Computer Modeling in
Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2022.022268

ARTICLE

Corpus of Carbonate Platforms with Lexical Annotations for Named Entity Recognition

Zhichen Hu[1], Huali Ren[2], Jielin Jiang[1], Yan Cui[4], Xiumian Hu[3] and Xiaolong Xu[1,*]

[1]School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, 210044, China

[2]Institution of Artificial Intelligence and Blockchain, Guangzhou University, Guangzhou, 515021, China

[3]School of Earth Sciences and Engineering, Nanjing University, Nanjing, 210023, China

[4]College of Mathematics and Information Science, Nanjing Normal University of Special Education, Nanjing, 210023, China

*Corresponding Author: Xiaolong Xu. Email: xlxu@ieee.org

Received: 01 March 2022 Accepted: 20 May 2022

ABSTRACT

An obviously challenging problem in named entity recognition is the construction of the kind data set of entities. Although some research has been conducted on entity database construction, the majority of them are directed at Wikipedia or the minority at structured entities such as people, locations and organizational nouns in the news. This paper focuses on the identification of scientific entities in carbonate platforms in English literature, using the example of carbonate platforms in sedimentology. Firstly, based on the fact that the reasons for writing literature in key disciplines are likely to be provided by multidisciplinary experts, this paper designs a literature content extraction method that allows dealing with complex text structures. Secondly, based on the literature extraction content, we formalize the entity extraction task (lexicon and lexical-based entity extraction) for entity extraction. Furthermore, for testing the accuracy of entity extraction, three currently popular recognition methods are chosen to perform entity detection in this paper. Experiments show that the entity data set provided by the lexicon and lexical-based entity extraction method is of significant assistance for the named entity recognition task. This study presents a pilot study of entity extraction, which involves the use of a complex structure and specialized literature on carbonate platforms in English.

KEYWORDS

Named entity recognition; carbonate platform corpus; entity extraction; english literature detection

1 Introduction

The data resources (e.g., carbonate platform literature) contain rich information about geological entities. Therefore, named entity recognition has a wide range of real-life applications, such as entity-relationship connection and object source inference [1]. How we translate these data into new understandings or knowledge to provide actionable data support for practical applications has become a hot topic of research [2].

In recent years, natural language processing has gained wide attention in academia and industry since Google proposed the Knowledge Graph. Information extraction is a pivotal technique in natural language processing, while named entity recognition (NER) is the most valuable part of the information extraction task [3]. As a result, some studies on named entity recognition have been accomplished. However, most of them are mainly devoted to the extraction of entity elements from texts by natural language processing techniques, i.e., the identification of commonly used entity categories (e.g., locations, names of people and organizations) [4]. In fact, the entity is a complex concept of knowledge cognition. In addition to the commonly used entity categories, disciplines across domains demand accurate and efficient extraction of disciplinary entity data used for research purposes and use them to build domain knowledge architecture structures so that disciplinary experts rapidly discover hidden clues between relevant knowledge points [5]. For example, the case "Early Cretaceous sedimentary evolution of the northern Lhasa terrane" has the substance entity "sedimentary evolution," the temporal entity "Early Cretaceous" and the location entity "the northern Lhasa". Therefore, these three entities need to be extracted for in-depth statement analysis. In this paper, we attempt to study the generic entity extraction approach.

Although named entity recognition corpora exist, they are primarily focused on Wikipedia (a structured text) [6]. However, since carbonate platform literature is an informal text with fewer restrictions on the style of writing papers, entity extraction in the discipline requires its corpus. Traditional entity kind extraction requires manual reading for extraction, which is time-consuming and prone to problems such as error propagation. In recent years, some systems within the discipline have been constructed [7]. However, the construction method still takes traditional construction methods and does not provide a data set construction process. With the rise of deep learning, some researchers have converted the named body recognition task into a sequence tagging task, which does not essentially solve the corpus generation problem [8].

By taking advantage of the combined lexicon and lexical model, we structure the entity extraction task with the carbonate platform literature and propose an end-to-end domain knowledge corpus construction. Based on our carbonate platform corpus, we investigate named body recognition models within the subject domain. The accuracy of the corpus is tested by three popular named entity recognition models, namely LSTM-CRF (Long and Short Term Memory Network and Conditional Random Fields), BILSTM-CRF (Bidirectional Long and Short Term Memory Network and Conditional Random Fields) and BERT-BILSTM-CRF (Bidirectional Encoder Representation from Transformers and Bidirectional Long and Short Term Memory Network and Conditional Random Fields), since the models finally Both add CRF layers, CRF is able to take into account the dependencies between tags and set linguistic specification rules, allowing it to address linguistic irregularities in various English texts. The contribution of this study is threefold:

1) In this paper, a combined lexicon and lexical model is proposed to construct the domain data set autonomously in the literature as a novel corpus construction method.
2) In this paper, we compare the accuracy of three models, LSTM-CRF, BILSTM-CRF and BERT-BILSTM-CRF, on this data set.
3) We share the construction data set source code and data set corpus, which can be downloaded from the following website, https://gitee.com/hu_zhichen/carbonate-platform-dataset. The shared code includes the training data set and labeled test data.

The article is organized as follows. Section 2 describes the corpus construction method and the work related to named corpora. Section 3 presents the lexicon and lexical model construction methods. Section 4 applies the proposed method to the constructed corpus to test its performance. Section 5 gives an in-depth experimental analysis. Section 6 emphasizes the main contributions of the paper with future work given.

2 Related Work

In the following summary, we first describe the existing named entity recognition corpus and previous methods for constructing the corpus. Second, this paper outlines the approach to named entity recognition.

