Journal of Ambient Intelligence and Humanized Computing (2024) 15:2105–2118
https://doi.org/10.1007/s12652-023-04740-4

ORIGINAL RESEARCH

Few-shot named entity recognition framework for forestry science metadata extraction

Yuquan Fan[1] · Hong Xiao[1] · Min Wang[2] · Junchi Wang[1] · Wenchao Jiang[1] · Chang Zhu[1]

Received: 6 June 2023 / Accepted: 8 December 2023 / Published online: 1 February 2024

Abstract
The effective utilization of accumulated forestry science papers is of paramount significance in enhancing our understanding of the current state of forests and the formulation of strategies for forest environmental preservation. However, the present challenge lies in the deficient richness of metadata associated with these pivotal documents, rendering their comprehensive exploitation a formidable endeavor. Metadata from forestry science papers serves as a foundational cornerstone for the efficient management and utilization of these scholarly documents, playing an indispensable role in the advancement of research within the domain of forestry science. Constructing a training corpus and extracting distant semantic relationships is challenging inherent, the utilization of named entity recognition (*NER*) technology for metadata entity identification in forestry science papers remains an unexplored avenue. To overcome these limitations, this paper creates a specialized training corpus and introduces a novel few-shot *NER* framework tailored specifically for metadata extraction from forestry science papers. Within this innovative framework, a data augmentation layer, employing word replacement (*WR*) and enhanced mixup (*EM*), effectively addresses the issue of suboptimal performance resulting from a scarcity of training data. The semantic comprehension layer incorporates a multi-granularity dilated convolution neural network (*MGDCNN*) to capture and extract distant semantic associations. Moreover, a meta-learning-based reweighting layer is introduced to mitigate the adverse effects of low-quality augmented examples on the model. Experimental results conclusively demonstrate the efficacy of the proposed framework, yielding *precision*, *recall*, and $F1$ of 91.08%, 88.96%, and 90.00%, respectively. Compared to traditional models, *precision*, *recall*, and $F1$ can be improved by up to 10.69%, 7.48%, and 9.07%, respectively.

Keywords Data augmentation · Reweighting · Forestry · Metadata · Named entity recognition (*NER*)

1 Introduction

As the dynamic evolution of forest ecosystems unfolds, scholarly attention to the domain of forestry science is increasingly accentuating its significance, resulting in a prolific emergence of scholarly papers. The effective utilization of this corpus of papers holds paramount importance in deepening our understanding of the contemporary forest landscape and formulating strategies for forest environmental conservation. Nonetheless, the current predicament pertains to the paucity of comprehensive metadata associated with these seminal works, posing considerable impediments to their systematic retrieval and application. Thus, the quest for an efficacious methodology for the extraction of metadata from forestry science papers assumes a position of profound significance.

1.1 The primary motivations

Forestry science data serves as a paramount repository, encapsulating the historical evolution and contemporary state of affairs pertaining to ecological construction and forestry preservation. Its pivotal role extends beyond mere documentation, as it underpins and propels the trajectory of innovation within the realm of forestry science and technology. Furthermore, it lends invaluable support to macro-level decision-making processes and drives the continuous evolution of the forestry industry (Wang et al. 2018; Jing 2022; Rubí et al. 2022). Amidst this landscape, forestry

✉ Wenchao Jiang
jiangwenchao@gdut.edu.cn

[1] School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China

[2] Chinamobile Park Construction and Development Company, Beijing 102206, China

science papers emerge as a prominent component of this wealth of data. Characterized by their substantial volume, decentralized storage, and wide-ranging disciplinary diversity, they pose a multifaceted challenge in terms of effective management and utilization. These inherent complexities, left unaddressed, obstruct the seamless dissemination of forestry knowledge and hinder the progressive development of forestry science (Ji et al. 2019).Thus, the efficient and judicious management and utilization of forestry science papers represent an exigent and compelling endeavor that warrants immediate attention.

In recent years, the landscape of named entity recognition (*NER*) technology (Kang et al. 2022; Lee et al. 2023) has undergone remarkable and accelerated transformation. This technology empowers the discernment of terminologies imbued with distinctive semantic significance embedded within textual data. Harnessing the prowess of *NER* technology to systematically extract metadata from the abstracts of forestry science papers not only streamlines their organization but also augments their reusability. This, in turn, fosters the broader advancement of forestry science, propelling it towards new frontiers of knowledge and innovation. Currently, various forestry science journals generate thousands or even tens of thousands of exemplary journal papers each year. However, these journal papers are often presented in a simple list, making it challenging to swiftly locate papers of genuine interest. Hence, we aspire to identify an expedient approach capable of autonomously extracting pertinent metadata from these journal papers. This metadata can significantly facilitate the rapid identification of target papers, thereby reducing the time expended on retrieval and expediting our knowledge acquisition process.

1.2 Innovation

Named entity recognition (*NER*) technology, originally introduced during the MUC-6 (Sundheim 1995), has evolved significantly and found extensive applications across diverse domains. In the contemporary landscape, deep learning-based *NER* methodologies have gained prominence. These methods frame *NER* as a sequence labeling task and harness the power of deep neural networks to automatically extract intricate features from input data, demonstrating robust generalization capabilities. Notably, deep learning-based *NER* methodologies have been widely adopted in domains such as e-commerce and healthcare (Qian et al. 2021; Ramachandran and Arutchelvan 2021; Ke et al. 2023). However, their utilization within the forestry domain remains at a nascent stage, with a predominant focus on specific subtopics, including forestry pests, diseases, or policies and regulations. To date, no endeavors have been undertaken to leverage *NER* technology for the extraction of metadata from forestry science papers. Forestry science papers present distinctive characteristics when compared to texts from other domains: (1) They encompass an abundance of proprietary nouns and domain-specific technical terminology, rendering conventional text annotation and part-of-speech tagging methodologies, prevalent in other domains, less directly applicable to the task of forestry metadata *NER*. (2) The domain of forestry science papers spans a wide spectrum of disciplines, potentially incurring substantial labeling costs. Existing *NER* technologies employed within the forestry domain primarily rely on extensively annotated corpora, which may exhibit high textual similarity, thereby diminishing reusability. In response to these challenges, this paper has undertaken the construction of a small-scale training corpus and introduced an innovative framework tailored for the extraction of metadata from forestry science papers.

1.3 Contributions

The research trajectory of this paper is illustrated in Fig. 1. The core components of the research center around our proposed framework, encompassing two enhanced data augmentation algorithms, a distinctive feature encoding structure, and a reweighting algorithm transferred from the computer vision domain.

The contributions of this paper are summarized as follows:

1. We propose an innovative few-shot *NER* framework expressly designed for metadata extraction from forestry science papers. This framework seamlessly integrates the formidable *ERNIE 3.0*, a large-scale pre-trained Chinese language model, with the multi-granularity dilated convolution neural network (*MGDCNN*) to realize comprehensive comprehension of lengthy forestry science texts. We enhance two data augmentation methodologies and harness them to alleviate the dependency on extensive corpora. Furthermore, the framework incorporates a reweighting mechanism to further enhance model performance and robustness.
2. We meticulously curate the abstract texts of forestry science papers and subject them to a rigorous process of data refinement, which involve expert domain-specific data cleansing and manual annotation. Subsequently, we assemble a dedicated training corpus tailored to meet the requirements of forestry metadata *NER*.
3. Extensive experiments have been meticulously executed, featuring a comparative evaluation of our framework against five deep learning-based methodologies. Additionally, we have conducted five sets of ablation experiments, conclusively demonstrating that our framework attains the pinnacle of performance in the context of forestry metadata *NER*.

