Neural Computing and Applications (2023) 35:17717–17731
https://doi.org/10.1007/s00521-023-08635-5

ORIGINAL ARTICLE

EduNER: a Chinese named entity recognition dataset for education research

Xu Li[1] · Chengkun Wei[1] · Zhuoren Jiang[2] · Wenlong Meng[1] · Fan Ouyang[3] · Zihui Zhang[4] · Wenzhi Chen[1]

Received: 25 August 2022 / Accepted: 2 May 2023 / Published online: 20 May 2023

Abstract
A high-quality domain-oriented dataset is crucial for the domain-specific named entity recognition (NER) task. In this study, we introduce a novel education-oriented Chinese NER dataset (EduNER). To provide representative and diverse training data, we collect data from multiple sources, including textbooks, academic papers, and education-related web pages. The collected documents span ten years (2012–2021). A team of domain experts is invited to accomplish the education NER schema definition, and a group of trained annotators is hired to complete the annotation. A collaborative labeling platform is built for accelerating human annotation. The constructed EduNER dataset includes 16 entity types, 11k+ sentences, and 35,731 entities. We conduct a thorough statistical analysis of EduNER and summarize its distinctive characteristics by comparing it with eight open-domain or domain-specific NER datasets. Sixteen state-of-the-art models are further utilized for NER tasks validation. The experimental results can enlighten further exploration. To the best of our knowledge, EduNER is the first publicly available dataset for NER task in the education domain, which may promote the development of education-oriented NER models.

Keywords Chinese named entity recognition · Dataset · Benchmark · Education

1 Introduction

Named entity recognition (NER) is a fundamental technique in NLP that can be used for generating structured knowledge from texts [50]. Especially in the domain of education, the application of NER technology is salient and elemental that can be used to facilitate a series of downstream tasks, e.g., enabling the structured analysis of text content to identify specific knowledge points [1], providing an automated pedagogical scaffolding for teaching and learning [6], supporting the insight into the topic

✉ Zhuoren Jiang
jiangzhuoren@zju.edu.cn

✉ Wenzhi Chen
chenwz@zju.edu.cn

Xu Li
lixu2019@zju.edu.cn

Chengkun Wei
weichengkun@zju.edu.cn

Wenlong Meng
mengwl@zju.edu.cn

Fan Ouyang
fanouyang@zju.edu.cn

Zihui Zhang
zhangzihui@zju.edu.cn

[1] College of Computer Science and Technology, Zhejiang University, 38 Zheda Rd., Hangzhou 310027, Zhejiang, China

[2] School of Public Affairs, Zhejiang University, 866 Yuhangtang Rd., Hangzhou 310058, Zhejiang, China

[3] College of Education, Zhejiang University, 866 Yuhangtang Rd., Hangzhou 310058, Zhejiang, China

[4] Information Technology Center, Zhejiang University, 866 Yuhangtang Rd., Hangzhou 310058, Zhejiang, China

knowledge of discussion forums [1], and assisting the automated writing assessments [51].

However, to the best of our knowledge, there is no education-oriented NER dataset publicly available. This problem could threaten the NER model performance in the education domain for the following reasons:

- From the *model* viewpoint, most machine learning models are built on the assumption that the training and testing data are from the same domain (the assumption of independent and identically distributed, I.I.D) [13]. Although a variety of methods have been proposed to address the cross-domain task for low-resource domains, creating a domain-target dataset is still the most direct and effective way to improve model performance. For instance, an existing research [33] has demonstrated that the state-of-the-art NER systems developed for a specific domain cannot keep their satisfactory performance on other domains. Therefore, a high-quality annotated dataset is one of fundamental requirements to build a domain-specific NER model.
- From the *domain* viewpoint, a dedicated dataset is urgently needed in education domain. First, although the open-domain dataset may contain some education-related data, education domain involves a large amount of domain-related terminologies that cannot be annotated in the open-domain. For instance, a series of off-the-shelf NER or Chinese word segmentation tools trained on open-domain datasets have failed to identify the correct entities in education domain. Figure 1 shows an example that for an educational related text, the off-the-shelf tools (Jieba,[1] LTP,[2] Spacy,[3] and Stanza[4]) provide wrong word segmentation and cannot identify any correct entity. Second, the rapid technological developments have already had a significant impact on education; a number of technology-related concepts are emerging. An up-to-date dataset is needed to capture these new concepts and knowledge in the education domain. For instance, there are 1150 technology-related entities and 825 artificial-intelligence-related entities in our constructed dataset (see Table 1). Third, many domain-oriented NER datasets are now available, such as e-commerce [35] and biomedicine [5, 19]. An education-oriented Chinese dataset can enrich the domain-oriented NER dataset family and support applications of low-resource language.