2.1 Named Entity Recognition Corpus

In the linguistic literature, Arabic language experts provide free and manual commentaries containing more than 7,000 hadiths. Based on Islamic themes, the simultaneous experts classified named entities into 20 types. In addition, Salah et al. [9] introduced a comprehensive statistical analysis to measure the factors that play an important role in the manual exegesis. To reduce the manual annotation time, Boroş et al. [10] proposed to build a multilingual medieval handwritten charter image corpus using text recognition combined with named body recognition techniques, which experimentally demonstrated the superior performance of the combined approach for name, date and location recognition. Faced with the evaluation of named entity recognition tasks for resource-poor languages like Punjabi, Kaur et al. [11] developed an annotated corpus of 200,000 words on their own. To evaluate the accuracy of the corpus, experiments were conducted using Hidden Markov Models, Maximum Entropy and Conditional Random Fields, respectively. The experiments showed a maximum F1 of 93.21%. Drovo et al. [12] proposed to generate a named body tag corpus for Bengali using a rule-based approach with machine learning, which consists of 10k words that have been manually annotated with seven tags. Tables in medical articles typically contain important information about research results. However, tables are unstructured leading to semantic complexity that ultimately fails to be read directly by applications. Wei et al. [13] used a manually annotated corpus to simultaneously identify biomedical entities in table headings in randomized controlled trial articles to construct a biomedical corpus.

2.2 Lexical Construction

Lexical construction, a natural language modelling is an essential issue in the theory of intelligence. The central object of natural language text description is the word. For languages with a high degree of grammar, the task of grammar recognition is complex and time-consuming, hence the need for automatic word replacement filtering based on lexicality. For this reason, Chetverikov et al. [14] proposed to combine word lexicality to approach the problem of homophones. Because of the multi-disciplinary nature of qua-disciplinary terms, which have different and diametrically opposed meanings when preceded and followed by different collocations, Takhom et al. [15] detected ambiguous interdisciplinary terms by using web-based text analysis. The NLTK (Natural Language Toolkit) technique incorporates lexical annotation features, Elbes et al. [16] collected news articles using SVM (Support Vector Machine) and NLTK, and classified the data set. In addition, the lexical screening function is used in essay writing activities, Contreras et al. [17] used NLTK to apply natural language processing algorithms to improve essay ratings. To instantly deal with emotions in social media, Jha et al. [18] proposed the conversion of text into emoticons to recognize emotions in text, where lexical annotation is the underlying logic of this technique.

2.3 *Named Entity Recognition*

Named entity recognition, as an essential sub-discipline of artificial intelligence, has a task assignment aiming to identify a set of predefined entity types that have defined token types. To address the problems of ambiguous entity recognition and little labeled data in the field of Chinese medicine, Qu et al. [19] constructed a named entity recognition model based on Bert-BILSTM-CRF and tested it on the corresponding data set. The results show that the model has the highest accuracy in recognizing drug names. The advancement of digital technology has led to the generation of a large amount the data. To rapidly improve entity classification, Tripathi et al. [20] used the K-means algorithm for entity classification. To improve the accuracy of Chinese named body recognition, Du et al. [21] proposed to embed the word position encoding of each word into the word vector to produce the Chinese entity boundary with high heel. The experimental results showed that the method improved the F1 value by about 1%. Facing the semantic disambiguation problem, Shah et al. [22] explored the use of different pooling mechanisms, used together with BILSTM-CRF architecture to obtain the whole sentence context information. Web texts contain rich security events, a novel web security entity recognition model using BILSTM-CRF was proposed by Ma et al. [23] for extracting security-related concepts and entities from unstructured texts.

3 Methods for Constructing the Lexicon and Lexical Models

Recently, there is no corpus and annotated data set publicly shared in the subject area of carbonate platforms in sedimentology [24]. However, the domain literature contains rich contextual information that is of significant research interest for subject experts to work on and summarize results over time; therefore, this section details corpus construction methods and named entity categories containing substance, location, and age [25]. As shown in Fig. 1, the model consists of three layers: a document excision layer, a lexical combination layer, and a corpus construction layer. In the document excision layer, the documents are disassembled into machine-readable text files according to requirements. The disassembled documents are then fed into the lexical combination layer, where the lexical properties are utilized to filter out the keywords that satisfy the lexicon requirements. Finally, the sentences in which the keywords are found are entered into the corpus construction layer, which ultimately generates the carbonate platform corpus.

3.1 *Sources of English Literature and Construction of Lexicon*

The data sources selected for this paper were mainly from publicly shared English literature. As shown in Table 1, we mainly collected literature information through three English-language websites (GeoScience, Sciencedirect and SpringerLink). The purpose of collating the literature is to systematically analyse the evolution of carbonate platforms over time, identify closely related substance features and location distributions under each epoch, to establish an evolutionary network of spatial-temporal objects.

However, since the literature is an informal text, there are specific typesetting styles for each journal literature that is impossible to standardize, on top of the fact that references, titles and authors do not provide valuable textual content. Therefore, for making each type of journal layout pattern, the machine is first used to mark the start and end characters before the abstract and after the conclusion. This minimizes the information interference caused by incorrect information. For example, in the journal “GEOLOGICAL JOURNAL”, the literature will not have abstract word starters, so it is necessary to add them. The reference will appear at the same time in the context of the description of the Spatial temporal substance that does not belong to the main description of the article, then it needs to be deleted.

