Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

CEREM: A segment-wise attention network for chinese highly aggregated semantic extraction

Bin Liu [a], Jiaqi Han [a], Zhenyu Zhang [a], Shijun Li [a], Haixi Zhang [a], Yijie Chen [a,*], Keqin Li [b]

[a] *College of Information Engineering, Northwest A&F University, Yangling, 712100, Shaanxi, China*
[b] *Department of Computer Science, State University of New York, New Paltz, 12561, New York, USA*

ARTICLE INFO

Keywords:
Chinese semantics
Information extraction
Segment information attention
Independent branches

ABSTRACT

The demand of large models for data has revitalized information extraction research, particularly for Chinese texts, where semantic isolation poses unique challenges. Existing methods often rely on Chinese word segmentation, but their capacity to capture full semantic meaning is constrained by polysemy, flexible word order, and other unique characteristics of the Chinese language. To address this limitation, we propose three-level semantic division and design CEREM, a prompt- and pointer-based IE network, to extract highly aggregated semantics. In our design, prompts unify multiple IE tasks while preserving semantic interactions, a Segment Information Attention mechanism implicitly aggregates the high-level semantics to enhance Chinese understanding, and an Independent Branches strategy decouples parameters to focus separately on the sub-tasks of start and end index prediction. We evaluate CEREM on four datasets–DiaKG, CMedCausal, Title2Event, and the self-constructed CAIT–covering named entity recognition (NER), relation extraction (RE), and event extraction tasks. CEREM achieves state-of-the-art performance: on CAIT, 88.59% F1 for NER and 71.82% for RE; on DiaKG, 81.77% for NER and 65.44% for RE; and for causal relation extraction on CMedCausal, 45.30% F1. These results demonstrate CEREM's effectiveness across domains and task types, highlighting its potential as a unified framework for Chinese information extraction.

1. Introduction

The popularity of large models has led to significant data demands, particularly the Chinese texts (Hui et al., 2024). Chinese information extraction tasks can extract essential knowledge from chaotic text data, effectively addressing the demand for high-quality Chinese texts data in large models (Zhang et al., 2019). However, unlike English texts, where words with complete semantics are naturally separated by spaces, the semantics of Chinese texts are characterized by their isolation, word segmentation, flexible word order, polysemy, and omission phenomena. These unique characteristics of the Chinese language bring huge challenges for Chinese information extraction tasks.

Chinese word segmentation task extracts key fragments from Chinese texts to solve the problem of semantic difficulty caused by a lack of segmentation (Li et al., 2023; Lin et al., 2023). Researchers have conducted many works on Chinese word segmentation to adopt

[*] Corresponding author.
E-mail addresses: liubin0929@nwsuaf.edu.cn (B. Liu), hanjiaqi@nwafu.edu.cn (J. Han), zhangzy0828@nwafu.edu.cn (Z. Zhang), lishijun@nwafu.edu.cn (S. Li), zh.haixi@nwafu.edu.cn (H. Zhang), yijiechen@nwafu.edu.cn (Y. Chen), lik@newpaltz.edu (K. Li).

https://doi.org/10.1016/j.ipm.2026.104617
Received 15 August 2025; Received in revised form 6 January 2026; Accepted 6 January 2026

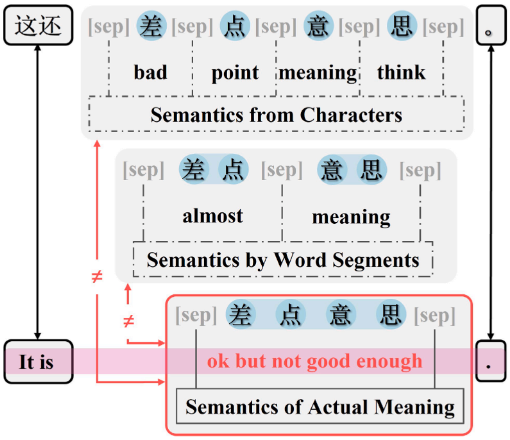

这还
[sep] 差 [sep] 点 [sep] 意 [sep] 思 [sep]
bad
point
meaning
think
Semantics from Characters
。
[sep] 差 点 [sep] 意 思 [sep]
almost
meaning
Semantics by Word Segments
≠
≠
[sep] 差 点 意 思 [sep]
It is
ok but not good enough
.
Semantics of Actual Meaning

Fig. 1. Understanding Chinese texts from three-level perspectives, including the semantics obtained from characters, word, and segment (high-level semantics with actual meaning, which corresponds to the English text in pink bar).

the Chinese pre-trained model such as T5 (Raffel et al., 2020), BERT-Base-Chinese (Devlin et al., 2019), and Chinese-RoBERTa-wwm-ext (Liu et al., 2019), which achieves great performance improvements in information extraction task. However, word segmentation is not sufficient to fully cope with the unique characteristics of Chinese Texts. Some related studies (Lee et al., 2024; Ponce et al., 2024) have shown that the segmentation task for Chinese language usually divides the text according to the most basic vocabulary, which still poses great difficulties for models to understand semantics. Analyzing the reasons, word segmentation typically generates lexical units that represent basic semantic elements, and these units often require further contextual integration within phrases with high-level structures to be fully comprehended. Using only lexical units that represent basic semantic elements still cannot avoid semantic isolation. The Semantic Isolation Index (SII)[1] is introduced to quantitatively characterize semantic isolation. It is a segmentation-based metric that measures the divergence between token-level compositional semantics and holistic entity meaning. Empirical results across multiple datasets consistently show that this divergence increases with entity length, indicating that longer entities exhibit stronger semantic isolation and are more likely to be interpreted as integrated semantic units rather than combinations of their parts.

To describe the complete semantics, as illustrated in Fig. 1, three levels of semantics are presented. The first level is the low-level semantics in the original Chinese characters, which are fragmented and isolated, making it difficult for the model to fully grasp the meaning of the Chinese text. The second level is the semantics of Chinese words derived from word segmentation which has represented a significant improvement, but there is still a gap from the actual meaning. The third level is the actual meaning of Chinese phrases, which are endowed by long-term social and cultural influences. It has already permeated into various fields, referred to as highly aggregated semantics. Understanding Chinese semantics only from the perspectives of the first and second levels is difficult to obtain the actual meaning of the text. The highly aggregated semantics is quite common in Chinese texts, with varying degrees of severity. Therefore, it is important for the model to understand Chinese texts from the perspective of highly aggregated semantics.

Although highly aggregated semantics have significant importance in Chinese texts, existing research on information extraction mainly focuses on the design of model input form and the design of task head with pre-trained language models (PLM) (Liu et al., 2019; Raffel et al., 2020), Some of researches insert question templates (Du & Ji, 2022; Li et al., 2020a; Silva et al., 2022), prompt information (Li et al., 2020b; Lu et al., 2022), or solid and levitated markers (Ye et al., 2022; Zhong & Chen, 2021) into the text as model input, which adopts additional information to promote the information extraction. Other researches develops a range of extraction techniques to represent the boundaries and categories of extracted information and use different structures to extract specific information, such as sequence labeling (Huang et al., 2015; Yu et al., 2020; Zheng et al., 2017), span-based methods (Jiang et al., 2020; Wang et al., 2021; Ye et al., 2022), token pair methods (Yan et al., 2023), word to NER (Li et al., 2022a), and generation-based methods (Hsu et al., 2022; Yan et al., 2021; Zeng et al., 2018). Due to the lack of emphasis on highly aggregated semantics, the difficulty of extracting Chinese information still needs to be further addressed.

In addition, for information extraction, existing researches employ parameter sharing across different sub-tasks to enhance semantic interactions among these sub-tasks (Wang et al., 2020; Wei et al., 2020). While parameter sharing often leads to parameter confusion due to the divergent objectives of different sub-tasks. To avoid parameter confusion, some methods utilize independent models for different sub-tasks (Ye et al., 2022; Zhong & Chen, 2021), which reduces interactions within semantics and brings in a lack of semantic richness. Both parameter confusion and insufficient semantic richness have a negative impact on the semantic quality, especially for Chinese information extraction tasks (Chen et al., 2025).

[1] A formal definition and cross-dataset analysis are provided in the Appendix.

In order to further explore Chinese semantic extraction and improve the performance of Chinese information extraction, this paper proposes a Chinese entity and relation extraction model (CEREM) for information extraction. Firstly, an architecture based on prompt and pointer networks for information extraction is proposed, which unifies multiple information extraction tasks to one task. Then, Segment Information Attention (SIA) is presented to enhance the understanding of Chinese texts. Finally, the Independent Branches (IB) strategy is conducted to make parameters focus on the semantics from every sub-task split from the single task for information extraction. The main contributions are summarized as follows:

- A novel three level semantic of Chinese Texts is presented and an information extraction network CEREM is proposed to extract different levels of Chinese semantics. CEREM is designed based on prompt and pointer networks to unify multiple information extraction tasks to one task.
- A novel attention structure termed segment information attention is presented to supplement highly aggregated semantics for the generation of the word embeddings.
- An independent branches structure is conducted to make parameters focus on the semantics from every sub-task split from the single task for entity and relation extraction. The structure alleviates the problem of ineffective utilization of Chinese semantics by resolving parameter confusion.

Extensive experiments are conducted on named entity recognition, relation extraction, and causal relationship extraction, using two publicly available Chinese texts datasets and one self-constructed Chinese texts dataset. The experimental results indicate that CEREM has achieved a signicantly performance outperforms the state-of-the-art models. Moreover, SIA can be inserted into various information extraction models to improve the quality of word embeddings.

2. Research objectives

In order to promote the development of information extraction research in Chinese text, this paper studies the different levels of semantics hidden in Chinese text, especially highly aggregated semantics. For that, this paper proposes a Chinese entity and relation extraction model CEREM for Chinese information extraction, details as follows: (1) To unify multiple information extraction tasks to one task and achieve effective information extraction, the proposed information extraction network is built on prompt and pointer networks. (2) To supplement highly aggregated semantics for the generation of the word embeddings, a novel attention structure termed segment information attention is presented. (3) To alleviate the problem of ineffective utilization of Chinese semantics by resolving parameter confusion, making parameters focus on the semantics from every sub-task, an independent branches structure is conducted. (4) To validate the impact of advanced aggregation semantics on the performance of information extraction tasks, extensive visualization experiments are conducted.

3. Related work

In recent years, information extraction has garnered significant attention in various research topics, involving named entity recognition, relation extraction, joint extraction, event extraction, and so on. According to the way of extracting information, existing methods can be roughly classified into three categories.