In this study, we introduce EduNER, a human-annotated NER dataset from scratch for education domain. Specifically, we collected text data focused on *learning science* and *educational technology*.[5] Abundant texts from multiple sources have been collected, including a textbook, journal papers, and education-related web pages. The collected documents span ten years (2012–2021). To ensure the quality of EduNER, a team of domain experts is invited to define and organize the education-specific NER schema. A group of trained annotators is hired to accomplish the annotation. Meanwhile, to simplify and accelerate the process of human annotation, we design and implement a collaborative corpus labeling platform.[6] The constructed EduNER dataset includes 16 entity types, $11k+$ sentences, and 35,731 entities. To better understand EduNER, a thorough statistical analysis is conducted; eight open-domain or domain-specific NER datasets are compared to summarize its distinctive characteristics. Sixteen state-of-the-art (SOTA) models are utilized for NER task validation. The experimental results demonstrate that EduNER is challenging for SOTA algorithms, indicating that EduNER could further be used to develop education-oriented NER models and facilitate practical education-related applications.

In summary, the contributions of this work are the following:

- We introduce EduNER; to the best of our knowledge, it is the first publicly available NER dataset for education domain.
- We invite a team of education domain experts to accomplish the NER schema definition and hire a group of trained annotators to complete the annotation. Two data correction strategies are proposed to ensure the labeling quality. An online collaborative labeling platform is built to facilitate the annotation process.
- We conduct extensive analyses and experiments to gain insight into EduNER, including conducting the thorough comparison analyses with 8 NER datasets and benchmarking 16 SOTA models for experimental validation. The empirical results can be used to enlighten future studies.

2 Related works

NER is an essential and fundamental task in NLP and can directly assist downstream tasks, e.g., Information Extraction, Question Answering, and Knowledge Graph Construction [18, 26, 50]. Normally, NER tasks require high-quality annotated datasets for training, which would cost

[1] https://github.com/fxsjy/jieba.

[2] https://github.com/HIT-SCIR/ltp.

[3] https://spacy.io/.

[4] https://stanfordnlp.github.io/stanza/.

[5] Educational technology is an interdisciplinary discipline consisting of education, computer science, and other disciplines [27].

[6] Collaborative Annotation Platform: http://openaied.cn/.

Off-the-shelf tools	Tokenize/NER	句　子: 认知主义理论开始影响教学设计 Sentence: Cognitivism theory has begun to influence instruction design
Gold Label	NER Entity	认知主义理论(cognitivism theory) / *** / 教学设计(instruction design)
Jieba v0.42	Tokenize	认知 / 主义理论 / 开始 / 影响 / 教学 / 设计 cognition / doctrine theory / has / begun / to / influence / instruction / design
	NER	< Jieba doesn't provide NER function. >
LTP v4.0	Tokenize	认知主义 / 理论 / 开始 / 影响 / 教学 / 设计 cognition doctrine / theory / has / begun / to / influence / instruction / design
	NER	✗
Spacy v3.1	Tokenize	认知主义 / 理论 / 开始 / 影响 / 教学 / 设计 cognition doctrine / theory / has / begun / to / influence / instruction / design
	NER	✗
Stanza v1.2.3	Tokenize	认知 / 主义 / 理论 / 开始 / 影响 / 教学 / 设计 cognition / doctrine / theory / has / begun / to / influence / instruction / design
	NER	✗

Fig. 1 The comparing results of off-the-shelf tools. As described in the table, *Cognitivism theory* is a candidate entity, but the Chinese characters will obtain different segmentation, e.g., *"cognition"*, "doctrine", and "theory" etc. The '*X*' means that the tool cannot recognize any entity

Table 1 Statistic of entities related to technical terms in EduNER dataset

	Model-related	Algorithm-related	Computer-related	AI-related	Technique-related
Entities	270	309	349	825	1150

significant labor and time to collect and annotate resources. In recent years, numerous NER datasets have been published in academia and industry; these datasets can be categorized into open-domain datasets and domain-oriented datasets [34].