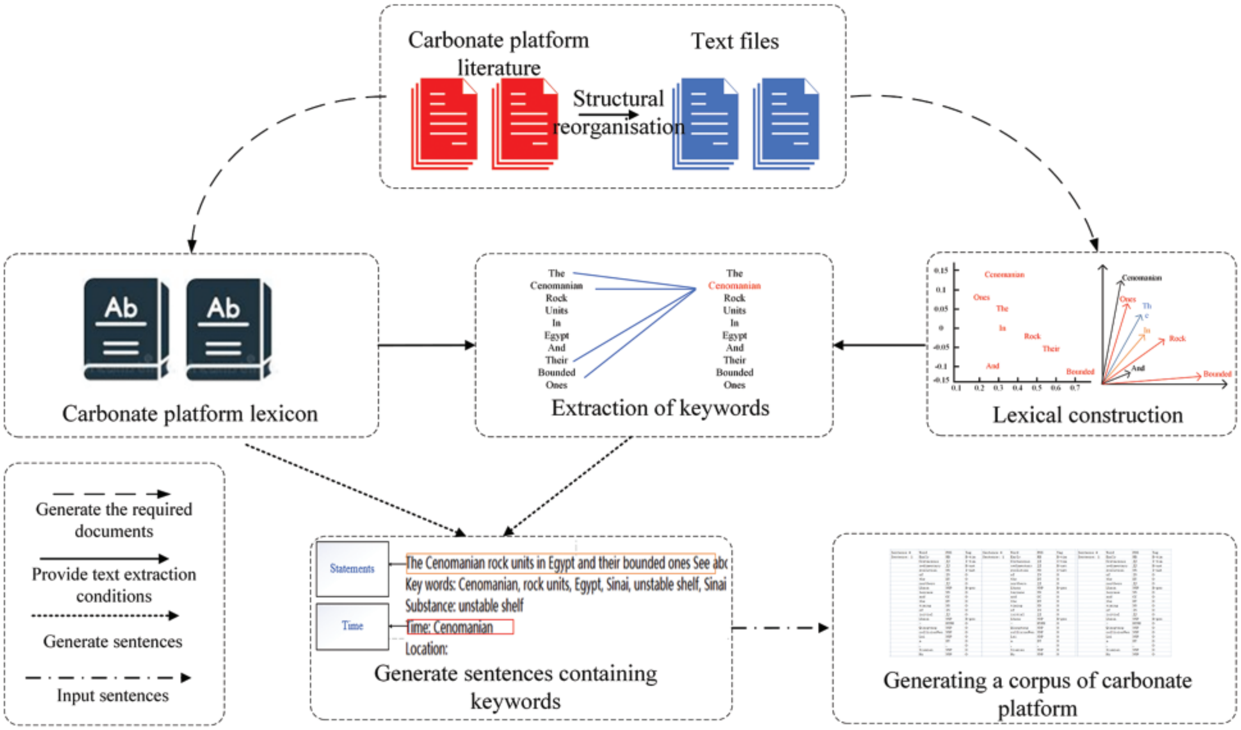
Carbonate platform literature
Text files
Structural reorganisation
Carbonate platform lexicon
The
Cenomanian
Rock
Units
In
Egypt
And
Their
Bounded
Ones
Extraction of keywords
Lexical construction
Generate the required documents
Provide text extraction conditions
Generate sentences
Input sentences
Statements
Time
The Cenomanian rock units in Egypt and their bounded ones See abc
Key words: Cenomanian, rock units, Egypt, Sinai, unstable shelf, Sinai
Substance: unstable shelf
Time: Cenomanian
Location:
Generate sentences containing keywords
Generating a corpus of carbonate platform

Figure 1: Architecture for implementing lexical and lexicon models

Table 1: Sources of literature

Source	Year of literature	Number of literature
GeoScience	2010–2021	500
Sciencedirect	2007–2021	600
SpringerLink	2016–2020	400
Geological Journal	2016–2020	500

3.2 Define Entity Lexicon

The construction of domain dictionaries enhances the efficiency of text search and lexical matching. The construction of lexicons enables the gradual enrichment of lexicons in a snowball construction mode through the collection of subject textbooks or literature keywords. At the same time, defining domain lexicon has a guiding role for information extraction. The descriptive statements in the data source are analyzed and disassembled to extract and determine the categories of carbonate platform knowledge entities. As shown in Table 2. For example, if the Cretaceous period appears in the dictionary, the researcher in turn requires to extract the words that connect before and after it, such as Early Cretaceous, or Late Cretaceous for accurate exploration.

Table 2: Classification of carbonate platforms and a set of examples

No.	Entity	Example
1	Cretaceous	Early Cretaceous, Lower Cretaceous, Middle Cretaceous
2	Formation	Duba Formation, Duoni Formation
3	Ramp	Carbonate Ramp, Shallow Ramp, Homoclinal Ramp
4	Aptian-Albiann	Aptian-Albian, Late AptianeAlbian, Late Aptiane early Albian
5	Barremiane	Upper Barremiane, Late Barremiane
6	Platforms	Ramp-Geometry Platforms, Inner Platforms
7	Rimmed	Rimmed Shelf, Rimmed Platforms

3.3 Define Entity Lexical

Lexicality is inherent to the expression of an individual word, either according to the grammatical structure of the word natively in the sentence, or according to the form of presentation in which the word is presented. In regular expressions [A−Z] stands for English characters, numeric characters are represented by [0–9], "?" indicates matching the previous expression once, "∗" means matching the previous expression any number of times, "^" shows matching from the current position, and "$" marks the end of the previous expression. In lexical definitions, where "NNP" stands for proper nouns, "CD" represents numbers, "NN" denotes nouns, "AT" indicates prepositions, and JJ is represented by the adjective. This paper combines both, constraining the particular forms of presentation within the inherent grammatical structure with regular expressions. The result is a lexical annotation method that is applicable to lexicons. As shown in Table 3, a total of 17 lexical features were defined for this order, subsequently regrouping these lexical features in a total of 12 combinations to satisfy the key information appearing in the utterance. For example, "∼" and numbers are defined as the lexical characteristic "VB" whereas featureless words are defined as "NNP", so that when the phrase "∼5 MA," the model will recognize the phrases that meet the time constraint for further filtering. In addition, the lexical definitions are made more flexible by the inclusion of regular expressions, so that the lexical definitions are adaptively modified in real time according to changes in lexicon type and extraction requirements.

