IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received January 10, 2021, accepted April 9, 2021, date of publication April 13, 2021, date of current version April 23, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3072927

Compressor Fault Diagnosis Knowledge: A Benchmark Dataset for Knowledge Extraction From Maintenance Log Sheets Based on Sequence Labeling

TAO CHEN[1], **JIANG ZHU**[1], **ZHIQIANG ZENG**[1], **(Member, IEEE), AND XUDONG JIA**[1,2]

[1]Faculty of Intelligent Manufacturing, Wuyi University, Jiangmen 529020, China

[2]College of Engineering and Computer Science, California State University, Northridge, Northridge, CA 91330, USA

Corresponding author: Zhiqiang Zeng (zhiqiang.zeng@outlook.com)

This work was supported in part by the Natural Science Foundation of Guangdong Province of China under Grant 2020A1515011468, and in part by the Guangdong University Scientific Research Project under Grant 2019KTSCX189.

ABSTRACT Compressor fault diagnosis requires expert knowledge. Using the sequence labeling technology, this expert knowledge can be automatically extracted from compressor maintenance log sheets. Previous studies indicate that sequence labeling methods often need a substantial amount of annotation data for knowledge extraction, Unfortunately, the annotation data are very scarce in the field of compressor fault diagnosis. In this paper, we introduce a benchmark dataset for extraction of knowledge suitable for air compressor fault diagnosis. First, we collected 11,418 pieces of information from air compressor maintenance log sheets. Fault description, service requests, causes and troubleshooting solutions were stored in a dataset for data preprocessing and masking. In addition, 6196 valid text pairs were developed after the "noises" in the raw dataset were cleaned. Second, five kinds of entities and sequences, such as equipment, faults, service requests, causes and troubleshooting solutions, were annotated by three subject experts. The annotation consistency was assessed with F1 scores. Furthermore, our proposed baseline model (or the BERT-BI-LSTM-CRF model) was compared against other five sequence labeling models (BI-LSTM-CRF, Lattice LSTM, BERT NER, ZEN, and ERNIE). The BERT-BI-LSTM-CRF model gives superior performance in extracting expert knowledge from the subject dataset. Although the baseline model is not the most cutting-edge model in the sequence labeling and named entity recognition fields, it indeed presents a great potential for compressor fault diagnosis. The dataset is available at https://github.com/chentao1999/CFDK.

INDEX TERMS Compressor fault diagnosis, dataset, named entity recognition, sequence labeling.

I. INTRODUCTION

Faults diagnosis, as a key component in modern manufacturing processes, plays a significant role in the reliability and safety of modern industrial systems. As numerous monitoring data are generated and deep learning technology is increasingly applied in manufacturing systems, engineers and researchers have captured significant attention to data-driven fault diagnosis over the past decade. The current deep learning methods used for fault diagnosis are normally supervised and thus require expert knowledge (such as fault types, equipment information, and fault causes) in real world industrial applications. This expert knowledge varies among manufacturing systems. An effective fault diagnosis approach therefore would require a massive set of expert knowledge which can be extracted from various maintenance log sheets of manufacturing systems.

A maintenance log sheet is a text-based document describing failures and pertaining findings, as well as numerical indicators related to resource consumption [1]. During the lifetime of a system, operators and maintenance experts address a wide range of malfunctions which may lead to the loss of productivity or create safety hazards. The interactions between the operators/experts and the system often result in numerous pieces of field experience or knowledge which are hidden in log sheets. These pieces of expert knowledge can be

The associate editor coordinating the review of this manuscript and approving it for publication was Zhe Xiao.

automatically learned or extracted through natural language preprocessing (NLP) technology, especially sequence labeling (SL) technology.

Sequence labeling, such as part-of-speech tagging, chunking, and named entity recognition (NER), is a category of fundamental tasks in NLP [2]. NER, which gets more attention from engineers and researchers, aims to locate and classify named entities mentioned in unstructured texts into predefined semantic types such as person names, organizations, locations, time expressions, quantities, monetary values, and percentages [3]. Entities are usually composed of nouns or noun phrases, while sequences are usually composed of one or more clauses. Sequence labeling can recognize sequences which are generally longer than entities. For fault diagnosis as an example, the semantic types we focus on are most likely equipment, faults, service requests, causes, and troubleshooting solutions. The current dominant technique for addressing sequence labeling problems consists of deep learning approaches and models which extract knowledge with successful case studies or applications [4]. Most deep learning-based sequence labeling models require a significant set of big annotated data in training.

In this paper, we use the sequence labeling technology to extract expert knowledge from maintenance log sheets for compressor fault diagnosis. We introduce a new dataset entitled the Compressor Fault Diagnosis Knowledge (CFDK) dataset. To our knowledge, this dataset is the first large-scale sequence tagging mass collected for mechanical fault diagnosis. The CFDK dataset contains 6,196 text pairs with more than 36k annotated entities and sequences, namely, equipment, faults, service requests, causes, and troubleshooting solutions. Figure 1 shows a sequence labeling example within the CFDK dataset. It is noted that references to equipment are entities (nouns or noun phrases), while other references, such as fault, service request, cause, and troubleshooting solutions are sequences (clauses or sentences).

放污阀 与 承受 压力 不 相 符合 。 请 公司 设计室 ， 质检 部门 给予 落实 处理 ！

The drain valve is inconsistent with the pressure. Please ask design office and quality inspection department to deal with it!

排污阀 Q11F-16P 球阀，压力 等级 不够 。客户 领取 一件 排污阀 赔付，旧件 寄回 。

The pressure rating of **the Q11F-16P ball valve of drain valve** is not enough. Send one drain valve to the customer for compensation and core return .

FIGURE 1. A CFDK data record (Chinese) and its translation (English) documenting equipment (bold font), fault (blue font), service request (red font), cause (yellow font), and troubleshooting solutions (green font) in a reference.

We also propose a deep learning model based on Bidirectional Encoder Representations from Transformers (BERT) [5] and Bidirectional Long Short-Term Memory with a Conditional Random Field layer (BI-LSTM-CRF) [6] for sequence labeling on the CFDK dataset. BERT is well known for its unique feature which pre-trains deep bidirectional representations from unlabeled texts by jointly conditioning on both left and right contexts in all layers. As a result, pre-trained BERT models can be fine-tuned with just one additional output layer to create good results for a wide range of NLP tasks [5]. In this paper, we decide to use the BI-LSTM-CRF model to fine-tune the pre-trained BERT model, because the BI-LSTM-CRF model is proven to be efficient in using both past and future input features, as well as sentence-level tag information [6]. The experimental results on the CFDK dataset indicate that the proposed BERT-BI-LSTM-CRF model achieves F1 measures of 91.25%, 62.55%, 76.44%, 49.44% and 42.15% on the equipment NER, fault, service request, cause, and troubleshooting solution sequence labeling tasks, respectively.

The main contributions of our work are summarized below:

1) We introduce a new dataset[1] collected from a compressor plant and manually annotated the dataset with equipment, fault, service request, cause, and troubleshooting solution labels. To the best of our knowledge, the annotated dataset is the first large scale one which can support the development and evaluation of compressor fault diagnosis systems. This dataset can be considered as a benchmark dataset for compressor fault diagnosis. It is noted that benchmark datasets are critical to assessing various systems and models in manufacturing industries. Using maintenance log sheets to build up a scalable benchmark dataset supports an AI way of extracting expert knowledge and establishing a common and reproducible base to compare state-of-the-art knowledge acquiring methods and algorithms.
2) We have conducted extensive experiments on the proposed dataset and found that compressor fault knowledge can be acquired by the proposed BERT-BI-LSTM-CRF model. The model is proven to be effective in extracting expert knowledge from text-based log sheets, diagnosing potential faults, and providing solutions to the potential faults. This model is language independent and can be easily applied to extract knowledge from maintenance log sheets in other languages. Additionally, the model indicates that knowledge extraction from massive maintenance log sheets paves a new direction of acquiring and using expert knowledge for fault diagnosis.
3) At present air compressor maintenance log sheets are only stored in various companies, without any standardized specifications. We provide an annotation specification to build a benchmark dataset from these log sheets. This benchmark dataset will play a common reference in the future to assess feasibility and rationality of compressor fault knowledge extraction models. In the end, this annotation specification will help engineers and researchers prove that the use of air compressor maintenance log sheets in Chinese and any other languages for fault diagnosis is feasible and effective.

