ENGINEERING Information Technology & Electronic Engineering
https://www.academax.com/EITEE; https://www.engineering.org.cn
https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=11344065
E-mail: eng.itee@zju.edu.cn

Engineering Journals
Available on IEEE Xplore®

Research Article
https://doi.org/10.1631/ENG.ITEE.2025.0047

DDiNER: domain dictionary-guided Chinese named entity recognition for complex industrial contexts

Ronghui LIU[1,2], Wei CUI[1,2]✉, Xiaojun LIANG[2], Weihua GUI[2,3]

[1]*School of Future Technology, South China University of Technology, Guangzhou 510641, China*

[2]*Pengcheng Laboratory, Shenzhen 518055, China*

[3]*School of Automation, Central South University, Changsha 410083, China*

Abstract: Accurate Chinese named entity recognition (NER) in the process industry is crucial for applications such as information extraction, knowledge graph construction, and intelligent decision-making. However, challenges, including ambiguous entity boundaries, semantic overlaps, and limited annotated data, significantly hinder performance. To address these issues, this study proposes DDiNER, a domain dictionary-guided Chinese NER framework that integrates a hierarchical industrial domain dictionary with bidirectional encoder representations from Transformers (BERT) via a hierarchical lexicon adapter (HLA), combined with bidirectional long short-term memory (BiLSTM) and conditional random field (CRF) layers for multilevel feature fusion. Experimental results show that DDiNER achieves superior performance, with average precision, recall, and F1-scores of 95.75%, 95.73%, and 95.74%, respectively, outperforming state-of-the-art models. Validation on an independent dataset confirms its robustness and strong capability in recognizing unseen and long-tail entities. This study provides an effective and scalable solution for industrial Chinese NER, with significant potential for downstream intelligent applications.

Key words: Named entity recognition (NER); Process industry; Domain dictionary; Hierarchical lexicon adapter (HLA)

1 Introduction

Named entity recognition (NER) aims to automatically extract specific entities from massive unstructured text and identify their corresponding categories (Gao et al., 2021; Liu P et al., 2022; Ehrmann et al., 2024), serving as a fundamental task for knowledge extraction and an essential foundation for downstream applications such as knowledge graph construction (Liu C and Yang, 2022; Zhong et al., 2024), information retrieval (Kumar and Starly, 2022), and question-answering systems (Hu Z and Ma, 2023; Prasanna et al., 2024). In the process industry domain, NER is a cornerstone of natural language processing (NLP), enabling the transformation of massive amounts of unstructured industrial text into structured machine-interpretable knowledge, which supports applications such as production optimization, safety monitoring, and intelligent decision-making. However, most of the current research is focused more on domains such as general purpose (Geng et al., 2023; Yang et al., 2024), finance (Zhang et al., 2023, 2024), agriculture (G et al., 2023; De et al., 2025), and medical (Hu Z and Ma, 2023; Hu Y et al., 2024), compared with the process industry domain, especially the Chinese process industry domain.

Compared with the general domain Chinese NER, process industry domain Chinese NER faces significantly greater challenges due to the complexity of domain knowledge, diversity of terminology, and heterogeneity of data sources. Texts in process industrial contexts, such as production records, equipment maintenance logs, and operational guidelines, often contain highly specialized terminologies, ambiguous abbreviations, compound expressions, and mixed data formats, making accurate entity recognition considerably difficult. There are three obvious challenges: (1) Process industrial texts often contain multi-word compound entities that describe specific equipment or phenomena, such as "连铸结晶器 (continuous casting crystallizer)," which combines "连铸 (continuous casting)" and "结晶器 (crystallizer)." Accurate identification of such entities requires precise handling of compound terms and correct Chinese word segmentation; otherwise, segmentation

✉ Wei CUI, aucuiwei@scut.edu.cn
Ronghui LIU, https://orcid.org/0009-0000-2965-2203
Wei CUI, https://orcid.org/0000-0003-1755-7887

CLC number: TP391.1

Received: Sept. 23, 2025; Revision accepted: Feb. 10, 2026;
Crosschecked: Mar. 5, 2026

errors can propagate and lead to recognition inaccuracies. (2) Many abbreviations and terminologies in process industry texts exhibit polysemy, meaning that their interpretations vary across operational contexts. For example, "CNC" may refer to "数控机床 (computer numerical control)" in mechanical manufacturing but represent "计算机数控" in other automation contexts, necessitating NER systems to incorporate domain knowledge and context-aware disambiguation strategies. (3) Unlike general-domain NER, process industry NER suffers from a shortage of large-scale and expert-annotated corpora due to the complexity and confidentiality of industrial data. This data scarcity makes it difficult to train deep learning models effectively and limits the performance of generic pre-trained language models on domain-specific tasks.

Most of existing NER methods have been dominated by deep learning techniques, beginning with sequence labeling models such as hidden Markov models (HMMs) and conditional random fields (CRFs) (McCallum and Li, 2003), and evolving towards more powerful neural architectures, including recurrent neural networks (RNNs) such as long short-term memory (LSTM) networks (Collobert et al., 2011). Among these, the bidirectional long short-term memory-conditional random field (BiLSTM-CRF) architecture (An et al., 2022) has been widely adopted, leveraging bidirectional LSTMs to capture contextual dependencies and a CRF layer to optimize sequence-level label predictions, which significantly improves recognition accuracy compared with traditional statistical approaches. Building upon these advances, Transformer-based encoders, including bidirectional encoder representations from Transformers (BERT) (Devlin et al., 2019), robustly optimized BERT approach (RoBERTa) (Liu YH et al., 2019), lexicon-enhanced BERT (LEBERT) (Liu W et al., 2021), and other domain-adaptive variants, have further boosted performance by producing highly contextualized semantic representations through large-scale pre-training and fine-tuning on annotated corpora. More recently, large language models (LLMs) have shown increasing potential in NER tasks (Hu Y et al., 2024; Zhou JY and Ma, 2025). Although LLMs are primarily designed for natural language generation, recent studies have repurposed them for information extraction. For instance, generative pre-trained Transformer-NER (GPT-NER) (Wang SH et al., 2023) reformulates NER as a text generation problem and introduces self-verification strategies to mitigate hallucination, achieving competitive results in low-resource scenarios. Similarly, prompt-based frameworks such as PromptNER (Ashok and Lipton, 2023) leverage the strong contextual understanding of LLMs to enable few-shot and cross-domain NER, demonstrating notable improvements in F1-scores with minimal training data.

However, despite these advancements, existing methods still exhibit significant limitations when applied to Chinese NER in process industry contexts. A fundamental challenge arises from the intrinsic characteristics of the Chinese language. Unlike alphabetic languages, Chinese texts lack explicit word boundaries, and word segmentation results are often ambiguous and non-unique. This issue is further amplified in process industry texts, where domain-specific entities are typically long compositional expressions and frequently appear as nested and overlapping character sequences. Under such conditions, deep learning-based models often fail to effectively incorporate domain-specific knowledge, leading to suboptimal recognition accuracy for rare and compound technical entities (Li ZZ et al., 2020; Hu ZT et al., 2024; Qu et al., 2024). Meanwhile, LLM-driven approaches also struggle with precise boundary detection, contextual disambiguation, and domain adaptation, largely due to their reliance on prompt engineering and limited access to specialized industrial corpora. As a result, models that rely on a single segmentation scheme or flat lexical representations frequently suffer from boundary errors and inadequate generalization to rare or unseen technical entities.

To address these challenges, we propose DDiNER, a domain dictionary-guided NER framework for process industry texts. DDiNER models domain knowledge as multi-granularity semantic concepts aligned with character-level representations. Domain terminology is organized into a hierarchical structure and associated with character positions, rather than treated as auxiliary word candidates. This design reflects the compositional, nested, and boundary-ambiguous nature of industrial Chinese entities.

Based on this modeling paradigm, DDiNER integrates hierarchical lexicon knowledge with character-level contextual representations through a hierarchical lexicon adapter (HLA). The framework enables selective fusion of coarse-grained conceptual cues and fine-grained lexical information, addressing flexible word boundaries in Chinese. By combining knowledge-guided structure with data-driven representation learning, DDiNER improves boundary detection, entity classification, and robustness to unseen entities in complex industrial scenarios. The main contributions of this study are summarized as follows:

1. We construct a hierarchical domain dictionary for the process industry. The dictionary supports semi-automatic annotation and serves as a structured knowledge source for modeling compositional and nested Chinese technical entities.

2. We propose DDiNER, a domain dictionary-guided Chinese NER framework. Through the HLA and a BiLSTM-CRF decoding architecture, the framework integrates multi-granularity lexicon knowledge into character-level representations, effectively handling boundary ambiguity and unseen entities.