Open-domain dataset Open-domain datasets are not limited to a specific domain. For example, CoNLL03 [40] is a popular NER dataset, which includes four entity types (person, location, organization, and miscellaneous names), and is extensively used as a benchmark dataset. The OntoNotes[7] project contains a large-scale annotated corpus with various text corpus (e.g., telephone conversations, newsgroups, broadcast news, broadcast conversation, weblogs, and religious texts). The OntoNotes has released five version datasets which are tested by a lot of NER models [22, 26]. The development of Chinese NER datasets has also attracted extensive attention. In recent years, scholars have constructed several Chinese open-domain datasets from social media, such as Weibo [31, 32], MSRA [17], and Resume [49]. These open datasets have significantly promoted the development of NER research.

Domain-oriented dataset Domain-oriented datasets are limited to specific domains. This kind of dataset requires a specific entity schema and fine-grained entity labels. Many biological and medical research-related datasets have been constructed to boost the health and disease diagnosis research. Kim et al. [14] develop GENIA, a NER corpus with 36 entity types for biology and clinic text mining. Tanabe et al. [39] create a biomedical corpus (GENETAG) for recognizing gene/protein names. Truong et al. [41] present a COVID-19 NER dataset for supporting epidemics application. Encouraged by previous works, in biomedicine domain, FSU-PRGE [9], NCBI-Disease [5], and BC5CDR [19] datasets are proposed one after another. Recently, Zheng et al. [50] build an electric power metering domain corpus. Based on the domain-oriented datasets, several models or applications are proposed and reported to achieve human-like performance in a specific domain, e.g., D3NER [3] and CMeKG project.[8]

While the NER models and dataset constructions have been rapidly developed in multiple domains, there is a lack of domain-oriented data resource to support community research and applications in the education domain. As aforementioned, the application of NER technology in the education domain can facilitate the structured analysis of textual content for teachers and students and support a series of downstream educational tasks [1, 6, 51]. Therefore, in education domain, there is an urgent need for constructing a high-quality domain-specific dataset.

[7] OntoNotes:https://catalog.ldc.upenn.edu/LDC2013T19.

[8] Medicine QA: http://cmekg.pcl.ac.cn/.

3 Corpus collection and construction

3.1 Corpus collection

The selection of sources for the text corpus should be representative, diverse, and correct. As for the education domain, the professional knowledge can be collected mainly from authoritative textbooks and academic papers. To guarantee the diversity and richness of our data sources, we also collected an amount of high-quality education-related web pages. Firstly, an authoritative book *modern educational technology (4th Edition)* [47] was chosen. While parsing the book, we obtained the full plain text content. Secondly, we crawled the education domain's representative papers from five high-level Chinese journals and twelve conferences. Among them, 1107 highly cited papers (ranging from 2012 to 2021) were retained. Thirdly, we manually selected credible and authoritative web pages as a corpus supplement, including discipline terminology dictionary, conferences reports, and domain experts' reports. Table 2 shows the statistic of the collected raw data, including: an authoritative book,[9] the education domain's representative papers from five high-level Chinese journals,[10] and twelve conferences[11], and seventy-five Web pages.

Table 2 Statistics of the sources of EduNER dataset

Source	# Documents	# Characters	Percentage (%)
Textbook	1	225,702	34.3
Papers	1107	385,189	58.5
Web pages	75	47,703	7.2
Total	1183	658,594	100

3.2 Named entity schema

The general entity types (e.g., *Date*, *Organization*, and *Location*) that defined for the open domain are inadequate for education domain. Therefore, a domain-oriented named entity schema is necessary for supporting the various scenarios of teaching or learning activities. To address this problem, we invited a team of three education domain experts to define the entity schema. Sixteen education-oriented entity types were clarified with these experts' guidance. Table 3 presents a detailed description of each entity type.