Pipeline-based Information Extraction Method employed distinct models to execute various sub-tasks of information extraction, with each model functioning independently. Yang et al. (2019) separately extracted event trigger and arguments to address overlapping arguments. Zhong and Chen (2021) proposed a pipeline model for named entity recognition and relation extraction by text slicing and label insertion methods. Based on this study, Ye et al. (2022) improved the strategy of inserting labels, further enhancing the performance of the model. Wang et al. (2023) proposed a concise approach using the fused features for the relation extraction task. The advantage of this method lies in the independence of the sub-tasks, which minimizes parameter confusion. Yan et al. (2025) introduced a modular pipeline, DocExtractNet, for receipt information extraction based on distinct sub-task modules. The advantage of this method lies in the independence of the sub-tasks, which minimizes parameter confusion. However, these approaches presented a problem: the lack of semantic information exchange between sub-tasks, resulting in low utilization of semantic information.

Joint-based Information Extraction Method utilizes a single integrated model with distinct components designed for various sub-tasks. Wei et al. (2020) proposed a joint information extraction method that completed named entity recognition and relation extraction tasks end-to-end. Wang et al. (2020) also conducted research in this direction, completing end-to-end information extraction by designing matching strategies between tokens and entities. Jia et al. (2023) designed binary factors and ternary factors to directly model interactions between not only a pair of instances but also triplets. Su et al. (2023) proposed a three-stage joint extraction model, which can tackle overlapping problems. Gui and Cui (2023) proposed a joint entity-relation extraction method AJE based on a dot-product attention mechanism. Shang et al. (2022a) cast joint extraction as a fine-grained triple classification problem and proposed a joint extraction model. Luo and Yu (2024) introduced ESGNet, a multimodal joint model incorporating entity semantic graphs, which captures latent semantic information from both text and image modalities to enhance extraction accuracy from Chinese resumes. Bölücü et al. (2024) introduced a weakly supervised learning framework with noise-robust training, which effectively improves the performance of joint extraction tasks under noisy annotation conditions. These methods allow the model to fully leverage the interdependencies among different sub-tasks during the learning process by sharing parameters. However, it may lead to parameter confusion.

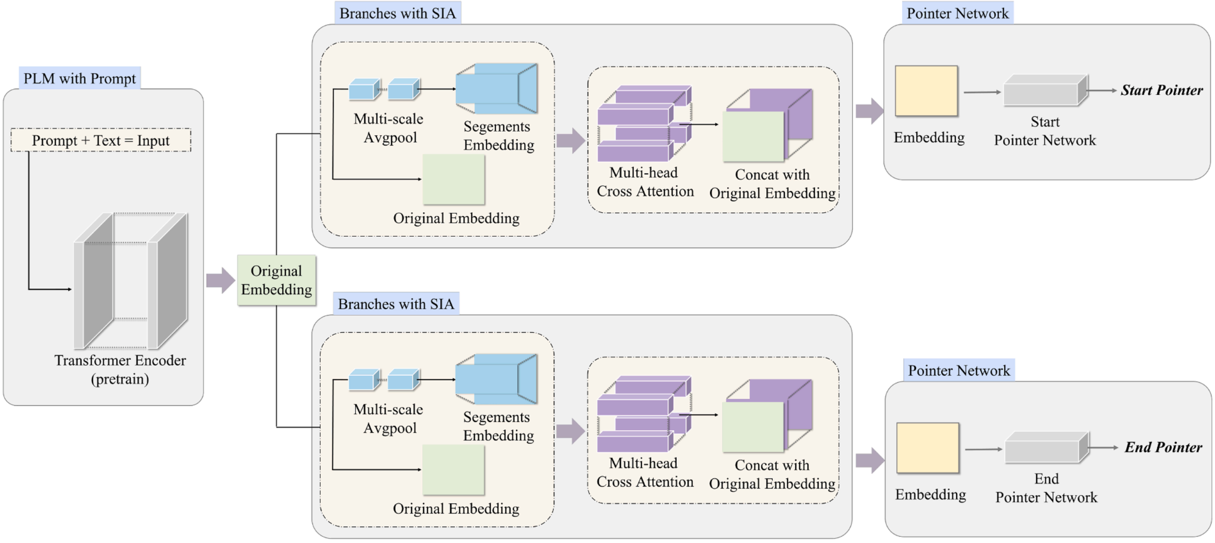


Fig. 2. The overview of the proposed CEREM. The "PLM with Prompt" provides information extraction targets and generates original word embeddings. The "Branches with SIA" achieves Chinese semantic enhancement and parameter decoupling. The "Pointer Network" indicates the start indices and the end indices of the extracted target entities.

Universal Information Extraction Method regards various information extraction sub-tasks as a unified task, thereby enabling completion using a singular structure. Li et al. (2020a) transformed event extraction into a multi-round question and answer task, abstracting events uniformly and completing event extraction tasks end-to-end. Lu et al. (2021) proposed an end-to-end universal event extraction model through generating paradigms of sequence to structure. Afterward, Lu et al. (2022) conducted further research and proposed a unified information extraction model based on prompt for various information extraction tasks. Yan et al. (2023) unified the information extraction task into a token pair classification task, proposing a unified token pair task head. Ping et al. (2023) proposed a new paradigm for universal information extraction that is compatible with any schema format and applicable to various IE tasks. In addition, He et al. (2025) proposed a hierarchical generation and multi-evidence alignment fusion model for multimodal entity and relation extraction, which further extends the universal information extraction paradigm to the multimodal domain by integrating hierarchical semantic generation and multi-source information fusion. The above methods effectively utilize the interdependencies among different sub-tasks and prevent parameter confusion to some extent. However, it requires the support of word embeddings with rich semantics.

4. Methodology

In this section, this paper first designs the unified architecture of information extraction, which achieves effective information extraction by leveraging semantic interaction among various sub-tasks. Subsequently, this paper presents Segment Information Attention to enhance the understanding of Chinese texts by highly aggregated semantics. Finally, this paper conducts a strategy called Independent Branches to avoid parameter confusion by adopting a fine-grained strategy for parameter decoupling. The overall architecture is depicted in Fig. 2.

4.1. The information extraction architecture

Firstly, this section illustrates the unified information extraction architecture, which adopts prompt and pointer networks to unify different tasks. To effectively extract information from Chinese texts and unify various information extraction tasks, the architecture is built on prompt techniques and pointer networks, which maintains semantic interaction and reduces parameter confusion. Inspired by (Lu et al., 2022), this paper enables the model to maintain semantic richness and unifies sub-tasks by a prompt component.

Specifically, the proposed model accepts a prompt and a text sequence as input and then generates embeddings using the encoder from a Chinese PLM by:

$$W = PLM(s \oplus x) \tag{1}$$

where s represents a prompt that controls what to spot, what to associate, and what to extract. x represents a text sequence. W represents embeddings of the input generated by a Chinese PLM. $\oplus$ is a splicing format, and the specific structure of input data is as follows:

$$s \oplus x = [[cls], p_1, p_2, \cdots, p_m, [sep], x_1, x_2, \cdots, x_n] \tag{2}$$

where $[cls]$ and $[sep]$ represent special tokens, used as separators in the input of the model. $p_1, p_2, \cdots, p_m$ is a sequence of prompt. $x_1, x_2, \cdots, x_n$ is a sequence of text.

Using prompt and pointer networks, information extraction tasks can be accomplished through the single task of target entity extraction. Given a sentence that contains several entities, there are relationships between these entities. For entity extraction, the prompt specifies an entity type, while for relation extraction, it combines an entity and a relation type. As the experiment in this paper includes entity extraction, relationship extraction, and causal relationship extraction, an example to describe how to use the proposed architecture to complete these extraction tasks is as follows.

For the sentence: "Steve Jobs was born in America in1999.″

- **Entity extraction**: The prompt is an entity category (For example, the prompt is "person", where the "person" is an entity category).
- **Relation extraction**: The prompt is a combination of an entity and a relation category (For example, the prompt is "Steve Jobs' birthplace", where the "Steve Jobs" is an entity and the "birthplace" is a relation category).
- **Causal relationship extraction**: The goal of this task is to extract causal triplets, including the extraction of head entities, tail entities, and causal relationships. Therefore, it is similar to entity and relationship extraction.

The proposed model extracts target entities (entities related to the prompt) each time. For completing information extraction, the model first uses each entity category as a prompt in turn to extract all entities. Then, all extracted entities are combined with all relation categories in turn to form prompts for extracting relations or arguments.

During the above process, Chinese PLM is adopted to generate embeddings. After that, the extraction of the target entities is completed using pointer networks, consisting of a pair of head and tail pointers. The head pointer and the tail pointer indicate the start indices and the end indices of the extracted target entities, respectively. Both the head and tail pointers are one-dimensional tensors, each with a length of the maximum sequence length.

In this architecture, sub-tasks are unified as a target entity extraction task, so that the parameters of the architecture can maintain semantic information from these different sub-tasks and interactions among these sub-tasks during training. Since there is only one unified task, the parameter confusion is significantly reduced.

4.2. Segment information attention

This section presents Segment Information Attention. The encoder used for generating embeddings typically relies on the computation of "attention mechanisms". By adaptively learning the relationships between tokens, relevant features are incorporated into each token to enrich the embeddings' global information. Existing attention mechanisms often compute the correlation from the perspectives of character and word segmentation, which are different from the actual semantics. Therefore, the introduction of highly aggregated semantics for enhancing the model's understanding of Chinese texts is crucial.

In fact, in the field of computer vision, Vision Transformer (ViT) (Dosovitskiy et al., 2021) has already proposed applications of this idea. Instead of using a pixel-based perspective, ViT divides an image into multiple patches, allowing attention mechanisms to be computed from complete local features. The difference between Chinese texts and images is that, even though the semantics of characters and words in texts may be isolated, they still play a crucial role in understanding the overall meaning of Chinese texts. Therefore, it is important to maintain the perspectives of character and word segmentation while also incorporating highly aggregated semantics.

In existing pre-trained encoders, the information of tokens uses the semantics of character and word segmentation. Performing attention calculations between tokens and segments, attaching the correlation features of tokens and segments to the tokens, and then integrating this with the pre-trained encoder constitutes a highly suitable structure. In detail, PLM generates word embeddings with the semantics of characters and words. Then, these embeddings are first pooled through multi-scale averaging for generating segment embeddings, which represent different segments in the text. Within a Chinese text, there are many phrases with highly aggregated semantics that vary in position and length. The multi-scale structure iteratively generates segment embeddings of different lengths, enabling coverage of these phrases. During training process, the proposed model adaptively selects the segments that require attention and disregards those that can be neglected, thereby enabling segment embeddings to effectively represent highly aggregated semantics of phrases. The average pooling is implemented by:

$$S_r(i,:) = \frac{1}{r}\sum_{m=0}^{r-1} W(i \cdot s + m, :) = t_{[i,(r+i)]} \tag{3}$$

where $W = \left[t_1, t_2, \cdots, t_L\right]^\top \in \mathbb{R}^{L\times d}$ is the original word embeddings, L is the length of sequence and d is the dimension. r is the kernel size that controls the window size for pooling, which is set to $2, 3 \ldots n$. s is the stride, which is set to 1. $S_r = \left[t_{[0,r]}, \quad t_{[1,(r+1)]}, \quad \cdots, \quad t_{[(L-r),L]}\right]^\top \in \mathbb{R}^{(L-r+1)\times d}$ is segment embeddings. Each token in S_r represents a text segment of length r, while each token in $W \in \mathbb{R}^{L\times d}$ represents only one character. Compared to the $W \in \mathbb{R}^{L\times d}$, the size differs in the sequence dimension but remains the same in the feature dimension.

