Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.37253

Domain Lexicon Enhanced Named Entity Recognition for Industrial Maintenance Field

Chen Cao

Department of Computer Science, University of Western Ontario, London, Canada
ccao55141@gmail.com

Abstract. With the increasing complexity of industrial maintenance scenarios, efficiently and accurately extracting key information from the text of maintenance work orders has become a critical challenge that needs to be solved in the field of maintenance operations. In order to help maintenance personnel identify the key information in the work order, this paper proposes a named entity recognition(NER) method based on a domain-lexicon-enhanced Bidirectional Encoder Representations from Transformers - Conditional Random Field (BERT-CRF). In the data processing stage, the dataset was divided into training, validation, and test sets using a fixed random seed of 42. For unstructured text data, this paper builds a NER model that integrates professional lexicon information in the industrial field to optimize the extraction effect of key entities. Complete the entry matching through the longest matching, concatenate and fuse the lexicon feature embedding with the Bidirectional Encoder Representations from Transformers (BERT) context vector and combine with the Conditional Random Field (CRF) layer learning label sequence constraints to realize the entity recognition of industrial maintenance work orders. The experimental results show that the Micro-F1 of the lexicon-enhanced BERT-CRF model reaches 93.57%, which is 1.98 percentage points higher than the BERT-CRF baseline, indicating that the domain lexicon can effectively improve the overall entity recognition performance on industrial maintenance work order texts.

Keywords: industrial maintenance work order, NER, domain lexicon, BERT, CRF

1. Introduction

With the continuous development of Industry 4.0 and the Industrial Internet, many enterprises have generated a large amount of data in the process of equipment operation and maintenance. These data are usually in semi-structured or unstructured form, covering technical texts such as equipment, components, maintenance operations, etc., which are an important basis for supporting maintenance analysis and decision-making [1]. Therefore, in the face of such a large and complex text data, it is crucial to extract valuable information from it efficiently. Traditional manual processing and classification of maintenance work orders are time-consuming, inefficient. Natural Language Processing (NLP), as an interdisciplinary discipline of computer science, artificial intelligence and linguistics, can automatically analyze text semantics and extract key information, which has been widely used in text analysis scenarios in the industrial field. Although NLP technology has been

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.37253

gradually applied to the analysis of industrial operation and maintenance text mining, existing research still faces practical problems such as the limitation of technical text data sets, the lack of consistency or integrity of maintenance records, and difficulties in processing specific technical languages in the field [1].

NER is a basic information extraction task in NLP. The core is to identify entities with specific meaning from the text and classify them into predefined categories. Common entity types include the names of people, place names, organization names, time, date and currency, etc., which are the core basis of text intelligent analysis. In the past research development, researchers have gradually developed NER from a method based on rules and lexicons to a method based on machine learning and deep learning, getting rid of the dependence on artificial feature construction, enhancing the ability of contextual information modeling, and improving the efficiency of identification. However, the study pointed out that NER in the vertical field still faces problems such as complex technical vocabulary, insufficient annotation data and limited cross-field generalization ability [2]. For example, the definition of some entities may be vague, the same entity may have different expressions, or some entities may be closely related to the context of the sentence [2].

With the development of deep learning, hybrid models and joint models based on neural networks have gradually become important directions for the research of NER. Although the deep learning model improves the ability of entity recognition, the recognition effect is highly dependent on the quantity and quality of the annotation corpus. In the scenario of lack of annotation data in the industrial field, the application of the model will still be limited [3]. In order to improve the recognition ability of technical terms, external domain dictionaries are introduced into BERT to provide vocabulary and boundary information [4], and at the same time, through the CRF layer modeling label sequence constraints [5]. Experiments on mining lifting equipment and fault texts show that in the industrial field, the performance of NER can be improved by introducing field-specific dictionaries through dictionary adapters [4]. This provides a relevant basis for the study of domain dictionary enhancement in industrial maintenance texts.

Based on the above research background, this paper takes the industrial operation and maintenance work order data set as the research object, constructs a BERT-CRF entity recognition model enhanced by domain lexicon, conducts research around the industrial maintenance work order scenario, verifies whether the domain lexicon can effectively improve the model recognition performance, and discusses the challenges and future research directions of NER in the industrial field.