3.3 Human annotation

Human annotation is a crucial step for building a high-quality NER dataset. In this study, five students from corresponding disciplines are hired to accomplish the NER annotation. Due to the expertise of the education domain, we require all annotators to have a corresponding disciplinary background. Through rigorous training by domain experts, all annotators are able to understand the well-defined named entity schema. The Cohen's kappa [2] was employed to examine the IAA (inter-annotator agreement). Overall, $\kappa = 0.82$, which indicates that all annotators have a good agreement [42]. For the inconsistent label, domain experts would recheck these entities and types and consequently modify ambiguous or redundant entity categories.

To simplify and accelerate the process of human annotation, we build an easy-to-use online annotation platform (see Fig. 2). This collaborative platform can support multiple users working simultaneously. The user interface is friendly for annotating the Chinese documents. To ensure that the annotation can be completed successfully, all annotators are also trained on how to use this annotation platform. Although utilizing an AI model for automatic data annotation could be a promising trend for future dataset construction tasks, such a solution has a prerequisite that only a model with high accuracy can be used to label the corpus automatically [48]. If the accuracy of the model is low, it will lead to more errors in the labeling results. In that case, human experts may need to dedicate additional time and effort to rectify these mistakes. Our annotation platform provides a simple and easy to follow procedure. First, the prepossessed corpus data is loaded into the labeling platform. Then, all annotators need to identify the entity type index for each character under candidate entity label. The platform supports variant NER label schema, such as BIO, BIOS, and BIOES. We follow the CoNLL label schema [37], BIO, where O is used to annotate all tokens that do not belong to any entities. (B-label) indicates that the token is the beginning of an entity (B-). (I-label) indicates that the token is contained or at the end of an entity (I-) [10]. After labeling is finished, the system can automatically transform the index to the corresponding label character. The output dataset is consistent with the CoNLL format, and the characters and labels are split by spaces.

Data correction To ensure the quality of labeling results, we employed two strict approaches: a two-phase labeling strategy and a regular expressions (regex)

[9] A textbook widely adopted in Chinese universities.

[10] Journals: *Open Education Research; e-Education Research; Modern Distance Education Research; Journal of Distance Education; Modern Educational Technology.*

[11] *Global Smart Education Conference; Global Chinese Conference on Computers in Education, etc.*

Table 3 Description of EduNER schema

Entity types	Description
ALG	Computer algorithm
BOO	Book, textbook
COF	Conference relate to Education domain
CON	Discipline concept
COU	Country
CRN	Course name, discipline domain, e.g., Math, Educational Psychology
DAT	Date
FRM	The typical architecture, framework, model applied in education situation
JOU	Journals
LOC	Address, state or province
ORG	All organization, e.g., association, committee, department, faculty
PER	Person name, person reference, e.g., student, teacher, etc.
POL	Policy, authoritative report, guidance document, etc.
TER	Discipline terminology
THE	Discipline theory
TOO	Technique, method, tool, platform, etc.

correction procedure. The two-phase labeling strategy entails annotating the corpus in two stages, with the second stage aimed at resolving issues encountered during the first phase. Due to the characteristics of the domain, we use authoritative textbooks as the source for annotation in the first phase. By analyzing the annotation results of the first phase, we found that our pre-defined knowledge types were not effectively covered (known as the long-tail problem). Therefore, we aim to address these problems in the second phase, for example, merging entity types, expanding the annotation corpus for sparse entity types in the first stage, and alleviating the data imbalance problem. For the two-phase labeling strategy, in the first phase, a batch of the corpus with 220,000 characters was manually labeled. Then, NER models [7, 12] were trained to evaluate the labeling performance on each entity type. In the second phase, we expanded the dataset based on the evaluation results in the first phase: If the models obtained poor performance on a certain entity type, then more raw data containing this entity type will be added to this dataset for labeling. This strategy can reduce the labeling error and keep the dataset as balanced as possible across different entity types. As shown in Table 4, in the first phase, the amounts of entity types are very imbalanced. With the correction of the second phase, the entity types are more balanced. A more detailed analysis will be provided in Sect. 4.2.