Table 3: Carbonate platform lexical definitions and combinations

Types of word performance	Lexical definitions	Lexical combinations
r′^-?[0–9]-(.[0–9]-)?$′	'CD'	["CD-NNP"] = "NNP"
r′(-|:|;)$′	':'	No requirement
r′\'∗$′	'MD'	No requirement
r′(The|the|A|a|An|an)$′	'AT'	No requirement
r′^±′	'JJ'	["CD-JJ"] = "NNP"
r′^ [A-Z].∗$′	'NNP'	["NNP-NNP"] = "NNP"
r′.∗ness$′	'NN'	["NN-NN"] = "NNI"
r′.∗ly$′	'RB'	No requirement
r′.∗s$′	'NNS'	No requirement
r′^∼?[0–9]′	'VB'	["VB-NNP"] = "NNP"

(Continued)

Table 3 (continued)

Types of word performance	Lexical definitions	Lexical combinations
r′.∗ing$′	‘VBG’	[“VBG-NNP”] = “NNP”
r′.∗ed$′	‘VBD’	[“VBD-NNP”] = “NNP”
r′and$′	‘CC’	[“CD-CC”] = “NNP”
r′.∗′	‘NN’	[“JJ-NN”] = “NNI”

3.4 Lexicon and Lexical Model

The entity data set is the core node in named entity recognition, where its extraction accuracy takes an influential part in pre-trained models. In this section, we present the proposed lexicon and lexical model for composed entity extraction in the field of a sedimentological carbonate platform, as shown in Fig. 2. The model has five main steps: text and lexicon input (Step 1), text segmentation (Step 2), lexical construction (Step 3), lexical filtering (Step 4) and output (Step 5).

Step 1. Based on the carbonate platform literature, the required literature and lexicon for Chapters 3.1 and 3.2 were entered into the model as initial data. Firstly, the abstract and acknowledgements are added consistently to the beginning and end of the article as a positioning point in the literature. Next, the key words from the subject structure are collected and stored in a thesaurus, which is classified into three lexicons on the basis of three categories: time, location and substance. Ultimately the model is batched to access the contents of the lexicons.

Step 2. First, read the text content between the abstract and the acknowledgments, remove images and table information, and delete special characters (e.g., Fig. number, name et al.), as shown in Algorithm 1. Next, the text is divided by removing the paragraph structure in the paper, using the English period as a segmentation marker. Finally, the lexicon words are read to match with the sentences in the newly divided text, the sentences containing the lexicon words in the upper and lower sentences are put into the extracted text, where the lexicon words are obtained by subject knowledge system extraction.

Step 3. Construct lexical categories that match with the lexicon, e.g., define the lexical category of special English words ending in “ed” as ‘VBD’ and the lexical category of no special English words as ‘NNP’, combining the two lexical categories to redefine the lexical category of ‘NNP’. For example, in the sentence ‘Cretaceous carbonate system of the Arabian Plate seems not to be specific for the rudist-barrier rimmed platform to intra-shelf basins . . . ’, ‘rimmed platform’ is consistent with the defined lexical feature, so the lexical construction is consistent with the actual labeling data set construction requirements.

Step 4. Since the lexical screening in Step 3 generates a large number of interfering phrases, which results in lower accuracy of the data set. Therefore, the extracted phrases based on lexical properties have to be lexicon matched twice. The phrases that match the filtered lexical properties but fail to match the lexicon range are removed. Afterwards, as shown in Algorithm 2, these phrases conform to both lexicon and lexical requirements.

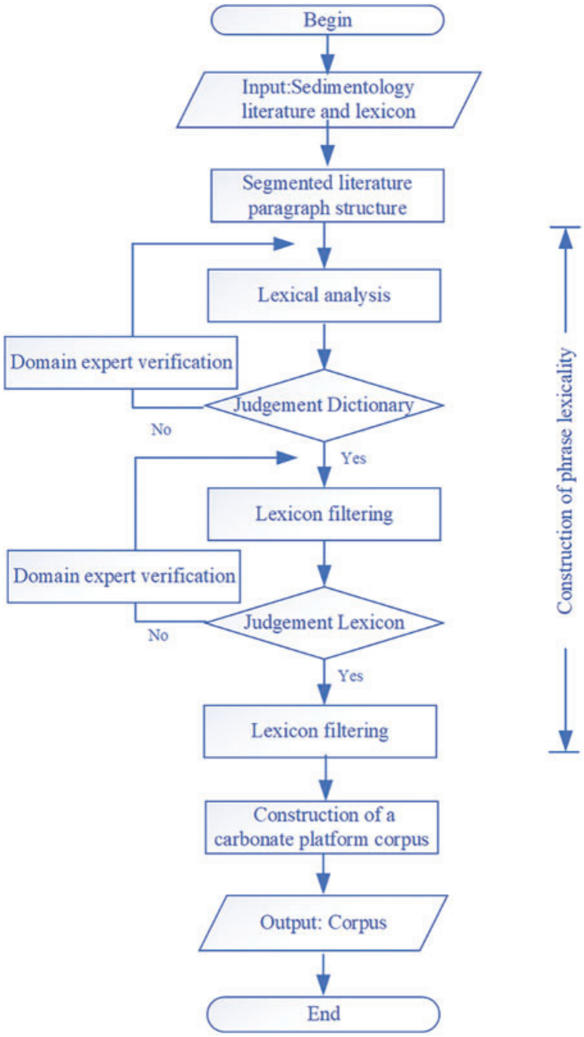
Begin
Input:Sedimentology literature and lexicon
Segmented literature paragraph structure
Lexical analysis
Domain expert verification
Judgement Dictionary
No
Yes
Lexicon filtering
Domain expert verification
Judgement Lexicon
No
Yes
Lexicon filtering
Construction of a carbonate platform corpus
Output: Corpus
End
Construction of phrase lexicality

Figure 2: Programming flowchart for the generation of the carbonate platform corpus

Algorithm 1: Text filtering

```
**Input:** input carbonate platform text
**Output:** output contains keyword sentences
**For** each index **in** text **do:**
   **If** text[index] find Fig. number or name et al. **then:**
      text[index] remove Fig. number or name et al.
      text[index] are stored in new text
   **End if**
**End for**
**For** each index **in** new text **do:**
   **If**  text[index] find keyword **then:**
```

(Continued)

Algorithm 1: (Continued)

```
   This.sentences ← 1
   If text[index – 1] find keyword(Another label categories) then:
     Next.sentences ← 1
   End if
  End if
  If This.sentences = 1 and line1geo = 1 then:
    output: new text[index]−new text[index – 1]
  End if
End for
```

Algorithm 2: Data quality control

```
Input: input carbonate platform lexicon, lexical properties, contains keyword sentences
Output: output removing redundant sentences
For each index in sentences do:
  For each indexLexicon in lexicon do:
     If sentences [index] find lexicon [indexLexicon] then:
         sentences [index] ← 1
     End if
  End for
End for
For each index in sentences do:
  For each properties in lexical do:
    If sentences [index] find lexical [properties] and properties != 'O' then:
        sentences [index] ← 2
    End if
  End for
End for
If This.sentences = 1 and line1geo = 2 then:
   output: sentences [index]
End if
```

Step 5. The phrases filtered out in Step 4. Annotate them with reference to the entity annotation pattern in named entity recognition, as shown in Table 4. The final corpus of carbonate platform entities was constructed. The corpus contains a total of 5008 relevant utterances and 270,000 entity annotations, which are stored in the form of tables to satisfy the number of training sets and test sets required for the named entity recognition model.