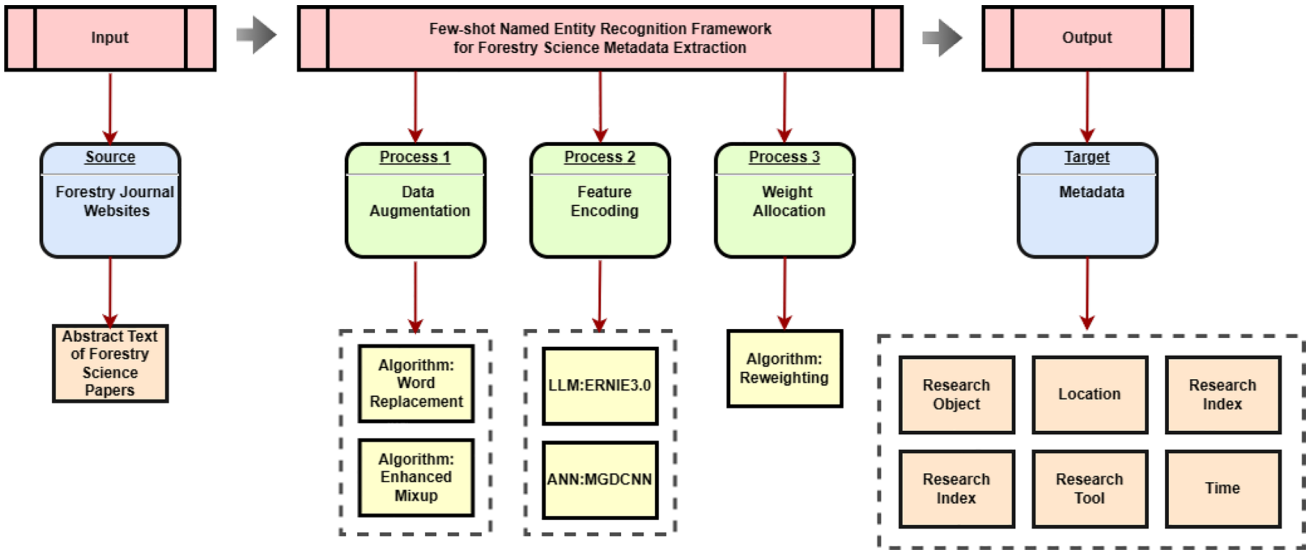

Fig. 1 Research framework. This figure illustrates the research trajectory throughout the entire document

Table 1 Abbreviations

Abbreviation	Description
NER	Named entity recognition
WR	Word replacement
EM	Enhanced mixup
MGDCNN	Multi-granularity dilated convolution neural network
CRF	Conditional random field
CNN	Convolutional neural network
FME	Forestry metadata extraction
BC	BiLSTM-CRF
BBC	BERT-BiLSTM-CRF
E1BC	ERNIE1.0-BiLSTM-CRF
E3BC	ERNIE3.0-BiLSTM-CRF
E3MC	ERNIE3.0-MGDCNN-CRF
E3WMC	ERNIE3.0-WR-MGDCNN-CRF
E3EMC	ERNIE3.0-EM-MGDCNN-CRF
E3WEMC	ERNIE3.0-WR-EM-MGDCNN-CRF

1.4 Sections

The remaining sections of the paper are organized as follows. Section 1 reviews some related work about metadata extraction from forestry science papers. Section 2 describes the overall architecture of our framework and presents the details. Section 3 presents the details of the corpus we constructed. And the performance evaluation of our framework is presented in Sect. 4 through comparative experiments and ablation experiments. Section 5 concludes this paper.

The abbreviations that may appear in the article along with their descriptions are presented in Table 1.

2 Related work

Early *NER* methods primarily consisted of rule-based approaches and feature-based approaches. Rule-based methods often achieve high accuracy when the formulated rules accurately reflect the text's characteristics. However, their efficacy heavily relies on linguistic expertise, and limited rules may struggle to comprehensively identify diverse entities. Subsequently, with the widespread adoption of statistical machine learning algorithms in natural language processing, feature-based methods outperformed rule-based methods. These approaches typically leverage a large number of manually defined features and utilize hidden Markov models (Patil et al. 2017) or conditional random fields (*CRF*) (Zhu et al. 2018) to train *NER* models on extensively annotated corpora. Feature-based methods employ statistical machine learning algorithms to learn knowledge from a vast amount of annotated data, eliminating the need for manually defined rules. However, these methods have shortcomings, including the necessity to manually define a feature set that reflects entity characteristics, with the method's performance depending on the discriminative power of the selected features. Additionally, there is a strong dependency on annotated corpora, requiring training on a large amount of manually labeled data, which constitutes a time-consuming and labor-intensive task.

In recent years, methods based on deep learning have been extensively applied in the field of natural language processing and have demonstrated notable success across various tasks. Compared to early statistical machine learning methods, deep learning methods exhibit significant advantages in terms of automatic feature learning, leveraging deep semantic knowledge, and alleviating issues related to data

sparsity. Notably, Huang et al. (2015) introduced the utilization of bidirectional long-short term memory (*BiLSTM*) in conjunction with*CRF* for *NER* tasks, achieving cutting-edge performance across diverse datasets. Subsequently, in 2016, Ma and Hovy (2016) augmented *NER* performance across several datasets by incorporating a convolutional neural network (*CNN*) into the *BiLSTM-CRF* architecture. Furthermore, with the advent of pre-trained language models and their formidable semantic comprehension capabilities, an escalating number of scholarly investigations have embraced the employment of pre-trained language models for *NER* undertakings in recent years. For instance, Gong et al. (2021) employed *BERT* (Devlin et al. 2019) for *NER* tasks pertaining to military literature, adeptly mitigating overfitting challenges stemming from a limited sample size through the application of logic fusion and differential fusion methodologies. In a similar vein, Zhao et al. (2022) harnessed *ALBERT* (Lan et al. 2019) for agricultural *NER* tasks, innovatively amalgamating Chinese character stroke features to enhance the model's capacity for semantic comprehension.

In the forestry domain, some scholars had also applied *NER* technology to forestry-related texts and documents for knowledge acquisition. For instance, Wang and Xiyou (2022) utilized Transformer for forestry disease analysis. Du (2020) utilized *BiLSTM-CRF* for forestry legal regulations analysis and for constructing a forestry legal regulations knowledge graph. And *BiLSTM-CNN-CRF* was utilized by (Dongmei and Wen 2019) to extract the context features in plant attribute texts. Zhang et al. (2023) utilized *BERT* to address challenges in entity recognition in apple disease and pest text. Although there have been some studies on *NER* technology in the forestry domain, there are currently no scholars who have applied *NER* technology to forestry metadata *NER* task. Zhang et al. (2022) conducted named entity recognition of kiwifruit disease and pest text based on a dictionary and attention mechanism. Moreover, these models that are commonly applied in the forestry domain do not perform well on forestry metadata *NER* task, as they usually require large-scale training corpus and involve short text lengths with high text similarity. Thus, a reliable method should be developed for conducting forestry metadata *NER* task. To address the issue of limited training corpus in some *NER* tasks across domains, researchers have proposed various methods of data augmentation. For example, Dai and Adel (2020) proposed a method known as token substitution that utilizes entity dictionaries to replace entity words in the training corpus. They obtained an expanded training corpus through this method. Guo et al. (2019) tried to generate pseudo examples at the feature representation level through the use of a mixup technology (Zhang et al. 2017). However, these methods generate pseudo examples in a relatively single manner, making it difficult to improve the generalization ability of deep learning-based models. In addition, data augmentation methods are bound to introduce some low-quality noisy examples, which may adversely affect the performance of these models. In the domain of computer vision, Ren et al. (2018) proposed a meta-learning-based reweighting mechanism for learning how to assign different weights to images of varying quality in the training set for image classification. It is valuable to explore the combination of this reweighting mechanism with data augmentation methods to further enhance the effectiveness of data augmentation.

