processes

MDPI

Article

A Study on Standardization of Security Evaluation Information for Chemical Processes Based on Deep Learning

Lanfei Peng, Dong Gao * and Yujie Bai

School of Information Science and Technology, Beijing University of Chemical Technology, No. 15, North Third Ring East Road, Beijing 100029, China; lanfei_peng@163.com (L.P.); baiyujie2021@163.com (Y.B.)
* Correspondence: gaodong@mail.buct.edu.cn; Tel.: +86-135-2284-8040

Abstract: Hazard and operability analysis (HAZOP) is one of the most commonly used hazard analysis methods in the petrochemical industry. The large amount of unstructured data in HAZOP reports has generated an information explosion which has led to a pressing need for technologies that can simplify the use of this information. In order to solve the problem that massive data are difficult to reuse and share, in this study, we propose a new deep learning framework for Chinese HAZOP documents to perform a named entity recognition (NER) task, aiming at the characteristics of HAZOP documents, such as polysemy, multi-entity nesting, and long-distance text. Specifically, the preprocessed data are input into an embeddings from language models (ELMo) and a double convolutional neural network (DCNN) model to extract rich character features. Meanwhile, a bidirectional long short-term memory (BiLSTM) network is used to extract long-distance semantic information. Finally, the results are decoded by a conditional random field (CRF), and then output. Experiments were carried out using the HAZOP report of a coal seam indirect liquefaction project. The experimental results for the proposed model showed that the accuracy rate of the optimal results reached 90.83, the recall rate reached 92.46, and the F-value reached the highest 91.76%, which was significantly improved as compared with other models.

Keywords: hazard and operability analysis; named entity recognition; neural network; deep learning

Citation: Peng, L.; Gao, D.; Bai, Y. A Study on Standardization of Security Evaluation Information for Chemical Processes Based on Deep Learning. *Processes* **2021**, *9*, 832. https://doi.org/10.3390/pr9050832

Academic Editors: Konstantinos Demertzis, Lazaros Iliadis, Nikos Tziritas and Panayotis Kikiras

Received: 30 March 2021
Accepted: 29 April 2021
Published: 10 May 2021

1. Introduction

One of the most difficult problems in chemical plant design is identification of hazards. Therefore, the industry has developed a range of hazard identification methods [1]. Several safety analysis techniques are available such as safety checklist analysis (SCA), preliminary hazard analysis (PHA), failure modes and effects analysis (FMEA), and hazard and operability analysis (HAZOP). Compared with traditional safety analysis methods, the HAZOP method has the following three advantages: First, the concept of system security is established instead of the concept of individual device security. Second, it is a systematic and well-integrated method, which is conducive to identifying a variety of potential hazards. Third, it is structurally good and easy to master. Therefore, since HAZOP was first developed, it has been widely used for safety assessment and risk analyses at home and abroad [2].

HAZOP is a process hazard analysis (PHA) technique based on systems engineering, which was first proposed in the 1970s for systematically identifying the possible deviations between the process and the expected intention, and therefore discover possible causes and consequences of hazards. Because of its excellent results and practicability, HAZOP has been a widely used method for the safety analysis and evaluation of continuous production systems [3]. The analysis process generally takes the form of a qualitative and discrete simulation process that answers various "what-if" questions, and the results are often stored in the form of tables. Specifically, to help the analysis team focus on the discussion, first, the complex process system is divided into a number of "subsystems", each called a "node". Then, for each node, the analysis team identifies possible accident scenarios

in terms of deviations with reference to a set of leading words. Guide words such as "more flow", "less flow" and "more temperature" are used as the inputs of a qualitative simulation. After a deviation is determined, simulation results are used to determine all possible states of plant units, and the expert group explores the possible causes and consequences, and therefore helps to find the potential hazards. For instance, "What are the consequences of higher temperature?" and "Are they dangerous?" [4]. For every pair of cause and consequence, safeguards must be identified that could prevent, detect, control, or mitigate the hazardous situation. Finally, if the safeguards are insufficient to solve the problem, recommendations must be considered [5].

Brainstorming is commonly used in the traditional HAZOP analysis process; however, a manual analysis method is a heavy workload, time-consuming, and relies heavily on expert experience [3]. Therefore, over the past 30 years, efforts to improve HAZOP have been devoted to the development of an automatic hazard and operability expert system including studies on computer-aided or computer-automated analysis methods. Nevertheless, whether it is manual or computer-aided analysis, there are still many shortcomings in the communication and storage of analysis results which include the following: (1) A manual HAZOP analysis report is difficult to manage and save, and there is no unified standard for communicating the analysis results made by different experts or for different projects. It is also difficult to transfer, reuse, and share information at different stages. (2) The computer-aided method also lacks a unified form of communication, it cannot share data between different types of software, and does not have the ability of automatic knowledge recognition; therefore, it is very difficult to reuse and share information between different types of software or the same software in different projects, teams, or stages [6,7]. In the face of a large number of unstructured data and information in the form of paper documents or electronic documents generated by HAZOP analysis, some automation technologies are urgently needed to help people quickly find the information they need in mass information sources.