3. We build a process industry-specific Chinese NER dataset covering four entity categories: process industry name, process technology, unit operation, and operating equipment. This dataset fills the gap of publicly available resources in the process industry and provides a benchmark for domain-adapted Chinese NER.

2 Related works

Chinese NER has undergone rapid development over the past two decades, progressing from early rule-based and statistical learning approaches to deep learning methods and, more recently, LLMs-driven frameworks (Seow et al., 2025). At the early stages, rule-based approaches dominated NER research by leveraging manually constructed dictionaries and handcrafted rules to identify entities from unstructured text (Liu P et al., 2022). Although these methods can achieve high

precision within narrow domains, they suffer from low recall and poor scalability due to the rapid emergence of new entities and the labor-intensive nature of rule construction (Wang XZ et al., 2022).

With the advent of annotated corpora, statistical learning-based methods became mainstream by formulating NER as a sequence labeling problem. Representative models include HMMs, maximum entropy models (MEs), support vector machines (SVMs), and particularly CRFs, leveraging contextual dependencies to improve sequence-level predictions (Zhang et al., 2023). These methods demonstrate significant performance improvements and enable broader applicability compared to rule-based approaches. However, their heavy reliance on manual feature engineering limits their adaptability across domains (Liu P et al., 2022).

The emergence of deep learning has revolutionized NER by enabling models to learn discriminative features automatically from data. Early architectures such as BiLSTM-CRF (An et al., 2022) combined character-level convolutional layers, bidirectional LSTM encoders, and CRF decoders to capture both local and global context, achieving state-of-the-art (SOTA) performance on general benchmarks. With the rise of pre-trained language models (PLMs), such as BERT (Devlin et al., 2019), RoBERTa (Liu YH et al., 2019), and LEBERT (Liu W et al., 2021), performance has been further improved by producing context-aware semantic embeddings fine-tuned on large-scale corpora. In domain-specific applications, these techniques have demonstrated strong results: For instance, in the mineral resource domain, Yu et al. (2022) applied a BERT-based NER framework to extract mineral-related entities from large-scale geological texts. In the financial domain, Zhang et al. (2023) proposed a BERT-CRF model enhanced with multi-channel attention to better capture complex semantic dependencies, while in engineering geology, Qiu et al. (2024) integrated domain-specific embeddings and multi-feature fusion to achieve superior accuracy.

Recently, LLMs such as GPT, chat generalized language model (ChatGLM), and LLM of Meta AI (LLaMA) have introduced a new paradigm for NER by reframing it as a prompt-driven generation task. For instance, GPT-NER (Wang SH et al., 2023) reformulates sequence labeling into a generation problem, employing self-verification mechanisms to reduce hallucination and improve reliability in low-resource settings. Similarly, PromptNER (Ashok and Lipton, 2023) leverages in-context learning to enable few-shot and cross-domain NER, significantly reducing reliance on large annotated datasets. Hu Y et al. (2024) proposed a prompt-engineering-based optimization strategy that improves clinical NER performance by enhancing entity boundary detection and contextual disambiguation. Zhou WX et al. (2023) introduced UniversalNER, which employs targeted knowledge distillation from large foundation models to improve open-domain NER accuracy across multiple languages and contexts. Similarly, Wang ZH et al. (2025) developed an LLM-driven NER framework that integrates domain priors and adaptive semantic constraints.

However, despite these advances, several challenges remain. Lu et al. (2025) highlighted that LLMs often struggle in token-level clinical NER, particularly when handling nested entities and ambiguous boundary decisions, suggesting that context-specific adaptations remain essential for high-precision applications. These findings collectively underscore both the promise and limitations of LLM-based NER approaches, motivating the development of domain-adaptive frameworks that can integrate expert knowledge, contextual modeling, and hierarchical resources—such as the proposed DDiNER framework—to address the unique challenges of Chinese NER in complex industrial contexts.

3 Methods

DDiNER is a domain-adaptive framework designed to enhance Chinese NER performance in process industry scenarios by integrating hierarchical domain knowledge with deep contextual modeling. As illustrated in Fig. 1, the framework operates through three tightly integrated components. First, a hierarchical domain dictionary is constructed from process industry corpora by combining term frequency-inverse document frequency (TF-IDF)-based keyword extraction with word to vector (Word2Vec)-driven semantic clustering, organizing lexicon entries into four representative categories: process industry name, process technology, unit operation, and operating equipment. This dictionary plays a dual role: It provides semi-automatic annotation guidance during dataset preparation and acts as an external knowledge source within the lexicon adapter. Second, the HLA fuses character-level embeddings with multiple matched lexicon embeddings at different semantic levels via a bilinear attention mechanism, facilitating deep interaction between textual context and domain-specific terminology. Finally, the BiLSTM-CRF decoding layer captures long-range sequential dependencies beyond self-attention and optimizes entity boundary detection by enforcing global label consistency.

To enhance readers' understanding, Table 1 summarizes the main notations used in different components of DDiNER, including domain dictionary construction, hierarchical lexicon fusion, and sequence modeling.

3.1 Domain dictionary construction

To effectively handle the complexity of domain-specific terminology in Chinese process industry texts, we construct a hierarchical domain dictionary that explicitly models the compositional nature of Chinese technical terms. In Chinese, industrial entities are often formed through recursive combinations of shorter character sequences, resulting in long nested expressions with flexible boundaries. Organizing such terms in a hierarchical structure enables the dictionary to reflect these compositional relationships and provides multi-granularity lexical knowledge for subsequent modeling. Specifically, the dictionary organizes domain terms into four representative categories: process industry name, process technology, unit operation, and operating equipment. As illustrated in Fig. 1 (left), the construction process consists of three main stages: word segmentation, keyword extraction, and word classification.

First, an industrial corpus is collected from multiple heterogeneous sources, including patents, technical manuals, production reports, and standard specifications. The raw corpus is then cleaned, normalized, and segmented using the Jieba

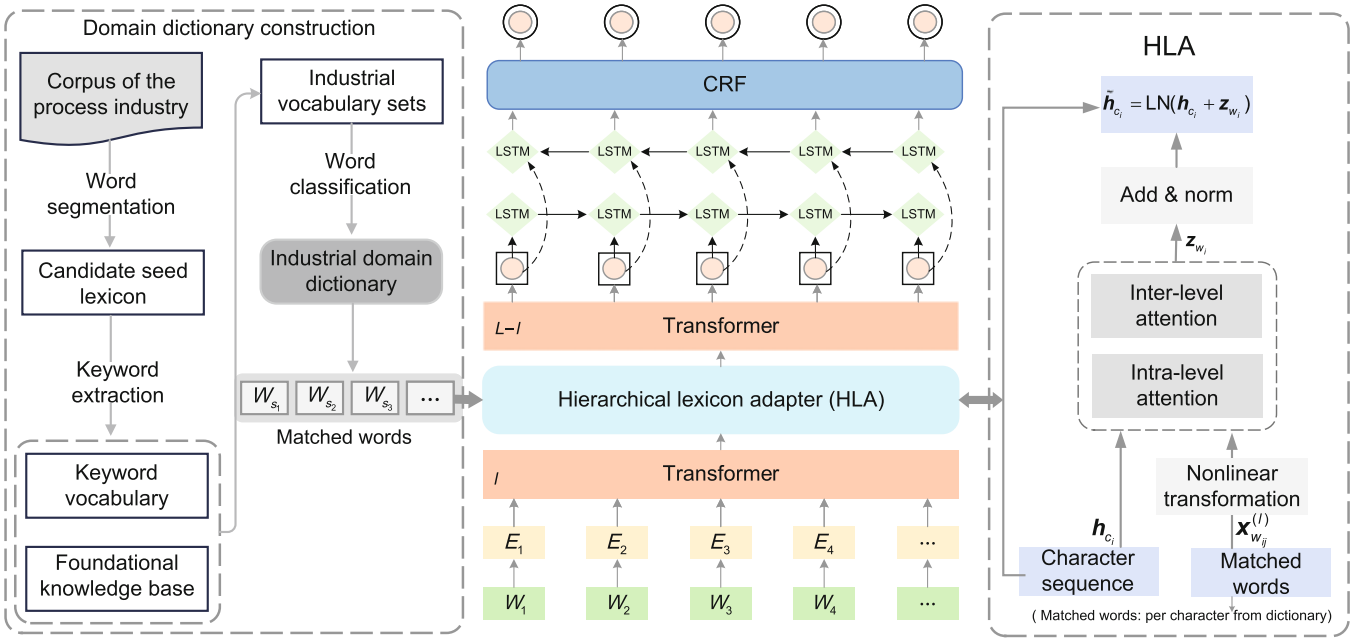

Fig. 1 Architecture of DDiNER. The framework integrates a hierarchical domain dictionary with a BERT-based encoder through an HLA. For each character, candidate domain terms at different semantic levels are matched and fused via attention mechanisms, and the resulting representations are decoded by a BiLSTM-CRF layer for structured sequence labeling. W_i denotes the i^{th} input token, E_i represents its corresponding embedding token, and W_{S_j} denotes the j^{th} candidate lexicon word in the matched lexicon set S obtained from the domain dictionary