In the attention mechanism, the correlation between tokens is calculated through matrix multiplication. For traditional self attention mechanisms, the correlation is calculated by $AS = W \cdot W^\top$, where $AS \in \mathbb{R}^{L\times L}$ is the attention score. Given a semantically complete text segment in a sentence that the index starts with c and ends with e, its correlation should be highlighted in the AS as:

$$AS = \begin{bmatrix} \cdots, & \cdots, & \cdots, & \cdots, & \cdots \\ \cdots, & \boldsymbol{t_{c,c}}, & \boldsymbol{\cdots}, & \boldsymbol{t_{c,e}}, & \cdots \\ \vdots, & \boldsymbol{\vdots}, & \boldsymbol{\ddots}, & \boldsymbol{\vdots}, & \vdots \\ \cdots, & \boldsymbol{t_{e,c}}, & \boldsymbol{\cdots}, & \boldsymbol{t_{e,e}}, & \cdots \\ \cdots, & \cdots, & \cdots, & \cdots, & \cdots \end{bmatrix}. \tag{4}$$

However, due to the scattered and isolated semantic meaning of Chinese characters, it is difficult to establish $(e-c)^2$ relationships completely and accurately. To address this issue, SIA is proposed. The correlation is calculated by $AS'_r = W \cdot S_r^{\top}$, where $AS'_r \in \mathbb{R}^{L\times(L-r+1)}$ can be expanded as:

$$AS'_r = \begin{bmatrix} t_{0,[0,r]}, & t_{0,[1,(r+1)]}, & \cdots, & t_{0,[(L-r),L]} \\ t_{1,[0,r]}, & t_{1,[1,(r+1)]}, & \cdots, & t_{1,[(L-r),L]} \\ \vdots, & \vdots, & \ddots, & \vdots \\ t_{(L-1),[0,r]}, & t_{(L-1),[1,(r+1)]}, & \cdots, & t_{(L-1),[(L-r),L]} \end{bmatrix} \tag{5}$$

where $t_{a,[c,e]}$ represents the correlation between character a and segment $[c,e]$. Similarly, for the above given example, the highlighted cases in AS are as follows:

$$AS'_r = \begin{bmatrix} \cdots, & \cdots, & \cdots \\ \cdots, & \boldsymbol{t_{c,[c,e]}}, & \cdots \\ \vdots, & \boldsymbol{\vdots}, & \vdots \\ \cdots, & \boldsymbol{t_{e,[c,e]}}, & \cdots \\ \cdots, & \cdots, & \cdots \end{bmatrix}. \tag{6}$$

On the one hand, correlation mining is conducted between characters and segments allows the full utilization of the highly aggregated semantics within segments, which is more likely to trigger a correlation response, making it highlighted and thereby facilitates the acquisition of more comprehensive and accurate correlations. On the other hand, compared to the self-attention mechanism, the number of relationships extracted by this method decreases exponentially (specifically, $e-c$, and the number of self-attention is $(e-c)^2$), making the computational process more focused and preventing the loss of correct correlations.

Furthermore, in order to reduce the negative impact of redundancy and useless correlations, multi-head and multi-scale structures were designed to complete the attention calculation process. In brief, the multi-head process is obtained through:

$$SW_r = Multi(Softmax\left(\frac{AS'_r}{\sqrt{d_k}}\right) \times S_r) \tag{7}$$

where SW_r is the new word embeddings that have semantics of characters, words, and phrases of length r. Due to $\frac{AS'_r}{\sqrt{d_k}} \in \mathbb{R}^{L\times(L-r+1)}$ and $S_r \in \mathbb{R}^{(L-r+1)\times d}$, and the $Softmax(\cdot)$ and $Multi(\cdot)$ do not affect the shape of the matrix, so $SW_r \in \mathbb{R}^{L\times d}$ remains consistent with the original word embeddings W in shape.

And the multi-scale process is obtained through:

$$W_{stack} = Stack(W, [SW_2, \cdots, SW_n], dim = -1) \tag{8}$$

After calculating by the $Stack(\cdot)$, $W, SW_2, \cdots, SW_n$ will be integrated into $W_{stack} \in \mathbb{R}^{L\times d\times n}$. The final embeddings with complete semantic can be extracted by a Linear layer through $FSW = W_{stack} \times T_{linear}$. The $T_{linear} \in \mathbb{R}^{n\times 1}$, so $FSW \in \mathbb{R}^{L\times d}$ remains consistent with the original word embeddings W in shape.

Multi-head structure enables the model to obtain correlations from different perspectives and improve the weight of correct correlations during the calculation process. Multi-scale structure also has this effect. and it can adapt to text segments of different scales to support complete correlation acquisition. The final embeddings SW possesses richer semantic information. While, it retains the same shape as the original word embeddings W, which enables it to be well ported to other jobs that require PLM to improve performance.

4.3. Independent branches

To alleviate the issue of parameter confusion brought by parameter sharing, this section conducts Independent Branches to decouple the parameters. DNN's parameters can be divided into those in shallow layers and those in deep layers. The parameters in shallow layers are close to the input and used to extract universal characteristics, while deep layer parameters are close to the output and are used to extract specialized characteristics. For example, CNN extracts universal features such as edges and textures by shallow layers, and further extracts specialized features by deep layers to perform classification or regression tasks (Zeiler et al., 2014).

Therefore, the SIA structure behind the encoder becomes the deep layers for extracting specialized features. While the encoder becomes the shallow layers for extracting universal features, as shown in Fig. 3. Considering that shallow layers exhibit minimal parameter confusion while deep layers are prone to confusion, the architecture employs a strategy of using two independent SIA branches to decouple deep layer parameters. These two branches are dedicated to extracting start and end indices, respectively.

Unlike other methods that adopt shared parameters to solve different sub-tasks, this paper treats start index extraction and end index extraction as different sub-tasks and decouples parameters based on them. This allows deep parameters to serve only a single

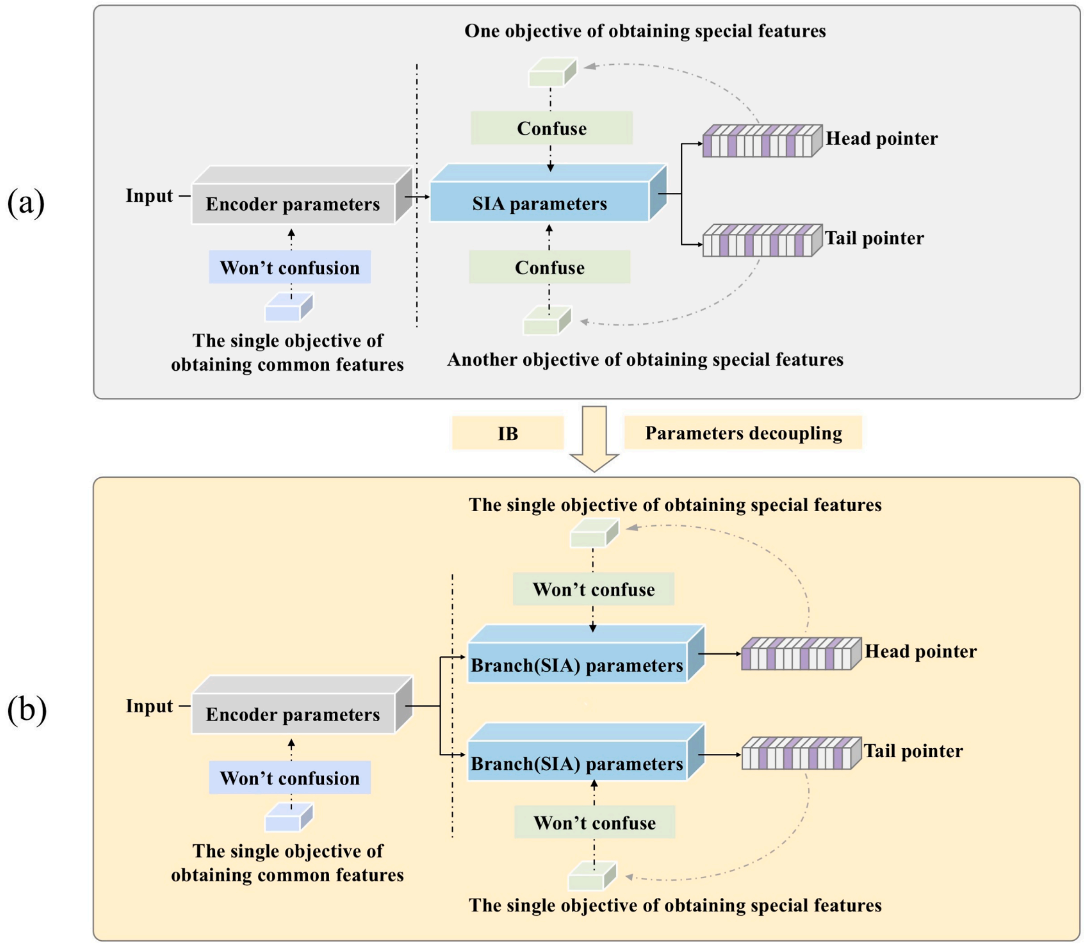

(a)
One objective of obtaining special features
Confuse
Head pointer
Input
Encoder parameters
SIA parameters
Tail pointer
Won't confusion
Confuse
The single objective of obtaining common features
Another objective of obtaining special features
IB
Parameters decoupling
(b)
The single objective of obtaining special features
Won't confuse
Branch(SIA) parameters
Head pointer
Input
Encoder parameters
Branch(SIA) parameters
Tail pointer
Won't confusion
Won't confuse
The single objective of obtaining common features
The single objective of obtaining special features

Fig. 3. The process of parameter decoupling. The unified information extraction task is divided into starting index extraction and ending index extraction tasks. It consists of two independent SIA branches.

task, with individual task objectives, enabling the parameters to focus on completing that task without being distracted. In detail, SIA is used to generate embeddings with complete semantics. For IB, two SIA structures are used in parallel, with original word embeddings as input, to generate two embeddings with complete semantics, which are applied to different sub-tasks respectively. This process can be obtained through:

$$\begin{cases} IB_{start} = FSW_{start} \\ IB_{end} = FSW_{end} \end{cases} \tag{9}$$

where FSW_{start} and FSW_{end} represent two independent SIA architectures used to complete different sub-tasks. Then the generated IB_{start} and IB_{end} are fed into the start pointer network and end pointer network to generate start and end pointers, which represents start and end indices, respectively. This process can be implemented by:

$$PN(IB) = IB \times P_{linear} \tag{10}$$

$$\begin{cases} Pointer_{start} = \mathbb{I}(PN_{start}(IB_{start}) \geq \tau) \\ Pointer_{end} = \mathbb{I}(PN_{end}(IB_{end}) \geq \tau) \end{cases} \tag{11}$$

where PN reduces the feature dimension to 1 via $P_{linear} \in \mathbb{R}^{d\times 1}$, and $\mathbb{I}$ means to set values greater than τ in the tensor to 1, otherwise set them to 0. $Pointer_{start} \in \mathbb{R}^{L\times 1}$ and $Pointer_{start} \in \mathbb{R}^{L\times 1}$ are the pointers used to represent start and end indices of entities. The number of elements with a value of 1 in these two is consistent. Take an example in the Fig. 4.