The regex correction procedure was designed to check and revise the labeling results. Due to the different understanding of the same entity in various contexts, the inconsistencies often occurred in the manual annotation. We summarized the most common errors in the labeling process and designed the corresponding regex to check the consistency. We also designed the corresponding regex to fix the different kinds of errors, e.g., the mix error of spaces

Fig. 2 A snapshot of the collaborative corpus labeling platform; the box (top) suggests the candidate entity, and the box (right) indicates the corresponding index of entity type. The annotator is asked to identify the correct type index of the candidate entity. The default index of character is set to 0

Table 4 Comparison of entity type amounts in two annotation phases. '–' indicates that this entity type was removed in final version. The amounts of entity types in the second phase are more balanced than in the first phase. ISB means the ISBN of book, and the PAP means the title of paper

	ALG	BOO	COF	CON	COU	CRN	DAT	FRM	ISB	JOU	LOC	ORG	PAP	PER	POL	TER	THE	TOO
The first phase	1	20	1	892	135	43	218	5	4	16	38	396	13	1,643	36	3,189	232	901
The second phase	309	251	205	4,477	947	1,195	2,048	376	–	173	385	2,354	–	8,205	326	9,398	715	4,367
Second/First ratio	309.00	12.55	205.00	5.02	7.01	27.79	9.39	75.20	–	10.81	10.13	5.94	–	4.99	9.06	2.95	3.08	4.85

and tabs. Overall, 159 regex expression items were designed to certify the annotation quality.

4 Data analysis of EduNER

4.1 Dataset scale

Dataset scale is an important measure of the dataset quality. After a series of preprocessing steps, we finally obtained 654,576 high-quality human-labeled characters. Table 5 shows the comparison of our dataset with six classic Chinese-related NER datasets. From the viewpoint of entity type amount, the proposed EduNER has the most entity types (16 types), while the other datasets have no more than 10 entity types (types $\leqslant 10$). It indicates that, compared with classic datasets, EduNER can express a richer and more detailed structured semantics. Meanwhile, form the viewpoint of oriented domain, EduNER is the only domain-oriented Chinese dataset. In summary, as a domain-oriented dataset, our dataset has a comparable size and covers the most entity types.

4.2 Dataset distribution

A balanced entity type distribution can benefit the NER model performance. As aforementioned, we designed a two-phase labeling strategy to overcome the unbalance problem. Table 4 shows the detailed comparison of entity distribution (instance amount under entity type) in two annotation phases. We can find that, in the first phase, there is only one instance of *ALG* type, while there are 3189 instances of *TER* type. The maximum ratio reached 3189 times. To further improve the balance of entity type distribution, in the second, more samples with rare entity type were added into the dataset. For example, the amount of instances of *ALG* entity type is increased to 309 times. As shown in Table 4, with correction of the second phase, the maximum ratio reduces to 54.32 (ratio of the number of *TER* entities to the number of *JOU* entities: 9398/173).

EduNER has been divided into three parts: train, dev, and test. Figure 3 shows the distribution of the *training* dataset. The *TER* type has the largest proportion of entities, followed by *PER*. The ratio of these two types of entities is twice as high as the following two entity types (*CON and TOO*).

4.3 Dataset feature

To further clarify our dataset's features, we compare features of our dataset and six Chinese-related NER datasets. The detailed information is shown in Table 6. We observe

Table 5 Comparing EduNER with other six Chinese-related NER datasets

Language	Dataset name	Year	Entity types	Domain	Tags	Size
Chinese	PeopleDaily	1998	Person name, location, and organization	Open-domain	3	$\approx 27.8k$
Multilingual	MSRA	2006	Person name, location, and organization	Open-domain	3	$\approx 48.5k$
Multilingual	OntoNotes	2011	Person, location, organization and geo-political	Open-domain	4	$\approx 24.4k$
Chinese	Weibo	2015	Person name, location, organization, and geo-political	Open-domain	4	$\approx 1.9k$
Chinese	Resume	2018	Country, educational institution, location, personal name, organization, profession, ethnicity background, and job title	Open-domain	8	$\approx 4.8k$
Chinese	CLUENER2020	2020	Address, book, company, game, government, movie, name, organization, position, and scene	open-domain	10	$\approx 13.4k$
Chinese	EduNER	2021	Date, person, algorithm, book, conference, country, course/discipline name, discipline concept, representative framework/model, journal, location, organization, policy, discipline terminology, discipline theory, and techniques/tools	education-oriented	16	$\approx 11.1k$