Table 1 Summary of main notations

Notation	Description
c_i	The i^{th} character in the input sentence
s_i	Input sentence
$\boldsymbol{h}_{c_i}$	Representation of character c_i from BERT
$\tilde{\boldsymbol{h}}_{c_i}$	Lexicon-enhanced character representation
$\mathcal{W}_i$	Set of lexicon words matched to character c_i
$\mathcal{W}_i^{(l)}$	Matched lexicon words at hierarchy level l
$\boldsymbol{x}_{w_{ij}}^{(l)}$	Embedding of the j^{th} matched word w at level l
$\boldsymbol{v}_{w_{ij}}^{(l)}$	Projected word representation at level l
$\boldsymbol{z}_{w_i}^{(l)}$	Intra-level aggregated lexicon representation
$\boldsymbol{z}_{w_i}$	Final hierarchical lexicon representation
$\boldsymbol{a}_{ij}^{(l)}$	Intra-level attention weight
$\boldsymbol{A}_{\text{intra}}^{(l)}$	Intra-level attention matrix
$\boldsymbol{A}_{\text{inter}}$	Inter-level attention matrix
$\boldsymbol{M}_1^{(l)}, \boldsymbol{M}_2^{(l)}$	Projection matrices for word embeddings
$\boldsymbol{f}_t$	Forget gate of LSTM at time step t
$\boldsymbol{i}_t$	Input gate of LSTM at time step t
$\boldsymbol{o}_t$	Output gate of LSTM at time step t
$\boldsymbol{C}_t$	Cell state of LSTM at time step t
$\boldsymbol{h}_t$	Hidden state of BiLSTM at time step t
$\boldsymbol{U}_{\text{f}}, \boldsymbol{U}_{\text{i}}, \boldsymbol{U}_{\text{C}}, \boldsymbol{U}_{\text{o}}$	Learnable weight matrices of LSTM gates

segmentation tool to obtain candidate terms for vocabulary construction.

Second, domain-specific terms are extracted from the processed corpus using a TF-IDF-based keyword extraction algorithm, which captures both high-frequency and rare terms but semantically important technical expressions. The detailed procedure is summarized in Algorithm 1.

Third, to enhance semantic organization, we adopt

Algorithm 1 TF-IDF-based domain keyword extraction

Require: preprocessed training corpus $\text{Cor} = \{\text{cor}_1, \text{cor}_2, \ldots, \text{cor}_n\}$, candidate term set $\mathcal{V}$, and top-k selection threshold k

Ensure: extracted keyword set $\mathcal{Z} = \{z_1, z_2, \ldots, z_k\}$

for each term $\in \mathcal{V}$ **do**
  step 1: compute TF
  $F_{\text{TF}}(\text{term}) = \dfrac{\text{count}(\text{term})}{\sum_{\text{term}' \in \text{cor}_i} \text{count}(\text{term}')}$
  step 2: compute IDF
  $F_{\text{IDF}}(\text{term}) = \ln\left(\dfrac{|\text{Cor}|}{|\{\text{cor}_i \in \text{Cor} : \text{term} \in \text{cor}_i\}| + 1}\right)$
  step 3: calculate the TF-IDF score
  $F_{\text{TF-IDF}}(\text{term}) = F_{\text{TF}}(\text{term}) F_{\text{IDF}}(\text{term})$
end for
step 4: rank and select the top-k keywords
sort all terms in $\mathcal{V}$ in descending order by $F_{\text{TF-IDF}}$
$\mathcal{Z} \leftarrow$ the first k terms from the sorted list
return $\mathcal{Z}$

Word2Vec-based embedding representations to compute similarity between extracted keywords and existing dictionary entries. The semantic similarity between two words with embedding vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is computed as

$$\cos\theta = \frac{\sum_{k=1}^{n} x_{1k} x_{2k}}{\sqrt{\sum_{k=1}^{n} x_{1k}^2} \sqrt{\sum_{k=1}^{n} x_{2k}^2}}, \tag{1}$$

where θ denotes the angle between two vectors, n is the vector dimension, and x_{1k} and x_{2k} represent the k^{th} components of vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively.

By clustering semantically related words based on cosine similarity, we organize the dictionary into a hierarchical structure, where higher-level nodes correspond to coarse-grained entity types (e.g., process names) and lower-level nodes represent fine-grained technical entities (e.g., hot acid leaching).

3.2 HLA with BERT

To better leverage domain-specific hierarchical knowledge for Chinese NER in the process industry, we enhance the original LEBERT (Liu W et al., 2021) architecture by introducing an HLA, enabling hierarchical lexicon fusion (HLF). Unlike the original LEBERT model, which integrates only flat lexicon features, the proposed HLA adopts a character-centric but lexicon-aware design that is particularly suitable for Chinese NER, where word boundaries are ambiguous and domain terms often span variable-length character sequences. By incorporating multi-level domain knowledge extracted from a hierarchical domain dictionary, the model jointly captures character-level context and lexicon semantics at different granularities.

Given an input sentence $S = \{c_1, c_2, \ldots, c_n\}$ consisting of n characters, the character sequence serves as the basic encoding unit of the model. Each character c_i corresponds to a position-wise representation learned by the BERT encoder, independent of whether it forms a standalone word in the lexicon. To address the flexible word boundary issue in Chinese, we retrieve a set of matched words for each character position by matching substrings that cover c_i against the hierarchical domain dictionary. These matched words provide auxiliary domain knowledge associated with the character position, rather than constituting an independently tokenized word sequence. The resulting candidate set is denoted as $\mathcal{W}_i = \{\mathcal{W}_i^{(1)}, \mathcal{W}_i^{(2)}, \ldots, \mathcal{W}_i^{(L)}\}$, where L denotes the number of hierarchical levels and each $\mathcal{W}_i^{(l)} = \{w_{i1}^{(l)}, w_{i2}^{(l)}, \ldots, w_{im_l}^{(l)}\}$ contains the matched words at level l with m_l denoting the number of matched lexicon words at the l^{th} hierarchical level.

3.2.1 Character-feature word matching

To integrate lexicon knowledge into character representations, we construct a prefix tree T from the hierarchical domain dictionary to support efficient substring matching. For each input sentence, all multi-character substrings are traversed and matched against the prefix tree T to retrieve potential feature words. Each character position c_i serves as an anchor to associate those matched feature words that cover its position, rather than being matched to the dictionary in isolation. For example, the input sentence "热酸浸出法" (hot acid leaching method) can be matched to four feature words: "热酸" (hot acid), "浸出" (leaching), "热酸浸出" (hot acid leaching), and "热酸浸出法" (hot acid leaching method). These feature words are then assigned to all characters they contain in the sentence. As illustrated in Fig. 2, the feature word "热酸" is assigned to the characters "热" and "酸," "浸出" is assigned to "浸" and "出," and "热酸浸出" is assigned to "热," "酸," "浸," and "出," and so on.

In practical implementation, the maximum number of feature words retained for a sentence is limited to five. This character-feature word matching procedure produces candidate lexicon sets for each character position.

3.2.2 Hierarchical word vector projection

For each level l, we project the matched word embeddings $\boldsymbol{x}_{w_{ij}}^{(l)}$ into a unified semantic space using level-specific transformation matrices:

$$\boldsymbol{v}_{w_{ij}}^{(l)} = \boldsymbol{M}_2^{(l)} \tanh\left(\boldsymbol{M}_1^{(l)} \boldsymbol{x}_{w_{ij}}^{(l)} + \boldsymbol{b}_1^{(l)}\right) + \boldsymbol{b}_2^{(l)}, \tag{2}$$

where $\boldsymbol{M}_1^{(l)}$ and $\boldsymbol{M}_2^{(l)}$ are the learnable projection matrices, $\boldsymbol{b}_1^{(l)}$ and $\boldsymbol{b}_2^{(l)}$ denote the bias vectors, and $\boldsymbol{v}_{w_{ij}}^{(l)} \in \mathbb{R}^{d_c}$ is the projected vector in the d_c-dimensional semantic space.