Compared to the parameter decoupling strategy of pipeline methods, the proposed parameter decoupling avoids parameter confusion to a greater extent. In addition, this paper unifies different tasks using prompts, so that the parameters retain the rich semantics of different extraction tasks. In this way, parameter confusion is avoided in a great measure without losing the rich semantics in texts.

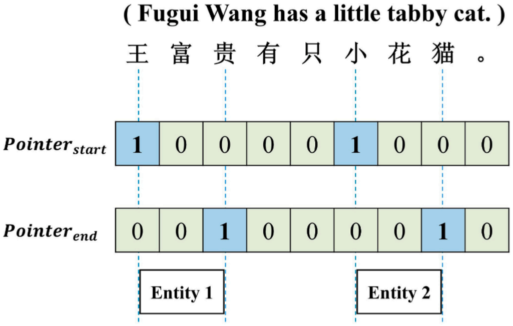


Fig. 4. Example of a practical application of the pointer network.

4.4. The dataflow of CEREM

In sum, the data flow of CEREM framework proceeds from prompt construction to final span decoding in a stage-wise manner. The concatenated input $s \oplus x$ (Eq. (1)) is encoded by the pre-trained language model to produce contextualized token embeddings $W = \mathrm{PLM}(s \oplus x)$, where $W \in \mathbb{R}^{L\times d}$ denotes the sequence length by embedding dimension. These token-level embeddings retain alignment with both the prompt tokens and the target text tokens and serve as the shared representation for all subsequent operations.

To enrich the semantic representation of compositional or long expressions, W is then processed by the Segment Information Attention module. SIA aggregates multi-scale, phrase-level features and fuses them back with the original token embeddings, producing an enhanced representation SW that keeps the same shape as W but encodes additional segment-aware context. Importantly, SIA is designed to be plug-and-play with the PLM output. It does not change embedding dimensionality or token alignment, which ensures compatibility with downstream heads.

After semantic enhancement, CEREM uses the Independent Branches strategy to decouple boundary predictions. As shown in Fig. 3, two parallel branches (one specialized for *start* positions and one for *end* positions) independently process the SIA-enhanced embeddings and produce task-specific hidden states IB_{start} and IB_{end}. Each branch is projected by a lightweight pointer head that scores tokens and, after thresholding, yields binary start/end pointers. Extracted spans are decoded from matched start–end pairs. By repeating forward passes with different prompts (entity-category prompts for NER, entity + relation prompts for RE, and causal prompts for causal triplet extraction), the same encoder and SIA/IB pipeline unify multiple subtasks while preserving task-specific precision through branch-level parameter decoupling.

5. Experiments

This article conducts experiments on multiple information extraction tasks using datasets from different fields to fully validate the effectiveness of the proposed method. To validate the effectiveness of the methods proposed in this paper for Chinese information extraction, this paper has selected several widely studied tasks in information extraction, including entity and relation extraction, and causal relationship extraction.

5.1. Datasets

The publicly available dataset Diakg (Chang et al., 2021) and the self-built dataset CAIT are chosen as the experimental datasets for entity and relation extraction. Diakg is a widely used Chinese medical text dataset specifically designed for diabetes-related research. It contains a comprehensive collection of clinical knowledge, treatment guidelines, and research findings on diabetes. The Diakg dataset is derived from 41 diabetes guidelines and consensus, which are from authoritative Chinese journals, including basic research, clinical research, drug usage, clinical cases, diagnosis and treatment methods, etc. This dataset covers the most extensive research content and hot areas in recent years, containing a total of 22,050 entities and 6890 relationships. CAIT is a Chinese information extraction dataset in the agricultural domain, primarily focusing on entity recognition and relation extraction related to crop diseases and pests. The dataset contains approximately 400 original sentences, which are expanded to 15,200 samples through prompt construction and negative sample generation. It covers 13 types of entity categories and 14 types of relation categories. On average, each sentence contains 7.75 entities and 7.00 relations, indicating a high level of information density and semantic complexity. With its strong domain specificity and clear structure, the CAIT dataset is well-suited for training and evaluating models in complex scenarios involving multiple entities and relations. The reason for choosing to use the CAIT dataset is that, on the one hand, compared with the existing public datasets, it has a higher density of entity relationships and a greater difficulty in extraction. On the other hand, this dataset is relatively small in scale and is more suitable for the rapid validation of model performance. Other details about the CAIT dataset will be described in the appendix. For causal relationship extraction, the publicly available dataset CMedCausal (Li et al., 2022b) is chosen as the experimental dataset. CMedCausal is a widely used Chinese medical texts dataset proposed by Alibaba, which contains rich knowledge of medical causality. The dataset covers various causal relations such as drug reactions, disease progressions, and treatment outcomes, serving as a valuable resource for training and evaluating models in this task.

Table 1
The key relevant experimental environment.

Indicator	Values
Operating System	Ubuntu 20.04.3 LTS
CPU	Intel Xeon Gold 5318Y @ 2.10GHz
GPU	NVIDIA A40
Number of GPUs	2
PyTorch	1.10.0
Transformers	4.22.1
Scikit-learn	1.1.2

It consists of 800 labelled training samples, 200 validation samples, 1000 test samples, and an additional 1000 unlabelled samples. CMedCausal comprises richly annotated clinical notes, with the longest text spanning 544 characters. For the event extraction task, the Title2Event(Deng et al., 2022) dataset is chosen, which is a large-scale Chinese dataset designed for open event extraction. It consists of over 40,000 real-world news titles collected from multiple domains, each annotated with one or more event triplets in the form of (subject, predicate, object). Different from traditional event datasets with fixed event types and predefined argument roles, Title2Event follows an open schema and focuses on extracting diverse predicates and entities directly from titles. Because Chinese news titles are usually short, condensed, and may contain multiple or incomplete events, this dataset presents unique challenges such as argument omission, predicate diversity, and event overlap. All annotations were manually verified through multiple rounds of quality control to ensure high reliability and coverage across different topics.

5.2. Implementation details

In the experiment, some of the related works with publicly available code and the method proposed in this paper are conducted in the same environment, where the CPU is an Intel Xeon Gold 5318Y CPU@2.10GHz, and the GPU is an NVIDIA A40. The operating system is Ubuntu 20.04.3 LTS. All models are evaluated using the Chinese PLM "Chinese-RoBERTa-wwm-ext" published by the Harbin Institute of Technology on Hugging Face, serving as the encoder. The experimental data for other related works are sourced from their respective papers. The details of the training and testing environment for the proposed model are shown in the Table 1.

As the method proposed in this paper relies on prompts, the model's ability to adapt to prompts needs to be strength. This paper applies the strategy of the generation of negative samples to better adapt to prompts and enhance the generalization of the model. In detail, many additional data are added to the dataset, where there prompts are unrelated to text or incorrect, and lable_list is an empty list. The mixing of a large number of negative samples helps improve the performance of information extraction tasks during the training process of the model. Especially, when testing and applying the model in practice, in order to extract complete information from the text, a comprehensive but redundant prompt set needs to be provided. The strategy of the generation of negative samples can effectively address this scenario.

During the experimental process, parameter freezing and thawing strategies are set in this paper. Considering that the parameters in the pre-trained model have been trained on a large amount of Chinese data, their parameters already contain rich and accurate knowledge. However, parameters that have not been trained in the model can lead to significant loss in the early stages of training. Adjusting the pre-trained model's parameters based on this loss could negatively impact the pre-trained model. Therefore, in this paper, we froze the pre-trained model during the initial training phase. The model was only unfrozen for overall training once its F1 score on the test set exceeded the threshold (default value is set to 0.2).

During the model performance evaluation phase. In the comparative experiments, Diakg, CAIT, and CMedCausal datasets are used for validation. The evaluation metrics for the relation extraction and causal relationship extraction tasks are the F1 score of triplet extraction, while the evaluation metric for the named entity recognition task is the F1 score of entity extraction. In the ablation experiments, the same three datasets are used for validation, and the evaluation metric employed is the F1 score of the unified information extraction task. In the pluggability experiments, the self-constructed dataset CAIT is used for validation. The evaluation metric for the relation extraction task is the F1 score of triplet extraction, and for the named entity recognition task, the evaluation metric is the F1 score of entity extraction.

In order to verify the effectiveness of the proposed methods in this paper, we have selected influential models in the field of information extraction in recent years.

5.3. Comparison for entity and relation extraction

The datasets Diakg and CAIT are used to validate performance on entity and relation extraction tasks. All models involved in the comparison are classified into three categories: Pipeline-based IE, Joint-based IE, and Universal IE. The experimental results are shown in the Table 2. It indicates that our method universally outperforms several previous studies on the two datasets.

Compared with advanced methods in Pipeline-based IE, the model proposed in this paper unifies multiple sub-tasks, allowing the embeddings to maintain the semantic interaction of information. The proposed model also uses the SIA structure to further enrich the embeddings, which promotes the model's understanding of semantics and has better generalization ability. Additionally, the parameter decoupling strategy conducted in this paper is more detailed than the Pipeline strategy, enabling the parameters to focus more precisely on specific problems without confusion. Compared with advanced methods in Joint-based IE, the model proposed

Table 2
Results of the comparative experiments, where NER represents the named entity recognition task, and RE represents the relation extraction task.