Entity names and their relationships form the main content of a document and by identifying them in a document, it is possible to understand the document to some extent [8]. Named entity recognition (NER) is an important basis for natural language processing tasks such as relation extraction and entity linking [9]. In this study, named entity recognition technology is used to process a HAZOP report to obtain detailed information. Consider the example, "naphtha channeling caused by abnormal internal leakage of steam generator". In this sentence, "steam generator" is equipment, and "naphtha" is material. These two are called named entities in information extraction research. Automatically identifying the various named entities in HAZOP documents can generate important information for sharing, reuse, etc. [10].

However, due to its unique characteristics, the chemical industry poses some new challenges to entity identification tasks, which include the following: (1) variability (HAZOP reports are produced by different experts and researchers, so there are many differences in the presentation); (2) domain specificity (the text of a HAZOP report is full of technical terms, abbreviations and so on, making it difficult to identify; and (3) complexity. The complexity challenge has several aspects as follows: (a) Different from the fixed input tag of an English NER task, Chinese often lacks clear word boundaries and there are multiple entities nested in the HAZOP text. For example, the "steam generator at the bottom of the desorption tower", where the "desorption tower" and the "steam generator" can be identified as two entities, respectively, but the combined entity "a steam generator which is at the bottom of the desorption tower" should be identified here according to context. (b) There are a lot of polysemy in HAZOP text, such as in "steam generator", "steam" is identified as equipment, but in "steam outlet", it is identified as material. (c) Compared with documents in English or other fields, HAZOP text tends to have a longer length, which makes recognition more difficult.

In order to solve the above-mentioned problems, the main task of this study is to identify the two entities of equipment and materials in the HAZOP document of a coal

indirect liquefaction project, in order to extract the most useful information in the document and achieve information standardization. First, in this paper, we briefly introduce and analyze the mission connotation, operation steps, important characteristics and improvement direction of HAZOP. In Section 2, we review the progress based on HAZOP studies and introduce the application status of named entity recognition in various fields. In Section 3, we describe the proposed deep learning model. In Section 4, the HAZOP data source and document characteristics used in the experiment are explained, followed by the details of the experimental environment configuration and training, and finally the significance and calculation method of the experimental indicators are introduced. In Section 5, we present the experimental results as well as analysis and discussion. Finally, in Section 6, we summarize the thesis.

2. Related Work

Studies on HAZOP improvements have mainly had two goals. One goal has been to expand the field, including the scope of hazard identification, considering quantitative analysis based on qualitative analysis, etc. The second goal has been to develop an automatic HAZOP. In this study, we focus on the latter.

Named entity recognition is an important task in NLP [11]. The purpose is to identify the entities in textual data and divide them into fixed categories, such as people names, place names, and organization names. Named entity recognition methods mainly include rule-based, statistics-based, and deep learning methods [12]. In recent years, named entity recognition has generally reached a high level, but there is still much room for progress. The development of vertical domain named entity recognition has mainly focused on medical, financial, news, chemistry, biology, and other fields. Therefore, progress in the field of entity recognition is briefly reviewed by referring to studies on entity recognition tasks in chemistry and biology.

The rule- or dictionary-based approach is the earliest method used in named entity recognition [13,14]. The rule-based method involves rule templates that are manually constructed by linguistic experts and feature selection includes statistics on punctuation marks, keywords, indicator words, direction words, position words, center words, etc., and matching between patterns and strings is used as the main method. Therefore, such systems mostly rely on the establishment of an information base and a dictionary [15]. Examples of dictionaries in chemistry and biomedicine domains are the Jochem dictionary [16], which is used to identify small molecules and drugs in the text, and the DrugBank dictionary for drugs. Hettne et al. [16] and Rebholz-Schuhmann et al. [17] are examples of dictionary-based systems used to extract drug names and molecules via string matching methods. Systems such as Peregrine [18], TaxonGrab [19], and LINNAEUS [20] have demonstrated the utility of dictionary matching approaches to identify disease and organism names. In general, the performance of the rule-based method is better than that of the statistical method when the language phenomenon can be accurately reflected by the extracted rules. However, these rules often depend on the specific language, domain, and text style, and the compilation process is time-consuming and difficult to cover all the language phenomena, which is particularly prone to errors and poor portability of the system. In addition, linguists need to rewrite the different rules for different systems. Another disadvantage of the rule-based approach is that the cost is too high; additionally, the rule-based method has problems such as long system construction cycle, poor portability and the need to build knowledge bases in different fields to improve the system identification performance.

The statistics-based method uses manually annotated corpus for training, which does not need extensive linguistic knowledge and can be completed in a short period of time. The advantage of this method is that it can be transplanted to a new field with little or no changes, and only one training with new materials is enough. Traditional statistics-based models include the maximum entropy model, hidden Markov model, conditional random field model, etc. [21]. For example, Po-Ting Lai et al. introduced the statistical principle-based approach (SPBA) for named entity recognition and participated in a Bio-Creative V.5

gene- and protein-related object (GPRO) task to evaluate the ability of SPBA for processing patent abstracts. In Bio-Creative V.5 GPRO task, this approach achieved an F-score of 73.73% on GPRO type 1 and an F-score of 78.66% on combining GPRO type 1 and 2 [22]. However, most statistics-based named entity recognition models have high training time complexity, high training cost, and strong dependence on the quality of corpus. In addition, due to the need for a clear normalization calculation, the CPU overhead is also relatively large [23,24].