3.2.3 Intra-level bilinear attention

Within each level l, we compute the attention weights between the character embedding $\boldsymbol{h}_{c_i}$ from the l^{th} Transformer layer and its candidate word vectors $\boldsymbol{v}_{w_{ij}}^{(l)}$:

$$\boldsymbol{a}_{ij}^{(l)} = \frac{\exp\left(\boldsymbol{h}_{c_i}^{\mathrm{T}} \boldsymbol{A}_{\text{intra}}^{(l)} \boldsymbol{v}_{w_{ij}}^{(l)}\right)}{\sum_{k=1}^{m_l} \exp\left(\boldsymbol{h}_{c_i}^{\mathrm{T}} \boldsymbol{A}_{\text{intra}}^{(l)} \boldsymbol{v}_{w_{ik}}^{(l)}\right)}, \tag{3}$$

where $\boldsymbol{A}_{\text{intra}}^{(l)}$ is the learnable intra-level attention matrix. The lexicon-enhanced representation for level l is obtained by

$$\boldsymbol{z}_{w_i}^{(l)} = \sum_{j=1}^{m_l} \boldsymbol{a}_{ij}^{(l)} \boldsymbol{v}_{w_{ij}}^{(l)}. \tag{4}$$

3.2.4 Inter-level attention fusion

To integrate multi-level semantic features, we introduce an inter-level attention mechanism to adaptively weight the

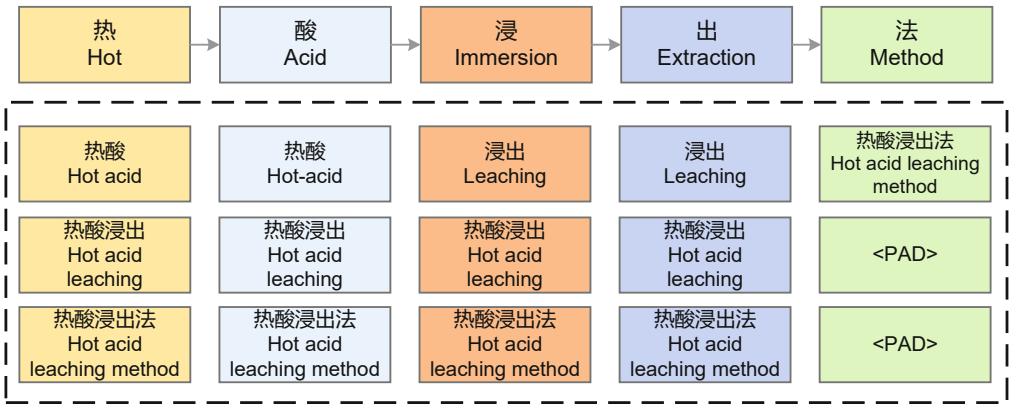

Fig. 2 Character word-pair sequence of a truncated Chinese sentence "热酸浸出法 (hot acid leaching method)." ⟨PAD⟩ denotes a padding token used to align candidate lexicon sequences of different lengths

contribution of each level:

$$\boldsymbol{\alpha}^{(l)} = \frac{\exp\left(\boldsymbol{h}_{c_i}^{\mathrm{T}} \boldsymbol{A}_{\mathrm{inter}} \boldsymbol{z}_{w_i}^{(l)}\right)}{\sum_{l=1}^{L} \exp\left(\boldsymbol{h}_{c_i}^{\mathrm{T}} \boldsymbol{A}_{\mathrm{inter}} \boldsymbol{z}_{w_i}^{(l)}\right)}, \tag{5}$$

where $\boldsymbol{A}_{\mathrm{inter}}$ is the inter-level attention matrix. The final hierarchical lexicon representation for character c_i is computed as

$$\boldsymbol{z}_{w_i} = \sum_{l=1}^{L} \boldsymbol{\alpha}^{(l)} \boldsymbol{z}_{w_i}^{(l)}. \tag{6}$$

3.2.5 Fusion with BERT representations

For each character position i (i denotes the position index in the sequence), we first obtain its contextual representation $\boldsymbol{h}_{c_i}$ from the l^{th} Transformer layer of the BERT encoder. In parallel, matched lexicon words $\mathcal{W}_i^{(l)}$ are retrieved at each hierarchy level via trie-based dictionary matching. Each matched word embedding $\boldsymbol{x}_{w_{ij}}^{(l)}$ is projected into a unified representation space and aggregated within the same level using intra-level attention to form $\boldsymbol{z}_{w_i}^{(l)}$. The level-specific representations are then fused across different hierarchy levels through inter-level attention, resulting in the final hierarchical lexicon representation $\boldsymbol{z}_{w_i}$.

The character-level and lexicon-level representations are integrated through a residual fusion followed by layer normalization:

$$\tilde{\boldsymbol{h}}_{c_i} = \mathrm{LN}\left(\boldsymbol{h}_{c_i} + \boldsymbol{z}_{w_i}\right), \tag{7}$$

where $\mathrm{LN}(\cdot)$ denotes the layer normalization, which stabilizes training by normalizing the fused features across the hidden dimension. The enhanced representation $\tilde{\boldsymbol{h}}_{c_i}$ is subsequently fed into the upper Transformer layers, enabling the model to jointly exploit contextual information from the character sequence and structured domain knowledge provided by the hierarchical lexicon.

3.3 BiLSTM-CRF decoding layer

The hierarchical lexicon-enhanced representations $\tilde{\boldsymbol{h}}_{c_i}$ obtained from the HLF are fed into a BiLSTM network to further model sequential dependencies and contextual semantics, as shown in Fig. 3. While the Transformer-based encoder captures global relationships through self-attention, the BiLSTM complements this by explicitly modeling forward and backward dependencies within local sequential contexts, which is particularly beneficial for recognizing complex process industry entities and accurately detecting entity boundaries.

3.3.1 BiLSTM contextual encoding

The BiLSTM comprises two LSTM layers that process the input sequence $\tilde{\boldsymbol{h}} = \{\tilde{\boldsymbol{h}}_{c_1}, \tilde{\boldsymbol{h}}_{c_2}, \ldots, \tilde{\boldsymbol{h}}_{c_n}\}$ in both forward and backward directions. For each time step t, the forward $\left(\overrightarrow{\boldsymbol{h}_t}\right)$ and backward $\left(\overleftarrow{\boldsymbol{h}_t}\right)$ hidden states are computed as

$$\overrightarrow{\boldsymbol{h}_t} = \mathrm{LSTM}_{\mathrm{fwd}}\left(\tilde{\boldsymbol{h}}_{c_t}, \overrightarrow{\boldsymbol{h}_{t-1}}\right), \quad \overleftarrow{\boldsymbol{h}_t} = \mathrm{LSTM}_{\mathrm{bwd}}\left(\tilde{\boldsymbol{h}}_{c_t}, \overleftarrow{\boldsymbol{h}_{t+1}}\right). \tag{8}$$

The final hidden state $\boldsymbol{h}_t$ at time step t is formed by concatenating the forward and backward states:

$$\boldsymbol{h}_t = \left[\overrightarrow{\boldsymbol{h}_t}, \overleftarrow{\boldsymbol{h}_t}\right]. \tag{9}$$

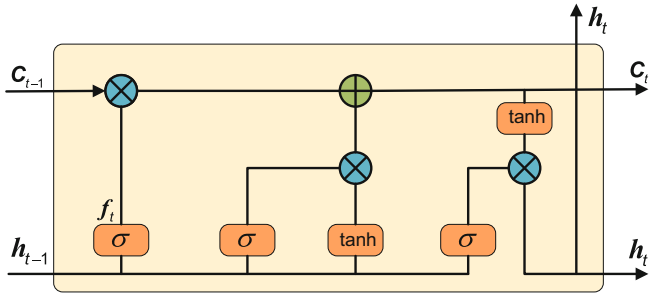

Fig. 3 Schematic of the LSTM unit displaying input, forget, and output gates with sigmoid and tanh functions essential for sequence memory management

Each LSTM unit maintains a memory cell $\boldsymbol{C}_t$ and uses three gates to control information flow: (1) The forget gate decides which information to discard; (2) The input gate controls how much new information to write into the cell state; (3) The output gate determines the next hidden state. Formally, the computations are defined as

$$\boldsymbol{f}_t = \sigma\left(\boldsymbol{U}_{\mathrm{f}}[\boldsymbol{h}_{t-1}, \tilde{\boldsymbol{h}}_{c_t}] + \boldsymbol{b}_{\mathrm{f}}\right), \tag{10}$$

$$\boldsymbol{i}_t = \sigma\left(\boldsymbol{U}_{\mathrm{i}}[\boldsymbol{h}_{t-1}, \tilde{\boldsymbol{h}}_{c_t}] + \boldsymbol{b}_{\mathrm{i}}\right), \tag{11}$$

$$\widetilde{\boldsymbol{C}}_t = \tanh\left(\boldsymbol{U}_{\mathrm{C}}[\boldsymbol{h}_{t-1}, \tilde{\boldsymbol{h}}_{c_t}] + \boldsymbol{b}_{\mathrm{C}}\right), \tag{12}$$

$$\boldsymbol{C}_t = \boldsymbol{f}_t \boldsymbol{C}_{t-1} + \boldsymbol{i}_t \widetilde{\boldsymbol{C}}_t, \tag{13}$$

$$\boldsymbol{o}_t = \sigma\left(\boldsymbol{U}_{\mathrm{o}}[\boldsymbol{h}_{t-1}, \tilde{\boldsymbol{h}}_{c_t}] + \boldsymbol{b}_{\mathrm{o}}\right), \tag{14}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \tanh(\boldsymbol{C}_t), \tag{15}$$

where $\sigma(\cdot)$ is the sigmoid activation function, $\widetilde{\boldsymbol{C}}_t$ is the candidate cell state of LSTM at time step t, $\boldsymbol{b}_{\mathrm{f}}$, $\boldsymbol{b}_{\mathrm{i}}$, $\boldsymbol{b}_{\mathrm{C}}$, and $\boldsymbol{b}_{\mathrm{o}}$ denote the corresponding bias vectors respectively, and $\boldsymbol{U}_{\mathrm{f}}$, $\boldsymbol{U}_{\mathrm{i}}$, $\boldsymbol{U}_{\mathrm{C}}$, and $\boldsymbol{U}_{\mathrm{o}}$ are learnable weight matrices respectively.