3 Method

In this section, we initially provide a comprehensive exposition of our framework, delineating its overarching structure. Subsequently, we provide an intricate elucidation of the implementation procedures and the multifaceted functionalities inherent within each constituent module comprising our framework.

The variables that may appear in the article along with their descriptions are presented in Table 2.

3.1 Framework

The few-shot *NER* framework for metadata extraction from forestry science papers is illustrated in Fig. 1. It is mainly divided into six parts. From left to right, they are: data augmentation layer based on word replacement (*WR*), encoding layer, data augmentation layer based on enhanced mixup (*EM*), semantic understanding layer, sequence labeling layer and reweighting layer. The input of the framework is the annotated forestry science papers' abstract texts. Firstly, these data examples are input to data augmentation layer

Table 2 Variables

Variable	Description
T	Forestry science paper's abstract text
L	Label sequence
A	Transition score matrix
l′	True label
Lx	All possible label sequence
F	Original forestry training dataset
N	The number of examples in forestry training dataset
G	GloVe based on Wikidata
D	Metadata entity dictionary
P	Pseudo forestry training dataset
$\widetilde{F}$	Forestry training dataset after preliminary data augmentation
$\hat{F}$	Forestry training dataset after secondary data augmentation

based on *WR*, which utilizes an entity dictionary and *GloVe* (Pennington et al. 2014) for preliminary data augmentation. Then, in the encoding layer, the preliminarily augmented data is further processed by *ERNIE3.0* (Sun et al. 2021) for feature representation. And the feature representation sequences are passed to data augmentation layer based on *EM* for secondary data augmentation. Then, the data after two rounds of augmentation is passed to the semantic understanding layer, which uses *MGDCNN* to obtain semantics at different distances. Finally, the optimal label sequences are obtained through the sequence labeling layer. And the reweighting layer is used to assign different weights to different examples for better adjustment of model parameters.

3.2 Encoding layer based on ERNIE3.0

We have employed *ERNIE3.0* for the encoding of abstract texts from forestry science papers. *ERNIE3.0*, introduced by Baidu's Natural Language Processing division in 2021, constitutes a pre-trained language model. It leverages a multi-layer Transformer-XL (Dai et al. 2019) as its foundational network structure and seamlessly integrates autoregressive and autoencoder networks. Additionally, it extends the capabilities of *ERNIE2.0* (Sun et al. 2020) by introducing both a general-purpose representation module and a task-specific representation module. Furthermore, it encompasses pre-training tasks operating at various levels, including word-level, sentence-structure level, and knowledge level, thereby endowing *ERNIE3.0* with enhanced proficiency in encoding lengthy textual content.

3.3 Semantic understanding layer based on MGDCNN

In order to enhance the amalgamation of contextual information, we have devised a semantic understanding layer. This layer employs an *MGDCNN* (multi-granularity dilated convolutional neural network) to derive encoded sequences corresponding to the abstract texts of forestry science papers that have been enriched with contextual information. These feature representation sequences of abstract texts from the domain of forestry science are subsequently fed into three distinct one-dimensional dilated convolutional neural networks, each characterized by different dilation factors. As a result, this architecture effectively captures contextual semantic information at varying spatial extents. The architectural representation of the *MGDCNN* is illustrated in Fig. 2.

3.4 Sequence labeling layer based on CRF

We choose a *CRF* as the sequence labeling layer in our framework. *CRF* can fully consider contextual relevance, ensuring the accuracy of sequence labeling. Formally, given a forestry science paper's abstract text T=$\{t_1, t_2, \ldots, t_n\}$and a label sequence L=$\{l_1, l_2, \ldots, l_n\}$, we can obtain the corresponding *CRF* evaluation score:

$$Score(T, L) = \sum_{i=0}^{n} A_{l_i, l_{i+1}} + \sum_{i=1}^{n} P_{i, l_i} \tag{1}$$

where A is the transition score matrix. $A_{l_i, l_{i+1}}$ represents the transition score from label i to label $i+1$. P_{i, l_i} represents the probability value of labeling the *i-th* Chinese character as l_i. Given the input T, the formula for predicting the probability of label l is:

$$P(l|T) = \frac{e^{Score(T,l)}}{\sum_{l' \in L_x} e^{Score(T,l')}} \tag{2}$$

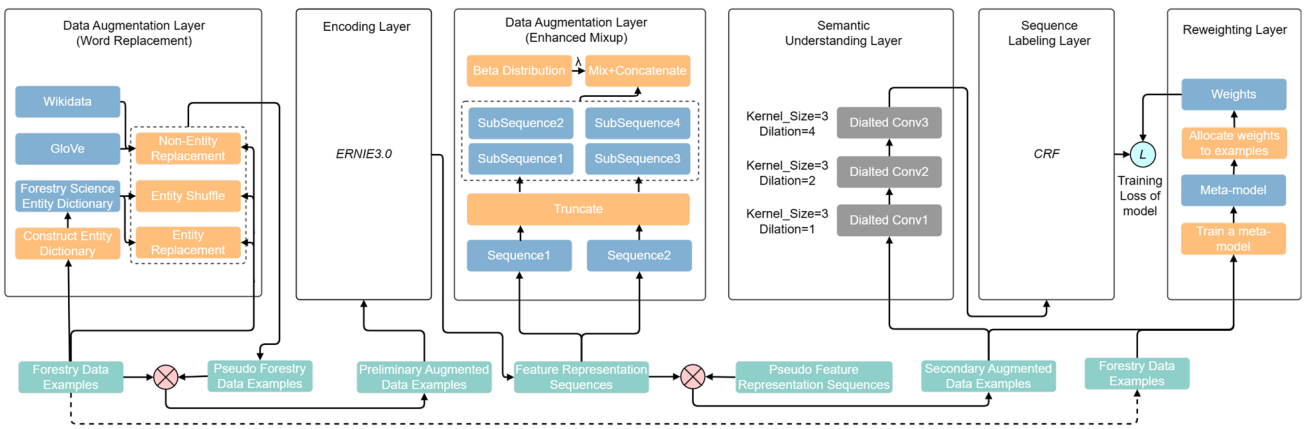

Fig. 2 A few-shot *NER* framework for metadata extraction from forestry science papers

where l' represents true label and L_x represents all possible label sequences. The objective of the model is to maximize $P(l|T)$, and the optimal sequence l^* is obtained by solving the maximum likelihood function.

$$l^* = \arg\max_{l \in L_x} (T, l') \tag{3}$$

3.5 Data augmentation layer based on WR and EM

Given the multidisciplinary nature and requisite specialization inherent in forestry science papers, the acquisition of extensive, well-annotated training datasets of high quality poses a formidable challenge. To address this challenge, we introduce a pair of data augmentation techniques, namely token substitution and mixup, and subsequently enhance their effectiveness to mitigate the likelihood of generating suboptimal pseudo examples. Consequently, we employ these refined data augmentation methodologies, denoted as "Word Replacement" and "Enhanced Mixup," in the context of the Named entity recognition (*NER*) task for forestry metadata. This strategic implementation serves to alleviate the imperative for copious manually-annotated instances.