[1] Available at https://github.com/chentao1999/CFDK

The rest of this paper is organized as follows: Section 2 describes previous research work relative to compressor fault diagnosis and sequence labeling based expert knowledge extraction. With a good understanding of the state-of-the-art practice in knowledge extraction and fault diagnosis, we present our CFDK dataset in Section 3. Section 4 explains a set of sequence labeling methods and the BERT-BI-LSTM-CRF model with which the CFDK dataset is used to pre-train and fine-tune the expert knowledge extraction process. The experimental setup and evaluation results of the BERT-BI-LSTM-CRF model and other models are reported in Section 5. The paper is concluded in Section 6 by summarizing the contributions of the research work and outlining the future research directions.

II. LITERATURE REVIEW

Compressor fault diagnosis (CFD) is a part of intelligent fault diagnosis (IFD) which refers to applications of machine learning theories in machine fault diagnosis [7]. CFD approaches can be divided into traditional machine learning approaches and deep learning approaches. Traditional machine learning approaches apply machine learning theories, such as support vector machine (SVM) [8], artificial neural network (ANN) [9], hidden Markov model (HMM) [10], hybrid method [11] etc. The above research has achieved certain results, but due to the complex structures of compressors, the diagnostic performance of these methods is not ideal [12]. The diagnosis accuracy is a concern since these traditional machine learning methods are not applicable to the increasingly growing data which require high generalization [7]. Furthermore, the traditional machine learning approaches are labor-intensive and time-consuming. They often involve three steps, i.e., data collection and preparation, artificial feature extraction, and health state recognition.

In recognizing the drawbacks of the traditional machine learning CFD approaches, researchers have recently explored deep learning CFD approaches to automatically capture, to some extent, useful features from collected raw data, and achieved good performances in compressor fault diagnosis. According to the structures of neural networks used in feature extraction, deep learning CFD approaches can be divided into convolutional neural network (CNN) approaches [12], deep belief network (DBN) approaches [13], stacked auto-encoder (AE) approaches [14], self-attention network, ResNet approaches, etc. Among all these approaches, the CNN and DBN approaches prevail in CFD applications. For example, Guo *et al.* [15] proposed a one-dimensional convolutional neural network (1DCNN) based compressor fault diagnosis model which took the differential pressure and temperature of each compressor stage as the input of 1DCNN. Using the characteristics of the CNN, the model automatically extracted features and classified various faults. Experimental results showed that the fault recognition rate of 1DCNN was very high. In addition, Tran *et al.* [13] presented a DBNs based approach to implement vibration, pressure, and current signals for fault diagnosis of the valves in reciprocating compressors. The experiments on the signals from a two-stage reciprocating air compressor under different valve conditions demonstrated that the proposed approach was highly reliable and applicable in fault diagnosis of industrial reciprocating machinery.

Most of the deep learning approaches predict possible faults according to the monitoring data or vibration data generated when compressors are running. Some researchers use data mining to evaluate the values in maintenance records to identify hardware failures and significant energy losses [16], [17]. There is still a lack of methods of using the natural language processing technology to automatically identify expert knowledge from text documents and use expert knowledge for real time fault prediction, detection, and diagnosis. In this paper, we propose a benchmark dataset and a baseline model for mining compressor fault diagnosis knowledge from maintenance log sheets based on the natural language processing technology, such as sequence labeling and NER. The extracted expert knowledge will form a foundation for future real time compressor fault diagnosis.

Sequence labeling is commonly recognized as a basic area in empirical NLP and refers to the automatic extraction of mentions in texts and the formation of predefined semantic types such as person, location, organization, time, event, clinical procedure, biological protein, and adverse drug reactions mentions [18]. Approaches to SL and NER are broadly classified into four main streams [4]:

1) Rule-based approaches, which rely on hand-crafted rules derived from domain-specific gazetteers and syntactic-lexical patterns [19].
2) Unsupervised and bootstrapped approaches [20], which mainly rely on clustering or boosting-based algorithms to extract named entities from similar clustered groups in texts.
3) Traditional supervised learning approaches, which rely on annotated data samples, carefully designed features and supervised machine learning algorithms (such as HMM [21], decision trees [22], maximum entropy models [23], SVM [24], conditional random fields (CRF) [25]) to train a model to recognize similar patterns from unseen data.
4) Deep learning approaches, which save significant efforts on designing features. The models developed from deep learning approaches can be trained from raw inputs in an end-to-end paradigm [26], [27].

According to the context encoder architectures of neural networks, deep-learning SL or NER approaches can be divided into CNN approaches [28], recurrent neural networks (RNN) approaches [29], attention based approaches [30], [31], transformer and pre-trained model based approaches [5], [32], etc. BERT, as mentioned before, is a transformer and pre-trained model designed to pre-train deep bidirectional representations from unlabeled text and process words in relation to all the other words in a sentence. It considers the full context of a word by looking at the words that are before and after it. Devlin *et al.* [5] applied BERT to the CoNLL-2003 NER task [33] by extracting the activations from one or more layers without fine-tuning any parameters of the BERT model. The experimental results show that the BERT method performs competitively with other state-of-the-art methods. The advantage of the BERT method is its unsupervised pre-training ability in handling large scale unstructured text by various NLP tasks. The drawback of the BERT model is that it requires a lot of computing resources. In fine-tuning a BERT model, a large amount of annotated data is required.

Most of these approaches are supervised learning approaches, which need a large amount of labeled data. Although some approaches, such as transfer learning, unsupervised or semi-supervised approaches can reduce the dependence on labeled data, they are not the mainstream approaches in practical applications. Therefore, it is necessary to develop a sequence labeling benchmark dataset, especially in the field of compressor fault diagnosis, for supervised learning.

III. CFDK DATA COLLECTION, PREPARATION, AND DATASET

In this section, we describe the process used to collect, prepare, and annotate the CFDK dataset.

A. DATA COLLECTION AND PREPARATION

A total of 11,418 raw records which documented the history of handling reciprocating compressors in maintenance log sheets from CIMC Intelligent Technology Co., Ltd. (whose parent company is a world leading supplier of logistics and energy equipment) were collected to build the CFDK dataset. These records were manually entered by engineers during the maintenance of compressors. Each record includes 45 columns, namely primary key, customer request information, record time, device information, operator information, fault description, cause analysis, troubleshooting solutions, work order number, etc. Most of them are in Chinese. The maintenance log sheets were saved in CSV format.

The raw records contained many pieces of incomplete, incorrect, coarse, abnormal or noise information about maintenance cases of compressors. The "noises" had to be pretreated and the information was required to be filtered before it was considered as "knowledge". The first thing we did was to remove noises by extracting useful information from the raw data. After consulting with domain experts in the CIMC Corporation, we selected 4 columns, namely fault description, service request, cause, and troubleshooting solution. Considering the fact that the fault description and cause fields were empty in some records and the service request field contains information about fault description, the research team thus combined the service request information and the fault description in the same record into one text. Following the same procedures, the team also merged the cause and the troubleshooting solution into another text. As a result, two texts were integrated to form a text pair, separated by a new line character.

The set of text pairs was further cleaned to ensure that each text pair fully contains necessary information for later use in the sequence labeling models. For example, if the first text of a text pair was empty or did not contain any meaningful information, the entire text pair was removed from the set. If the second text was empty or lacked any meaningful information, the text pair was kept. However, the second text was replaced by "None". The set of text pairs also went through a privacy check. Considering the privacy of people or companies, the research team replaced all the actual names of people and companies in the corpus by the words "person" and "company", respectively. At end, a total of 6196 text pairs were created.