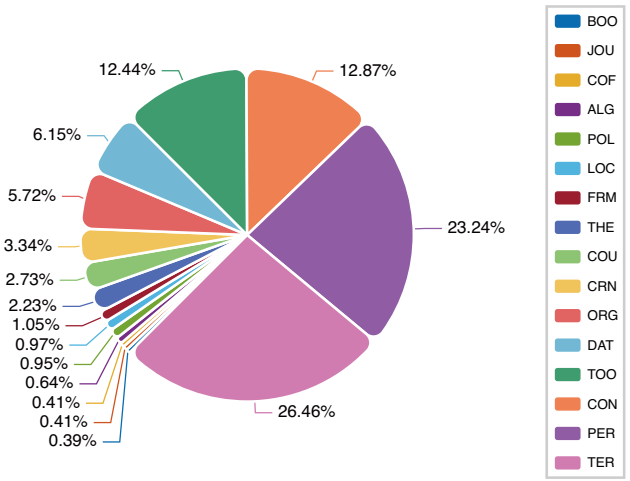

Fig. 3 Distribution of each entity type in the training dataset

the following: The average sentence length in EduNER is the longest. Meanwhile, EduNER and Resume achieve the highest entity density. This indicates that our dataset is relatively more complex for the NER task. For example, PeopleDaily dataset (train/dev) has almost twice the number of characters as our dataset, but the numbers of entities are almost the same. A higher density in entities could increase the task difficulty of the NER model. Therefore, EduNER brings a greater challenge to the performance of the NER model.

Two popular domain-oriented datasets are also compared. Table 7 reports detailed comparison of features. From the comparison, we can have following observations: Among these three datasets, our dataset has largest size, longest sentence length, and the second highest entity density. This also shows that EduNER has richer semantic features among domain-oriented datasets.

4.4 Long-tail phenomenon

The long-tail phenomenon of datasets is a widespread problem in machine learning datasets [25]. The long-tail problem implies that the different types of samples in a dataset are unevenly distributed (see Fig. 3), potentially affecting the performance of neural network models. The research community uses several strategies to address the long-tail problem. The first is to design novel loss functions or network architectures to mitigate the impact of the long-tail problem on the model's performance [21, 25, 30]. The second is to enlarge the dataset size to cover as many different entity types as possible. In our construction process, we adopt the second strategy to weaken the long-tail problem by expanding the sample size of different entity types in the second round of annotation stage (see Table 4). The final data annotation results indicate that our pre-defined entity types are not evenly distributed among the texts in the education domain. For example, the social science-related texts will contain fewer entity types of *ALG* and *TOO*. However, identifying these types of entities is critical for cross-disciplinary research in education. As a result of our dataset's specific target domain and application requirements, EduNER faces difficulty in fully addressing the long-tail problem, presenting new challenges for researchers in both computer science and education.

Table 6 Comparing the features of EduNER with other Chinese-related NER datasets, '–' indicates that these information are not available. Entity density = Number of entities/number of sentences

		Num. # sentence	Avg. # length	Num. # characters	Num. # entity	Entity density
Weibo	*train*	1,350	54.65	75,123	1,370	1.01
	dev	270	53.74	14,778	301	1.11
	test	270	54.97	15,110	308	1.14
Resume	*train*	3,821	32.49	127,919	13,343	3.49
	dev	463	30.00	14,352	1,488	3.21
	test	477	31.67	15,576	1,630	3.42
OntoNotes	*train*	15,724	31.28	507,626	13,372	0.85
	dev	4,301	46.62	204,805	6,950	1.62
	test	4,346	47.88	212,411	7,684	1.77
CLUENER2020	*train*	10,748	37.38	401,764	17,501	1.63
	dev	1,343	37.42	50,260	2,219	1.65
	test	1,345	37.90	50,972	–	–
Peopledaily	*train*	20,864	46.93	1,000,043	33,992	1.63
	dev	2,318	47.40	112,187	3,819	1.65
	test	4,636	47.28	223,832	7,707	1.66
MSRA	*train*	45,057	48.19	2,216,572	75,059	1.67
	dev	–	–	–	–	–
	test	3,442	50.15	176,042	6,190	1.80
EduNER (our)	*train*	8,419	59.75	511,436	28,113	3.34
	dev	1,015	56.68	58,546	3,093	3.05
	test	1,617	51.32	84,594	4,525	2.80