2. Related work

In recent years, the NLP research for industrial maintenance work orders has adopted methods such as semantic and syntactic text similarity and subject modeling to predict corrective measures, identify major maintenance themes, and extract key technical terms [1]. Among them, the research work related to this article mainly includes two categories, one is the recognition of named entities of industrial maintenance texts. For example, the operation and maintenance knowledge graph is automatically generated from the maintenance work order. First, mark the work order text and define the industrial entity classification criteria, and then extract the entity from the repair text with spelling errors, abbreviations and out-of-order short texts based on the NER model. In the follow-up processing stage, the messy colloquial fault descriptions are standardized and corrected, and all entities in the same work order are associated with each other, and structured data such as entities, relationships, structured time, cost and location are stored in the knowledge graph [6]. The second is to improve the model's understanding and extraction ability of industrial terms in combination with

DOI: 10.54254/2755-2721/2026.37253

field lexicons. When the labeling data in the industrial maintenance field is scarce and the data is confidential, rely on the manual construction of maintenance professional lexicons and domain ontologies, and design fixed syntax matching rules to extract entities. Use the domain lexicon to do the coarse screening of entity candidate tokens, narrow the scope of text retrieval, and then transform the entity recognition task into a question and answer fine-tuning task, fine-tune the pre-training language model on a small amount of annotation data, and finally build a maintenance domain ontology, define entity categories and semantic relationships between entities, and unify conceptual specifications [7].

In terms of information extraction from industrial maintenance work orders, some researchers proposed MaintIE, a fine-grained annotation schema and benchmark for maintenance short texts, and constructed a high-quality fine-grained corpus annotated by multiple annotators. This corpus provides a data foundation for NER research in industrial maintenance [8]. In terms of lexicon-enhanced NER, the LEBERT model proposes a method for integrating lexicon knowledge into BERT. This method is designed for Chinese sequence-labeling tasks and introduces external vocabulary information into BERT through a lexicon adapter. The experimental results show that lexicon knowledge can improve Chinese NER performance [9]. Although LEBERT was mainly evaluated on Chinese sequence-labeling datasets and its lexicon-matching method cannot be directly applied to English industrial term matching, it provides a reference for introducing domain lexicon information into BERT to solve the industrial maintenance work-order entity recognition task. In addition, another low-resource Chinese NER study incorporating domain lexicon knowledge into BERT showed that the vocabulary knowledge provided by the lexicon and the contextual information learned by BERT improved entity recognition performance [4]. However, the applicability of this method to industrial maintenance text still requires further study.

In summary, existing research has proved the feasibility of NLP technology in some industrial text analysis, but there are still some shortcomings. First of all, industrial maintenance texts contain a large number of highly specialized terms, technical terms, abbreviations and non-standard expressions, which makes it difficult to use common NLP methods and difficult to adapt between different technical fields [1]. Secondly, the original maintenance work order is usually short, containing technical terms, abbreviations, spelling mistakes and non-standard expressions. Even after a certain standardized processing, the text still has the characteristics of limited context and dense domain terms. It may be difficult to fully identify domain proprietary entities by relying on pre-trained language models alone [6]. Finally, some existing studies mainly use lexicons as external screening tools, while the construction of lexicons, rules and domain ontologies depends on manual experience. Therefore, the use of rules and lexicons alone cannot dynamically distinguish entity boundaries and categories in combination with contextual semantics, and there are recognition problems [7]. Therefore, in the industrial maintenance work order scenario, the relevant research on combining domain lexicon knowledge with context representation models is still relatively limited.

In response to the above problems, this study focuses on the English text of real industrial maintenance work orders and builds a domain lexicon-enhanced NER framework. The framework uses domain lexicons to provide technical vocabulary and their boundaries, integrates lexicon a priori knowledge into the BERT model, and combines the transfer relationship between CRF learning labels to complete entity sequence prediction. By integrating lexicon knowledge enhancement and context semantic characterization, this research systematically explores the impact of domain lexicons on the performance of industrial maintenance text entity recognition.