A Chinese word segmentation system[2] was executed over the desensitized corpus or the text pairs and a set of segmented words for the CFDK dataset were generated. The segmented words were further proofread manually, and the wrong segmentations were fixed.

It is noted that the above data collection and preparation can be conducted if maintenance log sheets are in other language. The only required language-specific software is the one for segmentation of words. NLTK[3] is a popular one for English word tokenization.

B. ANNOTATIONS ON THE CFDK DATASET

Before we started the annotations on the CFDK dataset, we did a literature review of NER annotations. We found that there have not been any applicable annotation specifications available for named entities or sequence labeling in the compressor fault diagnosis field. Most NER annotation specifications have been focused on common entity types, such as people, organizations, and locations [33], [34]. Some specifications have been related to symptoms, drugs, diseases, and other entities in the context of clinical applications, geopolitical entities in political science, and financial entities in business products. We developed our own specifications in this study for the CFDK applications. We believe that these specifications (described below) can be considered as the benchmark specifications:

1) Entities pertaining to equipment are primarily compressors, compressor components, and related supporting parts, such as crude oil engine, air supply line, etc.
2) Multiple devices in a series are annotated as one equipment entity. For example, "压缩机风机电机" (or "compressor's fan motor") is annotated as one device rather than three.
3) Devices with brackets are annotated as one equipment entity. For example, "安全阀(二级安全阀)" (or "safety valve (secondary safety valve)") is annotated as one device rather than two.
4) Individual device numbers or models are annotated as one equipment entity. A device followed with an equipment number or model is annotated as

[2]Jieba was used for segmentation: https://github.com/fxsjy/jieba
[3]Available at https://www.nltk.org/

TABLE 1. Statistics of the final corpus.

	Equipment	Fault	Request	Cause	Solution
Annotator 1	19,419	5526	6736	1989	4336
Annotator 2	19,145	5861	6562	1569	2470
Annotator 3 (final corpus)	19,218	5909	6595	1588	2786
$N_{a_1\&a_2}$	16,893	3051	4571	836	1258
$N_{a_1\&a_3}$	16,759	3072	4587	856	1399
$N_{a_2\&a_3}$	19,074	5716	6545	1532	2442
$F1_{a_1\&a_2}$	0.876	0.536	0.687	0.470	0.370
$F1_{a_1\&a_3}$	0.868	0.537	0.688	0.479	0.393
$F1_{a_2\&a_3}$	0.994	0.971	0.995	0.971	0.929

one equipment entity. Equipment number or model followed by a device is also annotated as one equipment entity. For example, "0107001", "v457", "v457 压缩机" ("v457 compressor" in English) and "冷却器14151002" (or "cooler 14151002" in English) each is annotated as one device, respectively.

5) Faults, service requests, causes, and troubleshooting solutions are independent text sequences which are not overlapping each other. When a fault description is embedded in a service request, the fault description is considered for annotation. For example, within the sentence of "用户需要公司派人前往处理漏气问题" (or "the user needs the company to send someone to deal with the air leakage problem" in English), the "漏气问题" (or "air leakage problem") is annotated as a fault description, and the whole sentence is not annotated as a service request.
6) Equipment entities may exist in the four types of sequences. In other words, equipment entities can be nested within the sequences of faults, service requests, causes, and troubleshooting solutions.
7) Fault description and service request always appear in the first part of a text pair. Cause and troubleshooting solution are always in the second part of the text pair.
8) Questions from customers are not annotated as service requests. Service requests with more than 100 words are annotated as multiple service requests. Fault descriptions, causes, and troubleshooting solutions follow the same rule.

The annotations on the CFDK dataset were completed by three expert annotators in this research following the above mentioned specifications and using a sequence labeling annotation tool.[4] The process of annotation involved two iterations. In the first iteration, the corpus was annotated by two annotators independently. In the second iteration, the third annotator integrated the results of the preceding iteration to create the final corpus. Inter-annotator consistency tests were performed to ensure that the annotation results have acceptable reliability. F1 score was used to evaluate the consistency of two annotators. The F1 score is defined as follows:

$$Precision = \frac{N_{a_1\&a_2}}{N_{a_1}} \tag{1}$$

$$Recall = \frac{N_{a_1\&a_2}}{N_{a_2}} \tag{2}$$

$$F1_{score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

where $N_{a_1\&a_2}$ refers to the number of identical annotations from both annotator a_1 and annotator a_2; N_{a_1} and N_{a_2} refers to the number of annotations from annotator a_1 and annotator a_2, respectively.

Table 1 shows statistics of the final corpus and the annotation consistency between any two annotators. Where $F1_{a_i\&a_j}$ refers to the F1 score calculated based on the annotation results of annotator a_i and a_j. It is noted from the table that the consistency of annotator 1 and 2 is not high, especially in the cause and troubleshooting solution sequences and both F1 scores are less than 50%. Annotator 3 carries out the final integration on the annotations from the former two. The consistency with annotator 2 is very high, but that with annotator 1 is low.

It is also noted from Table 1 that the annotation consistency ($F1$) of equipment is very high (more than 86%) for all three annotators, and that of faults, service requests, causes, and troubleshooting solutions is relatively low. One possible reason is that the equipment entities are nouns or noun phrases, while the other four types of sequences are clauses or sentences. The average length of equipment is much shorter than that of the other four types of sequences. In the final corpus, the average length of equipment, faults, service requests, causes and troubleshooting solutions is 4.57, 12.64, 18.28, 17.09, and 17.84 characters, respectively.

Figure 2 shows a portion of the annotated CFDK dataset in the BioNLP'11 shared task standoff format[5] (or the tab-delimited format) for the text pair. It is noted that each line in Figure 2 contains one annotation (with a given ID, a TAB character, and the body of the annotation) extracted from the CFDK data record as shown in Figure 1. The body of the annotation is further formed by a SPACE-separated triple combination (that is, sequence type, starting position of sequence, and the ending position of sequence). For example, the second annotation (as shown on the third line) consists of T2 (or the annotation ID), a TAB character, and the body of the annotation. The body of annotation is made of the fault sequence type, SPACE characters, 1, and 19. The English

[4]Available at https://brat.nlplab.org

[5]http://2011.bionlp-st.org/home/file-formats

T1	设备 1 4 (Equipment)	放污阀 (Drain valve)
T2	故障 1 19 (Fault)	放污阀 与 承受 压力 不 相 符合 (The drain valve is inconsistent with the pressure)
T3	需求 22 47 (Demand)	请 公司 设计室， 质检 部门 给予 落实 处理 (Please ask design office and quality inspection department to deal with it)
T4	设备 50 65 (Equipment)	排污阀 Q11F-16P 球阀 (Q11F-16P ball valve of drain valve)
T5	原因 50 75 (Cause)	排污阀 Q11F-16P 球阀， 压力 等级 不够 (The pressure rating of the Q11F-16P ball valve of drain valve is not enough)
T6	方案 78 100 (Solution)	客户 领取 一件 排污阀 赔付， 旧件 寄回 (Send one drain valve to the customer for compensation and core return)

FIGURE 2. A portion of the CFDK dataset in the BioNLP'11 shared task standoff format.