With the development of deep learning, neural networks such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have emerged in the field of named entities recognition [25,26]. Machine learning-based methods are mainly based on classification and sequence annotation. Since the latter can jointly consider the labeling results of adjacent words, more attention has been given to them. Collobert et al. [27] proposed an effective neural network model, which no longer needed a large number of artificial features and was able to learn word vectors from a large number of unlabeled texts, thus, contributing to the training of the model. Huang et al. [28] combined a bidirectional long short-term memory (BiLSTM) and a conditional random field (CRF) and proposed the BiLSTM-CRF model, in which BiLSTM effectively used the context of the current word, while the CRF layer used the label sequence information at the sentence level. Ma et al. [29] proposed the BiLSTM-CNN-CRF model, which, first, used CNN to learn the character-level representation, and then used BiLSTM-CRF for subsequent processing. Compared with traditional machine learning methods, deep learning has significant advantages in feature representation, i.e., it does not need or only needs a small number of features to achieve better performance. In fact, the BiLSTM-CRF model has become the mainstream method for sequence labeling nowadays [30].

Our main contribution is that we combine these neural network models for a NER task. We present a hybrid model of an embeddings from language models (ELMo) and a bidirectional long short-term memory (LSTM) and a DCNN that learn both character- and word-level features, presenting the first evaluation of such an architecture on a self-annotated HAZOP text dataset.

3. Model

Named entity identification tasks in the chemical safety field focus on the identification of material and equipment terms in hazardous and operational analysis documents. The implementation process involves, first, identifying these entities, and then classifying them to distinguish the material and equipment nouns.

In the process of Chinese named entity recognition, Chinese words are not separated by spaces as English words are, and therefore word segmentation is needed before formal training. However, if word segmentation is carried out directly, many words that are not in the existing word segmentation thesaurus may not be correctly segmented, resulting in poor recognition results. The general method is to use character vectors instead of word vectors as input for training, in order to avoid the errors in word segmentation. However, the character sequence cannot express enough semantic information, and therefore the recognition accuracy of Chinese NER is not as good as English. In order to make up for the word vector defects and obtain more semantic information, it is necessary to mine more features to enrich the character vector.

In this study, first, the HAZOP report text is represented by word embedding vectors, and then a DCNN is used to extract character level features and an ELMo is used to extract word level dynamic features. Entity features in the HAZOP text are extracted by a multi-feature fusion-based deep learning method and the extracted results are stitched together. After that, the output is sent to the BiLSTM network for training and the relationships between the long-distance words in the sentence are extracted. In the following step, the output of the CNNS and ELMo and the output of the BiLSTM network are merged and sent to the next BiLSTM network. Finally, the CRF network is used for sequence tagging to

obtain the various entities and the entity tags we need. The model framework is shown in Figure 1.

Figure 1. The overall structure of named entity recognition.

3.1. Word Embedding

A defining feature of a neural network language model is its representation of words as high dimensional real-valued vectors. Words are converted via a learned lookup table into real-valued vectors which are used as the inputs to a neural network, so that the computer can automatically extract and learn the features and information contained in the text [31]. This process is called textual vectorization. Each character in the preprocessed text sequence is transformed into a vector whose dimension represents the richness of semantic information contained in the character. The input of the model in the form of vectors avoids the criticism of traditional machine learning methods, and can also effectively express the semantic information and grammatical relationships in the text. In this study, pretrained external word vectors are used to accomplish text vectorization.

3.2. Embeddings from Language Models (ELMo) Model

As the applications of natural language processing become wider and more demanding, learning high-quality representations can be challenging due to the following: (1) the complex characteristics of word use (e.g., syntax and semantics); (2) the variations in word use across linguistic contexts (i.e., to model polysemy) [32]. Similarly, there is a large number of polysemous words in HAZOP text, and the same word often has different meanings in different sentences, for example, considering "alarm device" and "equipment alarm", here, the word "alarm" has different meanings. The ELMo language model can solve the above problems effectively.

The ELMo model, first proposed by Matthew E. Peters et al., in 2018, is a method to derive word representations based on the complete context of a sentence using the bidirectional LSTM language model (BILM), which is mainly used to extract dynamic features of text [33]. Different from the traditional method in which one word corresponds to one vector such as word2vec and GloVE, ELMo is a trained model that infers the word vector of each word through the multi-layer bidirectional LSTM structure.

As shown in Figure 2, ELMo is actually a combination of some network layers. The forward language model predicts the following text through the preceding text, and the backward language model predicts the preceding text through the latter text, and therefore rich dynamic features can be extracted. The output vector at the moment k of the j-layer of the forward LSTM language model is denoted as $\overrightarrow{h}_{k,j}$; correspondently, the output vector at the moment k of the j-layer of the backward LSTM language model is denoted as $\overleftarrow{h}_{k,j}$. For the bidirectional LSTM language model of each layer, $H_{k,j} = \left[\overrightarrow{h}_{k,j}, \overleftarrow{h}_{k,j}\right]$ is the output vector of each layer (particularly, when j = 0 is the input of LSTM layer). Then, the ELMo eigenvectors of text sequences can be obtained by weighting and combining the obtained vectors, using Formula (1) for the specific weighted combination as follows:

$$ELMO_k = \gamma \sum_{j=0}^{L} s_j H_{k,j}, \quad (1)$$

where γ is the scaling factor, and s_j is the weight of the vector of each layer.