DOI: 10.54254/2755-2721/2026.37253

3. Method

The text of industrial maintenance work orders usually contains a variety of professional information. Manual processing is time-consuming and error-prone, and the basic model is difficult to distinguish between low-frequency technical terms and the semantic characteristics of entities in the subdivision field. Therefore, this article introduces industrial lexicon on the basis of the BERT-CRF model, and provides technical terms and category information with the help of the lexicon. Uses BERT to excavate the contextual context characterization of tokens, and then completes the entity boundary recognition and entity category prediction through the CRF modeling of the dependency constraints between labels.

3.1. Model architecture

For the task of NER of industrial maintenance text, this paper builds a BERT-CRF model enhanced by domain lexicon. The model is mainly composed of four modules: domain lexicon matching, BERT context semantic coding, lexicon feature fusion and CRF sequence decoding. As shown in Figure 1, the model processing architecture includes repair text input, lexicon matching and BERT coding, feature fusion, CRF decoding and entity recognition result output.

When entering an industrial maintenance text: $X = (x_1, x_2, \ldots, x_n)$. The model first uses the domain lexicon to find technical terms such as equipment, status and maintenance activities in the text, and records the entity locations and categories that may correspond to these tokens. Secondly, input the original text into the BERT encoder to learn the meaning of each token in the current sentence. Then the information provided by the lexicon is integrated with the contextual information obtained by BERT, so that the model can refer to the industrial expertise in the lexicon while understanding the content of the sentence. The merged information calculates the scores of different BIO labels through the linear layer and then enters the CRF layer. CRF adjusts the prediction results according to the relationship between adjacent labels and outputs a complete BIO label sequence. Finally, according to the BIO label, the consecutive tokens are combined into a complete entity, and the category of the entity is determined, so as to obtain the entity recognition results of the industrial maintenance text.

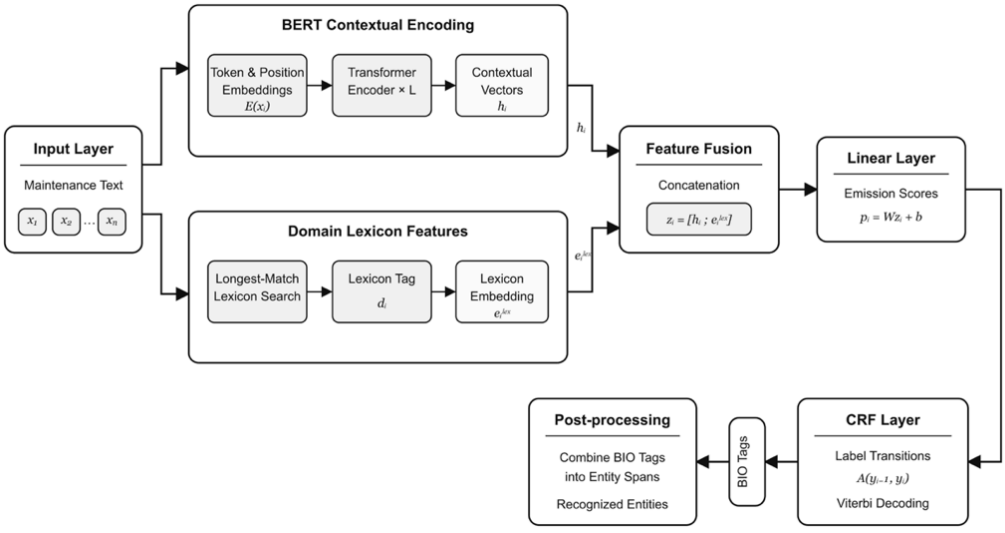

Figure 1. Domain lexicon-enhanced BERT-CRF model architecture (picture credit: original)

DOI: 10.54254/2755-2721/2026.37253

3.2. Lexicon construction and fusion

This paper builds a domain lexicon based on the five entity categories of PhysicalObject, State, Process, Activity and Property defined in the MaintIE corpus [8]. First of all, the lexicon entry only extracts the annotation entities of the training set, completely excludes the data of the validation set and the test set, and avoids the problem of data leakage. In the process of constructing the lexicon, normalize the entries, unify the case, delete blank characters and duplicate entries. The technical terms in the English industrial maintenance text may consist of multiple tokens, and there may be overlap between different entries. In order to resolve the conflict, this article adopts the longest matching strategy, giving priority to complete entries with greater coverage, and retaining industrial technical terms composed of multiple tokens as much as possible [10].