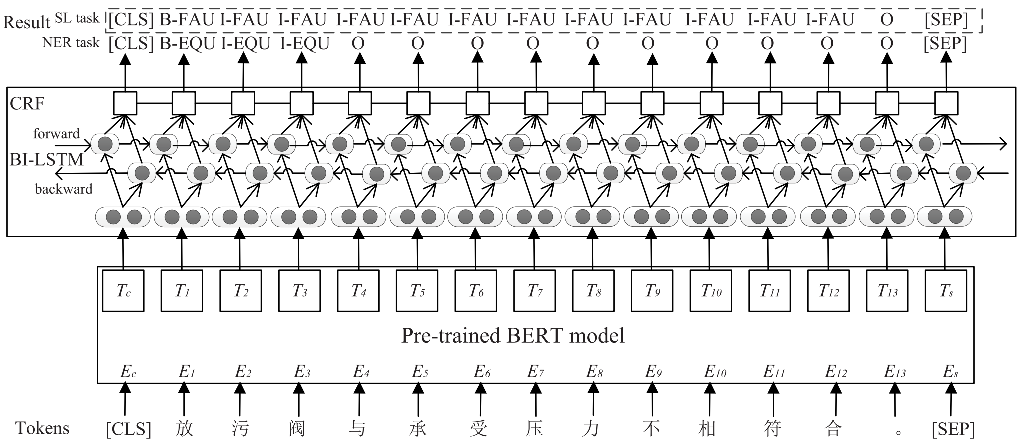


FIGURE 3. The architecture of the baseline model or the BERT-BI-LSTM-CRF model. The dotted line at the top of the figure indicates that the NER task and the sequence labeling tasks for fault, request, cause, and solution (or the SL tasks) are carried out separately.

translation in brackets is not part of the annotation. In this study, a total of 19218 equipment entities, 5909 fault descriptions, 6595 service requests, 588 fault causes, and 2786 troubleshooting solutions were created and stored in 310 separate files.

IV. BASELINE MODEL

Deep learning based sequence labeling models have become dominant and successively advanced their state-of-the-art performance for several years [6], [28], [29]. Recently, pre-trained models, such as ELMo [35], ULMFiT [36], BERT [5], GPT-3 [32] followed by fine-tuning on NER tasks have demonstrated substantial gains [37]. In this paper, we build our baseline model by using BERT for pre-training and BI-LSTM-CRF for fine-tuning the knowledge extraction process from the CFDK dataset.[6] An overview of the architecture of the baseline model is shown in Figure 3.

It is noted from Figure 3 that the pre-trained BERT module, siting at the bottom of the baseline model, uses a token sequence as input. A special classification token [CLS] is inserted at the beginning of the text sequence and a [SEP] token is inserted at the end of the text sequence [5].

[6]The source code for the baseline model (or the BERT-BI-LSTM-CRF model) is available at: https://github.com/yumath/bertNER

The pre-trained BERT module has a multi-layer bidirectional transformer [38] encoder stack. In each layer of the stack, the number of transformer encoder nodes is equal to the length of the input tokens. Each node fully connects with every transformer encoder node in the upper layer. The final hidden vector for the i^{th} input token (or T_i) is used as the input to the BI-LSTM-CRF model.

The BI-LSTM-CRF module is further divided into two sub-modules: the BI-LSTM network and the CRF network. The BI-LSTM network utilizes a bidirectional LSTM network to capture both left-to-right and right-to-left context features. The CRF network is a statistical modeling method which makes use of context information in predicting targeted tags. The BI-LSTM-CRF module is the most popular sequential labeling model before the rise of deep learning. The model outputs the most probabilistic BIO tag for each input token.

In this study, the fault entity recognition task (or the NER task in Figure 3) and the sequence labeling tasks for fault, service request, cause, and troubleshooting solution (or the SL task in Figure 3) were carried out separately. The CFDK dataset was converted from the BioNLP'11 shared task standoff format into the CoNLL-2002 shared task BIO format [39]. In the BIO format, all the input tokens are presented one Chinese character per line, while the end of the token sequence is presented by an empty line. Additionally, each

放 B-EQU	放 B-FAU
污 I-EQU	污 I-FAU
阀 I-EQU	阀 I-FAU
与 O	与 I-FAU
承 O	承 I-FAU
受 O	受 I-FAU
压 O	压 I-FAU
力 O	力 I-FAU
不 O	不 I-FAU
相 O	相 I-FAU
符 O	符 I-FAU
合 O	合 I-FAU
。 O	。 O

FIGURE 4. Two examples of the CFDK dataset in the CoNLL-2002 shared task BIO format.

line contains a tag which states whether its related token is inside a named entity or not. The tag also encodes the type of the named entity. Figure 4 is an example token sequence in the BIO format. Each Chinese character in the example sequence is tagged with other (O) or one of five entity/sequence types: Equipment (EQU), Fault (FAU), Service Request (REQ), Cause (CAU), and Troubleshooting Solution (SOL). The B-tag and I-tag indicate the beginning and the intermediate token of a named entity, respectively, while the O-tag indicates that this token is not part of a named entity.

Considering that equipment entities can exist in the sequence of fault, request, cause, and solution and the baseline model (or the BERT-BI-LSTM-CRF model) cannot deal with overlapping labels, we stored all the fault entities in the CFDK dataset in one BIO file and the other four sequences in another BIO file. The back-propagation through time (BPTT) [40] was used in this study to train the BI-LSTM-CRF model. The training error was back-propagated to fine-tune the parameters of the BERT module.

The baseline model, at its testing stage, produces the BIO label of each token in the input sequence. Take "放污阀与承受压力不相符合。" (or "The drain valve is inconsistent with the pressure." in English) as example, the equipment entity recognition task of the baseline model generates the BIO sequence of "B-EQU I-EQU I-EQU O O O O O O O O O O", while the fault sequence labeling task creates the BIO sequence of "B-FAU I-FAU I-FAU I-FAU I-FAU I-FAU I-FAU I-FAU I-FAU I-FAU I-FAU I-FAU O".

V. EXPERIMENTS

In this section, we describe the performance of different sequence annotation methods on the CFDK dataset. Other common datasets (such as CoNLL2003, BC2GM, CoNLL2000, PTB POS, Cora, etc.) could be used for assessing the performance of the baseline model against other sequence annotation models. However, these datasets are not specific to compressor fault diagnosis. We thus concentrate our focuses on the benchmark dataset and develop a most feasible and effective knowledge extraction model for fault diagnosis.

A. EXPERIMENTAL SETTINGS

A total of 6196 text pairs within the CFDK dataset were randomly divided into a training set and a test set. The statistics of the two sets are shown in Table 2.

TABLE 2. Statistics of the training and test sets for the baseline model.

	Training set	Test set
#text pairs	4,960	1,236
#sentences	17,832	4,663
#characters	410,907	118,175
#equipment entities	14,943	4,275
#fault sequences	4,795	1,114
#request sequences	5,114	1,481
#cause sequences	1,199	389
#solution sequences	2,290	496

Precision (P), recall (R), and $F1$ measure are computed as follows to evaluate the performance of the sequence labeling and NER tasks.

$$P = \frac{TP}{TP + FP} \tag{4}$$

$$R = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{6}$$

where TP refers to entities or sequences that are correctly identified; FP refers to the entities or sequences that are wrongly identified by the sequence labeling task or the NER task; FN refers to the entities or sequences not correctly identified in the test set. All of these metrics are calculated in terms of the whole entity or sequence, rather than the label of a single character. It is noted that $F1$ measure in Eq (6) and $F1$ score in Eq (3) are same in terms of its structure, but they represent different context.

The training and test sets were further used in this study to compare the baseline model against the following five sequence labeling models for compressor fault diagnosis knowledge extraction. These models are conventional methods and have achieved good results in NER or SL fields.

- BI-LSTM-CRF [6]: It combines a bidirectional LSTM network and a CRF network for sequence labeling. It relies on the bidirectional LSTM to keep track of dependencies in input sequences by efficiently working with both past and future input features or entities. It also takes advantage of the CRF network to do label predictions through its probabilistic graphical model.[7]
- Lattice LSTM [41]: It encodes a sequence of input characters as well as all potential words matching a lexicon. The application of this model for Chinese NER indicates that Lattice LSTM does not suffer from segmentation errors when it is compared with other word-based methods.[8]

[7] Source code is available at https://github.com/Determined22/zh-NER-TF

[8] Source code is available at https://github.com/jiesutd/LatticeLSTM

TABLE 3. System results of equipment NER task and SL task for fault, request, cause, and solution sequence (% is omitted for conciseness. The best results are highlighted in boldface and the second best results are in italics).