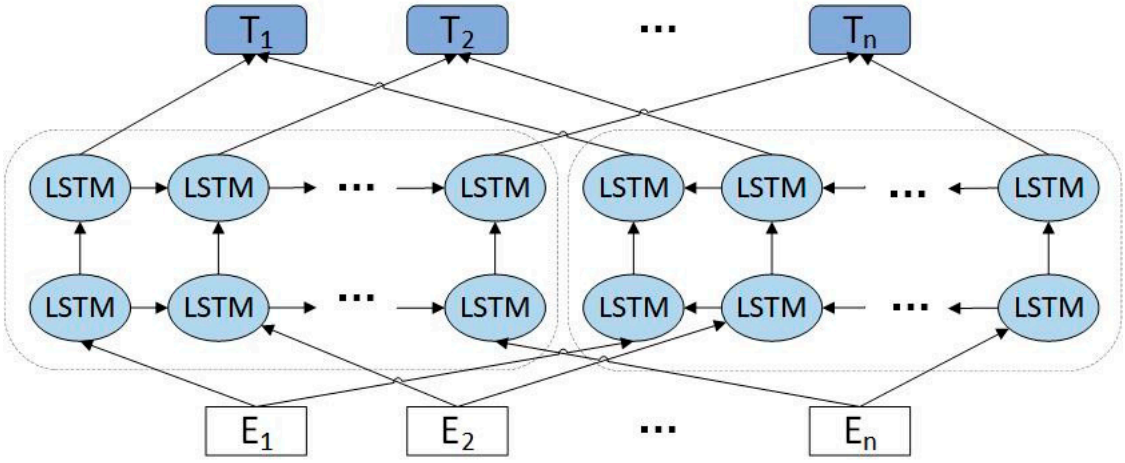

Figure 2. A schematic of the embeddings from language models (ELMo) language model, which consists of two layers of a bidirectional long short-term memory (LSTM) network that can extract richer semantic features.

3.3. Extracting Character Features Using a Double Convolutional Neural Network

Convolutional neural networks are feedforward neural networks, which have mainly been used in image processing, computer vision, and other fields in the early stage. Yoon Kim proposed a Text-CNN (Text-Convolutional Neural Network) in a convolutional neural network for sentence classification [34], which was the first time a convolutional neural network had been applied to a text classification task. The core of CNN is that it can capture a local correlation. Text-CNN uses multiple convolution cores of different sizes to extract the key information in the sentence (similar to the multi-window N-gram model), then, uses max-pooling to select the most influential high-dimensional classification features, and then uses the full connection layer with dropout to extract the text depth features, and finally connects to Softmax for classification. Text-CNN is a traditional model in the field of text classification. According to the characteristics of the CNN network, the algorithm is easy to parallelize. Max-pooling ensures that sentences with different lengths can be turned into fixed-length representations after the pooling layer; however, it also causes a problem, since the global max-pooling loses structural information, and therefore it is difficult to find complex modes such as turning relations in the text. Therefore, the network framework, in this study, has been improved on the basis of Text-CNN. A dual CNN network is used to extract local features of data. The size of the convolution kernel is the same, and the size of max-pooling is different. Therefore, different convolution and pooling methods are

used to extract the most important features in a sentence. Each CNN network obtains a vector matrix containing the local features of a sentence through convolution and pooling operations. Then, the feature vectors extracted from different CNN networks are stitched together to obtain a more effective feature matrix. The CNN model is shown in Figure 3. The convolution and pooling methods of each CNN network are shown in Table 1.

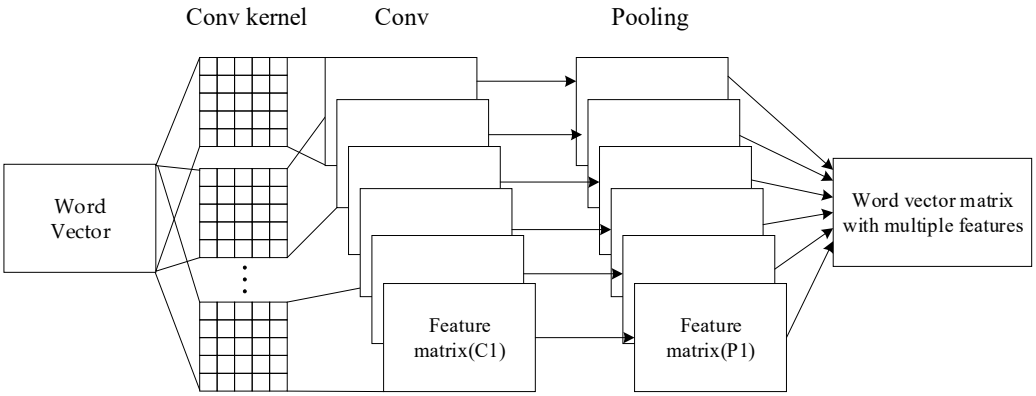

Figure 3. The convolutional neural network extracts character features from each word. The character embedding and the character type feature vector are computed through lookup tables. Then, they are concatenated and passed into the convolutional neural network (CNN).