This paper refers to the research idea of integrating domain dictionary features into the NER model in existing studies, adopts lexicon label embedding and feature concatenation to introduce domain lexicon information into the proposed BERT-CRF model [11]. For the *i-th* token, the lexicon matching module generates a lexicon label d_i to record whether the token appears in the lexicon, and the entity position and category it may correspond to. For example, when pump bearing is matched as PhysicalObject, pump is marked as B-PhysicalObject, bearing is marked as I-PhysicalObject, and tokens that are not matched by the lexicon are labeled as O.Because the model needs to use numerical values for calculation, lexicon tags cannot be directly entered into the subsequent network. This paper converts the lexicon label of the token into a numerical dictionary feature vector through a trainable embedding layer. As shown in formula (1):

$$e_i^{lex} = E_{lex}[d_i] \tag{1}$$

i represent the position of the token in the input text, d_i represents the lexicon label corresponding to the i token, E_{lex} is the lexicon feature embedding matrix, and e_i^{lex} is the lexicon vector corresponding to the i token.

Then, the lexicon feature vector and the context vector output by BERT will be stitched together:

$$z_i = [h_i;\ e_i^{lex}] \tag{2}$$

In formula (2), h_i represents the context vector generated by the i token, and z_i represents the characteristics of the merged token.

The integrated features calculate the predicted score of each BIO label through the linear layer:

$$p_i = Wz_i + b \tag{3}$$

In formula (3), w and b are the weight matrix and bias vector of the linear layer respectively, and p_i represents the emission score of the i element corresponding to each BIO label, which is used as the input of the CRF layer.

3.3. BERT contextual encoding

At the same time, BERT is a pre-trained language model based on Transformer. It can read the content before and after a token at the same time and judge the meaning of the token in the current context according to the whole sentence [12]. Enter a text $X = (x_1, x_2, \ldots, x_n)$ composed of n tokens. After BERT coding, get the context vector corresponding to each token:

DOI: 10.54254/2755-2721/2026.37253

$H = (h_1, h_2, \ldots, h_n) = BERT(X)$. Among them, h_i represents the context semantic vector of the i -th token. Traditional token vectors generally provide the same representation for the same token, while BERT will generate different representations in combination with the sentence in which the token is located. Therefore, when the same token has different meanings in different maintenance texts, BERT can judge according to the tokens before and after it. When processing English texts, BERT will also split some uncommon tokens into several sub-tokens. For example, when an industrial term does not appear completely in BERT's token list, BERT can divide it into several parts that have been learned to obtain the basic representation of the term [12]. However, industrial maintenance texts contain more technical terms, and some terms appear less often in training data. It is difficult to fully capture the semantic characteristics of such proper nouns by relying only on BERT pre-training and fine-tuning [13]. Therefore, the addition of domain lexicons can supplement the model with industrial technical terms and entity category information.

3.4. CRF sequence decoding

After feature fusion and linear layer processing, the model will have a score of different BIO labels for each token. If the model only selects the label according to the score of each token, there may be results that do not meet the BIO rules. For example, in the absence of B-PhysicalObject, I-PhysicalObject appears directly after it, which means that the entity's internal label appears before the entity begins. Conditions can also consider the relationship between the label score of the current token and the label before and after, so as to adjust the unreasonable label combination [5].

For the candidate label sequence $Y = (y_1, y_2, \ldots, y_n)$, CRF calculates its overall score, which is equal to the sum of all emission scores plus the sum of all adjacent transition scores:

$$s(X, Y) = \sum_{i=1}^{n} P_{i,y_i} + \sum_{i=2}^{n} A_{y_{i-1},y_i} \tag{4}$$

As shown in formula (4), X represents the input text, n represents the number of tokens contained in the input text, and y_i represents the candidate label corresponding to the i token. P_{iyi} comes from the front linear layer output, indicating that the ith position corresponds to the emission score of the label y_i, and A_{y_{i-1},y_i} is the score of the previous label jumping to the current label.