Model	Metric	Equipment	Fault	Request	Cause	Solution
BI-LSTM-CRF	Precision	83.67	53.80	*76.34*	44.42	*36.63*
	Recall	85.72	52.59	70.03	45.90	37.57
	F1	84.68	53.19	73.05	45.15	37.10
Lattice LSTM	Precision	88.03	52.22	73.94	47.20	33.27
	Recall	89.25	51.63	73.83	43.35	38.03
	F1	88.64	53.19	73.89	45.19	35.49
BERT NER	Precision	**89.99**	57.95	70.95	44.04	33.63
	Recall	**93.08**	60.34	**79.85**	**50.52**	**45.69**
	F1	**91.51**	*59.12*	*75.14*	*47.06*	*38.74*
ZEN	Precision	83.33	46.34	63.05	44.93	27.41
	Recall	87.32	55.28	70.49	42.27	33.67
	F1	85.28	50.42	66.56	43.56	30.22
ERNIE 1.0	Precision	85.62	46.47	64.43	*47.44*	24.78
	Recall	88.89	*62.34*	*75.61*	43.29	*44.33*
	F1	87.22	53.25	70.72	45.27	31.79
BERT-BI-LSTM-CRF	Precision	*89.69*	**61.56**	**77.38**	**49.62**	**44.18**
	Recall	*92.87*	**63.57**	75.52	*49.25*	40.30
	F1	*91.25*	**62.55**	**76.44**	**49.44**	**42.15**

- BERT NER [5]: It places a full connection layer on the top of the final hidden layer of the BERT encoder to fine-tune the pre-trained model for NER and sequence labeling tasks.[9]
- ZEN [42]: It is a BERT-based Chinese text encoder enhanced by N-gram representations, where different combinations of characters are considered during training. The potential word or phrase boundaries are explicitly pre-trained and fine-tuned with the BERT's character encoder.[10]
- ERNIE 1.0 [43]: It uses named entity-level and phrase-level masking strategies to enhance language representation models. Entity-level strategy masks entities which are usually composed of multiple words. Phrase-level strategy masks phrases and combines words (which stand together in each phase) to form a conceptual unit.[11]

A single NVIDIA GeForce GTX 2080Ti GPU with 11 GB of RAM was employed for training the above models. For the BERT pre-trained purpose, we used the "chinese_L-12_H-768_A-12"[12] model for the Chinese NER and sequence labeling tasks. This model has 12 layers, 768 hidden nodes, 12 heads, and 110M parameters. For the fine-tuning purpose, we set the maximum sequence length and the training batch size to be 128 and 32, respectively. For other model parameters, we adopted the default settings provided from the BI-LSTM-CRF and BERT specifications.

B. MODEL COMPARISONS

Precision, recall, and $F1$ measure of the five benchmark models and the BERT-BI-LSTM-CRF model (proposed by this study) are shown in Table 3. These metrics are the system results obtained from the equipment NER task and the SL tasks for fault, request, cause, and solution sequences. It is noted that the baseline model or the BERT-BI-LSTM-CRF model has the best performance.

The BERT-BI-LSTM-CRF model, in comparison with BERT NER model, improves the F1 measure by 3.38%, 1.30%, 2.38%, and 3.41% on the SL task for the fault, request, cause, and solution sequences, respectively. This implies that using the BI-LSTM-CRF module to fine-tune the BERT pre-trained module is effective in sequence labeling for CFDK extraction. With regards to the equipment NER task, the BERT-BI-LSTM-CRF model achieves its F1 measure of 91.25%, which is 0.26% lower than the best F1 measure in Table 3 and outperforms the other four models by a good margin. It indicates that using the pre-trained BERT module can improve the performance of NER. This also demonstrates that using the BI-LSTM-CRF module to fine-tune the BERT pre-trained module may not necessarily improve the performance of the baseline model. One reason for this finding may be that the named entities within the CFDK dataset are typically shorter than the sequences. In the CFDK dataset, the average length of the equipment entities is only 4.57 characters, while the average length of the fault, request, cause, and solution sequences is 12.64, 18.28, 17.09 and 17.84 characters, respectively. This is also the reason why the precision, recall, and F1 measure of the models in the equipment category are much higher than the findings for the fault, request, cause, and solution categories. For all the models in Table 3, the shorter the sequence length, the better the effect of the models. Compared with the other five models, the BERT-BI-LSTM-CRF model performs better on longer sequences.

The BERT-BI-LSTM-CRF model, in comparison with the traditional deep learning NER model, namely the BI-LSTM-CRF model, improves the F1 measure by 6.57%, 9.36%, 3.39%, 4.29%, and 5.05% on all the five tasks, respectively.

[9] Source code is available at https://github.com/xuanzebi/BERT-CH-NER
[10] Source code is available at https://github.com/sinovation/ZEN
[11] Source code is available at https://github.com/PaddlePaddle/ERNIE
[12] https://storage.googleapis.com/bert_models/2018_11_03/chinese_L-12_H-768_A-12.zip

This indicates that using the pre-trained BERT model can improve the performance of the traditional deep learning NER model.

When we compare the BERT-BI-LSTM-CRF model with the Lattice LSTM model for CFDK extraction, we can conclude that the Lattice LSTM model, although it is improved from the BI-LSTM-CRF model, only performs well in the NER task, but not in the sequence labeling tasks. The ZEN and ERNIE models do not perform well for the CFDK dataset.

It is also noted from Table 3 that the sequence labeling (SL) task for the request sequence performs well by all the models when we compare with the SL task for other sequences. A possible reason could be that the sentence patterns in the request sequence are less diversified. In addition, the performance of the SL task for the fault sequence is higher than that for the cause and solution sequences. This could be caused by the fact that the fault sequence is shorter than the cause and solution sequences.

C. COMPARISON OF EQUIPMENT ENTITIES OF DIFFERENT LENGTHS

In order to investigate the influence of entity length on NER, we have also conducted the equipment NER task with different entity lengths. The equipment entities in this sensitivity study were divided into three subsets according to the length of the entities: short entities (entities with ≤ 5 characters), medium entities (entities with >5 characters and ≤ 10 characters) and long entities (entities with >10 characters). Using the three subsets, the BERT-BI-LSTM-CRF and the five other models produced the $F1$ measure as shown in Figure 5.

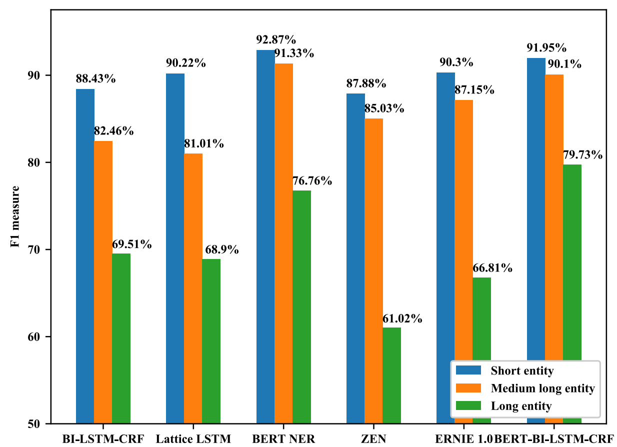


FIGURE 5. Comparisons of the equipment entities of different lengths ($F1$ Measure).

It is noted from Figure 5 that the longer the equipment entities, the worse the performance of the entity named recognition from all the models. When we look at the long entity set, the BERT-BI-LSTM-CRF model achieves an F1 measure of 79.73%, which is 2.97% higher than the second best F1 measure (from the BERT NER model) and 18.71% higher than the worst F1 measure (from the ZEN model). This finding indicates that the BERT-BI-LSTM-CRF model has a very good recognition effect on long entities. For the short and medium entities, fine-tuning the BERT pre-trained model through a complex NER model does not improve the NER performance. This may be caused by the fact that the BERT pre-trained models are designed to be large enough to absorb information during the pre-training process first, and fine-tune on very narrow task distributions second [32].