Table 1. CNN network convolution and pooling.

Model	Convolution Kernel	Convolution Padding	Pooling Padding	Pooling	Pooling Window Size
CNN1	[2,120,1,100]	VALID [1]	SAME [1]	Max-pooling	[1,2,1,1]
CNN2	[2,120,1,100]	VALID	VALID	Max-pooling	[1,99,1,1]

[1] "VALID" and "SAME" are two different padding methods. "VALID" means only ever drops the right-most columns (or bottom-most rows) while "SAME" means pad evenly left and right, but if the amount of columns to be added is odd, it will add the extra column to the right.

3.4. Bidirectional Long Short-Term Memory (BiLSTM) Model

A convolutional neural network can be used to capture local features, but it cannot preserve long-distance relationships. Therefore, the recurrent neural network can encode sequences of any length into vectors of fixed size and capture the structural characteristics of the sequences at the same time. In order to solve the long-term dependence problem (such as gradient vanishing or gradient explosion) that exists in a traditional RNN, LSTM is further improved to effectively solve this problem.

In this study, a bidirectional LSTM network with long short-term memory units convert word features into named entity tag scores. The extracted features of each word are fed into the forward and backward LSTM networks. The output of each network at each time step is decoded by a linear layer and a logarithmic soft maximum layer into the logarithmic probability of each label category. Then, the two vectors are simply added together to produce the final output [35].

3.5. CRF Model

After training the BiLSTM model, the classifier is used to obtain valid sequence output results, but this classification method has shortcomings. Because BiLSTM is not limited, the label of each entity is determined by its probability value, which may appear as an error result, for example, tag the sentence in the form of B-I-O(Begin-Inside-Outside), B represents the beginning of the entity, I represents the middle or end of the entity, O represents non-entity, and the first word at the beginning of a sentence may be "O" or "B". If the label of the training result is "I", the label is an error, and the CRF layer learns

the constraint features of the sentence, that is, the label of the first word of the sentence cannot be I. In the B-Label1, I-Label2, and I-Label3 entity labels, Label1, Label2, and Label3 should be the same type of labels and other features. With these constraints, the model can significantly improve the accuracy of the final label prediction.

4. Experiment

4.1. Dataset and Experimental Setup

4.1.1. Dataset

The dataset used in the experiment, in this study, comes from the HAZOP analysis report of the preliminary design of the oil synthesis device for the Shenhuaning Coal 4 million tons/year indirect coal liquefaction project. The scope of this analysis project involves nine units in the oil synthesis unit, including a Fischer-Tropsch synthesis unit, a catalyst reduction unit, a wax filtration unit, a tail gas decarbonization unit, a fine desulfurization unit, a synthetic water treatment unit, a liquid intermediate material tank farm unit, a low temperature oil washing unit, and a deoxidized water and condensate refining station unit, with a total of 102 nodes, from which 5000 effective sentences are extracted. A great deal of information can be mined by analyzing the HAZOP report, for example, in the text of the report, "purified gas enters the subsequent process, causing overpressure of light oil-water separator and the equipment was damaged". In this sentence, "purified gas" is a material, and "light oil-water separator" is an equipment, which are called named entities in the information extraction of HAZOP documents, and the relationship between the two entities is that a deviation in the state of "purified gas" results in damage to the "light oil-water separator". The purpose of automatically mining this information from the HAZOP documents is to automatically identify and classify all kinds of named entities which are closely related to the technological process in the text, and therefore clarify the relationships between the entities and facilitate the subsequent reuse and sharing of information.

Data are tagged by domain experts. The data labels use the BIO method, where "B" represents the beginning of the entity, "I" represents the middle or end of the entity, and "O" represents the non-entity part. There are two main types of entities, i.e., materials and equipment. The materials are labeled as "MAT" and the equipment is labeled as "EQU". The specific labeling example is shown in Table 2. The training data are divided into training, validation, and test sets in the form of 8:1:1. The word vectors used in the experiment are 100-dimensional word vectors trained in Wikipedia. In order to improve the accuracy of recognition, in this study, we use the Jieba word segmentation tool for sentence segmentation. Word boundary information is added to the training data. The head of the word is represented by one, the word is represented by two, the tail of the word is represented by three, and the single word is represented by zero, so that the neural network model can extract more features.

Table 2. A concrete example of data labeling.

Text	**分**	**离**	**器**	**故**	**障**
Tagging	B-EQU (Begin of the equipment)	I-EQU (Inside of the equipment)	I-EQU (Inside of the equipment)	O (non-entity)	O (non-entity)

In Table 2, "text" is a general sentence randomly extracted from the HAZOP analysis report. It means "separator has failed", which is expressed in five words in Chinese. The first three characters mean "separator", which is an equipment, and the last two characters mean "failure", which is the non-entity part. In the task of labeling data, such sentences are segmented and classified, and finally each word has its own label, as shown in "tagging".

4.1.2. Experimental Setup

The experiments, in this study, were completed in the Windows 10 environment, using the open-source deep learning framework TensorFlow1.4.0 and the programming language python 3.5.2.