Then, in the model reasoning stage, solve the label sequence with the highest score through the Viterbi algorithm, traverse all possible label sequences, and obtain the one with the highest score as the prediction output:

$$Y^* = arg\,max_Y\, s(X, Y) \tag{5}$$

In the formula (5), Y represents the collection of all possible label sequences corresponding to the input text X, s (X, Y) represents the overall score of the candidate label sequence Y, and Y* represents the optimal label sequence finally selected by the model.

Therefore, CRF not only considers the prediction results of a single token but also uses the connection between labels to improve the integrity of entity boundary prediction.

DOI: 10.54254/2755-2721/2026.37253

4. Experiments

4.1. Experimental setting

This experiment uses the Gold corpus of the real industrial maintenance work order MaintIE public data set, which is jointly completed by multiple field labelers. It contains a total of 1076 English maintenance work order short texts and provides annotation information such as entity boundaries and entity categories [8]. In the experimental preprocessing stage, first delete 2 completely duplicate texts, and finally 1,074 valid experimental data are obtained. This article maps entity labels into five entity categories: PhysicalObject, Activity, State, Process and Property. In response to the similar entity nesting problem in the data set, because the BIO annotation paradigm is not compatible with the nested structure, this article uniformly retains the entity span with a larger span and more complete semantics, and completes the standard BIO format conversion. The experimental fixed random seeds are 42. As shown in Table 1, the training set, validation set and test set are divided according to the ratio of 8:1:1. To avoid the entity information in the validation set and test set being used for model training in advance, all entries in the domain lexicon are only extracted from the training set annotation entities.

Table 1. Dataset partition and entity counts

Set	Texts	Tokens	PhysicalObject	Activity	State	Process	Property
Train set	859	4,917	1,357	612	312	116	28
Validation set	107	610	170	74	41	15	3
Test set	108	581	167	77	39	15	4

All experiments in this article are carried out in the central processing unit (CPU) environment, and the software versions are Python 3.9.23, PyTorch 2.8.0 and Transformers 4.57.6. Based on Python and PyTorch, the pre-trained encoder chooses bert-base-uncased. The experiment sets three groups of BERT, BERT-CRF and lexicon enhancement BERT-CRF models. The maximum sequence length of the model is set to 64, and the batch size is set to 16. This article adopts a layered learning rate strategy, pre-training BERT to set the learning rate of 2×10^{-5}, and randomly initialize new modules such as lexicon embedding, linear layer and CRF, which will use a larger learning rate of 1×10^{-4} to accelerate parameter convergence, the weight decay is 0.01, and the lexicon feature vector dimension is 32. The model uses the AdamW optimizer, and the maximum number of training epochs is set to 15 rounds. When the validation set F1 does not improve for 4 consecutive rounds, end the training early and save the highest model of the validation set F1. The experimental results are only correct when both the entity boundary and the entity category are predicted correctly. The main evaluation indicators include precision P, recall rate R and F1. This paper takes the Micro-F1 calculated by all entities as the main indicator, and outputs the F1 of each entity category and its average Macro-F1 to observe the recognition effect of the model on different entity categories.

4.2. Experimental process

This article first constructs a domain lexicon. The entries are extracted from the annotation entities of the training set. After lowercase, removing redundant blank characters, and deduplication, 790 domain entries are finally obtained. Among them, there are 552 PhysicalObject entries, 122 Activity

DOI: 10.54254/2755-2721/2026.37253

classes, 91 State classes, 19 Process classes, and 6 Property classes. After the lexicon construction is completed, multi-token terms use the longest matching to alleviate the conflict of long and short entries. When an original token is divided into multiple sub-tokens, this article uses the context characteristics of the first sub-token to align with the original BIO label. Then use the same experimental settings to train BERT, BERT-CRF and lexicon enhancement BERT-CRF three groups of models in turn. For the BERT model, input the context characteristics of BERT output into the linear layer, calculate the score of each token corresponding to each BIO label, and independently select the label with the highest score. The BERT-CRF model also first calculates the label score through the linear layer and then enters the results into the CRF layer to complete the sequence prediction without using domain lexicon characteristics. The lexicon enhancement BERT-CRF model adds an additional lexicon matching module in the coding stage. After inputting text tokens, the module performs lexicon matching for each token to generate lexicon labels and converts it into lexicon feature vectors through trainable embedding matrix. The lexicon vector and BERT context vector are stitched and integrated, the emission score of each BIO label is calculated by the linear layer, and then the constraint relationship between the learning labels is sent to the CRF layer, and the entity prediction results are output. The training process is set to stop early to prevent overfitting and finally use the highest model of F1 in the respective validation set to evaluate on the test set.