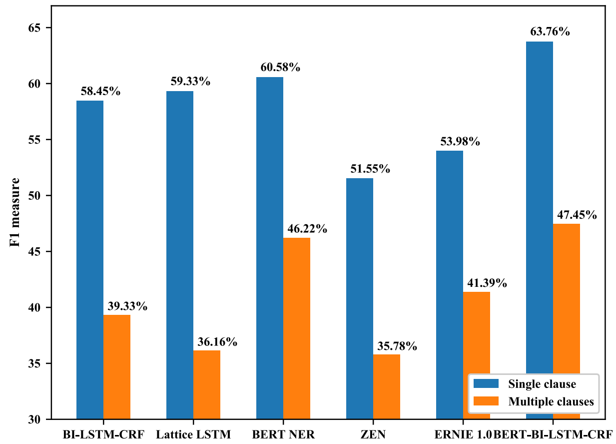


FIGURE 6. Comparison of fault description sequences of different clause numbers.

TABLE 4. Recognition results of a complex equipment entity.

Model	Result
BI-LSTM-CRF	断路器（ic65N-D/3P+iOF (Circuit breaker（ic65N-D/3P+iOF)
Lattice LSTM	断路器（ic65N-D/3P+iOF订货号A9F19310+ (Circuit breaker（ic65N-D/3P+iOF order No. A9F19310+)
BERT NER	断路器（ic65N-D/3P+iOF订货号A9F19310+A9A26924） (Circuit breaker（ic65N-D/3P+iOF order No. A9F19310+）
ZEN	断路器 (Circuit breaker)
ERINE 1.0	断路器（ic65N-D/3P+iOF订货号A9F19310+ A9A26924） ((Circuit breaker（ic65N-D/3P+iOF order No. A9F19310+）
BERT-BI-LSTM-CRF	断路器（ic65N-D/3P+iOF订货号A9F19310+A9A26924） (Circuit breaker（ic65N-D/3P+iOF order No. A9F19310+）

D. COMPARISON OF FAULT SEQUENCES OF DIFFERENT CLAUSE NUMBERS

Different from the equipment entities which are composed of noun or noun phrase, the fault sequences contain multiple phrases, clauses, and even sentences. Each fault sequence normally has a long and complex list of characters, which imposes significant challenges on the NER and SL tasks. The research team assessed the recognition effect of the BERT-BI-LSTM-CRF model on the fault sequences. Two subsets (one with single clauses and the other one with multiple clauses) were formed in this study from the fault sequences. For example, "安全阀漏气故障" (or "air leakage of safety valve") is a single clause fault sequence, while "气阀弹簧及阀片频繁损坏, 尤其是三级进排气阀" (or "valve springs and valve plates are frequently damaged,

TABLE 5. Recognition results of a multiple clauses fault sequence.

Model	Result
BI-LSTM-CRF	四级活塞环后，排气量有所提升，但与技术要求仍存在约百分之十的差距 (After the four stage piston ring, the displacement has increased, but there is still a gap of about 10% with the technical requirements)
Lattice LSTM	四级活塞环后，排气量有所提升，但与技术要求仍存在约百分之十的差距 (After the four stage piston ring, the displacement has increased, but there is still a gap of about 10% with the technical requirements)
BERT NER	与技术要求仍存在约百分之十的差距 (There is still a gap of about 10% with the technical requirements)
ZEN	与技术要求仍存在约百分之十的差距 (There is still a gap of about 10% with the technical requirements
ERINE 1.0	与技术要求仍存在约百分之十的差距 (There is still a gap of about 10% with the technical requirements)
BERT-BI-LSTM-CRF	更换新式四级活塞环后，排气量有所提升，但与技术要求仍存在约百分之十的差距 (After the replacement of the new four stage piston ring, the displacement has increased, but there is still a gap of about 10% with the technical requirements)

especially the third stage intake and exhaust valves'') is a multiple clause fault sequence. We have tested the BERT-BI-LSTM-CRF and five other models on these two datasets. The results are shown in Figure 6.

All the models yield $F1$ measure over 50% from the single clause sequences and less than 50% from the multiple clause fault sequences. This shows that the number of clauses has a significant impact on the models. The BERT-BI-LSTM-CRF model gives its superior performance on both the single clause and multiple clause sequences. This indicates that the BERT-BI-LSTM-CRF model has a very good recognition effect on single clause and multiple clause sequences.

E. CASE STUDY

In order to further examine the recognition effect of the baseline model, in this section, we introduce two typical cases for compressor fault diagnosis knowledge extraction from the CFDK dataset. Two text sequences, one containing a complex equipment entity and one containing a multiple clauses fault sequence, are entered into the above six models. The recognition results of these models are shown in Table 4 and Table 5, respectively.

``断路器 (ic65N-D/3P + iOF 订货号 A9F19310 + A9A26 924)'' (or ``circuit breaker(ic65N-D/3P+iOF order No. A9F1 9310+)'' in English) is a long and complex equipment entity in the CFDK dataset. It contains a total of 39 Chinese characters, English characters, numbers and a variety of symbols. Table 4 shows the recognition results of the BERT-BI-LSTM-CRF and five other models. It can be seen from the table that only the BERT NER and BERT+BiLSTM+CRF models correctly identify the entity. The BI-LSTM-CRF and Lattice LSTM models do not correctly identify the English and digital parts, while the ZEN model only recognizes the initial Chinese word ``断路器 (circuit breaker)''.

Table 5 shows the recognition results of a multiple clauses fault description sequence: ``更换新式四级活塞环后，排气量有所提升，　但与技术要求仍存在约百分之十的差距 (or ``After the replacement of the new four stage piston ring, the displacement has increased, but there is still a gap of about 10% with the technical requirements'' in English). It can be seen that only the BERT-BI-LSTM-CRF model accurately identifies the fault description, other models identify only a part of it. This shows the superiority of the BERT-BI-LSTM-CRF model in identifying long sequences.

VI. CONCLUSION AND FUTURE WORK

Considering that most deep learning approaches predict compression faults through monitoring data or vibration data generated when compressors are running, this paper provides a new direction for compression fault diagnosis. The natural language processing (NLP) technology is explored in this study to automatically identify expert knowledge from text documents and use the expert knowledge for future real time fault prediction, detection, and diagnosis. In this paper, we propose a benchmark dataset and a baseline model for mining compressor fault diagnosis knowledge from maintenance log sheets based on the Sequence labeling (SL) and Named Entity Recognition (NER) technologies.

This paper presents the benchmark dataset developed from compressor maintenance log sheets. This dataset, to our knowledge, is the first one for compressor fault diagnosis knowledge extraction. In this dataset, 19218 equipment entities, 5909 fault sequences, 6595 service request sequences, 1588 cause sequences, and 2786 solution sequences are annotated under new annotation specifications customized for compressor diagnosis. The annotation method and the specifications used in this paper are applicable to those for maintenance log sheets in other languages.

The BERT-BI-LSTM-CRF model and the other five benchmark models (that is, the BI-LSTM-CRF, Lattice LSTM, BERT NER, ZEN, and ERNIE models) were used in this study to assess the model performance in named entity recognition and sequence labeling. The experiment results show that the BERT-BI-LSTM-CRF model (or the baseline model) is the best one in extracting expert knowledge from the subject dataset. The experiment results also demonstrate that the BERT-BI-LSTM-CRF model gives superior performance on sequence labeling tasks. The results also demonstrate that the baseline model is suitable for the benchmark dataset in knowledge extraction. Although the baseline model is not the most cutting-edge model in the sequence labeling and NER fields, it indeed presents a potential for compressor fault diagnosis.