3.5 Carbonate Platform Data Set

The carbonate platform corpus, conforming to the BIO international labeling rules, where "B" represents the beginning position of a sentence that matches the labeled phrase, "I" the middle position, and "O" the remaining words that will not match the label. In the corpus, "tim" was defined as the time label, "geo" as the location label and "nat" as the substance label. Therefore, "B-tim" represents the first word at the beginning of the time phrase, "I-time" for the rest of the time phrase. "B-nat" represents the first word at the beginning of the substance phrase, "I-nat" for the rest of the substance phrase. "B-geo" represents the first word at the beginning of the location phrase, "I-geo" for the rest of the location phrase. As shown in Fig. 3, the percentage of labels counted is 14,677 time labels, accounting for 36% of valid labels; 17,787 substance labels, accounting for 44% of valid labels; 8322 location labels, accounting for 20% of valid labels. In total, valid labels account for 15% of the total labels. Apparently, the number of valid tags satisfies the minimum number available for the corpus. However, the "O" label accounts for the majority of the ratio in the labels, therefore the effect of changing the label on the model needs to be reduced. For this reason this paper introduces CRF models into both models to adjust for the interference of "O" on the predicted values.

Table 4: Example sentences with annotation

No.	Entity	Example
1	Early Cretaceous sedimentary evolution of the northern Lhasa terrane and the timing of initial Lhasa.	<B-tim > Early <I-tim > Cretaceous <B-nat > sedimentary <I-nat > evolution <O > of <O > the <O > northern <B-geo > Lhasa <O > terrane <O > and <O > the <O > timing <O > of <O > initial <B-geo > Lhasa.
2	Lower Cretaceous strata in the Baingoin basin of the northern Lhasa terrane record initial collision.	<B-tim > Lower <I-tim > Cretaceous <O > strata <O > in <O > the <B-geo > Baingoin <O > basin <O > of <O > the <O > northern <B-geo > Lhasa <O > terrane <O > record <O > initial <O > collision.
3	Abundant volcanic clasts, detrital zircons Yielding Cretaceous ages.	<O > Abundant <B-nat > volcanic <I-nat > clasts <O > , <B-nat > detrital zircons <I-nat > Yielding <B-tim > Cretaceous <O > ages.

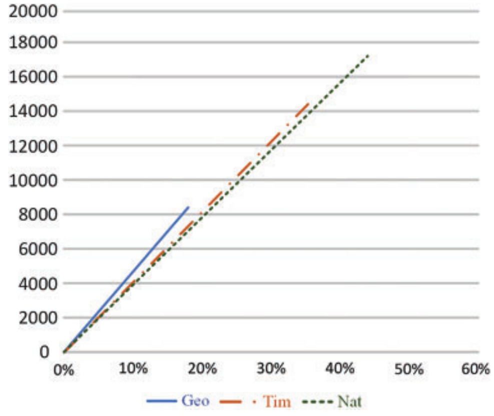

Figure 3: Percentage of each label in the corpus

4 The Proposed Methods

The information extraction method proposed in this paper mainly consists of LSTM-CRF, BILSTM-CRF and BERT-BILSTM-CRF, which are upgraded and simplified in a recursive relationship. BILSTM, compared to LSTM, not only takes into account the above information, but also combines the below information. The BERT pre-trained language model uses a bi-directional transformer as the feature extractor, which allows for longer contextual information and improved feature extraction compared to the first two models. The CRF model combines the features of maximum entropy and the hidden Markov model, which is able to consider not only the state of the previous moment, but also multiple states before and after it, taking into full consideration the dependencies between the labels, making CRF have better label sequential properties. Three models were experimented with python 3.6, 64 GB of RAM, Core i5 CPU and RTX 3090 GPU.

4.1 LSTM-CRF for Named Entity Recognition

The LSTM-CRF model has four layers, namely the embedding layer, the LSTM layer and the softmax layer. For each sentence, the LSTM language takes the corresponding word sequence as input. The embedding layer is used to embed each word into a continuous space where semantically similar words are placed close to each other, with the LSTM layer used to encode the contextual information of each word. After that, the LSTM layer returns a hidden sequence. Then, the hidden representation is normalized to the probability distribution of the input sentences using a softmax function. Finally, the entities contained in the descriptive text are determined according to the CRF model.

4.2 BILSTM-CRF for Named Entity Recognition

In the BiLST-CRF sequence annotation model, combined with the characteristics of LSTM-CRF, a reverse LSTM layer is added to the original LSTM layer to form a BILSTM layer, and the BILSTM layer is used to integrate and extract the features of the input sequence. The BILSTM layer combines the word vectors of contextual information, and the final output is the probability distribution matrix of the label category of each character in the sequence. Finally, the CRF layer is used to determine the feasible label sequence according to the probability distribution matrix.

4.3 BERT-BILSTM-CRF for Named Entity Recognition

BERT-BILSTM-CRF is used for named entity extraction in carbonate mesa, integrating the strengths of the first two models, as shown in Fig. 4. The model has 5 main layers: input layer, BERT layer, BILSTM layer, CRF layer and output layer. The BERT pre-trained model is used to obtain the features of the input sentence text and convert each character into a word vector and a character vector. In the BILSTM layer, in order to train the parameters, the output is the score of all labels corresponding to each word, and the probability distribution result of the label sequence is calculated according to the CRF model.