4.3. Experimental results

The running results of the three groups of models are shown in Table 2, and the Micro-F1 of the BERT model is 91.21%. After adding CRF, the Micro-F1 of the BERT-CRF model increased to 91.60%, an increase of 0.39 percentage points to prove that CRF can effectively model label transition constraints, correct local unreasonable annotations, and slightly improve the stability of sequence annotations. On this basis, after adding the domain lexicon, the precision, recall rate and Micro-F1 of the model are 93.11%, 94.04% and 93.57% respectively, which are 1.97, 1.99 and 1.98 percentage points higher than BERT-CRF. The result shows that the constructed domain lexicon can provide a priori information of industrial technical terms for the model, make up for the problem of insufficient characterization of professional terms in the vertical field of pre-training model, and improved the Micro-F1 on the test set to identify maintenance work order entities.

Table 2. Recognition results of different models (%)

Model	Precision	Recall	Micro-F1	Macro-F1
BERT	91.36	91.06	91.21	95.39
BERT-CRF	91.15	92.05	91.60	95.54
Lexicon-enhanced BERT-CRF	93.11	94.04	93.57	94.77

As shown in Table 3, after adding the domain lexicon, the F1 value of the PhysicalObject category increased from 86.31% to 91.02%, an increase of 4.71 percentage points. However, some categories decreased. The F1 value of the Activity category decreased from 98.06% to 97.44%, the Process category decreased from 93.33% to 86.67%, and the Macro-F1 of its model decreased from 95.54% to 94.77%. This is related to the small number of samples of some entity categories themselves, and the existence of polysemy or category conflicts in the lexicon may also interfere with the judgment of the model. First, there are few test samples in categories such as Process and Property, the data distribution is uneven, and the robustness of the model is limited. Second, there is a widespread phenomenon of multiple meanings and cross-category semantics in terms in the

DOI: 10.54254/2755-2721/2026.37253

industrial field, and the forced matching of lexicons will interfere with a small number of samples. It may further interfere with the model judgment.

Table 3. F1 scores by entity type (%)

Entity type	Test entities	BERT-CRF F1	Lexicon-enhanced BERT-CRF F1
PhysicalObject	167	86.31	91.02
Activity	77	98.06	97.44
State	39	100.00	98.73
Process	15	93.33	86.67
Property	4	100.00	100.00

In summary, the domain lexicon enhancement strategy can optimize the recognition effect of entities in the field of industrial maintenance, especially improve the recognition accuracy and integrity of core high-frequency entities, and effectively make up for the lack of knowledge of general pre-training models in the vertical field. However, limited by the sparse data of small sample categories and the strong ambiguity of domain terms, the lexicon enhancement still has slight interference with a small number of low-frequency entities.

5. Conclusion

In general, this paper studies the data characteristics of the industrial maintenance work order text, and constructs a lexicon in the field of industrial maintenance from the training set annotation entities, so as to improve the entity recognition effect of the maintenance work order text with the help of this lexicon. For the traditional BERT-CRF NER model, when processing industrial maintenance texts, in the face of the problem of insufficient entity recognition effect of a large number of technical terms, this paper proposes a lexicon to enhance the BERT-CRF NER model. The model stitches and integrates the domain lexicon features with the BERT context semantic features and uses the sequence constraints between the CRF layer learning labels to complete the recognition tasks of various entities in the industrial maintenance work order. Comparative experiments are carried out based on real and public industrial maintenance data sets. The experimental results show that the lexicon enhancement scheme improves the Micro-F1 on the test set of the model, which can provide reference for information extraction tasks under industrial low-resource scenarios.