3.3.2 CRF-based sequence labeling

Although the BiLSTM captures bidirectional contextual information, it does not explicitly model the dependencies between adjacent output labels. In the process industry NER, such label dependencies are crucial; for example, an entity tag B-Process is often followed by an I-Process tag rather than an unrelated label. To address this, we integrate a CRF layer on top of the BiLSTM outputs to jointly optimize label prediction across the entire sequence.

Given the BiLSTM output sequence $\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_n]$, the CRF assigns a score to a label sequence $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n]$ as

$$s(\boldsymbol{H}, \boldsymbol{y}) = \sum_{i=1}^{n} \left(O_{i,\boldsymbol{y}_i} + T_{\boldsymbol{y}_{i-1},\boldsymbol{y}_i}\right), \tag{16}$$

where $O_{i,\boldsymbol{y}_i}$ is the emission score for predicting label $\boldsymbol{y}_i$ at position i, and $T_{\boldsymbol{y}_{i-1},\boldsymbol{y}_i}$ is the learnable transition score from label $\boldsymbol{y}_{i-1}$ to $\boldsymbol{y}_i$.

The conditional probability of a label sequence $\boldsymbol{y}$ is defined as

$$p(\boldsymbol{y}|\boldsymbol{H}) = \frac{\exp(s(\boldsymbol{H}, \boldsymbol{y}))}{\sum_{\tilde{\boldsymbol{y}} \in \boldsymbol{Y}_H} \exp(s(\boldsymbol{H}, \tilde{\boldsymbol{y}}))}, \tag{17}$$

where $\boldsymbol{Y}_H$ denotes the set of all possible label sequences and $\tilde{\boldsymbol{y}}$ denotes a candidate label sequence. The optimal label

sequence $\hat{\boldsymbol{y}}$ is obtained via

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}\in\boldsymbol{Y}_H} s(\boldsymbol{H}, \boldsymbol{y}). \tag{18}$$

4 Experimental analysis

To comprehensively evaluate the effectiveness of the proposed DDiNER framework, we conduct extensive experiments on a self-constructed industrial corpus. This section presents the overall experimental design, including corpus construction, hierarchical dictionary generation, and dataset annotation. We further describe the model implementation details, hyperparameter configurations, and evaluation metrics. Finally, we report the experimental results and analyze the recognition performance of DDiNER across different entity categories. The experimental findings demonstrate the advantages of integrating hierarchical domain knowledge with deep contextual representations, validating the capability of DDiNER to achieve accurate and robust NER in complex process industry contexts.

4.1 Experimental data

4.1.1 Corpus

To construct a high-quality corpus for Chinese NER in the process industry, we use Selenium 4 in Python to collect patent data from the IPRDB database. A total of 6277 patents are initially retrieved, and after a rigorous manual screening process, 5120 representative patents are retained. The dataset includes essential information such as patent titles and abstracts. All abstracts are segmented into sentences, resulting in an initial industrial corpus comprising 14 941 sentences, which serves as the foundation for subsequent experiments.

Following an 8:2 partition strategy, 11 953 sentences are allocated for model training and evaluation, while the remaining 2988 sentences are reserved for constructing the hierarchical domain dictionary described in Section 3.1.

4.1.2 Domain dictionary

Based on the methodology introduced in Section 3.1, TF-IDF-based keyword extraction and Word2Vec-driven clustering are applied to the reserved 2988 sentences. The extracted keywords are analyzed according to the corpus characteristics and the requirements of the NER task, and then organized into four semantic categories: process industry name layer, process technology layer, unit operation layer, and operating equipment layer.

The resulting hierarchical domain dictionary contains a total of 850 specialized terms, which are distributed as follows: 36 in the process industry name layer, 245 in the process technology layer, 337 in the unit operation layer, and 232 in the operating equipment layer. This dictionary plays a dual role in our framework, serving as both the foundation for semi-automatic dataset annotation and the external knowledge source for the HLA within the DDiNER model.

4.1.3 Datasets

To train the DDiNER model effectively, the corpus is annotated using the widely adopted BIO labeling scheme. Each character in the text is assigned one of three labels: B_X to mark the beginning of an entity of type X, I_X for characters inside the entity, and O for non-entity characters. Four entity types are defined based on the constructed dictionary: process industry name (NAM), process technology (TEC), unit operation (OPE), and operating equipment (EQU). Table 2 summarizes the entity classes, their semantic meanings, and label formats.

Table 2 Entity types and the corresponding labels

Type	Description	Label
NAM	Process industry name	B_NAM, I_NAM
TEC	Process technology	B_TEC, I_TEC
OPE	Unit operation	B_OPE, I_OPE
EQU	Operating equipment	B_EQU, I_EQU
O	Others	O

The annotated corpus is divided into training, validation, and test sets using an 8:1:1 ratio to ensure balanced representation across entity types. Detailed dataset statistics are summarized in Table 3.

Table 3 Dataset partitioning and entity distribution

Dataset	Number of distributions				
	Sentence	NAM	TEC	OPE	EQU
Training	9586	24 398	2681	585	262
Validation	1778	3046	363	74	36
Test	1709	2994	343	78	25

4.2 Evaluation metrics

To comprehensively evaluate the performance of the proposed DDiNER framework, three standard evaluation metrics are adopted: precision (P), recall (R), and the F1-score (F1). The true positive (TP) signifies the number of positively identified samples, true negative (TN) denotes the number of accurately predicted negative instances, false positive (FP) characterizes the number of erroneously predicted positive cases, and false negative (FN) represents the number of instances of misclassification of positive samples as negative.

Precision represents the proportion of correctly predicted positive samples out of all samples identified as positive. The calculation for P is expressed as

$$P = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}). \tag{19}$$

Recall is the percentage of all correctly predicted positive samples out of the total number of actual positive samples. The calculation for R is illustrated as

$$R = \mathrm{TP}(\mathrm{TP} + \mathrm{FN}). \tag{20}$$

F1-score serves as a comprehensive evaluation metric that considers both precision and recall. The calculation for F1 is presented as

$$\mathrm{F1} = (2PR)/(P + R). \tag{21}$$

4.3 Experiment settings

The experiments are conducted on a workstation equipped with Ubuntu 22.4, 32 GB RAM, an NVIDIA GeForce RTX 3080 Ti GPU (24 GB), and Python 3.8. The software environment includes the Transformers library (v3.4.0), PyTorch (v1.7.0+cu102), and other supporting dependencies.

The proposed DDiNER framework integrates three core components: (1) an HLA with BERT for lexicon-aware contextual representation learning; (2) a bidirectional LSTM network for sequential context modeling; (3) a CRF layer for globally optimized label decoding. To ensure stable and efficient training, each module is assigned an independent learning rate following the principles of differential optimization. Specifically, the HLA with BERT is fine-tuned with a learning rate of 1×10^{-6}, while the BiLSTM and CRF layers are trained with learning rates of 1×10^{-5} and 1×10^{-4}, respectively. This hierarchical learning rate strategy mitigates gradient vanishing in deeper layers and accelerates convergence in task-specific layers.

All experiments are conducted on our self-constructed process industry datasets. The maximum number of training epochs is set to 100; however, an early stopping strategy is applied, terminating training if the F1-score on the validation set fails to improve for three consecutive epochs. To further enhance generalization and mitigate overfitting, dropout regularization is incorporated into the BiLSTM layers, and weight decay is applied across all model parameters.

Finally, to ensure the robustness and reliability of DDiNER, stratified data sampling is performed to balance entity type distributions, particularly addressing long-tail categories within the process industry. To ensure robust and unbiased evaluation under limited and imbalanced data conditions, we employ a 10-fold cross-validation strategy, which allows each sample to be used for both training and validation while reducing variance in performance estimation.