Algorithm 1 The procedure of *WR*

Input: Original forestry training dataset F, the number of examples in forestry training dataset N, $GloVe$ based on Wikidata G
Output: Metadata entity dictionary D, pseudo forestry training dataset P, forestry training dataset after preliminary data augmentation $\widetilde{F}$
Initialize entity dictionary D
Initialize pseudo training dataset P
for $n = 1$ **to** N **do**
  $e \leftarrow$ ExtractEntities(F)
  add e into D
end for
for $n = 1$ **to** N **do**
  sample training example f
  $p_1 \leftarrow$ ReplaceEntities(f, D)
  $p_2 \leftarrow$ ShuffleEntities(f, D)
  $p_3 \leftarrow$ ReplaceNonEntities(f, D, G)
  add p_1, p_2, p_3 into P
end for
$\widetilde{F} \leftarrow$ add P into F

Token substitution involves the utilization of entity dictionaries to replace entities within the original corpus, thereby extending the training dataset. In pursuit of heightened diversity among generated pseudo examples and the enhancement of the overall robustness and generalization capabilities of our framework, we have extended the concept of token substitution. In this regard, we introduce a novel data augmentation technique denoted as "Word Replacement" (*WR*). The intricate procedure for implementing *WR* is elucidated in Algorithm 1. Algorithm 1 has a time complexity of O(n^2) and a space complexity of O(n).

The *WR* methodology unfolds in three distinct phases: entity replacement, entity shuffle, and non-entity replacement. Entity replacement entails the meticulous construction and utilization of a specialized entity dictionary meticulously curated from forestry-related examples. This dictionary serves as a repository of entities categorized by type, allowing for the random replacement of original entities with entities of the same type. This infusion of diversity adds a layer of complexity to the training data, further enhancing the framework's adaptability.Additionally, the entity shuffle component operates by systematically randomizing the order of words belonging to entities of identical types within the training corpus. This shuffling process introduces variability while preserving semantic integrity, thereby contributing to the framework's resilience.Moreover, the non-entity replacement facet of *WR* leverages the capabilities of the *GloVe* model to identify lexically analogous words from the extensive corpus of Wikipedia data. These semantically congruent words are subsequently used to replace non-entity terms in the corpus. This step not only enhances the vocabulary diversity but also aligns the corpus with a broader contextual understanding.In essence, the extension of token substitution through the introduction of the *WR* methodology embodies our commitment to comprehensively bolstering the quality and adaptability of our framework. Through the strategic manipulation of training data, we endeavor to harness the

full potential of machine learning to address the unique challenges posed by forestry-related research and further elevate the precision and efficacy of our models.

In contrast to the word replacement (*WR*) technique applied directly to the original forestry examples, the mixup methodology generates pseudo-training instances at the level of feature representation. Conventional mixup techniques typically involve linear interpolation applied to the entire feature sequence, resulting in the generation of a single pseudo feature sequence for each pair of feature sequences. To overcome this limitation and increase the volume of pseudo-training instances, we introduce an innovative mixup method termed "Enhanced Mixup" (*EM*). *EM* achieves this by amalgamating and concatenating subsequences originating from a pair of feature sequences. The procedural details of *EM* are delineated in Algorithm 2. Algorithm 2 has a time complexity of O(n^2) and a space complexity of O(n).

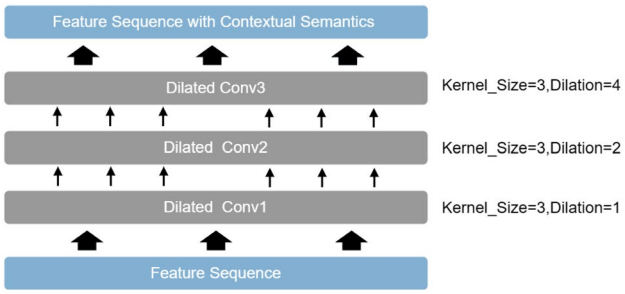

Fig. 3 Multi-granularity dilated convolution neural network

Algorithm 2 The procedure of *EM*

Input: Forestry training dataset after preliminary data augmentation $\widetilde{F}$, Baidu pre-trained language model *ERNIE3.0*, batch size m, training epoch E, augmentation rate *times*
Output: Forestry training dataset after secondary data augmentation $\hat{F}$
Initialize pseudo forestry training example sequences P_s
for $e = 1$ **to** E **do**
  $batchData \leftarrow$ SamepleBatch(F, m)
  times $\leftarrow$ len($batchData$) / 2
  for $time = 1$ **to** $times$ **do**
    $(T_1, L_1), (T_2, L_2) \leftarrow$ SampleExamplePair($batchData$)
    s_1=(e_{11},e_{12},...,e_{1n_1}),s_2=(e_{21}, e_{22}, ..., e_{2n_2})$\leftarrow$Embedding(*ERNIE3.0*, T_1, T_2)
    s_{11}=(e_{11},e_{12},...,e_{1m_1}) $\leftarrow$ (e_{11},e_{12},...,e_{1n_1}), ($m_1 = n_1/2$)
    s_{12}=($e_{1(m_1+1)}$,$e_{1(m_1+2)}$,...,e_{1n_1}) $\leftarrow$ (e_{11},e_{12},...,e_{1n_1})
    s_{21}=(e_{21},e_{22},...,e_{2m_2}) $\leftarrow$ (e_{21},e_{22},...,e_{2n_2}), ($m_2 = n_2/2$)
    s_{22}=($e_{2(m_2+1)}$,$e_{2(m_2+2)}$,...,e_{2n_2}) $\leftarrow$ (e_{21},e_{22},...,e_{2n_2})
    $\lambda \leftarrow Beta(\alpha, \alpha)$, ($\lambda \in [0, 1]$, $\alpha > 0$)
    s_{31}, $s_{32} \leftarrow \lambda s_{11} + (1-\lambda)s_{21}$, $\lambda s_{12} + (1-\lambda)s_{22}$
    $\widetilde{s_1} \leftarrow$ concatenate(s_{31}, s_{12})
    $\widetilde{s_2} \leftarrow$ concatenate(s_{31}, s_{22})
    $\widetilde{s_3} \leftarrow$ concatenate(s_{11}, s_{32})
    $\widetilde{s_4} \leftarrow$ concatenate(s_{12}, s_{32})
    add $\widetilde{s_1}$, $\widetilde{s_2}$, $\widetilde{s_3}$, $\widetilde{s_4}$ into P_s
  end for
end for
$\hat{F} \leftarrow$ add P_s into Embedding(*ERNIE3.0*, $\widetilde{F}$)

Given an example pair (T_1, L_1) and (T_2, L_2) (T represents the text sequence and L represents the entity label sequence) randomly sampled from forestry examples. And we obtain their vector representations s_1 and s_2 through *ERNIE*3.0. Then, both s_1 and s_2 are split into two subsequences, resulting in four subsequences: s_{11}, s_{12}, s_{21} and s_{22}. We mix s_{11} and s_{21}, s_{12} and s_{22} to obtain pseudo subsequences s_{31} and s_{32}. After that, we concatenate s_{31} with s_{12} or s_{22} and concatenate s_{32} with s_{11} or s_{21}. Finally, we obtain four pseudo feature sequences $\widetilde{s_1}$, $\widetilde{s_2}$, $\widetilde{s_3}$ and $\widetilde{s_4}$. And the *CRF* evaluation scores of these pseudo feature sequences also undergo similar truncation and concatenation operations.