In this experiment, there are only three types of entity recognition, i.e., equipment entity, material entity, and non-entity, which are marked in the corpus as B-EQU(Begin of the equipment), I-EQU(Inside of the equipment), B-MAT(Begin of the material), I-MAT(Inside of the material), and O(non-entity), and the simple processing performed when reading the corpus ensures that the labeled form read in the training expectation is correct and avoids affecting the training of the neural network due to data labeling errors. In this experiment, the length of each sentence is set to 100 characters; sentences with fewer than 100 characters are supplemented with zero, if there are more than 100 characters, the characters after 100 characters are removed. The dimensions of all character vectors are 100, and for each layer of BiLSTM, the number of neurons is 100, the number of iterations is 1000, the learning rate is 0.005, and the dropout value is 0.5.

4.2. Experimental Evaluation Indicators

In this experiment, the precision (P), recall (R), and F-value (F) scores are used to evaluate the effect of the model. The calculation formulae for the three evaluation indicators are as follows:

$$P = \frac{n}{M} \times 100\%, \tag{2}$$

$$R = \frac{n}{N} \times 100\%, \tag{3}$$

$$F = \frac{2PR}{P+R} \times 100\%, \tag{4}$$

In Formulas (2)–(4), n represents the number of correctly identified entities, M represents the number of identified entities, and N represents the number of all entities in the verification set or test set.

5. Results

Four different models of BiLSTM-CRF, CNN-BiLSTM-CRF, DCNN-BiLSTM-CRF, and ELMo-DCNN-BiLSTM-CRF were trained in the same configured experimental environment, and the precision, recall and F-values of each model were calculated. The experimental results are shown in Table 3.

Table 3. Experimental results of three models.

Model	P (%)	R (%)	F (%)
BiLSTM-CRF	70.29	76.14	73.10
CNN-BiLSTM-CRF	83.46	81.54	82.49
DCNN-BiLSTM-CRF	87.90	89.31	88.60
ELMo-DCNN-BiLSTM-CRF	90.83	92.46	91.64

The traditional BiLSTM-CRF method has a precision of 90.40 in the hazard and operability analysis data, a recall rate of 76.14, and an F-value of 73.10. Using the same training parameters, the CNN-BiLSTM-CRF method improves the accuracy by 13.17, the recall rate by 5.40, and the F-value by 9.39. The precision of the DCNN-BiLSTM-CRF model is improved by 4.44 as compared with the CNN-BiLSTM-CRF method, the recall rate is flat, and the F-value is increased by 6.11. Compared with the DCNN-BiLSTM-CRF model, the accuracy of the ELMo-DCNN-BiLSTM-CRF model proposed in this study is improved to 90.83, the recall rate increased to 90, and the F-value reached 91.64. It shows that the proposed method proposed has a good identification effect on the HAZOP data.

The experimental results as well show that the identification accuracy of the equipment is always higher than the material, and the identification effect of the two kinds of entities is listed in Table 4.

Table 4. Material and equipment identification rates.

Entity Category	P (%)	R (%)	F (%)
Equipment	92.59	93.52	93.05
Material	86.08	83.95	85.00

Considering the influence of the size of the CNN network convolution kernel on the effect of feature extraction, seven groups of experiments were designed to compare the recognition results of the ELMo-DCNN-BiLSTM-CRF model using convolution kernels of different sizes. A convolution kernel with a length of 120 and an uncertain height was adopted, and the experimental results are shown in Table 5.

Table 5. Experimental results of different convolution kernel sizes.

Size	P (%)	R (%)	F (%)
2	91.02	90.46	90.74
3	91.25	90.87	91.06
4	91.19	90.25	90.72
5	91.23	90.66	90.95
6	91.65	91.08	91.36
7	91.51	91.70	91.61
8	91.53	91.91	91.72

It can be concluded from the experimental results that, when the window is small, the model F-value of 3 * 120 convolution kernel is higher, and when the window is large, the model with 8 * 120 convolution value has a better effect.

The two entities that need to be identified in this dataset are equipment and materials. Since some equipment are always in different positions within the overall equipment, the name of the equipment usually appears as a combination noun in the report to show the difference, which makes the length of nouns relatively long. However, the material does not need to be in the form of a combined noun, so the length of the noun is shorter. Therefore, for material nouns, the recognition effect of convolution kernel height 3 is better than that of convolution kernel height 2, and then the recognition effect drops slightly. For all entities, the recognition effect is gradually optimized with a further increase in the height of the convolution kernel. When the size of the convolution kernel is 8 * 120, the recognition effect of the network model is the best, that is, the accuracy reaches 91.53, the recall rate is 91.91, and the F-value is 91.72.

In addition, in our model, the identification results of some entities are quite different from those of other models. Table 6 lists specific examples for comparison.

Table 6. An example of model accuracy improvement.

Text	脱	醇	水	自	分	水	器	流	出
BLSTM-CRF	B-EQU	I-EQU	I-EQU	I-EQU	I-EQU	I-EQU	I-EQU	O	O
ELMo-DCNN-BLSTM-CRF	B-MAT	I-MAT	I-MAT	O	B-EQU	I-EQU	I-EQU	O	O

In Table 6, the text “脱醇水自分水器流出 ” means “dealcoholized water flows out from the water separator”, where “脱醇水 ” refers to “dealcoholized water” and “分水器 ” refers to “water separator”, which should be identified as “MAT” and “EQU”, respectively.