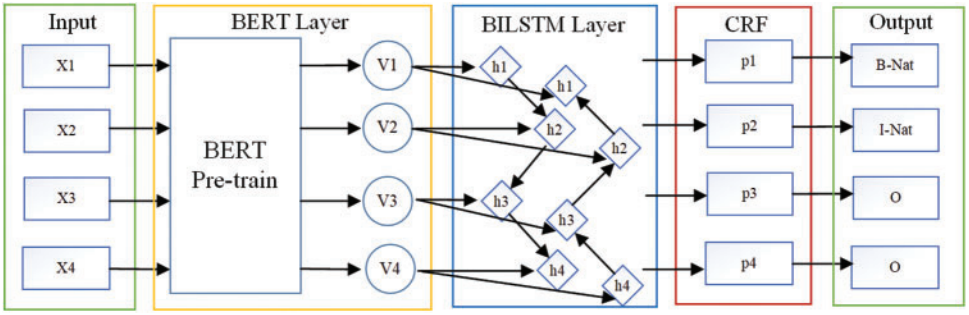

Figure 4: Architecture of the BERT pre-trained language model

Input layer. The input layer contains a series of text segments, and the sentence is split into each word, that is, each word represents a vector consisting of character embeddings or word embeddings. For example the input "Early Cretaceous Carbonate Platform".

Bert layer. BERT uses a bidirectional Transformer as an encoder to better understand the meaning of words and the rich syntactic and semantic information in sentences. The vector of the input layer is input, and the BERT decoding structure can be calculated in parallel to improve the calculation efficiency.

BILSTM layer. BILSTM combines the BERT long utterance computational capability to obtain contextual information and learn the dependencies between contexts. BILSTM wraps for bidirectional LSTM, captures contextual information completely, extracts deep semantic features, and finally connects two LSTM networks to the same CRF layer.

CRF layer. As the output layer of the model, CRF generates the sequence annotation results of the text. The output is the label of each unit in the sentence. For example, the beginning of a sentence can only be "B" or "O", not "I". And the set of labels must be standardized, e.g., B-tim must end with "O" or "I-tim" and not with "I-geo" or "I-Nat".

Output layer. The last layer reduces the annotated sequence based on the output of the CRF model to extract the corresponding entity labels with the highest probability. The final input statement corresponds to the output "B-tim I-tim B-nat I-nat"

5 Experimental Results

In this section, LSTM-CRF, BILSTM-CRF and BERT-BILSTM-CRF are selected to test the corpus. Because of the large number of O labels in the data set, a CRF layer is added to all of the named entity recognition models for applying constraints to ensure that the predicted labels are reasonable.

5.1 The Evaluation Metric

In the experiments of this paper, we have selected three common classification metrics: precision (Pre), recall (Rec), and F1 score (F1). The precision metric reflects the ratio of the number of correctly recognized samples to the actual number of samples, which is the ratio of the number of relevant carbonate platform entities extracted to the total number of entities extracted, as shown in Eq. (1). The recall metric reflects the ratio of the number of correctly identified samples to the number of samples that should have been recognized, i.e., the ratio of the number of relevant carbonate platform entities extracted to the total number of entities that should have been extracted, as shown in Eq. (2). Generally speaking, in the process of evaluating a model, assessing the model using only precision or recall is not a comprehensive assessment of the merit of the model, therefore precision and recall are combined to obtain the F1 score as the actual scoring criterion for the model. Thus, as shown in Eq. (3), F1 is the weighted summed average of precision and recall, representing the balance between the first two metrics. The ultimate results of named entity recognition are compared with the data set annotation for assessing the quality of named entity recognition.

$$Precision(P) = \frac{TP}{TP + FP}, \tag{1}$$

$$\mathrm{Re}call(R) = \frac{TP}{TP + FN}, \tag{2}$$

and

$$\mathrm{F1} = 2\frac{PR}{P + R}, \tag{3}$$

where TP is the positive sample predicted by the model as positive class, i.e., the positive sample predicted correctly, FP is the negative sample predicted by the model as positive class, i.e., the positive sample predicted incorrectly, and FN is the positive sample predicted by the model as negative class, i.e., the negative sample predicted incorrectly.

5.2 Model Parameter Setting

Parameter settings can have a significant impact on the performance of a model. In named entity recognition, three model parameters are uniformly defined to compare the performance of the corpus, such as epochs and batch_size to determine the parameters of the model. Where batch_size is the amount of data accessed by the model at one time and epochs is the number of iterations of the model, while a time function is introduced in this article to calculate the time consumption for each set of tests. Using the BILSTM-CRF model as an example, the model mainly contains should input gates, output gates, hidden gates and bias, so the overall number of parameters is defined as N, which is given by the following Eq. (4):

$$N = 8 * Hidden\_size * (Input\_size + Bias + Output\_size), \tag{4}$$

For the parameters in the model, Output_size and Hidden_size are practically the same. Ultimately, this reduces to two matrices, which map the inputs and outputs respectively, with the dimension of U being hidden∗Input and the dimension of V being hidden∗hidden. Therefore, the space complexity of the simplified model is Eq. (5), referred to as Eq. (6).

$$N = 8(Hidden * Input + Hidden * Hidden + Hidden), \tag{5}$$

$$Space \sim o(8(nm + n^2 + n)), \tag{6}$$

where *n* is the hidden_size and *m* is the input_size. The final time complexity of the model is the number of operations on the number of parameters of the model, which is given by the Eq. (7).

$$Time \sim o(Epochs\ *\ 8(nm + n^2 + n)) \sim o(n^2). \tag{7}$$

As shown in Table 5, when epochs are increased to 20 rounds and batch_size is increased to 32, the time consumption is not significantly changed compared to the last two sets of tests, although the overall precision is the highest. Therefore, this paper defines epochs as 20 and batch_size as 32. Dropout is introduced to prevent over-fitting during model computation and the dropout ratio is 0.1.