3.6 Reweighting layer based on meta-learning

Although data augmentation techniques undeniably mitigate the need for manual annotation, they unavoidably introduce a subset of lower-quality pseudo examples. The presence of such suboptimal pseudo examples carries the potential to detrimentally affect the performance of the model. In response to this challenge, we have conceived a reweighting layer, which employs a meta-learning-driven reweighting mechanism to dynamically allocate weights to the augmented forestry examples. The intricate process

of this reweighting layer is visually depicted in Fig. 3. The process initiates with the application of the word replacement (*WR*) and enhanced mixup (*EM*) algorithms to augment the original forestry data examples. Subsequently, we initialize a meta-model tailored for the augmented data. This meta-model engages in iterative learning, continually adapting to the augmented data by leveraging small batches of the original forestry data examples. In the end, we obtain the weights of the augmented data examples.

Algorithm 3 elucidates the intricacies of the procedure for reweighting forestry examples. Algorithm 3 has a time complexity of $O(n^2)$ and a space complexity of O(n). The determination of weights for various examples is achieved through the training of a meta-model, which acquires the capability to judiciously allocate weights to different examples. Concretely, we treat the original forestry training examples as the meta-dataset, using them as guiding references for the refinement of model parameters. Subsequently, we proceed to reallocate weights to the augmented forestry training examples, guided by the loss generated through the batch meta-data. In instances where the data distribution and gradient-descent trajectory of the augmented forestry training examples closely mirror those of the original forestry training examples, our model excels in fitting these examples, endowing them with elevated weights. This strategic approach amplifies the adaptability and precision of our framework, enabling a nuanced distinction between high-quality and suboptimal examples, thereby augmenting its overall efficacy and performance.

Algorithm 3 The procedure of reweighting

Input: Original forestry training dataset F, Forestry training dataset after secondary data augmentation $\hat{F}$, initial model parameters θ^0, initial learnable weights of examples w_L, batch size m, training epochs E
Output: Weights of training examples w
for $e = 1$ **to** E **do**
  $b_1 \leftarrow \text{SampleBatch}(\hat{F}, m)$
  $b_2 \leftarrow \text{SampleBatch}(F, m)$
  $\mathcal{L}_1 \leftarrow \text{CalculateLoss}(b_1, \theta^e)$
  $\nabla\theta^e \leftarrow \text{BackwardAD}(\mathcal{L}_1, \theta^e)$
  $\hat{\theta}^e \leftarrow \theta^e - \beta\nabla\theta^e$
  $\mathcal{L}_2 \leftarrow \text{CalculateLoss}(b_2, \hat{\theta}^e)$
  $\nabla w_L \leftarrow \text{BackwardAD}(\mathcal{L}_2, w_L)$
  $w' \leftarrow \max(-\nabla w_L, 0)$
  $w \leftarrow w' / (\sum_j w' + \delta(\sum_j w'))$
  $\hat{\mathcal{L}}_1 \leftarrow \text{CalculLateLoss}(b_1, \theta^e, w)$
  $\nabla\theta^e \leftarrow \text{BackwardAD}(\hat{\mathcal{L}}_1, \theta^e)$
  $\theta^{e+1} \leftarrow \text{OptimizerStep}(\theta^e, \nabla\theta^e)$
end for

3.7 Formulation

For the data augmentation layer, our emphasis is particularly on the use of *WR* and *EM*. *WR* is employed for data augmentation at the word-level. It is characterized by the following core formulas:

$$p = \begin{cases} ReplaceEntities(f, D) \\ ShuffleEntities(f, D) \\ ReplaceNonEntities(f, D, G) \end{cases} \tag{4}$$

where f represents the data sampled from the original dataset. D represents the entity dictionary constructed from the original dataset. G represents the *GloVe* based on Wikipedia data. And p represents the generated pseudo-data examples.

Diverging from *WR*, *EM* performs data augmentation at the feature-level. Firstly, we obtain example pairs through random sampling. And then pseudo feature sequences are generated through truncation, mixup and concatenation operations. The core formulas of *EM* are as follows:

$$\overline{T} = \begin{cases} ERNIE3.0(T) = (e_1, e_2, \ldots, e_n) \\ \overline{e_i} = \lambda e_{1i} + (1 - \lambda) e_{2i}, i \in [1, n] \end{cases} \tag{5}$$

where n is obtained by taking the maximum length of the two parts of feature sequences used in mixup, and λ is obtained through a $Beta(\alpha, \alpha)$ with $\alpha > 0$, and $\lambda \in (0, 1)$.

For the reweighting layer, we migrate commonly used reweighting algorithm from the computer vision domain to the natural language processing domain and incorporate them into our research. Assuming that we currently possess a forestry training dataset $\hat{F}=(T_i, L_i)_{i=1}^N$ that has undergone secondary data augmentation (where N represents the number of

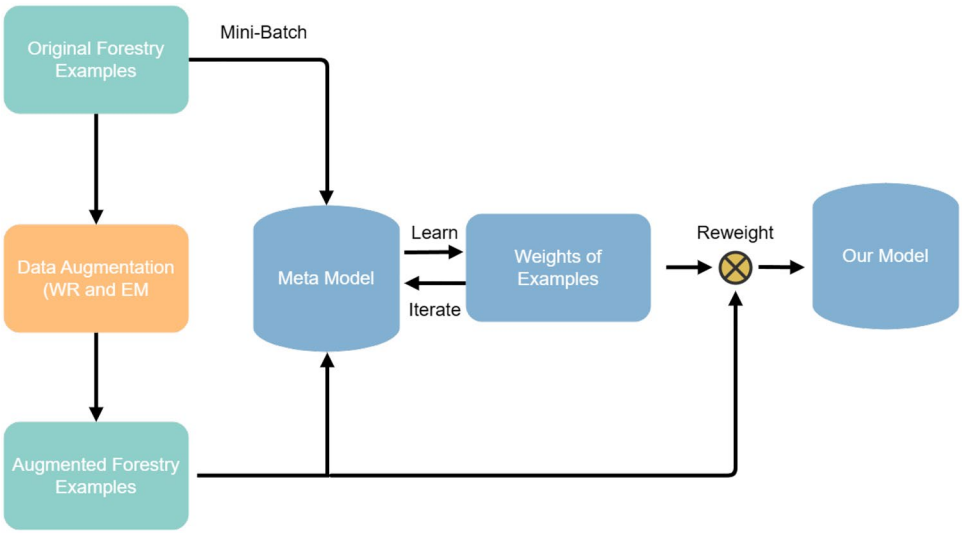

Fig. 4 The process of reweighting layer based on meta-learning

Table 3 Detailed information of *FME*

Corpus	Type	Number
FME	*Char*	313 k
	Word	134 k
	Sentence	17 k
	Example	1546

samples in this dataset), the optimization objective we seek can be formulated as the following equation:

$$\theta^*(w) = \arg\min_{\theta} \sum_{i=1}^{N} w_i \mathcal{L}(f(T_i;\theta), L_i) \tag{6}$$

where w_i represents the learnable weight for the loss of the ith training example in $\hat{F}$. $\mathcal{L}$ represents loss function and f represents the forward process. The optimal parameter w is obtained by calculating the following equation on the meta-dataset $\mathrm{M}=(T_i, L_i)_{i=1}^{n}$:

$$w^* = \arg\min_{w} \frac{1}{n} \mathcal{L}(f(T_i;\theta^*(w)), L_i) \tag{7}$$

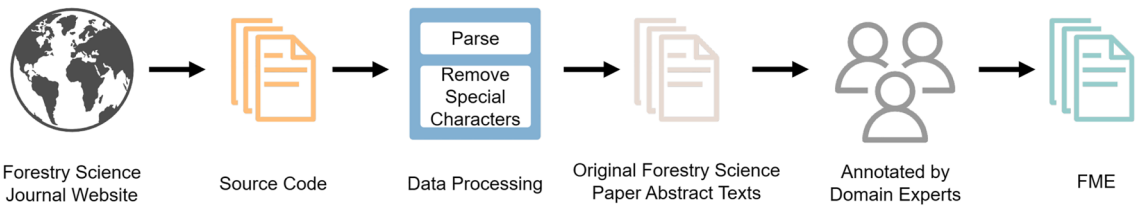

Fig. 5 The process of corpus acquisition

Table 4 Labels and description examples of some entities

Category	Type	Description	Example
Research object	Object	The main object of forestry science research	Larch
Location	Loc	Related locations	Ordos
Research index	Index	Indexes for measuring forestry science research	Vegetation biomass
Research tool	Tool	Tools used for forestry science research	Electron spin resonance spectrometer
Research method	Method	Methods used for conducting forestry science research	Marker-assisted selection
Time	Time	Related time	1999

4 Data collection and pre-processing

4.1 Data collection

Unlike other domains, currently, in the forestry domain, there is a lack of corpus and annotation methods that can be used for metadata extraction. Therefore, without affecting the normal operation of the forestry science journal website, we obtain 1546 forestry science paper abstract texts by web crawling and parsing the webpage source code. These texts are then analyzed and annotated by domain experts with rich forestry science knowledge to ensure the quality of the annotated corpus. Then we get a small-scale forestry training corpus named *FME* (forestry metadata extraction). The specific process of obtaining *FME* is shown in Fig. 4, and the detailed information is presented in Table 3. Subsequent access to *FME*'s data files can be obtained by visiting the following link: https://github.com/FFFyyq/ner4fme.

4.2 Data annotation

We use the web-based text annotation tool Doccano to annotate the forestry science papers' abstract texts and convert the annotated texts into the standard annotation format BIOES. Based on the characteristics of the forestry science papers' abstract texts and expert opinions, we select and annotated 6 categories, namely: research object, location, research index, research tool and research method. These categories can fully describe the metadata information of forestry science papers. Table 4 shows the labels and description examples of some entities, and Fig. 5 shows the quantity and distribution of each entity.

5 Experiments

5.1 Experiment settings

To ensure the approximate distribution similarity between the training data and test data, *FME* are divided into a training set, a validation set and a test set in a 6:2:2 ratio. Then, we use Pytorch to construct a metadata entity recognition model for forestry science papers and conduct training and testing to obtain the best prediction results. Algorithm 4 shows the procedure of our experiment. Algorithm 4 has a time complexity of $O(n^2)$ and a space complexity of O(n).

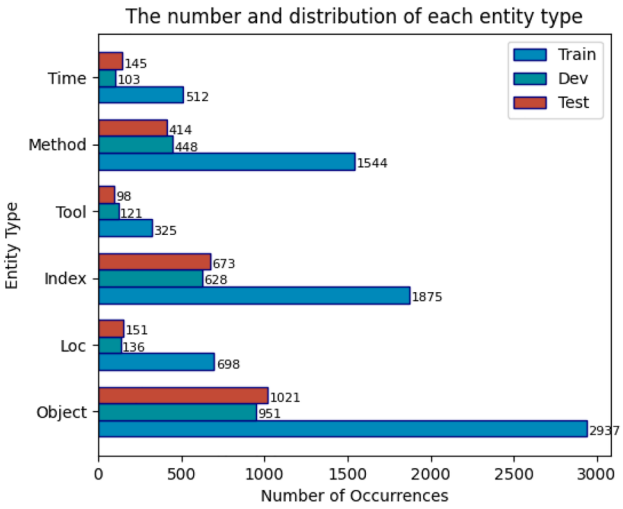

Fig. 6 The number and distribution of each entity

Table 5 Adjustable parameters

Parameter	Value
The datasets' split ratio	6:2:2
Batch size	16
Training epoch	10
Learning rate of ERNIE3.0	2e−5
Learning rate of other modules	1e−3
Weight decay	1e−4
Dropout	0.5
Optimizer	AdamW

Table 6 Experimental environment

Experimental environment	Environment configuration
Operating system	Ubuntu 22.04.2 LTS
CPU	13th Gen Intel(R) Core(TM) i7-13700
GPU	NVIDIA GeForce RTX 4090
Memory	128 G
Deep learning framework	Pytorch
CUDA	11.5

Algorithm 4 The procedure of the experiment

Input: Training corpus FME, split ratio r, initial model parameters θ^0, batch size m, training epochs E
Output: Update model parameters θ^E
$s_{train}, s_{val}, s_{test} \leftarrow \text{split}(FME, r)$
$s_{train'} \leftarrow \text{WR}(s_{train})$
for $e = 1$ **to** E **do**
  $P_s \leftarrow \text{EM}(s_{train'}, m)$
  $S_{train'} \leftarrow \text{Embedding}(s_{train'})$
  $S \leftarrow P_s \cup \text{S}_{train'}$
  $w \leftarrow \text{Reweighting}(s_{train}, s_{train'}, S)$
  $\mathcal{L} \leftarrow \text{CalculateLoss}(s_{train'}, S, w)$
  $\nabla\theta^t \leftarrow \text{BackwardAD}(\mathcal{L}, \theta^t)$
  $\theta^t \leftarrow \text{OptimizerStep}(\theta^t, \nabla\theta^t)$
end for

Our model employs the pre-trained language model *ERNIE3.0* and *MGDCNN* with different dilation factors as the main architecture. For the parameters of *ERNIE3.0*, the learning rate is set to 2e−5. For parameters of other modules, the learning rate is set to 1e−3 with a weight decay of 1e−4 and a dropout ratio of 0.5 are used. Table 5 displays the adjustable parameters.

Additionally, we use the AdamW optimizer to update the trainable parameters and set the batch size to 16 for training in the environment shown in Table 6.