4.4 Results

This subsection presents a comprehensive evaluation of the proposed DDiNER framework on the industrial NER task. We first report the main experimental results, followed by validation experiments on unregistered entities to assess generalization ability. Comparative experiments are then conducted against several SOTA and domain-specific NER models to demonstrate DDiNER's superiority. Finally, ablation studies are performed to quantify the contributions of individual components within the framework.

4.4.1 Main experimental results

Table 4 summarizes the overall performance of the proposed DDiNER model on the industrial NER dataset, evaluated under a 10-fold cross-validation setting. For each entity type, precision, recall, and F1-score are summarized using the mean (Mean) and standard deviation (Std) across folds to reflect effectiveness and stability.

The results show that DDiNER achieves a mean precision of 95.75%, a mean recall of 95.73%, and a mean F1-score of 95.74%, with low standard deviations of approximately 0.27, 0.28, and 0.26, respectively. These findings indicate that the model provides high recognition accuracy and strong consistency across different folds. Among the entity types, EQU exhibits the highest overall performance, achieving a mean F1-score of 98.07% with the lowest variance (Std = 0.18), demonstrating the model's robustness in accurately identifying equipment-related terms. NAM and OPE also achieve competitive performance, with mean F1-scores of 97.74% and 96.71%, respectively, reflecting DDiNER's ability to effectively handle these categories. In contrast, TEC shows a relatively lower mean F1-score of 93.52% and a slightly higher variance (Std = 0.35), highlighting the increased difficulty of accurately recognizing more diverse and context-sensitive technical expressions.

Table 4 Main experimental results. Mean and Std are computed over a 10-fold cross-validation setting

Entity type	Precision		Recall		F1-score	
	Mean (%)	Std	Mean (%)	Std	Mean (%)	Std
NAM	97.75	0.23	97.73	0.25	97.74	0.21
TEC	93.92	0.34	93.13	0.37	93.52	0.35
OPE	96.01	0.29	97.42	0.28	96.71	0.27
EQU	98.13	0.19	98.01	0.22	98.07	0.18
O	92.94	0.32	92.36	0.30	92.65	0.31
Average	95.75	0.27	95.73	0.28	95.74	0.26

4.4.2 Comparative experiments

To evaluate the effectiveness and robustness of DDiNER, we compare it with several SOTA Chinese NER models and representative domain-specific approaches on the constructed industrial dataset. All models are trained and evaluated under identical experimental settings, following the hyperparameter configurations recommended in their original studies. Performance is assessed using precision, recall, and F1-score, reported as the Mean and Std over a 10-fold cross-validation setting. An entity prediction is considered correct only when both its span and type exactly matched the gold annotation.

The baseline models are briefly summarized below.

1. Flat-lattice Transformer (FLAT) (Li XN et al., 2020): a lexicon-aware flat-lattice Transformer that encodes character spans via head–tail indices to capture absolute positions and handle overlapping entities.

2. LEBERT (Liu W et al., 2021): a lexicon-enhanced BERT that injects word features via adapter modules between Transformer layers to enrich character-level representations.

3. Fusion glyph network (FGN) (Xuan et al., 2021): a feature-gated network that fuses morphological cues and semantic features to enhance Chinese character representations for sequence labeling.

4. Longer entity attention (LEA) (Gong et al., 2022): a length-enhanced attention architecture tailored to long and complex entities through length-aware attention weighting.

5. In-context NER (Chen et al., 2023): a contextual-learning method leveraging pre-trained language models to recognize novel entity types with minimal task-specific supervision.

6. GPT-NER (Wang SH et al., 2023): a framework that leverages LLMs for NER by framing it as a prompted text

generation task with self-verification.

7. Data augmentation-BERT-BiLSTM-CRF (DA-BERT-BiLSTM-CRF) (Niu and Hou, 2024): a data-augmentation-based pipeline that combines BERT embeddings with BiLSTM-CRF decoding to improve generalization in low-resource scenarios.

Table 5 reports the comparative results. DDiNER attains the highest mean precision, recall, and F1-score, while also exhibiting the lowest fold-to-fold variance, indicating superior stability on complex industrial texts. Specifically, DDiNER outperforms all baselines in terms of mean F1-score, with a margin of 1.05 percentage points over the second-best LEA, and exhibits the lowest standard deviations across metrics, reflecting strong fold-level stability. These results suggest that integrating a hierarchical domain dictionary with a BERT encoder via an HLA, followed by BiLSTM-CRF decoding, is particularly effective for handling nested boundaries, polysemous terminology, and long-tail entities in process industry texts.

4.4.3 Ablation experiments

To examine the contribution of individual components in DDiNER, we conduct a series of ablation experiments by selectively removing key modules from the full architecture while keeping all other settings unchanged. We first perform module-level ablations by excluding the HLA, BERT encoder, BiLSTM layer, or CRF decoding layer. Performance is evaluated using the F1-score under the same experimental protocol.

As shown in Table 6, the full DDiNER model achieves an F1-score of 95.74%. Removing the HLA module results in a decrease to 93.54%, confirming the importance of lexicon-guided knowledge integration. Excluding the BERT encoder leads to the most significant performance drop (86.92%), highlighting the critical role of contextualized semantic representations. Removing the BiLSTM or CRF layer reduces the F1-score to 92.34% and 91.83%, respectively, indicating their contributions to sequential dependency modeling and structured prediction. To further assess the independent effects of hierarchical lexicon design and fusion mechanisms, we conduct fine-grained ablations following the module-level analysis. Specifically, we evaluate a variant without hierarchical organization (flattened lexicon) (92.21%), as well as variants removing inter-level attention (91.48%) and replacing bilinear attention with a simplified dot-product formulation (94.37%). The results show that removing hierarchical structures leads to a noticeable performance degradation, while simplifying the fusion mechanisms consistently reduces recognition accuracy.

4.4.4 Performance on challenging Chinese entity types

To further examine whether DDiNER effectively addresses Chinese-specific NER challenges, we evaluate its performance on several representative difficult entity types, including unseen entities, rare entities, long entities, and nested and overlapping expressions. These cases are particularly common in industrial Chinese texts due to implicit word boundaries, strong term compositionality, and evolving domain terminology.

First, we evaluate generalization on unseen entities using an independent dataset constructed from a production manual of a zinc smelting factory, where entity mentions do not appear in the training corpus. On this dataset, DDiNER achieves F1-scores of 95.32% for process industry name (NAM), 87.28% for process technology (TEC), 88.92% for unit operation (OPE), and 93.65% for operating equipment (EQU), with an overall average F1-score of 91.29%. Compared with in-domain test results, performance degradation is more evident for TEC and OPE, reflecting the increased difficulty of recognizing previously unseen and compositionally complex technical expressions. Nevertheless, the model maintains stable performance across all categories, indicating strong robustness to domain shift and vocabulary expansion.

To analyze rare and long entities, we further group entity instances according to their frequency and character length. Rare entities (occurring fewer than five times in the training set) achieve an average F1-score of 90.84%, while long entities consisting of five or more characters reach an F1-score of 92.17%. Although these scores are lower than those of frequent or short entities, the performance gap remains limited, suggesting that hierarchical lexicon guidance provides effective semantic cues even when surface forms are sparse or extended. Finally, we examine nested and overlapping entities, which are common in expressions such as zinc electrowinning process, where multiple semantic categories partially overlap. On manually identified nested entity cases, DDiNER achieves an F1-score of 89.76%. Error analysis indicates that most remaining errors arise from fine-grained boundary ambiguities between closely related process and operation terms, rather than complete entity misses.

Table 5 Comparative results on the industrial NER dataset. Mean and Std are computed over a 10-fold cross-validation setting

Baseline model	Precision		Recall		F1-score	
	Mean (%)	Std	Mean (%)	Std	Mean (%)	Std
FLAT	93.24	0.42	92.97	0.47	93.10	0.44
LEBERT	88.25	0.33	88.63	0.31	88.44	0.32
FGN	90.02	0.52	89.26	0.49	89.64	0.51
LEA	94.32	0.39	95.07	0.41	94.69	0.36
In-context NER	93.95	0.38	94.42	0.36	94.18	0.34
GPT-NER	94.15	0.35	94.03	0.37	94.09	0.33
DA-BERT-BiLSTM-CRF	89.63	0.48	89.31	0.45	89.47	0.46
DDiNER (our model)	95.75	0.27	95.73	0.26	95.74	0.26

Table 6 Ablation study results of DDiNER measured by F1-score

Model configuration	F1-score (%)
DDiNER (full model)	95.74
Without the HLA module	93.54
Without the BERT encoder	86.92
Without the BiLSTM layer	92.34
Without the CRF layer	91.83
Without hierarchy	92.21
Without inter-level attention	91.48
Without bilinear attention	94.37

5 Discussion

This study proposes DDiNER, a domain dictionary-guided Chinese NER framework in the process industry, addressing the challenges of entity boundary ambiguity, semantic overlap, and limited annotated data in industrial contexts. By integrating hierarchical lexical knowledge with contextual semantic representations, DDiNER demonstrates superior recognition performance, outperforming SOTA models across multiple evaluation metrics. The discussion below highlights the key innovations, practical implications, limitations, and future research directions.