Table 5: Model parameter setting

Epochs	Batch_size	Time	Precision
10	8	83 s	97.14%
15	16	66 s	97.58%
20	32	50 s	98.47%
25	64	37 s	98.13%
30	128	43 s	97.73%
35	256	36 s	96.14%

5.3 Accuracy of Models in the Corpus

The experimental results are shown in Table 6. In the experiments, we can see that the accuracy of these three models reaches more than 90% in the training set. Compared with the HMM (Hidden Markov Model), CRF (Conditional Random Fields), BILSTM benchmark methods, the corpus designed by this method has been able to obtain excellent performance results in named entity recognition models. Among them, the precision of BILSTM-CRF recognition is the highest, which can reach 98.47%; the recall rate basically reaches 98%; the F1 score of causal relationship is 98.70%, while the F1 scores of the other three relationship types all reach 91% or more, as can be seen from Table 5, the average values of precision, recall rate and F1 score of these three models are 94.52%, 95.38% and 95.0%, indicating that the models have an important role in the validation of the corpus, suggesting that the corpus has an important disciplinary utility. However, BERT-BILSTM-CRF performs the worst owing to the fact that the pre-trained model is used directly for target detection though. However, it has a large number of parameters and a large training set requirement that limits the application scenario which makes it impossible to apply directly to the carbonate platform discipline. Therefore, through multiple comparison experiments, the modified entity extraction method effectively solves the problems of insufficient labeled data and blurred entity boundaries in the carbonate platform discipline. It provides a favourable value for the extraction of data set construction.

Table 6: Effect of naming entity recognition

Model	Recall	Precision	F1 score
HMM	73.22%	73.49%	73.30%
CRF	88.43%	88.37%	88.32%

(Continued)

Table 6 (continued)

Model	Recall	Precision	F1 score
BILSTM	94.52%	88.05%	90.27%
BERT-BILSTM-CRF	91.22%	90.00%	91.80%
LSTM-CRF	96.22%	95.11%	95.66%
BILSTM-CRF	**98.72%**	**98.47%**	**98.70%**

As shown in Fig. 5, the BILSTM-CRF model with the best results was chosen in this paper, which outperformed the other models in the training set due to the small number of label types in the database and the large amount of each valid label. However, in the test set, significant data fluctuations are visible, which are attributed to the failure of the corpus to make all label definitions completely correct, for example, defining time or substance phrases as invalid labels “O”.

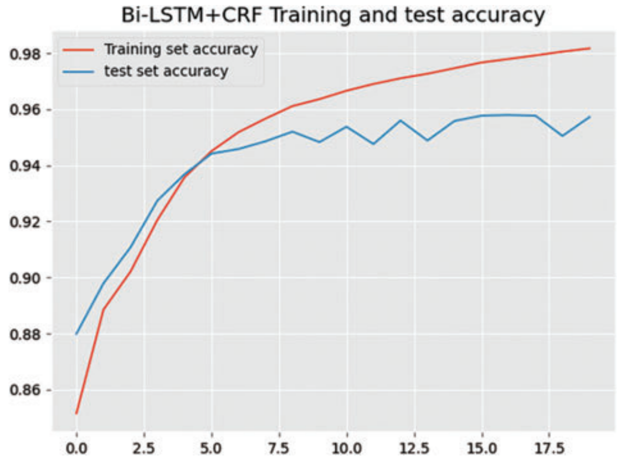

Figure 5: The accuracy of BILSTM-CRF in a carbonate platform corpus

5.4 Applications of the Corpus

Through named entity recognition experiments, we demonstrated that the corpus exhibited better performance. To further analyse the performance of the corpus, a representative set of examples was selected for further analysis, as shown in Table 7. In this experiment, enter the experimental statement, “Lithostratigraphic and sequence stratigraphic context of the Lista Formation and Mey Sandstone Member within the Paleocene of the Central North Sea”. The Blistm-crf model extracts substance, time and location entities from the statement, with the substance entity extracted as “Lithostratigraphic, sequence stratigraphic context, Lista Formation, Mey Sandstone Member”, the time entity extracted as “Paleocene” and the location entity extracted as “Central North Sea”, where the “Lista formation” and the “Mey sandstone member” do not appear in the corpus, but the model is still able to identify and extract them in the statement.

Table 7: Recognition effects based on corpus

Original sentence	Extracted results

Lithostratigraphic and sequence stratigraphic context of the Lista Formation and Mey Sandstone Member within the Paleocene of the Central North Sea.	Nat:{Lithostratigraphic, sequence stratigraphic context, Lista Formation, Mey Sandstone Member}
	Tim:{Paleocene}
	Geo:{Central North Sea }

Red-bed member are similar to those in the roughly contemporaneous Nima Basin of central Tibet, which is documented to have been at high paleoelevation by late Oligocene time.	Nat:{red-bed member}
	Tim:{ late Oligocene}
	Geo:{Nima Basin }

U-Pb isotope compositions of pyrite types in the Proterozoic Black Reef, Transvaal sequence.	Nat:{isotope composition, pyrite types}
	Tim:{ Proterozoic black reef}
	Geo:{Transvaal sequence }

Crystallographic and chemical variations during pyritization in the upper Barremian and lower Aptian dark claystone from the lower Saxonian basin.	Nat:{Crystallographic, chemical variations, dark claystone}
	Tim:{upper Barremian, lower Aptian}
	Geo:{Lower saxonian basin}

6 Conclusion

Corpus construction and detection remains a research challenge for specialized lexical entity recognition within domain disciplines. In this paper, a lexicon- and lexical-based model construction scheme is designed with a deep learning approach chosen to detect the accuracy of the prediction. Experiments show that the scheme needs further exploration in the face of more classification criteria. In the future, we will attempt to add more kinds of entities to detect the generalizability of the reformulation scheme.

Acknowledgement: Thanks to Dr. Hongyue Wu for providing the benchmark models of HMM and CRF, in addition to his valuable suggestions on the overall revision of the paper.