We use *precision*, *recall*, and $F1$ as evaluation metrics. Let N denote the number of metadata entity categories. For the ith metadata entity category, *precision*, *recall* and $F1$ are defined as follows:

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \times 100\% \tag{8}$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \times 100\% \tag{9}$$

$$F1_i = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \times 100\% \tag{10}$$

where TP_i represents the number of correctly identified entities of the ith metadata entity category by our model. FP_i represents the number of incorrectly identified entities of the ith metadata entity category by our model. FN_i represents the number of incorrectly identified entities of the ith metadata entity category by our model. The overall *precision*, *recall* and $F1$ are defined as follows:

Table 7 Results of comparative experiments

Models	Precision (%)	Recall (%)	F1 (%)
BC	80.39	81.48	80.93
BBC	82.85	82.37	82.61
E1BC	85.85	83.38	84.60
E3BC	86.51	84.95	85.72
E3MC	87.53	85.97	86.74
Ours	91.08	88.96	90.00

Table 8 Results of the comparative experiments on different sampling sets

Models	30%			50%		
	Precision (%)	Recall(%)	F1 (%)	Precision (%)	Recall (%)	F1 (%)
E3MC	52.94	63.46	57.73	62.31	68.89	65.44
Ours	67.47 (+14.53)	71.79 (+8.33)	69.57 (+11.84)	75.80 (+13.49)	76.58 (+7.69)	76.17 (+10.73)
Models	70%			100%		
	Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)	F1 (%)
E3MC	80.60	81.21	80.89	87.53	85.97	86.74
Ours	86.01 (+5.41)	84.33 (+3.12)	85.16 (+4.27)	91.08 (+3.55)	88.96 (+2.99)	90.00 (+3.26)

Table 9 Results of ablation experiments

Models	Precision (%)	Recall (%)	F1 (%)
Baseline	86.58	85.38	85.98
E3MC	87.53	85.97	86.74
E3WMC	89.60	86.26	87.87
E3EMC	88.90	86.18	87.52
E3WEMC	90.15	87.90	89.00
Ours	91.08	88.96	90.00

$$Precision = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + \sum_{i=1}^{N} FP_i} \times 100\% \tag{11}$$

$$Recall = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + \sum_{i=1}^{N} FN_i} \times 100\% \tag{12}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \times 100\% \tag{13}$$

5.2 Experimental results

To validate our framework has better performance in metadata extraction task from forestry science papers, we compare our framework with 5 deep learning methods on *FME*. These methods include *BiLSTM-CRF(BC)*, *BERT-BiLSTM-CRF(BBC)*, *ERNIE1.0-BiLSTM-CRF(E1BC)*, *ERNIE3.0-BiLSTM-CRF(E3BC)* and *ERNIE3.0-MGDCNN-CRF(E3MC)*. Each experiment is repeated 5 times and the average results are presented in Table 7.

As is shown in Table 7, it can be seen that utilizing pre-trained language models such as *BERT* and *ERNIE* for forestry *NER* task can lead to varying degrees of performance improvement. For example, *BBC* has improved *precision*, *recall* and *F*1 by 2.46%, 0.89% and 1.68% respectively compared with *BC*. Furthermore, *E1BC* achieves better performance than *BC*, with improvements in *precision*, *recall* and *F*1 by 5.46%, 1.9% and 3.67%. This indicates that pre-trained language models are better able to capture the relationships between words and contexts and obtain feature representations of words. It can be observed that both *E1BC* and *E3BC* have improved results on *FME* compared with *BBC*. *E1BC* has improved *precision*, *recall* and *F*1 by 3%, 1.01% and 1.99% respectively compared with *BBC*. In addition, *E3BC* has improved *precision*, *recall* and *F*1 by 3.66%, 2.58% and 3.11% respectively compared with *BBC*. This indicates that *ERNIE* is more suitable for our task Fig. 6.

E3BC achieves higher *precision*, *recall* and *F*1 than *E1BC* on *FME* by 0.66%, 1.57% and 1.12%, respectively. This indicates that *ERNIE3.0* has stronger language understanding capabilities compared with *ERNIE1.0*. Furthermore, to better capture long-distance semantics, *BiLSTM* is replaced with *MGDCNN*. And it can be seen that *E3MC* achieves better performance than *E3BC*, with three metrics improving by 1.02%. Based on *E3MC*, we further add data augmentation layers and a reweighting layer to obtain a complete framework. And the complete framework achieved further improvements on *FME* of 3.55%, 2.99%, and 3.26% in *precision*, *recall* and *F*1, respectively. In summary, our framework combine the advantages of *ERNIE3.0* and *MGDCNN* and utilized data augmentation and reweighting mechanism to achieve state-of-the-art performance on *FME*.

To validate that our framework can perform forestry metadata *NER* task with few training examples, we conducted experiments on *FME* with different sampling ratios and on different sampling sets. We sampled *FME* at rates of 30%, 50%, and 70%, resulting in three different sampling sets. And we conducted a set of comparative experiments on each of these sampling sets to compare the performance of our framework with *ERNIE3.0-MGDCNN-CRF(E3MC)*, which removes the data augmentation layers and reweighting layer. Table 8 presents the experimental results.

As is shown in Table 8, our framework outperforms *E3MC* on different sizes of *FME*. This indicates that our framework can improve the performance on the forestry metadata *NER* task with few training examples. In addition, our framework performs better when there are fewer training examples. When the training examples are 30% of the original corpus, our model outperforms *E3MC* with an improvement of 14.53%, 8.33%, and 11.84% in *precision*, *recall*, and *F*1, respectively. As the sampling ratio increases, the performance improvement of our framework gradually decreases, but it still shows some improvement.

To further validate the effectiveness of our framework, we conduct ablation experiments on *FME* to demonstrate the effectiveness of each module in our framework. We adapt *ERNIE3.0-CRF* as the baseline model. We gradually add data augmentation layers, a semantic understanding layer and reweighting layer, while keeping the basic parameters consistent. We compare *Baseline*, *ERNIE3.0-MGDCNN-CRF(E3MC)*, *ERNIE3.0-WR-MGDCNN-CRF(E3WMC)*, *ERNIE3.0-EM-MGDCNN-CRF(E3EMC)*, *ERNIE3.0-WR-EM-MGDCNN-CRF(E3WEMC)* with our framework(*ERNIE3.0-WR-EM-MGDCNN-RW-CRF*) on *FME*. In these models, *WR* and *EM* represent two types of data augmentation layers in our framework, while *RW* represents the reweighting layer. Table 9 shows the results of the ablation experiments.

As is shown in Table 9, the addition of the modules in our framework based on *Baseline* leads to varying degrees of improvement in the model performance. Compared with Baseline, the addition of a semantic understanding layer based on *MGDCNN* lead to an improvement of 0.95%,

0.59% and 0.76% in *precision*, *recall* and *F*1, respectively. This suggests that utilizing *MGDCNN* for extracting contextual semantics at different distances has a certain promoting effect on understanding long forestry science abstract texts.