However, due to the conditions of the data set, the experiment only uses English maintenance short text for verification. Lexicon entries are all derived from training set annotation entities, and the recognition of terms not covered by the lexicon is limited. At the same time, the unevenness of category samples and the appearance of low-frequency entity categories will affect the performance decline. Therefore, this model is more suitable for industrial operation and maintenance text analysis scenarios with some annotation corpus. The domain lexicon needs to be updated with business data to continuously adapt to new business terms. In terms of scalability, the current model effect is constrained by the quality of the lexicon and the ambiguity of the entry. Follow-up research can collect more real industrial maintenance work orders to improve the lexicon matching strategy. At the same time, it can expand the knowledge base of the external industrial field and reduce the model's dependence on manually labeling entities to build lexicons. It can also be considered to construct the entity extraction results into a knowledge graph in the maintenance field, realize work

DOI: 10.54254/2755-2721/2026.37253

order information retrieval, fault statistics analysis, and further expand the business value of industrial maintenance work order data.

References

[1] Sundaram, S., & Zeid, A. (2025). Technical language processing for prognostics and health management: Applying text similarity and topic modeling to maintenance work orders. Journal of Intelligent Manufacturing, 36(3), 1637–1657. https://doi.org/10.1007/s10845-024-02323-4.

[2] Ding, J., Li, W., Liu, X., & Chen, X. (2024). A review of named entity recognition research. Computer Engineering & Science, 46(7), 1296–1310.

[3] Nasar, Z., Jaffry, S. W., & Malik, M. K. (2022). Named entity recognition and relation extraction: State-of-the-art. ACM Computing Surveys, 54(1), 1–39. https://doi.org/10.1145/3445965.

[4] Dang, X., Wang, L., Dong, X., Li, F., & Deng, H. (2023). Improving low-resource Chinese named entity recognition using bidirectional encoder representation from transformers and lexicon adapter. Applied Sciences, 13(19), 10759. https://doi.org/10.3390/app131910759.

[5] Han, D., Wang, Z., Li, Y., Ma, X., & Zhang, J. (2024). Segmentation-aware relational graph convolutional network with multi-layer CRF for nested named entity recognition. Complex & Intelligent Systems, 10(6), 7893–7905. https://doi.org/10.1007/s40747-024-01551-8.

[6] Stewart, M., Hodkiewicz, M., Liu, W., & French, T. (2024). MWO2KG and Echidna: Constructing and exploring knowledge graphs from maintenance data. Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, 238(5), 920–932. https://doi.org/10.1177/1748006X221131128.

[7] Dixit, S., Mulwad, V., & Saxena, A. (2021). Extracting semantics from maintenance records. arXiv:2108.05454.

[8] Bikaun, T. K., French, T., Stewart, M., Liu, W., & Hodkiewicz, M. (2024). MaintIE: A fine-grained annotation schema and benchmark for information extraction from maintenance short texts. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, 10939–10951. https://aclanthology.org/2024.lrec-main.954.

[9] Liu, W., Fu, X., Zhang, Y., & Xiao, W. (2021). Lexicon Enhanced Chinese Sequence Labeling Using BERT Adapter. Proceedings of ACL-IJCNLP 2021, 5847–5858. https://aclanthology.org/2021.acl-long.454.

[10] Alves, D., Thakkar, G., & Tadić, M. (2025). UNER: Universal named-entity recognition framework. In I. Marenzi, S. Gottschalk, E. Müller-Budack, M. Tadić, & J. Winters (Eds.), Event Analytics across Languages and Communities. Springer, Cham. 3-15. https://doi.org/10.1007/978-3-031-64451-1_1.

[11] Chen, X., Ouyang, C., Liu, Y., & Bu, Y. (2020). Improving the Named Entity Recognition of Chinese Electronic Medical Records by Combining Domain Dictionary and Rules. International Journal of Environmental Research and Public Health, 17(8), 2687. https://www.mdpi.com/1660-4601/17/8/2687.

[12] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019, 4171–4186. https://doi.org/10.18653/v1/N19-1423.

[13] Du, P., Gao, X., & Xiang, Y. (2026). KG-APC: Knowledge graph-guided adaptive prototype correction for few-shot entity recognition in industrial maintenance information systems. Electronics, 15(15), 3275. https://doi.org/10.3390/electronics15153275.