Models	Type	CAIT		Diakg	
		NER	RE	NER	RE
CasRel (Wei et al., 2020)	Joint-based	-	60.29%	-	56.90%
TPLinker (Wang et al., 2020)	Joint-based	-	67.87%	-	59.89%
PLMarker (Ye et al., 2022)	Pipeline-based	86.35%	68.53%	80.53%	63.49%
UIE (Lu et al., 2022)	Universal	87.66%	65.48%	81.69%	64.97%
UTC-IE (Yan et al., 2023)	Universal	85.48%	71.81%	80.38%	63.58%
PP-UIE-0.5b (PaddleNLP, 2025)	Universal	81.99%	56.58%	79.45%	54.23%
PP-UIE-1.5b (PaddleNLP, 2025)	Universal	79.89%	59.38%	63.68%	**67.72%**
PP-UIE-7b (PaddleNLP, 2025)	Universal	83.77%	68.71%	81.07%	60.70%
PP-UIE-14b (PaddleNLP, 2025)	Universal	87.50%	**72.76%**	**83.49%**	65.55%
CEREM	Universal	**88.59%**	71.82%	81.77%	65.44%

Table 3
Results of the comparative experiments on causal relationship extraction task.

Models	Precision	Recall	F1
TSBN (Jiang & Zhao, 2022)	45.92%	41.64%	43.23%
DRP (Liang et al., 2022)	-	-	42.58%
OneRel (Shang et al., 2022b)	46.80%	37.10%	41.40%
PRGC (Zheng et al., 2021)	41.10%	20.70%	27.50%
UIE (Lu et al., 2022)	38.24%	40.53%	39.35%
PP-UIE-0.5b (PaddleNLP, 2025)	13.92%	36.50%	20.16%
PP-UIE-1.5b (PaddleNLP, 2025)	28.99%	42.86%	34.59%
PP-UIE-7b (PaddleNLP, 2025)	32.92%	**52.64%**	40.51%
PP-UIE-14b (PaddleNLP, 2025)	40.12%	43.80%	41.88%
CEREM	**48.69%**	42.35%	**45.30%**

in this paper effectively avoids parameter confusion through parameter decoupling, which is a common issue in Joint methods. Meanwhile, by using the SIA structure, the model enhances its ability to understand Chinese texts by incorporating the perspective of highly aggregated semantics. Compared with advanced methods in Universal IE, the proposed model achieves the best overall performance, obtaining the highest NER score of 88.59% on CAIT and competitive results on Diakg. Although the large-scale PP-UIE-14B model achieves comparable performance on some RE metrics, CEREM still exhibits superior generalization and semantic understanding with a more compact architecture. These results further confirm the effectiveness and robustness of the proposed approach.

5.4. Comparison for casual relationship extraction

The dataset CMedCausal is used to validate the performance on causal relationship extraction tasks. The relevant work for comparison are TSBN, DRP, OneRel, PRGC, UIE, and PP-UIE. The experimental results are shown in the Table 3. It indicates that our method universally outperforms several previous studies on the CMedCausal datasets.

For the causal relationship extraction tasks, the proposed CEREM model consistently achieves the highest precision and F1-score among all compared methods, demonstrating its strong overall effectiveness and stability. Although PP-UIE-7B attains the highest recall, CEREM still outperforms all PP-UIE variants in terms of precision and comprehensive performance, reflecting its better balance between accuracy and coverage. Compared with previous studies, the proposed model unifies multiple sub-tasks, enabling the embeddings to preserve semantic interaction and contextual consistency. In addition, the introduced Segment Information Attention structure further enriches the semantic representations, enhancing the model's understanding of complex causal semantics and improving generalization capability. Furthermore, the parameter decoupling strategy allows the parameters to focus more precisely on task-specific representations, reducing interference among sub-tasks and further improving model robustness.

5.5. Comparison for event extraction

The Title2Event dataset is used to validate the performance on event extraction tasks, including trigger extraction, argument extraction, and event triplet extraction. The model is trained with a batch size of 128, for 100 epochs, using a multi-scale kernel size of 5. Comparative methods include EventGLM-gwn (EGLM), ST-Seq2SeqMRC (ST-Seq2Seq), ST-SpanMRC (ST-Span), SeqTag, Unsuper, and UIE, and the results are summarized in Table 4.

For the event extraction tasks, CEREM demonstrates robust and consistent performance across trigger extraction, argument extraction, and event triplet extraction. It achieves a precision of 80.2 and an F1-score of 74.8 in trigger extraction, confirming its strong ability to accurately identify event-triggering expressions. In argument extraction, the model attains precision, recall, and F1-scores

Table 4
Results of the comparative experiments on event extraction task. P, R, F1 stand for precision, recall, and F1-score, respectively.

	Trigger Ex.			Argument Ex.			Triplet Ex.		
Methods	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
EGLM	70.4	**70.7**	70.5	58.5	58.3	58.4	50	50.2	50.2
ST-Seq2Seq	-	-	-	57.9	58.6	58.2	49.8	50.1	49.9
ST-Span	-	-	-	60.1	54.9	57.4	44.5	44.8	44.7
SeqTag	69.5	69.9	69.7	50.8	51.2	51	41.1	41.3	41.2
Unsuper	21	32	25.4	12	15.5	13.5	4.5	6.8	5.4
UIE	**80.9**	67.5	73.6	**67.5**	62.9	65.1	**57.8**	51.2	54.3
CEREM	80.2	70.1	**74.8**	65.3	**67.3**	**66.3**	56.3	**54.1**	**55.7**

Table 5
The pluggability experiments for the proposed SIA module, which significantly improves the performance of existing methods.

Models	SIA	NER	REL
CasRel (Wei et al., 2020)	-	-	60.29%
	✓	-	**62.84%**
TPLinker (Wang et al., 2020)	-	-	67.87%
	✓	-	**71.85%**
UIE (Lu et al., 2022)	-	87.66%	65.48%
	✓	**89.66%**	**71.86%**
UTC-IE (Yan et al., 2023)	-	85.48%	71.81%
	✓	**87.09%**	**73.29%**

of 65.3, 67.3, and 66.3, respectively, revealing its effectiveness in capturing event-related entities and their semantic roles. Moreover, CEREM obtains an F1-score of 55.7 for event triplet extraction, illustrating its robustness in integrating multiple event components into coherent structural representations.

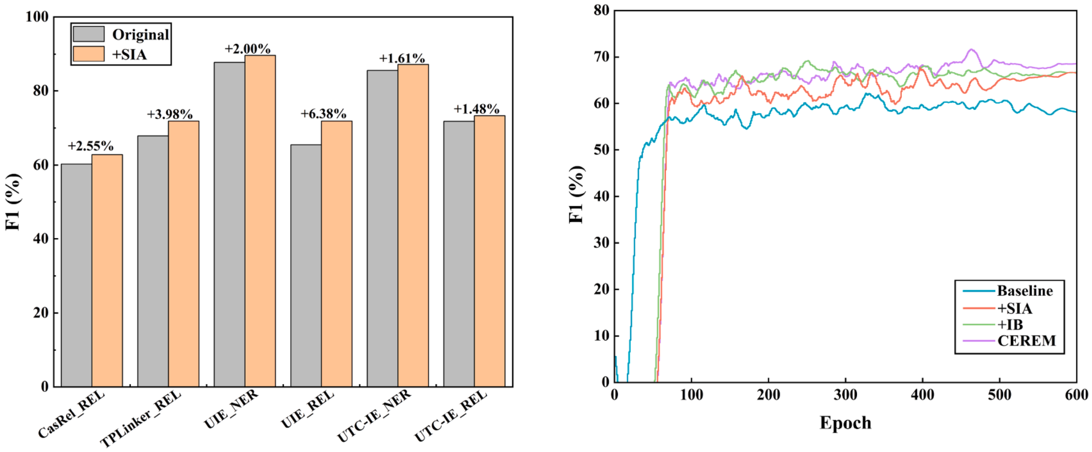
Distinct from traditional models that process event elements separately, CEREM models the entire event structure in a holistic and interaction-aware manner. This design allows the contextual embeddings to dynamically capture dependencies among triggers, arguments, and their roles within a unified semantic space. The SIA mechanism facilitates fine-grained interaction across segments, effectively distinguishing overlapping or nested event patterns. In parallel, the IB design contributes to stable optimization by selectively refining event-related representations without cross-task interference. Through this synergy of global semantic modeling and modular parameter learning, CEREM achieves a deeper comprehension of complex event semantics and delivers superior robustness and generalization in diverse event extraction scenarios. The above results demonstrate that CEREM is not confined to entity and causal relation extraction but can also be effectively extended to more complex tasks such as event extraction. Given its unified architecture and semantic modeling capability, CEREM exhibits strong potential for cross-domain generalization and can be readily adapted to various application fields with minimal structural modification, highlighting its scalability.

5.6. *Ablation studies*

5.6.1. *Pluggability of the SIA*

The presented SIA is a structure used to enrich the semantics of embeddings without changing the shape of the embeddings. Therefore, SIA is a portable structure that can be widely inserted into other models. In the comparative experiment, although our method outperforms the relevant advanced works, the performance improvement is not significant. In order to further verify the effectiveness of our method and to verify the pluggability of SIA, pluggability experiments are conducted. The experimental results are shown in Table 5.

This paper successfully inserts SIA in several advanced models and achieved improvements in named entity recognition and relation extraction tasks in all of them. The application method is just inserting SIA after the encoder of the model. Since SIA does not change the shape of embeddings, embeddings generated by SIA can still be smoothly applied to subsequent calculations of the model. Furthermore, SIA enriches the semantics of embeddings, which can provide a better foundation for subsequent information extraction. Inserting SIA in several advanced models consistently resulted in significant improvements in both named entity recognition and relation extraction tasks. In order to clearly demonstrate the performance improvement effect brought by SIA to several related studies, the experimental results are visualized as shown in the Fig. 5(a). The experimental results demonstrate that the SIA presented in this paper possesses excellent pluggability. And it is orthogonal to other works, which can further improve the performance of information extraction. It can be easily applied to other information extraction models and enhance their ability in Chinese information extraction tasks by enriching the semantics of Chinese.



(a) Comparison of pluggability of SIA.

(b) F1 curves of ablation experiments.

Fig. 5. Performance comparison of the proposed components with baseline.

Table 6
The performance improvement brought by the proposed components and the corresponding resource overhead. GPU Memory is measured with batch size 1.

Order	SIA	IB	F1			Paras	FLOPs	Memory
			CAIT	Diakg	CMedCausal			
1	-	-	66.55%	56.38%	63.75%	85.65M	21.76B	1947MB
2	-	✔	71.36%	65.03%	66.82%	+6.3M	+1.61B	+80.5MB
3	✔	-	70.93%	64.94%	67.01%	+15.76M	+4.04B	+202MB
4	✔	✔	**73.14%**	**70.04%**	**67.88%**	117.17M	29.84B	2391MB

5.6.2. Module analysis of SIA and IB

This section conducts ablation studies to assess the improvements attributed to the contributions proposed in this paper and provides an analysis of the experimental results. Four models are introduced to compare performance. The first model is the baseline, which is combined by prompt and pointer networks. The second model only uses the IB structure, which replaces two SIA modules with two multi-head cross-attention modules in CREAM. The third model only uses one SIA structure, which is inserted after the PLM. The whole structure is depicted in Fig. 3 (a). The fourth model is the proposed CEREM (Fig. 3 (b)).

