakraborty P, Dutta S, Sanyal DK, Das PP. Joint entity and relation extraction from scientific documents: role of linguistic information and entity types. In: 2nd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE 2021) Co-Located with JCDL 2021; 2021; USA.

33. Bi M, Liu X, Zhang Q, Yang Z. Machine reading comprehension combined with semantic dependency for Chinese zero pronoun resolution. Artif Intell Rev. 2023;56(8):7597–612. doi:10.1007/s10462-022-10364-5.