Based on *E3MC*, applying *WR* for data augmentation alone leads to improvements of 2.07%, 0.29%, and 1.13% in *precision*, *recall*, and *F*1, respectively. In addition, applying *EM* for data augmentation alone leads to an improvement of 1.37%, 0.21%, and 0.78% in *precision*, *recall*, and *F*1, respectively. Furthermore, applying both *WR* and *EM* for data augmentation simultaneously leads to an improvement of 2.62%, 1.93% and 2.26%, respectively. This suggests that these two data augmentation methods have a positive effect on solving the problem of insufficient forestry training corpus. And applying both methods simultaneously can achieve better results. Comparing *E3WEMC* and *Ours*, it can be observed that under the premise of applying data augmentation, introducing reweighting layer leads to an improvement of 0.93%, 1.06% and 1.00% in *precision*, *recall*, and *F*1, respectively. This indicates that the reweighting layer can guide our model to learn the weights between different examples and avoid the damage caused by low-quality examples to the model performance.

Based on the comprehensive analysis, each module in our framework can fully play its role. And our framework can be applied to metadata extraction from forestry science papers and achieve the best performance.

6 Conclusion and further discussions

In the pursuit of enhancing the management and reutilization of forestry science papers, this study delves into metadata named entity recognition (*NER*) within the context of forestry science papers. To circumvent the challenge posed by the scarcity of accessible corpora, we meticulously assemble and annotate a corpus comprising 1546 abstract texts, thereby establishing an experimental linguistic resource. Moreover, we introduce a novel few-shot *NER* framework tailored for the extraction of metadata from forestry science papers, thereby mitigating the performance limitations typically associated with constrained training corpora. Leveraging the training corpus we construct, our model obtain exceptional performance metrics, with *precision*, *recall*, and *F*1 scores reaching 91.08%, 88.96%, and 90.00%, respectively.

Knowledge graph technology, known for its effectiveness in elucidating intricate relationships among entities, has been found multifarious applications across diverse domains. In our forthcoming researching, we will envisage the creation of a comprehensive knowledge graph dedicated to the metadata of forestry science papers. This initiative will entail the fusion of advanced graph computing methodologies with sophisticated recommendation algorithms, synergistically geared toward addressing the exigencies associated with the effective management and reutilization of forestry science papers.

Acknowledgements This study was funded by Guangdong Basic and Applied Basic Research Fund Project (Grant/ Award Number 2020B1515120010), Key Technology Project of Foshan City (Grant/Award Number 1920001001367), Guangdong Science and Technology Plan Project (Grant/Award Number 2019B010139001), Guangdong Natural Science Fund Project (Grant/Award Number 2021A1515011243), and Guangzhou Science and Technology Plan Project (Grant/Award Number 201902020016).

References

Dai X, Adel H (2020) An analysis of simple data augmentation for named entity recognition. arXiv:2010.11683

Dai Z, Yang Z, Yang Y, Carbonell J, Le Quoc V, Salakhutdinov R (2019) Transformer-xl: attentive language models beyond a fixed-length context. arXiv:1901.02860

Devlin J, Chang MW, Lee K, Toutanova K (2019) Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT, vol 1, p 2

Dongmei LI, Wen TAN (2019) Research on named entity recognition method in plant attribute text. J Front Comput Sci Technol 13(12):2085

Du H (2020) Research and construction of a forestry law and regulation q &a system integrating knowledge graph. Beijing Forestry University

Gong Y, Mao L, Changliang L (2021) Few-shot learning for named entity recognition based on bert and two-level model fusion. Data Intell 3(4):568–577

Guo H, Mao Y, Zhang R (2019) Augmenting data with mixup for sentence classification: an empirical study. arXiv:1905.08941

Huang Z, Xu W, Yu K (2015) Bidirectional lstm-crf models for sequence tagging. arXiv:1508.01991

Ji P, Xiao Y, Hou R (2019) Exploration and practice of forestry science data management. J Agric Big Data 1(03):46–56

Jing S (2022) Thoughts and countermeasures on strengthening scientific data management in the era of big data. China Soft Sci 09:50–54

Kang Y, Sun L, Zhu R, Li M (2022) A review of deep learning chinese named entity recognition research. J Huazhong Univ Sci Technol (Natural Science Edition) 50(11)

Ke J, Wang W, Chen X, Gou J, Gao Y, Jin S (2023) Medical entity recognition and knowledge map relationship analysis of Chinese emrs based on improved bilstm-crf. Comput Electr Eng 108:108709

Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2019) Albert: a lite bert for self-supervised learning of language representations

Lee C-S, Wang M-H, Reformat M, Huang S-H (2023) Human intelligence-based metaverse for co-learning of students and smart machines. J Ambient Intell Humaniz Comput 14(6):7695–7718

Ma X, Hovy E (2016) End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. Association for Computational Linguistics, pp 1064–1074

Patil NV, Patil AS, Pawar BV (2017) Hmm based named entity recognition for inflectional language, pp 565–572

Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference

on empirical methods in natural language processing (EMNLP), pp 1532–1543
Qian H, Liu N, Wang J, Zhichao W, Zhang X, Liu Q, Zhao Y, Feng X (2021) An overlapping sequence tagging mechanism for symptoms and details extraction on Chinese medical records. Comput Electr Eng 91:107019
Ramachandran R, Arutchelvan K (2021) Named entity recognition on bio-medical literature documents using hybrid based approach. J Ambient Intell Humaniz Comput 1–10
Ren M, Zeng W, Yang B, Urtasun R (2018) Learning to reweight examples for robust deep learning. In: International conference on machine learning. PMLR, pp 4334–4343
Rubí JNS, de Carvalho PHP, Gondim PRL (2022) Forestry 4.0 and industry 4.0: use case on wildfire behavior predictions. Comput Electric Eng 102:108200
Ruidan Wang, Jing Yang, Menxu Gao, Wang C (2018) Reflections on strengthening and standardizing scientific data management in china. China Sci Technol Resour Guide 50(02):1–5
Sundheim BM (1995) Named entity task definition, version2.1. In: Proc. sixth message understanding conf. (MUC-6)
Sun Y, Wang S, Li Y, Feng S, Tian H, Hua W, Wang H (2020) Ernie 2.0: a continual pre-training framework for language understanding. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 8968–8975
Sun Y, Wang S, Feng S, Ding S, Pang C, Shang J, Liu J, Chen X, Zhao Y, Lu Y, Liu W, Wu Z, Gong W, Liang J, Shang Z, Sun P, Liu W, Ouyang X, Yu D, Tian H, Wu H, Wang H (2021) Ernie 3.0: large-scale knowledge enhanced pre-training for language understanding and generation. arXiv:2107.02137
Wang Q, Xiyou S (2022) Research on named entity recognition methods in Chinese forest disease texts. Appl Sci 12(8):3885
Zhang H, Cisse M, Dauphin YN, Lopez-Paz D (2017) mixup: beyond empirical risk minimization. arXiv:1710.09412
Zhang L, Nie X, Zhang M, Gu M, Geissen V, Ritsema CJ, Niu D, Zhang H (2022) Lexicon and attention-based named entity recognition for kiwifruit diseases and pests: a deep learning approach. Front Plant Sci 13:1053449
Zhang Y, Pu P, Huang L, Qian B, Liu Y (2023) Chinese named entity recognition of apple diseases and pests based on iterative dilated convolution, pp 1810–1815
Zhao P, Wang W, Liu H, Han M (2022) Recognition of the agricultural named entities with multifeature fusion based on albert. IEEE Access 10:98936–98943
Zhu H, Yang L, Ding W (2018) Chinese weibo named entity recognition based on topic tags and crf. J Central China Normal Univ (Natural Science Edition)

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com