As shown in Table 6, both innovation points proposed in this paper have shown significant improvements compared to the baseline. The complete model proposed in this paper achieves the best evaluation metric value. SIA structure satisfies the baseline model's dependency on semantics, leading to improved performance of information extraction tasks. Additionally, the conducted IB provides effective parameter decoupling, allowing parameters to focus on different sub-tasks and thus achieving performance gains. The complete model demonstrates optimal performance, indicating that the methods proposed in this study can be effectively integrated to enhance information extraction capabilities. The F1 curve on the dev dataset during the model training process is plotted in Fig. 5(b), where the curve has undergone smooth spline interpolation. The improvement brought by each innovation point can be intuitively observed from the F1 trend.

In addition, parameters, FLOPs, and GPU memory were used to validate the computational overhead of the SIA module. The results are shown in Table 6. Compared with the baseline model, the SIA module introduces only a slight increase in model complexity, accounting for 18.40% of the parameter count and 18.57% of the FLOPs of the baseline. Although the SIA and IB components incur non-negligible computational overhead, this cost is justified by their significant performance improvement and SIA's unique capability to address the core issue of semantic isolation in Chinese, which cannot be achieved merely by scaling up the model. Therefore, the additional computation introduced by SIA and IB can be considered a reasonable and effective trade-off between efficiency and accuracy.

As shown in Table 7, the introduction of the Segment Information Attention module results in a moderate decrease in throughput and an increase in batch-level latency under different batch sizes. This additional computational cost mainly arises from the segment-aware attention computation and multi-scale semantic aggregation. Nevertheless, the efficiency degradation remains limited under small batch settings, which are representative of latency-sensitive inference scenarios. When considered together with the consistent performance improvements reported in the preceding experiments, these results indicate that SIA achieves a reasonable efficiency-performance balance. By focusing computation on semantically meaningful token-segment interactions rather than exhaustive token-level correlations, SIA improves semantic representation quality while maintaining practical computational efficiency.

Table 7
Throughput and latency of different model variants under various Batch sizes.

Batch	Models	Throughput (samples/sec)	Average delay (ms)
1	Baseline	9.09	109.97
	IB	7.06	141.72
	SIA	6.80	147.03
	IB + SIA	5.24	190.96
2	Baseline	9.95	200.96
	IB	8.30	240.89
	SIA	7.84	255.23
	IB + SIA	6.43	311.07
4	Baseline	10.13	394.88
	IB	8.37	477.86
	SIA	7.89	507.16
	IB + SIA	6.45	620.38

Table 8
Performance of CEREM on the CAIT dataset under different maximum window sizes k.

k	NER(%)	RE(%)
3	92.70	74.70
4	92.02	76.59
5	92.47	76.04
6	92.43	77.44
7	93.26	74.75

Table 9
Performance comparison of UTC-IE and CEREM on NER and RE tasks under different random seeds on the CAIT dataset.

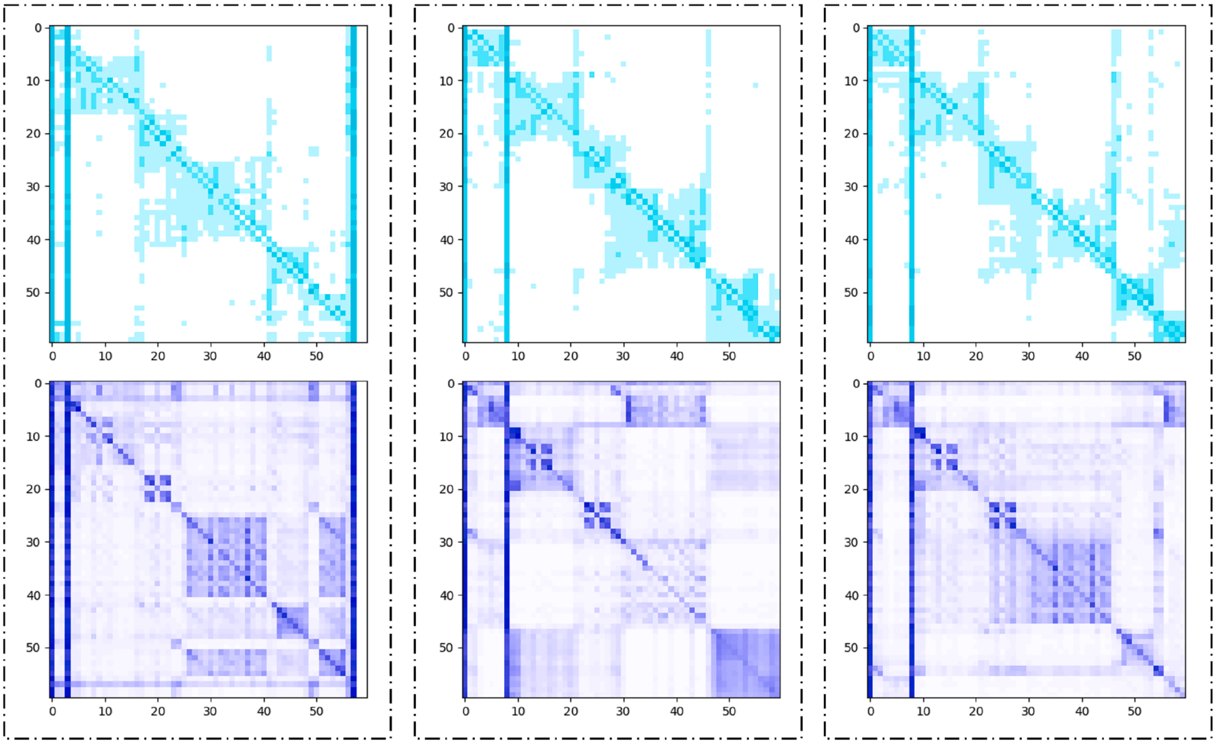
Seed	NER(%)		RE(%)	
	UTC-IE	CEREM	UTC-IE	CEREM
9	85.48	**88.59**	71.81	**71.82**
42	87.40	**91.86**	**73.20**	72.50
567	85.71	**90.66**	70.96	**72.32**
Mean ± Std	86.20 ± 1.05	**90.37 ± 1.65**	71.99 ± 1.13	**72.21 ± 0.35**

5.6.3. Robustness and parameter sensitivity analysis

In this section, we analyze the robustness and parameter sensitivity in the proposed model. Firstly, to investigate the impact of the maximum window size k on CEREM's performance, we conduct a systematic ablation study on the CAIT dataset. All other hyperparameters are kept fixed (random seed = 1234, batch size = 128, epochs = 600), while the maximum window scale k is varied from 3 to 7. This analysis aims to assess the robustness of the Segment Information Attention mechanism across different segment scales and to provide guidance for selecting an appropriate k value in practice.

As shown in Table 8, CEREM maintains consistently strong performance across all tested window sizes. For the NER task, the F1 scores fluctuated narrowly between 92.02% (k = 4) and 93.26% (k = 7), while for the RE task, the F1 scores ranged from 74.70% (k = 3) to 77.44% (k = 6). Although minor variations exist, no single window size demonstrates a decisive advantage across both tasks, indicating that CEREM is relatively insensitive to the choice of maximum window scale within the tested range. These findings suggest that the SIA mechanism is capable of effectively capturing semantic information across different segment lengths, maintaining stable performance even when the maximum window size varies. Considering both computational efficiency and semantic coverage, we selected k = 5 as the default configuration. This value provides a balanced trade-off, covering the majority of typical phrase lengths in Chinese text while avoiding unnecessary computational overhead. Overall, the parameter sensitivity analysis confirms the robustness of CEREM and validates the flexibility of the SIA mechanism to adapt to different segment scales.

Moreover, to ensure the reliability of the improvements presented, we perform a multi-run assessment on the CAIT dataset. For CEREM and the UTC-IE baseline, each model was trained and evaluated multiple times using different random seeds, with three independent runs per model. For each run, F1 scores were recorded in both the Named Entity Recognition and Relation Extraction tasks. This setup enables the computation of mean performance, mean deviation, paired differences, 95% confidence intervals (CI), and statistical significance, allowing a comprehensive evaluation of both effectiveness and stability. The experimental results are shown in Table 9.



TextA TextB TextC

Fig. 6. The attention scores of the model after training with/without SIA on CAIT dataset. Three examples are as input, where light blue images without SIA and dark blue images with SIA. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The results demonstrate that CEREM consistently outperforms UTC-IE on the NER task, achieving an average improvement of approximately 4.17%, which is statistically significant (t(2) = 7.586, p = 0.0169; 95% CI = [1.81, 6.54]). This confirms CEREM's ability to effectively capture entity semantics. For the RE task, the improvement achieved by CEREM is relatively small (0.22%) and not statistically significant (t(2) = 0.370, p = 0.7471; 95% CI = [-2.38, 2.82]), indicating that part of the observed gain may fall within the range of random variation. Nevertheless, CEREM exhibits notably lower variability across runs, especially for RE (Â ± 0.35 compared with Â ± 1.13 for UTC-IE), which highlights the stability and reliability of its predictions. Although the RE improvement on the CAIT dataset is limited, CEREM delivers clearly stronger gains on other datasets such as DiaKG, where the absolute improvement reaches 1.86%. Combined with the substantially reduced variance, these results suggest that the advantages of CEREM are reflected not only in average F1 scores but also in prediction stability and cross-domain robustness. Overall, CEREM integrates NER and RE through prompt-based and pointer-based modeling in a stable and scalable manner, providing both competitive performance and consistent, theoretically meaningful results.

Finally, to assess the model's robustness against semantic isolation, we conduct a length-based evaluation on the CAIT dataset using the named entity recognition (NER) task. To quantify the prevalence of long entities, we introduce the Long Entity Ratio, defined as the proportion of entities whose length exceeds six Chinese characters. Long entities typically contain multiple lexical units and represent highly aggregated semantics, which makes them more susceptible to semantic isolation. A cross-dataset statistical summary of long-entity proportions under this length threshold is provided in the Appendix to quantify the prevalence of long entities across datasets and substantiate the six-character threshold as a meaningful boundary for identifying semantically isolated units. This analysis supports the consistency and general applicability of the definition across different datasets. All entities are grouped by length, and the performance of CEREM is compared across different length intervals, including 1, 2, 3, 4, 5, 6, 7, and at least 8 characters. For fairness, entities longer than eight characters were combined into a single group due to their relatively small number. Table 10 depicts the proportion of entities of different lengths in the CAIT dataset alongside model performance within each group. The results show a general downward trend in performance as entity length increases, confirming that longer entities pose greater challenges for extraction due to their complex compositional semantics. Nevertheless, CEREM maintains strong and stable performance across all length groups and achieves particularly significant gains for long entities. Specifically, CEREM attains an overall F1-score of 88.95%, outperforming UIE (86.45%) and all PP-UIE variants. Notably, its relative advantage becomes more evident with increasing entity length, demonstrating its superior resilience to semantic aggregation and contextual sparsity.