5.1 Key innovations and contributions

This work centers on modeling domain knowledge in Chinese NER as a set of multi-granularity semantic concepts aligned with character-level representations. Instead of viewing lexicon entries merely as word candidates or auxiliary matching units, DDiNER organizes domain terminology into a hierarchical semantic structure and explicitly associates semantic concepts at different abstraction levels with each character position. This modeling center is particularly suited to industrial Chinese texts, where technical entities are often compositional, nested, and boundary-ambiguous, and where informative semantic cues may arise at different levels of granularity depending on context.

Within this framework, the proposed fusion strategies enable effective hierarchical knowledge integration. Inter-level attention supports selective information aggregation across semantic granularities, allowing higher-level conceptual cues and lower-level lexical details to be dynamically weighted according to context. Bilinear attention further enhances this process by modeling richer feature interactions between character representations and lexicon embeddings than simple dot-product operations. Together, these mechanisms ensure that hierarchical lexicon knowledge is selectively filtered and effectively incorporated into character-level representations.

From a methodological perspective, DDiNER adopts a hierarchical and intervenable knowledge integration paradigm that distinguishes it from representative lexicon-based Chinese NER models such as SoftLexicon (Ma et al., 2020), FLAT (Li XN et al., 2020), LEBERT (Liu W et al., 2021), and lightweight lexicon-enhanced Transformer (LLET) (Ji and Xiao, 2024). As summarized in Table 7, prior approaches generally incorporate lexicon information as flat auxiliary features, lattice nodes, or lightweight word-level enhancements, implicitly assuming uniform contributions across matched terms. In contrast, DDiNER encodes lexicon knowledge as structured semantic evidence and injects it into information flow through explicit intra- and inter-level interactions. By enabling higher-level semantic concepts to guide lower-level boundary detection and type inference, the hierarchical design embeds a task-specific inductive bias that reflects the compositionality and nested structure of industrial Chinese entities. Consequently, the performance gains observed in comparative experiments arise from this architecture-level inductive bias rather than from lexicon scale or feature accumulation alone.

5.2 Practical impact and deployment considerations

DDiNER offers a practical and scalable solution for Chinese NER in industrial settings, where accurate identification of process-specific entities is essential for tasks such as knowledge graph construction, safety monitoring, and intelligent decision support. By leveraging hierarchical domain dictionaries, the framework effectively bridges unstructured industrial texts and structured semantic representations, enabling robust extraction of domain knowledge for downstream applications including production optimization and fault analysis.

In real-world industrial environments, texts such as production logs, maintenance records, and operational manuals often exhibit noisy formatting, implicit domain assumptions, and continuously evolving terminology. In these scenarios, purely data-driven NER models may struggle with newly emerging or low-frequency entities. The hierarchical design of DDiNER mitigates this issue by allowing higher-level semantic concepts to guide entity boundary detection and type inference even when exact lexical matches are unavailable. This knowledge-guided mechanism improves robustness under domain shift and highlights the practicality of combining hierarchical lexicon modeling with incremental dictionary updates for long-term deployment.

5.3 Limitations and future work

Despite the strong performance of DDiNER, several limitations merit further investigation. First, the construction and maintenance of the hierarchical domain dictionary still require domain expertise, and updating lexicon entries as industrial terminology evolves remains a non-trivial task. Although the hierarchical design allows partial semantic matching at higher abstraction levels, more efficient strategies are needed to support scalable and timely dictionary refinement in dynamic industrial environments.

In addition, the current framework focuses on lexicon-level knowledge integration and does not explicitly model finer-grained internal structures of Chinese characters, such as radicals, glyph patterns, or sub-character components. Recent studies suggest that incorporating such information introduces non-trivial challenges in representation granularity, alignment, and fusion. Threshold analyses of Hanzi image and component recognition have shown that effective character and component modeling depends on carefully defined visual and structural boundaries across writing systems (Li WG et al., 2025). Multimodal approaches grounded in the six-writings

Table 7 Structure-level comparison between representative lexicon-based Chinese NER models

Model	Lexicon structure	Basic unit	Fusion strategy	Boundary assumption
SoftLexicon	Flat	Character	Pooling-based fusion	Implicit word boundary
FLAT	Lattice	Word span	Lattice position encoding	Explicit span modeling
LEBERT	Flat	Word adapter	Adapter-based additive fusion	Implicit word boundary
LLET	Flat	Character	Lightweight Transformer fusion	Implicit word boundary
DDiNER (ours)	Hierarchical	Character-centric	Intra-/Inter-level attention	Multi-granularity boundary

theory further demonstrate that pictophonetic and structural encodings can enhance Chinese language models, but require principled mechanisms to integrate visual, phonetic, and semantic cues (Weigang et al., 2024). Similarly, radical-based token representations and glyph-augmented pre-training methods have shown benefits for Chinese language modeling, while also highlighting the complexity of sub-character alignment and fusion (Sun et al., 2021; Qin et al., 2025). Building on these insights, our future work will focus on representing, aligning, and fusing fine-grained sub-character features of Chinese scripts, such as glyph and structural cues. By integrating these sub-character representations with DDiNER's existing hierarchical lexicon fusion framework, this direction may facilitate the development of a full-stack and multi-level Chinese semantic understanding model that spans sub-character, character, and lexicon-level concepts.

6 Conclusions

This paper presents DDiNER, a domain dictionary-guided Chinese NER framework in the process industry, explicitly addressing Chinese-specific challenges such as the absence of explicit word boundaries and the prevalence of long, compositional, and nested technical entities. The core modeling principle of DDiNER is to represent domain knowledge as multi-granularity semantic concepts aligned with character-level representations. Based on this principle, the framework organizes industrial terminology into a hierarchical structure and selectively integrates these semantic concepts into a BERT-based encoder through an HLA, followed by a BiLSTM-CRF decoder for structured prediction. Experimental results show that DDiNER achieves a mean precision of 95.75%, recall of 95.73%, and F1-score of 95.74%, consistently outperforming SOTA baselines. Additional evaluations on independent datasets further confirm its robustness and strong generalization ability, particularly for rare and unseen entities. Overall, DDiNER provides an effective and scalable solution for Chinese industrial NER and offers a solid foundation for downstream applications such as industrial knowledge graph construction and intelligent industrial analysis.

Acknowledgments

This work was supported by the Major Key Project of Pengcheng Laboratory (No. PCL2023A09), the National Natural Science Foundation of China (Nos. 62103207 and 52471291), and the Guangdong Basic and Applied Basic Research Foundation (No. 2023A1515240044).

Author contributions

Wei CUI designed the research. Ronghui LIU processed the data and drafted the paper. Xiaojun LIANG helped organize the paper. Weihua GUI revised and finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declaration on the use of generative AI tools

During the preparation of this work, the authors used ChatGPT to improve language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

An Y, Xia XY, Chen XL, et al., 2022. Chinese clinical named entity recognition via multi-head self-attention based BiLSTM-CRF. *Artif Intell Med*, 127:102282. https://doi.org/10.1016/j.artmed.2022.102282

Ashok D, Lipton ZC, 2023. PromptNER: prompting for named entity recognition. https://doi.org/10.48550/arXiv.2305.15444

Chen JW, Lu YJ, Lin HY, et al., 2023. Learning in-context learning for named entity recognition. Proc 61st Annual Meeting of the Association for Computational Linguistics, p.13661-13675. https://doi.org/10.18653/v1/2023.acl-long

Collobert R, Weston J, Bottou L, et al., 2011. Natural language processing (almost) from scratch. *J Mach Learn Res*, 12:2493-2537.