Table 6 shows the recognition results of the two models. The BLSTM-CRF model mistakenly identifies two entities as one entity and classifies them as "EQU", while the recognition result of the model proposed in this paper is correct. It shows that the identification result of the ELMo-DCNN-BLSTM-CRF model is more accurate.

6. Discussion

In the first experiment, four models were trained under the same experimental environment. It can be seen that BiLSTM, as a variant of RNN, can effectively solve the text sequence labeling problem and is commonly used in a NER task. However, there are still shortcomings especially in feature extraction. Therefore, the second model is combined with a convolutional neural network. The CNN can extract rich character features, significantly improving the F-value. In this study, a DCNN network is designed based on Text-CNN. The experimental results show that the accuracy is improved, and the training speed is also accelerated. After adding an ELMo language model, the F-value reaches a higher score. Combined with the example in the fourth experiment (Table 6), it clearly shows that the model proposed in this study can effectively improve the entity recognition difficulty caused by plenty of polysemous words and complex terms in petrochemical HAZOP text. Therefore, it makes the reuse and sharing of Chinese HAZOP document information more convenient and automatic and contributes to information standardization.

7. Conclusions

In view of the fact that a large amount of text data cannot be reused in hazard analysis and operability in the petrochemical industry, in this study, a new model is proposed to perform a named entity recognition task of a Chinese HAZOP document. The aim of this study is to identify and to classify materials and devices in HAZOP texts. For special problems such as a large number of polysemous words in texts of the chemical industry and lack of available annotated data, we establish the ELMo-DCNN-BiLSTM-CRF deep learning framework, and extract feature information with a pretrained ELMo language model, which can effectively solve the problem of polysemous words and is also suitable for entity recognition of small sample data. At the same time, a convolutional neural network with double-layer convolutional kernels of different sizes is used to extract word-level feature information, and then the obtained vectors are spliced to capture sequence information through the bidirection long short-term memory network, and finally the CRF layer is used to decode. The experimental results show that the proposed method is better than the existing BiLSTM-CRF and CNN-BiLSTM-CRF models in terms of accuracy rate, recall rate, and F-value. The proposed model achieves good results on a hazard and operability entity identification task, and provides a new idea for data reuse as well as intelligence and operational analysis in the chemical industry.

Author Contributions: Conceptualization, writing—original draft preparation and supervision, L.P.; writing—review and editing, Y.B.; review, supervision, and funding acquisition, D.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China (61703026).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available in this article.