The above results indicate that the Segment Information Attention mechanism plays a key role in alleviating semantic isolation by capturing internal segment dependencies within long entities. The length-based robustness analysis provides direct empirical evidence

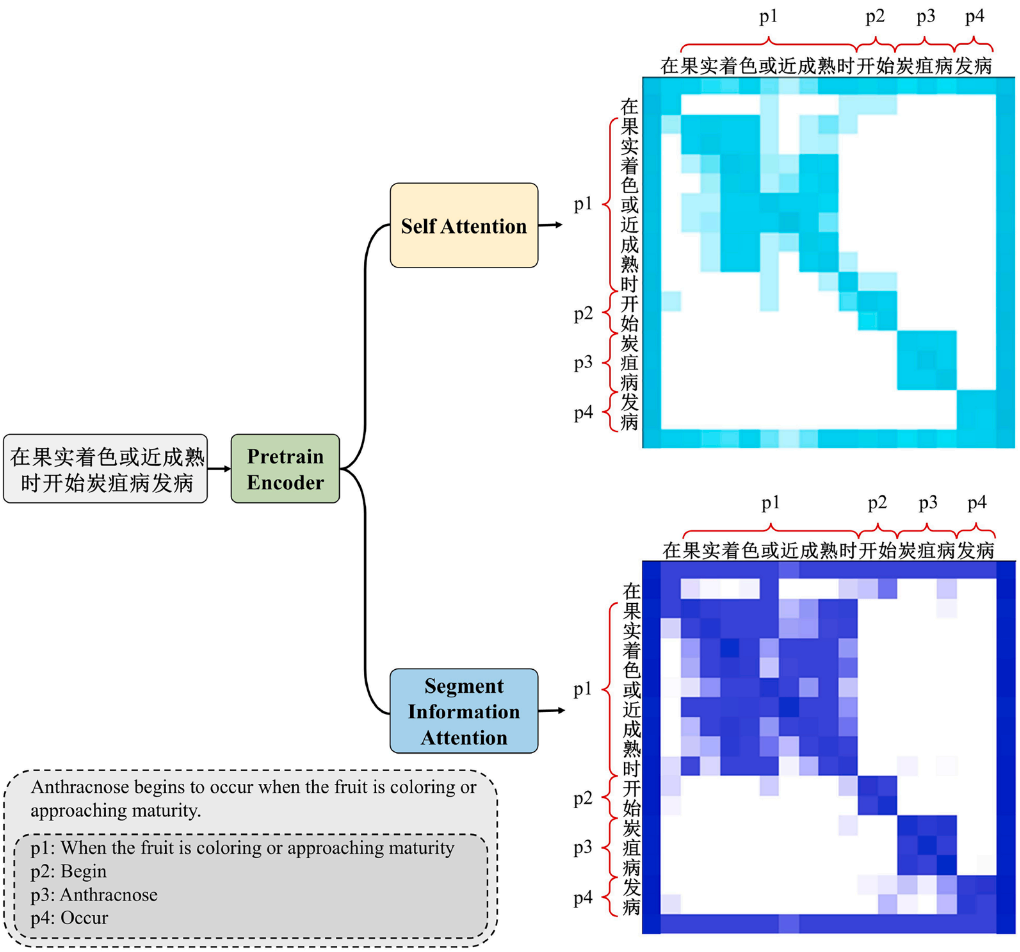

在果实着色或近成熟时开始炭疽病发病
Pretrain Encoder
Self Attention
Segment Information Attention
p1
p2
p3
p4
在果实着色或近成熟时开始炭疽病发病
Anthracnose begins to occur when the fruit is coloring or approaching maturity.
p1: When the fruit is coloring or approaching maturity
p2: Begin
p3: Anthracnose
p4: Occur

Fig. 7. A comparison example of SIA and self-attention. The same color grading method was used for SIA and self attention to demonstrate the difference in their attention scores.

Table 10
Comparison of model performance (F1%) on the CAIT dataset for entities of different lengths. Prop. represents the percentage of entities of each length within the CAIT dataset.

Length	Prop.(%)	UIE(%)	PPUIE-0.5b(%)	PPUIE-1.5b(%)	PPUIE-7b(%)	PPUIE-14b(%)	CEREM(%)
1	1.84	100.00	81.82	100.00	100.00	90.00	**100.00**
2	54.61	91.60	90.05	86.70	90.04	94.41	**94.76**
3	12.40	93.33	93.43	92.75	90.00	94.03	**95.45**
4	11.73	84.06	80.00	81.63	83.22	85.71	**86.13**
5	2.85	**90.91**	65.12	45.90	66.67	68.18	80.00
6	3.52	68.42	75.00	76.60	87.50	**89.36**	72.22
7	1.51	80.00	60.87	73.68	73.68	80.00	**80.00**
≥8	11.56	59.50	46.88	50.38	53.85	60.29	**63.87**
All	100	86.45	81.99	79.89	83.77	87.50	**88.95**

that CEREM not only performs reliably on short, compositional entities but also generalizes effectively to complex, semantically entangled expressions. Collectively, these results reinforce CEREM's robustness and interpretability in Chinese information extraction tasks, particularly in scenarios involving highly aggregated semantics.

5.7. Visualization: The effectiveness of SIA

As shown in Fig. 6, the attention scores after applying SIA exhibit distinct "patch-like" patterns. These "patches" mainly gather on the diagonal and are almost all square in shape. This phenomenon occurs because, when a token is in a text segment with complete semantics, SIA strengthens the correlations between this token and all other tokens within the segment. This suggests that SIA facilitates the model in achieving a holistic semantic understanding of complete text segments, enabling it to get the actual meaning of the segment.

In fact, it is also observed from the attention scores without the SIA that weak "patch-like" patterns emerge. However, due to the absence of the perspective of highly aggregated semantics, the responses between tokens and segments are relatively weak, resulting in phenomena such as blurred and incomplete boundaries in the "patches". In contrast, the SIA treats segments as wholes and captures correlations between tokens and segments. When a token is deemed relevant to a specific segment, SIA enhances the correlation between that token and all other tokens within the segment, leading to significant "patch-like" patterns. This phenomenon indicates that the perspective of highly aggregated semantics is fully leveraged by SIA, improving the model's understanding of Chinese semantics.

In order to present this phenomenon more intuitively, we selected a piece of data from the CAIT dataset for display, and the results are shown in the Fig. 7. Among them, p1-p4 are semantically complete text spans. For shorter words (p2, p3, p4), SIA and self attention can establish good correlations. However, for longer segments (p1), SIA establishes clearer and more accurate correlations (this is reflected in its patch-like characteristics and precise span on p1.). This ability can promote SIA to better understand the meaning of Chinese text. Moreover, this paper introduces two quantitative metrics to ensure that our interpretability: attention entropy and cluster distance. The calculation processes of these two quantitative metrics are described in detail in the appendix. The attention entropy of SIA is 1.7284, which is significantly lower than that of Self Attention (2.9137), representing a 68.5% reduction. This indicates that SIA encourages a more concentrated attention distribution. More importantly, the inter-cluster distance of the SIA model reaches 3.6479, 2.5 times larger than the baseline model (1.4620). This quantitatively demonstrates that the "block-style" attention patterns generated by SIA can better separate different semantic units, effectively alleviating the semantic isolation problem in Chinese. The quantitative analysis further validates the effectiveness of the SIA mechanism. The detail calculation principles are explained in the Appendix.

6. Discussion

The designed model demonstrates consistent superiority across datasets from multiple domains, which strongly indicates its robust cross-domain generalization capability. Since highly aggregated semantics are also present in other languages, this finding may, to some extent, validate the broader relevance of the methodology presented in this paper. Nevertheless, CEREM faces significant limitations. The quadratic computational complexity of its attention mechanism, coupled with the fixed input length constraint of its underlying pre-trained language model, leads to performance degradation when processing long documents containing complex discourse structures or nested semantic dependencies. While the Segment Information Attention mechanism partially mitigates this issue by aggregating segment-level context, it cannot fully capture long-range dependencies spanning multiple sentences or paragraphs. This constraint may restrict the model's scalability in real-world applications such as document-level information extraction, legal text analysis, or long-form medical reporting.

Beyond these architectural constraints, linguistic and domain diversity in Chinese presents further challenges and opportunities for extending CEREM. The framework has been evaluated primarily on standard Mandarin corpora, which exhibit relatively uniform lexical and syntactic characteristics. In practice, however, substantial linguistic variation exists–from regional dialects to domain-specific registers in fields such as law, finance, and agriculture. These variations introduce deviations in vocabulary, prosody, and contextual semantics that challenge the model's underlying assumptions. To improve generalization and robustness, future research could explore theoretically grounded adaptation strategies, including domain-informed fine-tuning, cross-dialect transfer learning, and multi-level semantic alignment. Such efforts would not only broaden CEREM's applicability across diverse linguistic contexts but also contribute to a more nuanced computational understanding of language variation in complex, low-resource environments.

In sum, while CEREM provides an effective and interpretable solution for Chinese semantic extraction, addressing its limitations in long-sequence modeling and extending its applicability to dialectal and domain-specific contexts represent key avenues for future research.

7. Conclusion

In this paper, we propose an information extraction architecture to extract highly aggregated semantics within Chinese texts. Firstly, a prompt-based unified information extraction network is proposed, which promotes semantic interaction and solves the parameter confusion issue. Secondly, a designed attention mechanism enrichs extracted semantics from the perspective of highly aggregated semantics. Finally, a parameter decoupling structure is conducted to make parameters focus on individual tasks. Experimental results indicate that the proposed network effectively addresses highly aggregated semantics and has a significant extent to guide the research of information extraction tasks on Chinese texts.

CRediT authorship contribution statement

Bin Liu: Methodology, Funding acquisition; **Jiaqi Han:** Writing – original draft, Data curation; **Zhenyu Zhang:** Validation, Formal analysis; **Shijun Li:** Investigation, Data curation; **Haixi Zhang:** Supervision, Methodology; **Yijie Chen:** Supervision, Methodology, Funding acquisition, Formal analysis, Conceptualization; **Keqin Li:** Supervision, Resources, Methodology.