De S, Sanyal DK, Mukherjee I, 2025. Fine-tuned encoder models with data augmentation beat ChatGPT in agricultural named entity recognition and relation extraction. *Expert Syst Appl*, 277:127126. https://doi.org/10.1016/j.eswa.2025.127126

Devlin J, Chang MW, Lee K, et al., 2019. BERT: pre-training of deep bidirectional Transformers for language understanding. Proc Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p.4171-4186. https://doi.org/10.18653/v1/N19-1423

Ehrmann M, Hamdi A, Pontes EL, et al., 2024. Named entity recognition and classification in historical documents: a survey. *ACM Comput Surv*, 56(2):1-47. https://doi.org/10.1145/3604931

G V, Kanjirangat V, Gupta D, 2023. AGRONER: an unsupervised agriculture named entity recognition using weighted distributional semantic model. *Expert Syst Appl*, 229:120440. https://doi.org/10.1016/j.eswa.2023.120440

Gao C, Zhang X, Han MT, et al., 2021. A review on cyber security named entity recognition. *Front Inform Technol Electron Eng*, 22(9):1153-1168. https://doi.org/10.1631/FITEE.2000286

Geng RS, Chen YP, Huang RZ, et al., 2023. Planarized sentence representation for nested named entity recognition. *Inform Process Manag*, 60(4):103352. https://doi.org/10.1016/j.ipm.2023.103352

Gong S, Xiong X, Liu YF, et al., 2022. A Transformer-based longer entity attention model for Chinese named entity recognition in aerospace. 5th Int Conf on Advanced Electronic Materials, Computers and Software Engineering, p.348-355. https://doi.org/10.1109/aemcse55572.2022.00077

Hu Y, Chen QY, Du JC, et al., 2024. Improving large language models for clinical named entity recognition via prompt engineering. *J Am Med Inform Assn*, 31(9):1812-1820. https://doi.org/10.1093/jamia/ocad259

Hu Z, Ma XN, 2023. A novel neural network model fusion approach for improving medical named entity recognition in online health expert question-answering services. *Expert Syst Appl*, 223:119880. https://doi.org/10.1016/j.eswa.2023.119880

Hu ZT, Hou W, Liu XX, 2024. Deep learning for named entity recognition: a survey. *Neur Comput Appl*, 36(16):8995-9022. https://doi.org/10.1007/s00521-024-09646-6

Ji ZC, Xiao YL, 2024. LLET: lightweight lexicon-enhanced Transformer for Chinese NER. IEEE Int Conf on Acoustics, Speech and Signal Processing, p.12677-12681. https://doi.org/10.1109/ICASSP48485.2024.10448408

Kumar A, Starly B, 2022. "FabNER": information extraction from manufacturing process science domain literature using named entity recognition. *J Intell Manuf*, 33(8):2393-2407. https://doi.org/10.1007/s10845-021-01807-x

Li WG, Ramos RM, Brom PC, et al., 2025. Threshold study for Hanzi image recognition: defining character and component limits in Chinese, Japanese, and Korean script processing. *Int J Asian Lang Process*, 35(1):2450011. https://doi.org/10.1142/S2717554524500115

Li XN, Yan H, Qiu XP, et al., 2020. FLAT: Chinese NER using flat-lattice Transformer. Proc 58th Annual Meeting of the Association for Computational Linguistics, p.6836-6842. https://doi.org/10.18653/v1/2020.acl-main.611

Li ZZ, Feng DW, Li DS, et al., 2020. Learning to select pseudo labels: a semi-supervised method for named entity recognition. *Front Inform Technol Electron Eng*, 21(6):903-916. https://doi.org/10.1631/FITEE.1800743

Liu C, Yang SW, 2022. Using text mining to establish knowledge graph from accident/incident reports in risk assessment. *Expert Syst Appl*, 207:117991. https://doi.org/10.1016/j.eswa.2022.117991

Liu P, Guo YM, Wang FL, et al., 2022. Chinese named entity recognition: the state of the art. *Neurocomputing*, 473:37-53. https://doi.org/10.1016/j.neucom.2021.10.101

Liu W, Fu XY, Zhang Y, et al., 2021. Lexicon enhanced Chinese sequence labeling using BERT adapter. Proc 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int Joint Conf on Natural Language Processing, p.5847-5858. https://doi.org/10.18653/v1/2021.acl-long.454

Liu YH, Ott M, Goyal N, et al., 2019. RoBERTa: a robustly optimized BERT pretraining approach. https://doi.org/10.48550/arXiv.1907.11692

Lu QH, Li R, Wen A, et al., 2025. Large language models struggle in token-level clinical named entity recognition. *AMIA Annu Symp Proc*, 2024:748-757.

Ma RT, Peng ML, Zhang Q, et al., 2020. Simplify the usage of lexicon in Chinese NER. Proc 58th Annual Meeting of the Association for Computational Linguistics, p.5951-5960. https://doi.org/10.18653/v1/2020.acl-main.528

McCallum A, Li W, 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. Proc 7th Conf on Natural Language Learning at HLT-NAACL, p.188-191. https://doi.org/10.3115/1119176.1119206

Niu PY, Hou C, 2024. Named entity recognition in Chinese rice breeding questions based on text data augmentation. *Trans Chin Soc Agri Mach*, 55(8):333-343 (in Chinese). https://doi.org/10.6041/j.issn.1000-1298.2024.08.030

Prasanna KSL, Lokesh S, Chandramouli G, et al., 2024. BERT-QA: empowering intelligent question answering with NLP and entity recognition. 3rd Int Conf on Applied Artificial Intelligence and Computing, p.1006-1010. https://doi.org/10.1109/icaaic60222.2024.10575242

Qin H, Li M, Wang L, et al., 2025. A radical-based token representation method for enhancing Chinese pre-trained language models. *Electronics*, 14(5):1031. https://doi.org/10.3390/electronics14051031

Qiu QJ, Tian M, Huang Z, et al., 2024. Chinese engineering geological named entity recognition by fusing multi-features and data enhancement using deep learning. *Expert Syst Appl*, 238:121925. https://doi.org/10.1016/j.eswa.2023.121925

Qu XY, Gu YJ, Xia QR, et al., 2024. A survey on Arabic named entity recognition: past, recent advances, and future trends. *IEEE Trans Knowl Data Eng*, 36(3):943-959. https://doi.org/10.1109/tkde.2023.3303136

Seow WL, Chaturvedi I, Hogarth A, et al., 2025. A review of named entity recognition: from learning methods to modelling paradigms and tasks. *Artif Intell Rev*, 58(10):315. https://doi.org/10.1007/s10462-025-11321-8

Sun ZJ, Li XY, Sun XF, et al., 2021. ChineseBERT: Chinese pretraining enhanced by glyph and pinyin information. Proc 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int Joint Conf on Natural Language Processing, p.2065-2075. https://doi.org/10.18653/v1/2021.acl-long.161

Wang SH, Sun XF, Li XY, et al., 2023. GPT-NER: named entity recognition via large language models. Findings of the Association for Computational Linguistics: NAACL 2025, p.4257-4275. https://doi.org/10.18653/v1/2025.findings-naacl.239

Wang XZ, Li JH, Zheng Z, et al., 2022. Entity and relation extraction with rule-guided dictionary as domain knowledge. *Front Eng Manag*, 9(4):610-622. https://doi.org/10.1007/s42524-022-0226-0

Wang ZH, Chen HR, Xu G, et al., 2025. A novel large-language-model-driven framework for named entity recognition. *Inform Process Manag*, 62(3):104054. https://doi.org/10.1016/j.ipm.2024.104054

Weigang L, Marinho MC, Li DL, et al., 2024. Six-writings multimodal processing with pictophonetic coding to enhance Chinese language models. *Front Inform Technol Electron Eng*, 25(1):84-105. https://doi.org/10.1631/FITEE.2300384

Xuan ZY, Bao R, Jiang SY, 2021. FGN: fusion glyph network for Chinese named entity recognition. Proc 5th China Conf on Knowledge Graph and Semantic Computing: Knowledge Graph and Cognitive Intelligence, p.28-40. https://doi.org/10.1007/978-981-16-1964-9

Yang K, Yang ZW, Zhao SW, et al., 2024. Uncertainty-aware contrastive learning for semi-supervised named entity recognition. *Knowl-Based Syst*, 296:111762. https://doi.org/10.1016/j.knosys.2024.111762

Yu YQ, Wang YZ, Mu JQ, et al., 2022. Chinese mineral named entity recognition based on BERT model. *Expert Syst Appl*, 206:117727. https://doi.org/10.1016/j.eswa.2022.117727

Zhang H, Wang XY, Liu JX, et al., 2023. Chinese named entity recognition method for the finance domain based on enhanced features and pretrained language models. *Inform Sci*, 625:385-400. https://doi.org/10.1016/j.ins.2022.12.049

Zhang H, Dang YP, Zhang YZ, et al., 2024. Chinese nested entity recognition method for the finance domain based on heterogeneous graph network. *Inform Process Manag*, 61(5):103812. https://doi.org/10.1016/j.ipm.2024.103812

Zhong L, Wu J, Li Q, et al., 2024. A comprehensive survey on automatic knowledge graph construction. *ACM Comput Surv*, 56(4):94. https://doi.org/10.1145/3618295

Zhou JY, Ma ZL, 2025. Named entity recognition for construction documents based on fine-tuning of large language models with low-quality datasets. *Autom Constr*, 174:106151. https://doi.org/10.1016/j.autcon.2025.106151

Zhou WX, Zhang S, Gu Y, et al., 2023. UniversalNER: targeted distillation from large language models for open named entity recognition. https://doi.org/10.48550/arXiv.2308.03279