Acknowledgments: Our special thanks go to Dong Gao, School of Beijing University of Chemical Technology, for his insightful suggestions and to Yao Xiao, FangGuo Li, and Yao Jia for their help in named entity recognition.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Parmar, J.; Lees, F. The propagation of faults in process plants: Hazard identification. *Reliab. Eng.* **1987**, *17*, 277–302. [CrossRef]
2. Taylor, R.J. Automated HAZOP revisited. *Process. Saf. Environ. Protect.* **2017**, *111*, 635–651. [CrossRef]
3. Dong, G.; Yao, X.; Beike, Z.; Xin, X.; Chongguang, W. Researching on HAZOP Information Standardization Based on Knowledge Ontology. *Prog. Chem. Ind.* **2020**, *39*, 2510–2518. (In Chinese)
4. Wang, S.M.; Holden, T.; Fan, C.C.; Wilhelmij, G.P. An intelligent simulation architecture for hazard and operability analysis of large-scale process plant. In Proceedings of the IEE Colloquium on Model Building Aids for Dynamic System Simulation, Coventry, UK, 9 September 1991; IET: London, UK.
5. Dunjó, J.; Fthenakis, V.; Vílchez, J.A.; Arnaldos, J. Hazard and operability (HAZOP) analysis. A literature review. *J. Hazard. Mater.* **2010**, *173*, 19–32. [CrossRef] [PubMed]
6. Mushtaq, F.; Chung, P. A systematic Hazop procedure for batch processes, and its application to pipeless plants. *J. Loss Prev. Process. Ind.* **2000**, *13*, 41–48. [CrossRef]
7. Bollacker, K.D.; Evans, C.; Paritosh, P.; Sturge, T.; Taylor, J. Freebase: A collaboratively created graph database for structuring human knowledge. In Proceedings of the Sigmod Conference (2008), Vancouver, BC, Canada, 10–12 June 2008.
8. Zhang, Y.; Zhou, J.F. A trainable method for extracting Chinese entity names and their relations. In Second Chinese Language Processing Workshop. In Proceedings of the Second Chinese Language Processing Workshop, Hong Kong, China, 8 October 2000; Zhang, Y., Zhou, J.F., Eds.; Association for Computational Linguistics (ACL): Stroudsburg, PA, USA, 2000; pp. 66–72.
9. Shijia, E.; Xiang, Y. Chinese Named Entity Recognition with Character-Word Mixed Embedding. In Proceedings of the The 26th ACM International Conference on Information and Knowledge Management (CIKM), Singapore, 6–10 November 2017.
10. Jinfeng, Y.; Qiubin, Y.; Yi, G.; Zhipeng, J. A survey of named entity recognition and entity relation extraction in electronic medical records. *J. Autom.* **2014**, *40*, 1537–1562. (In Chinese)
11. Reimers, N.; Gurevych, I. Alternative Weighting Schemes for ELMo Embeddings. *arXiv* **2019**, arXiv:1904.02954.
12. Nadeau, D.; Sekine, S. A survey of named entity recognition and classification. *Lingvist. Investig.* **2007**, *30*, 3–26. [CrossRef]
13. Fan, T.; Tong, F.; Luo, Z.; Zhao, D. A deep network based integrated model for disease named entity recognition. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Kansas, MO, USA, 13–16 November 2017.
14. Wu, Y.; Jiang, M.; Lei, J.; Xu, H. Named Entity Recognition in Chinese Clinical Text Using Deep Neural Network. *Stud. Health Technol. Inform.* **2015**, *216*, 624–628. [PubMed]
15. Zhen, S.; Huilin, W. A review of named entity recognition. *Mod. Libr. Inf. Technol.* **2010**, 42–47. (In Chinese)
16. Hettne, K.M.; Stierum, R.H.; Schuemie, M.J.; Hendriksen, P.J.M.; Schijvenaars, B.J.A.; Van Mulligen, E.M.; Kleinjans, J.; Kors, J.A. A dictionary to identify small molecules and drugs in free text. *Bioinformatics* **2009**, *25*, 2983–2991. [CrossRef] [PubMed]
17. Rebholz-Schuhmann, D.; Kirsch, H.; Arregui, M.; Gaudan, S.; Riethoven, M.; Stoehr, P. EBIMed—Text crunching to gather facts for proteins from Medline. *Bioinformatics* **2007**, *23*, e237–e244. [CrossRef] [PubMed]
18. Schuemie, M.J.; Schuemie, M.J.; Jelier, R.; Kors, J.A. Peregrine: Lightweight gene name normalization by dictionary lookup. In *Proc of the Second BioCreative Challenge Evaluation Workshop*; Erasmus University Medical Center: Rotterdam, The Netherlands, 2007; pp. 131–135.
19. Koning, D.; Sarkar, I.N.; Moritz, T. TaxonGrab: Extracting Taxonomic Names from Text. *Biodivers. Inform.* **2005**, *2*, 79–82. [CrossRef]
20. Gerner, M.; Nenadic, G.; Bergman, C.M. LINNAEUS: A species name identification system for biomedical literature. *BMC Bioinform.* **2010**, *11*, 85. [CrossRef]
21. Berger, A.L.; Pietra, S.A.; Della Pietra, V.J. A maximum entropy approach to natural language processing. *Comput. Linguist.* **1996**, *22*, 39–71.
22. Lai, P.-T.; Huang, M.-S.; Yang, T.-H.; Hsu, W.-L.; Tsai, R.T.-H. Statistical principle-based approach for gene and protein related object recognition. *J. Chem.* **2018**, *10*, 64–69. [CrossRef]
23. Nguyen, N.-V.; Nguyen, T.-L.; Thi, C.-V.N.; Tran, M.-V.; Nguyen, T.-T.; Ha, Q.-T. Improving Named Entity Recognition in Vietnamese Texts by a Character-Level Deep Lifelong Learning Model. *Vietnam J. Comput. Sci.* **2019**, *6*, 471–487. [CrossRef]
24. Ritter, A.; Clark, S.; Etzioni, O. Named Entity Recognition in Tweets: An Experimental Study. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK, 27–31 July 2011.
25. Bi, M.; Zhang, Q.; Zuo, M.; Xu, Z.; Jin, Q. Bi-directional LSTM Model with Symptoms-Frequency Position Attention for Question Answering System in Medical Domain. *Neural Process. Lett.* **2020**, *51*, 1185–1199. [CrossRef]
26. Pandey, S.K.; Janghel, R.R. Recent Deep Learning Techniques, Challenges and Its Applications for Medical Healthcare System: A Review. *Neural Process. Lett.* **2019**, *50*, 1907–1935. [CrossRef]
27. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.
28. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv* **2015**, arXiv:1508.01991.
29. Ma, X.; Hovy, E. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv* **2016**, arXiv:1603.01354.
30. Pei, Y.; Zhihao, Y.; Ling, L.; Hongfei, L.; Jian, W. Chemical drug named entity recognition based on attention mechanism. *Comput. Res. Dev.* **2018**, *55*, 1548–1556. (In Chinese)
31. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. *Efficient Estimation of Word Representations in Vector Space*; Computer Science: San Diego, CA, USA, 2013.

32. Mikolov, T.; Yih, W.T.; Zweig, G. Linguistic Regularities in Continuous Space Word Representations. In Proceedings of the 2013 Conference of The North American Chapter of The Association for Computational Linguistics: Human Language Technologies, Atlanta, GA, US,, 9–14 June 2013.
33. Deping, C.; Bo, W.; Hong, L.; Fang, F.; Run, W. Geological entity recognition based on ELMO-CNN-BILSTM-CRF model. *Geoscience* **2021**, 1–22. (In Chinese)
34. Kim, Y. *Convolutional Neural Networks for Sentence Classification*; EMNLP: Doha, Qatar, 2014.
35. Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; Dyer, C. Neural Architectures for Named Entity Recognition. *arXiv* **2016**, arXiv:1603.01360.