The expert knowledge extracted from the maintenance log sheets using the natural language processing (NLP) technology will form a foundation for the future real time compressor fault diagnosis. There is still a long way to go to make the baseline model adopted by the compressor industry. As the NER and SL methods considered in this paper use the fine-tuning mechanism to update the weights of a pre-trained model by training on a supervised dataset specific to the desired task, they need a large dataset for every task [32]. In the future, a massive set of maintenance log sheets should be collected and significant efforts in annotating the log sheets should be considered to train the baseline model. Eventually, the baseline model can be employed as the industry standard model for real time prediction and diagnosis of compressor faults.

As discussed in the Model Comparisons section, the precision, recall, and F1 measure of the models on long and complex sequences especially in the cause and solution categories is very low (less than 50% in F1 measure). One possible solution is to use more powerful sequence models, such as FLAT [44] and Soft Lexicon (LSTM) + BERT [45] to fine-tune larger pre-trained models, such as RoBERTa-wwm-ext-large[13] and ELECTRA-large.[14] Another potential solution is to use large-scale compressor maintenance corpus to train a domain special pre-trained model, but this method requires hundreds of millions of compressor maintenance corpus, which is difficult to obtain.

In addition, the knowledge from domain experts is usually more useful than knowledge extracted from log sheets. We plan to develop a hybrid mechanism to integrate the two domains of knowledge for fault diagnosis.

ACKNOWLEDGMENT

The authors would like to thank the editor-in-chief, the associate editor, and the reviewers for their insightful comments and suggestions.

REFERENCES

[1] V. Ebrahimipour and S. Yacout, "Ontology-based schema to support maintenance knowledge representation with a case study of a pneumatic valve," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 45, no. 4, pp. 702–712, Apr. 2015, doi: 10.1109/TSMC.2014.2383361.

[2] Z. Ye and Z.-H. Ling, "Hybrid semi-Markov CRF for neural sequence labeling," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Short Papers)*, vol. 2, 2018, pp. 235–240. [Online]. Available: https://www.aclweb.org/anthology/P18-2038

[3] K. Shaalan, "A survey of arabic named entity recognition and classification," *Comput. Linguistics*, vol. 40, no. 2, pp. 469–510, Jun. 2014, doi: 10.1162/COLI_a_00178.

[4] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," 2018, *arXiv:1812.09449*. [Online]. Available: http://arxiv.org/abs/1812.09449

[13] Available at https://github.com/ymcui/Chinese-BERT-wwm/

[14] Available at https://github.com/ymcui/Chinese-ELECTRA

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

[6] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015, *arXiv:1508.01991*. [Online]. Available: http://arxiv.org/abs/1508.01991

[7] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, and A. K. Nandi, "Applications of machine learning to machine fault diagnosis: A review and roadmap," *Mech. Syst. Signal Process.*, vol. 138, Apr. 2020, Art. no. 106587.

[8] H. Cui, L. Zhang, R. Kang, and X. Lan, "Research on fault diagnosis for reciprocating compressor valve using information entropy and SVM method," *J. Loss Prevention Process Industries*, vol. 22, no. 6, pp. 864–867, Nov. 2009.

[9] Y. Tang, "Fault diagnosis based on nonlinear complexity measurefor reciprocating compressor," *J. Mech. Eng.*, vol. 48, no. 3, p. 102, 2012.

[10] J. Qiu, W. Liang, L. Zhang, X. Yu, and M. Zhang, "The early-warning model of equipment chain in gas pipeline based on DNN-HMM," *J. Natural Gas Sci. Eng.*, vol. 27, pp. 1710–1722, Nov. 2015.

[11] S. Zhang, M. Xu, Y. Li, H. Zhao, and R. Wang, "Valve fault diagnosis of a reciprocating compressor based on hybrid method," *J. Vibrat. Shock*, vol. 35, no. 11, pp. 167–173, Jun. 2016. [Online]. Available: https://kns.cnki.net/kcms/detail/detail.aspx?FileName=ZDCJ201611027&DbName=CJFQ2016

[12] H.-B. Yang, J.-A. Zhang, L.-L. Chen, H.-L. Zhang, and S.-L. Liu, "Fault diagnosis of reciprocating compressor based on convolutional neural networks with multisource raw vibration signals," *Math. Problems Eng.*, vol. 2019, pp. 1–7, Jul. 2019.

[13] V. T. Tran, F. AlThobiani, and A. Ball, "An approach to fault diagnosis of reciprocating compressor valves using Teager–Kaiser energy operator and deep belief networks," *Expert Syst. Appl.*, vol. 41, no. 9, pp. 4113–4122, Jul. 2014.

[14] Y. Liu, L. Duan, Z. Yuan, N. Wang, and J. Zhao, "An intelligent fault diagnosis method for reciprocating compressors based on LMD and SDAE," *Sensors*, vol. 19, no. 5, p. 1041, Feb. 2019.

[15] F.-Y. Guo, Y.-C. Zhang, Y. Wang, P. Wang, P.-J. Ren, R. Guo, and X.-Y. Wang, "Fault detection of reciprocating compressor valve based on one-dimensional convolutional neural network," *Math. Problems Eng.*, vol. 2020, pp. 1–10, Apr. 2020.

[16] S.-V. Oprea, A. Bara, D. Preotescu, and L. Elefterescu, "Photovoltaic power plants (PV-PP) reliability indicators for improving operation and maintenance Activities. A case study of PV-PP Agigea located in Romania," *IEEE Access*, vol. 7, pp. 39142–39157, 2019.

[17] T. Gunda, S. Hackett, L. Kraus, C. Downs, R. Jones, C. McNalley, M. Bolen, and A. Walker, "A machine learning evaluation of maintenance records for common failure modes in PV inverters," *IEEE Access*, vol. 8, pp. 211610–211620, 2020.

[18] V. Yadav and S. Bethard, "A survey on recent advances in named entity recognition from deep learning models," in *Proc. 27th Int. Conf. Comput. Linguistics*. Santa Fe, NM, USA: Association for Computational Linguistics, Aug. 2018, pp. 2145–2158. [Online]. Available: https://www.aclweb.org/anthology/C18-1182

[19] S. Zhang and N. Elhadad, "Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts," *J. Biomed. Informat.*, vol. 46, no. 6, pp. 1088–1098, Dec. 2013, doi: 10.1016/j.jbi.2013.08.004.

[20] M. Collins and Y. Singer, "Unsupervised models for named entity classification," in *Proc. Joint SIGDAT Conf. Empirical Methods Natural Language Process. Very Large Corpora, (EMNLP)*, P. Fung and J. Zhou, Eds. College Park, MD, USA: Association for Computational Linguistics, Jun. 1999, pp. 1–11. [Online]. Available: https://www.aclweb.org/anthology/W99-0613/

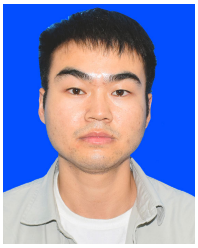
[21] D. M. Bikel, R. Schwartz, and R. M. Weischedel, "An algorithm that learns what's in a name," *Mach. Learn.*, vol. 34, nos. 1–3, pp. 211–231, 1999, doi: 10.1023/A:1007558221122.

[22] G. Szarvas, R. Farkas, and A. Kocsor, "A multilingual named entity recognition system using boosting and C4.5 decision tree learning algorithms," in *Proc. 9th Int. Conf. Discovery Sci., (DS)* in Lecture Notes in Computer Science, vol. 4265, L. Todorovski, N. Lavrac, and K. P. Jantke, Eds. Barcelona, Spain: Springer, Oct. 2006, pp. 267–278, doi: 10.1007/11893318_27.