Funding Statement: This research is supported by the National Natural Science Foundation of China under Grant No. 42050102, the National Science Foundation of China (Grant No. 62001236), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant No. 20KJA520003).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Sun, L., Sun, Y., Ji, F., Wang, C. (2020). Joint learning of token context and span feature for span-based nested NER. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28,* 2720–2730. DOI 10.1109/TASLP.6570655.

2. Zhang, Y., Wang, Y., Yang, J. (2020). Lattice LSTM for Chinese sentence representation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28,* 1506–1519. DOI 10.1109/TASLP.6570655.

3. Tang, Z., Wan, B., Yang, L. (2020). Word-character graph convolution network for Chinese named entity recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28,* 1520–1532. DOI 10.1109/TASLP.6570655.

4. Yu, J., Jiang, J., Xia, R. (2019). Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE Transactions on Audio, Speech, and Language Processing, 28,* 429–439.

5. Xu, J., He, H., Sun, X., Ren, X., Li, S. (2018). Cross-domain and semisupervised named entity recognition in Chinese social media: A unified model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(11),* 2142–2152. DOI 10.1109/TASLP.2018.2856625.

6. Guven, Z. A., Unalir, M. O. (2021). Improving the BERT model with proposed named entity recognition method for question answering. *2021 6th International Conference on Computer Science and Engineering*, pp. 204–208. Nanjing.

7. Fei, H., Zhang, Y., Ren, Y., Ji, D. (2021). Optimizing attention for sequence modeling via reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems, 31(8),* 3612–3621. DOI 10.1109/TNNLS.2021.3053633.

8. Yilmaz, S. F., Kaynak, E. B., Koç, A., Dibeklioğlu, H., Kozat, S. S. (2021). Multi-label sentiment analysis on 100 languages with dynamic weighting for label imbalance. *IEEE Transactions on Neural Networks and Learning Systems*, 1–13. DOI 10.1109/TNNLS.2021.3094304.

9. Salah, R. E., Zakaria, L. Q. B. (2018). Building the classical arabic named entity recognition corpus. *2018 Fourth International Conference on Information Retrieval and Knowledge Management*, pp. 1–8. Kota Kinabalu.

10. Boroş, E., Romero, V., Maarand, M., Zenklová, K., Křečková, J. (2020). A comparison of sequential and combined approaches for named entity recognition in a corpus of handwritten medieval charters. *2020 17th International Conference on Frontiers in Handwriting Recognition*, pp. 79–84. Dortmund.

11. Kaur, A., Khattar, S. (2021). A systematic exposition of punjabi named entity recognition using different machine learning models. *2021 Third International Conference on Inventive Research in Computing Applications*, pp. 1625–1628. Coimbatore.

12. Drovo, M. D., Chowdhury, M., Uday, S. I., Das, A. K. (2019). Named entity recognition in bengali text using merged hidden markov model and rule base approach. *2019 7th International Conference on Smart Computing & Communications*, pp. 1–5. Wuhan.

13. Wei, Q., Zhou, Y., Zhao, B., Hu, X., Mei, Q. (2020). Named entity recognition from table headers in randomized controlled trial articles. *2020 IEEE International Conference on Healthcare Informatics*, pp. 1–2. Victoria.

14. Chetverikov, G., Puzik, O., Tyshchenko, O. (2018). Analysis of the problem of homonyms in the hyperchains construction for lexical units of natural language. *2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies*, pp. 356–359. Niagara Falls.

15. Takhom, A., Boonkwan, P., Hoppe, H. U., Ikeda, M., Usanavasin, S. (2018). A study of lexical ambiguity in large forum discussions for multidisciplinary knowledge engineering. *2018 Thirteenth International Conference on Knowledge, Information and Creativity Support Systems*, pp. 1–6. Pattaya.

16. Elbes, M., Aldajah, A., Sadaqa, O. (2019). P-Stemmer or NLTK stemmer for arabic text classification? *2019 Sixth International Conference on Social Networks Analysis, Management and Security*, pp. 516–520. Granada.

17. Contreras, J. O., Hilles, S., Abubakar, Z. B. (2018). Automated essay scoring with ontology based on text mining and NLTK tools. *2018 International Conference on Smart Computing and Electronic Enterprise*, pp. 1–6. Shah Alam.

18. Jha, N. K. (2018). An approach towards text to emoticon conversion and vice-versa using NLTK and WordNet. *2018 2nd International Conference on Data Science and Business Analytics*, pp. 161–166. Changsha.

19. Qu, Q., Kan, H., Wu, Y., Gao, Y. (2020). Named entity recognition of TCM text based on bert model. *2020 7th International Forum on Electrical Engineering and Automation*, pp. 652–655. Hefei.
20. Tripathi, S. P., Rai, H. (2018). SimNER-an accurate and faster algorithm for named entity recognition. *2018 Second International Conference on Advances in Computing, Control and Communication Technology*, pp. 115–119. Allahabad.
21. Du, Y., Zhao, W. (2020). Named entity recognition method with word position. *2020 International Workshop on Electronic Communication and Artificial Intelligence*, pp. 154–159. Shanghai.
22. Shah, S., Ramteke, J. (2021). Context aware named entity recognition with pooling. *2021 International Conference on Communication Information and Computing Technology*, pp. 1–4. Mumbai.
23. Ma, P., Jiang, B., Lu, Z., Li, N., Jiang, Z. (2020). Cybersecurity named entity recognition using bidirectional long short-term memory with conditional random fields. *Tsinghua Science and Technology, 26(3),* 259–265. DOI 10.1109/TST.5971803.
24. Nowshin, N., Ritu, Z. S., Ismail, S. (2018). A crowd-source based corpus on bangla to English translation. *2018 21st International Conference of Computer and Information Technology*, pp. 1–5. Dhaka.
25. Lertpiya, A., Chaiwachirasak, T., Maharattanamalai, N., Lapjaturapit, T., Chalothorn, T. et al. (2018). A preliminary study on fundamental Thai NLP tasks for user-generated web content. *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing*, pp. 1–8. Pattaya.