Data availability

The authors do not have permission to share data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (No. grantnumGS501100001809623762226), the Natural Science Foundation of Shaanxi Province (No. Natural Science Foundation of Shaanxi Province2025JC-YBQN-888), Shaanxi's Key Research and Development Program (No. 2024NC-ZDCYL-05-05), Xi'an Key Technology Research Projects for Key Agricultural Industry Chains (No. 2024JH-NYZD-0027)' and the Yangling Demonstration Zone Science and Technology Plan Project (No. 2024NY-14).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.ipm.2026.104617

References

Bölücü, N., Rybinski, M., Dai, X., & Wan, S. (2024). An adaptive approach to noisy annotations in scientific information extraction. *Information Processing & Management*, *61*(6), 103857.

Chang, D., Chen, M., Liu, C., Liu, L., Li, D., Li, W., Kong, F., Liu, B., Luo, X., Qi, J. et al. (2021). Diakg: An annotated diabetes dataset for medical knowledge graph construction. In *China conference on knowledge graph and semantic computing* (pp. 308–314). Springer.

Chen, Z., Hao, J., Sun, H., Zhao, L., Li, J., Qian, Q., Peng, Q., Wang, X., Cong, S., Shen, L. et al. (2025). MedscaleRE-PF: A prompt-based framework with retrieval-augmented generation, chain-of-thought, and self-verification for scale-specific relation extraction in chinese medical literature. *Information Processing & Management*, *62*(6), 104278.

Deng, H., Zhang, Y., Zhang, Y., Ying, W., Yu, C., Gao, J., Wang, W., Bai, X., Yang, N., Ma, J. et al. (2022). 2event: Benchmarking open event extraction with a large-scale chinese title dataset. arXiv preprint arXiv:2211.00869.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. (pp. 4171–4186). Minneapolis, Minnesota.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International conference on learning representations*.

Du, X., & Ji, H. (2022). Retrieval-augmented generative question answering for event argument extraction. (pp. 4649–4666). Abu Dhabi, United Arab Emirates.

Gui, W., & Cui, A. (2023). Aje: Attention mechanism for entity-relation joint extraction. In *Journal of physics: Conference series* (pp. 12020–12027). Virtual, Online (*vol. 2504*).

He, X., Li, S., Zhang, Y., Li, B., Xu, S., & Zhou, Y. (2025). The more quality information the better: Hierarchical generation of multi-evidence alignment and fusion model for multimodal entity and relation extraction. *Information Processing & Management*, *62*(1), 103875.

Hsu, I.-H., Huang, K.-H., Boschee, E., Miller, S., Natarajan, P., Chang, K.-W., & Peng, N. (2022). DEGREE: A data-efficient generation-based event extraction model. (pp. 1890–1908). Seattle, United States.

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging.

Hui, B., Yang, J., Cui, Z., Yang, J., Liu, D., Zhang, L., Liu, T., Zhang, J., Yu, B., Lu, K., Dang, K., Fan, Y., Zhang, Y., Yang, A., Men, R., Huang, F., Zheng, B., Miao, Y., Quan, S., Feng, Y., Ren, X., Ren, X., Zhou, J., & Lin, J. (2024). Qwen2.5-coder technical report.

Jia, Z., Yan, Z., Han, W., Zheng, Z., & Tu, K. (2023). Modeling instance interactions for joint information extraction with neural high-order conditional random field. (pp. 13695–13710). Toronto, Canada.

Jiang, Y., & Zhao, J. (2022). Medical causality extraction: A two-stage based nested relation extraction model. In *China health information processing conference* (pp. 73–85). Springer.

Jiang, Z., Xu, W., Araki, J., & Neubig, G. (2020). Generalizing natural language analysis through span-relation representations. (pp. 2120–2133). Online.

Lee, J., Moon, H., Lee, S., Park, C., Eo, S., Ko, H., Seo, J., Lee, S., & Lim, H. (2024). Length-aware byte pair encoding for mitigating over-segmentation in korean machine translation. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the association for computational linguistics: ACL 2024* (pp. 2287–2303). Bangkok, Thailand: Association for Computational Linguistics.

Li, F., Peng, W., Chen, Y., Wang, Q., Pan, L., Lyu, Y., & Zhu, Y. (2020a). Event extraction as multi-turn question answering. (pp. 829–838). Online.

Li, H.-W., Lin, Y.-J., Li, Y.-T., Lin, C., & Kao, H.-Y. (2023). Improved unsupervised chinese word segmentation using pre-trained knowledge and pseudo-labeling transfer. (pp. 9109–9118). Singapore.

Li, J., Fei, H., Liu, J., Wu, S., Zhang, M., Teng, C., Ji, D., & Li, F. (2022a). Unified named entity recognition as word-word relation classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, (p. 10965–10973).

Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., & Li, J. (2020b). A unified MRC framework for named entity recognition. (pp. 5849–5859). Online.

Li, Z., Chen, M., Ma, Z. et al. (2022b). Cmedcausal: Chinese medical causal relationship extraction dataset. *Journal of Medical Informatics*, *43*(12), 23–27.

Liang, T., Yuan, S., Zhou, P., Fu, H., & Wu, H. (2022). Domain robust pipeline for medical causal entity and relation extraction task. In *China health information processing conference* (pp. 57–65). Springer.

Lin, C., Lin, Y.-J., Yeh, C.-J., Li, Y.-T., Yang, C., & Kao, H.-Y. (2023). Improving multi-criteria chinese word segmentation through learning sentence representation. (pp. 12756–12763). Singapore.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, *abs/1907.11692*, 1–13.

Lu, Y., Lin, H., Xu, J., Han, X., Tang, J., Li, A., Sun, L., Liao, M., & Chen, S. (2021). Text2event: Controllable sequence-to-structure generation for end-to-end event extraction. (pp. 2795–2806). Online.

Lu, Y., Liu, Q., Dai, D., Xiao, X., Lin, H., Han, X., Sun, L., & Wu, H. (2022). Unified structure generation for universal information extraction. (pp. 5755–5772). Dublin, Ireland.

Luo, S., & Yu, J. (2024). Esgnet: A multimodal network model incorporating entity semantic graphs for information extraction from chinese resumes. *Information Processing & Management*, *61*(1), 103524.

PaddleNLP, C., (2025). PaddleNLP: An Easy-to-use and High Performance NLP Library, https://github.com/PaddlePaddle/PaddleNLP.

Ping, Y., Lu, J., Gan, R., Wang, J., Zhang, Y., Zhang, P., & Zhang, J. (2023). UniEX: An effective and efficient framework for unified information extraction via a span-extractive perspective. (pp. 16424–16440). Toronto, Canada.

Ponce, D., Etchegoyhen, T., Calleja, J., & Gete, H. (2024). Split and rephrase with large language models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 11588–11607). Bangkok, Thailand: Association for Computational Linguistics.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67.

Shang, Y.-M., Huang, H., & Mao, X. (2022a). Onerel: Joint entity and relation extraction with one module in one step. *Proceedings of the AAAI Conference on Artificial Intelligence*, (pp. 11285–11293).

Shang, Y.-M., Huang, H., & Mao, X. (2022b). Onerel: Joint entity and relation extraction with one module in one step. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 11285–11293). (*vol. 36*).

Silva, R. J., Gedela, K., Marr, A., Desmet, B., Rose, C., & Zhou, C. (2022). QA4IE: A quality assurance tool for information extraction. (pp. 4497–4503). Marseille, France.

Su, Y., Wang, P., Cui, S., Xu, F., & Ishdorj, T.-O. (2023). Bije: A joint extraction model for biomedical information extraction. In *Lecture notes in computer science* (pp. 119–130). Zhengzhou, China (*vol. 14088 LNCS*).

Wang, Y., Sun, C., Wu, Y., Zhou, H., Li, L., & Yan, J. (2021). UniRE: A unified label space for entity relation extraction. (pp. 220–231). Online.

Wang, Y., Wang, Y., Peng, Z., Zhang, F., & Yang, F. (2023). A concise relation extraction method based on the fusion of sequential and structural features using ERNIE. *MATHEMATICS*, *11*(6), 1439–1458.

Wang, Y., Yu, B., Zhang, Y., Liu, T., Zhu, H., & Sun, L. (2020). TPLinker: Single-stage joint extraction of entities and relations through token pair linking. (pp. 1572–1582). Barcelona, Spain.

Wei, Z., Su, J., Wang, Y., Tian, Y., & Chang, Y. (2020). A novel cascade binary tagging framework for relational triple extraction. (pp. 1476–1488). Online.

Yan, H., Gui, T., Dai, J., Guo, Q., Zhang, Z., & Qiu, X. (2021). A unified generative framework for various NER subtasks. (pp. 5808–5822). Online.

Yan, H., Sun, Y., Li, X., Zhou, Y., Huang, X., & Qiu, X. (2023). UTC-IE: A unified token-pair classification architecture for information extraction. (pp. 4096–4122). Toronto, Canada.

Yan, Z., Ye, Z., Ge, J., Qin, J., Liu, J., Cheng, Y., & Gurrin, C. (2025). Docextractnet: A novel framework for enhanced information extraction from business documents. *Information Processing & Management*, *62*(3), 104046.

Yang, S., Feng, D., Qiao, L., Kan, Z., & Li, D. (2019). Exploring pre-trained language models for event extraction and generation. (pp. 5284–5294). Florence, Italy.

Ye, D., Lin, Y., Li, P., & Sun, M. (2022). Packed levitated marker for entity and relation extraction. (pp. 4904–4917). Dublin, Ireland.

Yu, B., Zhang, Z., Shu, X., Liu, T., Wang, Y., Wang, B., & Li, S. (2020). Joint extraction of entities and relations based on a novel decomposition strategy. In *Frontiers in artificial intelligence and applications* (pp. 2282–2289). Santiago de Compostela, Online, Spain (*vol. 325*).

Zeiler, Matthew, D., FergusR., (2014). Visualizing and understanding convolutional networks. In *Computer vision – ECCV 2014* (pp. 818–833). Cham: Springer International Publishing.

Zeng, X., Zeng, D., He, S., Liu, K., & Zhao, J. (2018). Extracting relational facts by an end-to-end neural model with copy mechanism. (pp. 506–514). Melbourne, Australia.

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. (2019). ERNIE: Enhanced language representation with informative entities. (pp. 1441–1451). Florence, Italy.

Zheng, H., Wen, R., Chen, X., Yang, Y., Zhang, Y., Zhang, Z., Zhang, N., Qin, B., Xu, M., & Zheng, Y. (2021). Prgc: Potential relation and global correspondence based joint relational triple extraction. arXiv preprint arXiv:2106.09895.

Zheng, S., Wang, F., Bao, H., Hao, Y., Zhou, P., & Xu, B. (2017). Joint extraction of entities and relations based on a novel tagging scheme. (pp. 1227–1236). Vancouver, Canada.

Zhong, Z., & Chen, D. (2021). A frustratingly easy approach for entity and relation extraction. (pp. 50–61). Online.