[23] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman, "NYU: Description of the MENE named entity system as used in MUC-7," in *Proc. 7th Message Understand. Conf., Conf. Held Fairfax*, Richmond, VA, USA: MUC, May 1998, pp. 1–6. [Online]. Available: https://www.aclweb.org/anthology/M98-1018/

[24] P. McNamee and J. Mayfield, ‘‘Entity extraction without language-specific resources,’’ in *Proc. 6th Conf. Natural Lang. Learn. COLING*, 2002, pp. 1–4. [Online]. Available: https://www.aclweb.org/anthology/W02-2020/
[25] A. McCallum and W. Li, ‘‘Early results for named entity recognition with conditional random fields, feature induction and Web-enhanced lexicons,’’ in *Proc. 7th Conf. Natural Lang. Learn. (HLT-NAACL)*, 2003, pp. 188–191. [Online]. Available: https://www.aclweb.org/anthology/W03-0430/
[26] Y. Chen, G. Wang, Q. Zheng, Y. Qin, R. Huang, and P. Chen, ‘‘A set space model to capture structural information of a sentence,’’ *IEEE Access*, vol. 7, pp. 142515–142530, 2019.
[27] M. Al-Smadi, S. Al-Zboon, Y. Jararweh, and P. Juola, ‘‘Transfer learning for arabic named entity recognition with deep neural networks,’’ *IEEE Access*, vol. 8, pp. 37736–37745, 2020.
[28] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, ‘‘Natural language processing (almost) from scratch,’’ *J. Mach. Learn. Res.*, vol. 12 pp. 2493–2537, Aug. 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=2078186
[29] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, ‘‘Neural architectures for named entity recognition,’’ in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2016, pp. 260–270, doi: 10.18653/v1/n16-1030.
[30] A. Zukov-Gregoric, Y. Bachrach, P. Minkovsky, S. Coope, and B. Maksak, ‘‘Neural named entity recognition using a self-attention mechanism,’’ in *Proc. IEEE 29th Int. Conf. Tools With Artif. Intell. (ICTAI)*, Nov. 2017, pp. 652–656, doi: 10.1109/ICTAI.2017.00104.
[31] Q. Zhang, J. Fu, X. Liu, and X. Huang, ‘‘Adaptive co-attention network for named entity recognition in tweets,’’ in *Proc. 32nd AAAI Conf. Artif. Intell., (AAAI), 30th Innov. Appl. Artif. Intell. (IAAI), 8th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*, New Orleans, LA, USA, Feb. 2018, pp. 5674–5681. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16432
[32] T. B. Brown *et al.*, ‘‘Language models are few-shot learners,’’ 2020, *arXiv:2005.14165*. [Online]. Available: http://arxiv.org/abs/2005.14165
[33] E. F. T. K. Sang and F. De Meulder, ‘‘Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition,’’ in *Proc. 7th Conf. Natural Lang. Learn. (HLT-NAACL)*, 2003, pp. 142–147. [Online]. Available: https://www.aclweb.org/anthology/W03-0419/
[34] J. Aguilar, C. Beller, P. McNamee, B. Van Durme, S. Strassel, Z. Song, and J. Ellis, ‘‘A comparison of the events and relations across ACE, ERE, TAC-KBP, and FrameNet annotation standards,’’ in *Proc. 2nd Workshop EVENTS, Definition, Detection, Coreference, Represent.*, 2014, pp. 45–53. [Online]. Available: https://www.aclweb.org/anthology/W14-2907
[35] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, ‘‘Deep contextualized word representations,’’ in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., (Long Papers)*, vol. 1, 2018, pp. 2227–2237, doi: 10.18653/v1/n18-1202.
[36] J. Howard and S. Ruder, ‘‘Universal language model fine-tuning for text classification,’’ in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2018, pp. 328–339. [Online]. Available: https://aclanthology.info/papers/P18-1031/p18-1031
[37] X. Li, J. Feng, Y. Meng, Q. Han, F. Wu, and J. Li, ‘‘A unified MRC framework for named entity recognition,’’ in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 5849–5859. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.519/
[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, ‘‘Attention is all you need,’’ in *Proc. Adv. Neural Inf. Process. Syst., Annu. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 5998–6008. [Online]. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need
[39] E. F. T. K. Sang, ‘‘Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition,’’ in *Proc. 6th Conf. Natural Lang. Learn. (COLING)*, 2002, pp. 142–147. [Online]. Available: https://www.aclweb.org/anthology/W02-2024
[40] M. Bodén, ‘‘A guide to recurrent neural networks and backpropagation,’’ in *The Dallas Project*, Dec. 2001, pp. 1–20. [Online]. Available: http://axon.cs.byu.edu/~martinez/classes/678/Papers/RNN_Intro.pdf
[41] Y. Zhang and J. Yang, ‘‘Chinese NER using lattice LSTM,’’ in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2018, pp. 1554–1564. [Online]. Available: https://www.aclweb.org/anthology/P18-1144
[42] S. Diao, J. Bai, Y. Song, T. Zhang, and Y. Wang, ‘‘ZEN: Pre-training chinese text encoder enhanced by N-gram representations,’’ 2019, *arXiv:1911.00720*. [Online]. Available: http://arxiv.org/abs/1911.00720
[43] Y. Sun, S. Wang, Y. Li, S. Feng, X. Chen, H. Zhang, X. Tian, D. Zhu, H. Tian, and H. Wu, ‘‘ERNIE: Enhanced representation through knowledge integration,’’ 2019, *arXiv:1904.09223*. [Online]. Available: http://arxiv.org/abs/1904.09223
[44] X. Li, H. Yan, X. Qiu, and X. Huang, ‘‘FLAT: Chinese NER using flat-lattice transformer,’’ in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 6836–6842. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.611
[45] R. Ma, M. Peng, Q. Zhang, Z. Wei, and X. Huang, ‘‘Simplify the usage of lexicon in Chinese NER,’’ in *Proc. 58th Annu. Meeting Assoc. Comput., Linguistics*, 2020, pp. 5951–5960. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.528

TAO CHEN was born in Shiyan, Hubei, China, in 1981. He received the B.Eng. degree in communication engineering from the Nanjing Institute of Communication Engineering, China, in 2003, the M.Eng. degree in computer application from Wuyi University, Guangdong, China, in 2013, and the Ph.D. degree in computer application from the Harbin Institute of Technology, China, in 2018. In 2018, he joined the Faculty of Intelligent Manufacturing, Wuyi University, China, as a Lecturer. He is the author of one book, 17 articles, and one invention. His research interests include natural language processing, deep learning, knowledge acquisition and reasoning, and sentiment analysis.



JIANG ZHU was born in Fengdu, Chongqing, China, in 1997. He received the B.S. degree in electronic information science and technology from Inner Mongolia University for the Nationalities, in 2018. He is currently pursuing the M.S. degree in pattern recognition and intelligent system with Wuyi University, Guangdong, China. His research interests include natural language processing, named entity recognition, and data mining.

ZHIQIANG ZENG (Member, IEEE) was born in Shaoguan, Guangdong, China, in 1989. He received the Ph.D. degree from the South China University of Technology, Guangzhou, China, in 2018. Since then, he has been a Lecturer with the Faculty of Intelligent Manufacturing, Wuyi University, China. He is the author of a number of high-level articles. His research interests include machine learning, data mining, industrial big data, and production scheduling.

XUDONG JIA received the B.S. and M.S. degrees from Beijing Jiaotong University, in 1983 and 1986, respectively, the second M.S. degree from the University of Toronto, Canada, in 1992, and the Ph.D. degree from the Georgia Institute of Technology, in 1996. He is currently a Visiting Scholar with the Faculty of Intelligent Manufacturing, Wuyi University, China. He is also a Professor and the Associate Dean with the College of Engineering and Computer Science, CSUN. His research interests include intelligent transportation systems (ITS) standards, geographic information system (GIS) applications in transportation, traffic safety, transportation information systems, travel demand management, and air quality. He is an Associate Editor for the IEEE Intelligent Transportation Systems Society and the IEEE Open Journal of Intelligent Transportation Systems.