Table 7 Comparing the features of EduNER with other domain-oriented NER datasets. Entity density = Number of entities/number of sentences

	NCBI			BC5CDR			EduNER		
	train	*dev*	*test*	*train*	*dev*	*test*	*train*	*dev*	*test*
Num. # sentence	5424	923	940	4560	4581	4797	8,419	1,015	1,617
Avg. # sentence length	25.02	25.97	26.06	25.91	25.64	26.01	59.75	56.68	51.32
Num. # character	141,124	24,891	25,436	122,728	122,032	129,546	511,436	58,546	84,594
Num. # entity	5134	787	960	9385	9593	9802	28,113	3,093	4,525
Entity density	0.95	0.85	1.02	2.06	2.09	2.04	3.34	3.05	2.80

5 Experiments

Experiment setting. We conducted extensive experiments to validate our dataset and reveal the existing models' potentials and challenges. Firstly, we randomly split EduNER into three subsets. We kept the proportion at 78.13% for training (TRAIN), 8.94% for validation (DEV), and 12.92% for testing (TEST).

Second, our experiments were carried out on two high-performance servers, with Ubuntu 20.04, 2 Tesla V100 GPUs with 32 G memory. As our experiments include models with many different architectures and involve a large number of parameters, the hyperparameter settings keep consistent with the original paper. Details of experimental code and a sampled dataset can be found in the GitHub repository https://github.com/anonymous-xl/eduner.

Third, for pre-trained word embedding, we use the Chinese pre-trained character or word embeddings, e.g., *ctb.50d*, *gigaword_chn.all.a2b.bi.ite50*, and *gigaword_chn.all.a2b.uni.ite50* in line with [46]. As pre-trained language model, we use the *bert-base-chinese*.[12]

[12] Chinese BERT: https://huggingface.co/bert-base-chinese.

5.1 Benchmark models

Currently, deep learning-based approaches show the most promising performance [29]. Therefore, in this study, we validate our benchmark with deep learning-based models. To build a comprehensive validation, we conduct thorough experiments to compare the performance of various SOTA models in recent years.

Classic deep learning model Bidirectional LSTM [12] as a classic deep learning model to solve sequence labeling tasks, like NER, attracts extensive attention and can be a fundamental baseline model. The architecture fully leverages the bidirectional contextual information to represent the input text and uses the CRF algorithm [15] to enhance the decoding capacity by learning the constraints within tags.

Pre-trained language models To improve the representation of input text, the pre-trained language model is proposed to represent the characters or word contextual embedding. It can differentiate the embedding of the same word in various contexts. To this end, the method significantly improves task performance. BERT [4] as the representative pre-trained language model is extensively integrated into various NLP tasks, including the NER task.

Chinese-oriented models The diverse characteristics of languages would affect NER task performance. For instance, English has natural segmentation, while Chinese does not. As there are no clear word separators, vague word boundaries will cause a lot of boundary ambiguity and increase the difficulty of locating the boundaries of Chinese entity [23]. Thereby, a number of Chinese-oriented models are proposed to resolve this problem. Zhang and Yang [49] firstly propose the lattice architecture to optimize the Chinese NER task. Inspired by this, a series of models for Chinese NER have been developed:

- *TENER* [45], a novel NER architecture is adopting an adapted Transformer Encoder to model the character-level and word-level features. By incorporating the information of direction-aware, distance-aware, and unscaled attention, the model is proved effective in improving performance.
- To alleviate injecting erroneous information due to wrong matched lexicon words, Hu and Wei [11] propose a novel *SLK-NER* architecture with the second-order lexicon knowledge for each character, including the semantic and boundary information of lexicon words.
- Ma et al. [26] propose a simple but effective lexicon fusion method, *SoftLexicon*, which incorporates the lexicon information by categorizing the potential words.
- Li et al. [22] propose the Flat-Lattice Transformer for Chinese NER, which converts the lattice structure into a flat structure consisting of spans. *FLAT* can fully leverage the lattice information and has an excellent parallelization ability.
- Liu et al. [24] introduce the *LEBERT* model, which considers integrating the lexicon features into the bottom layer of BERT effectively.
- Wu et al. [44] use a two-stream Transformer to integrate the multi-metadata embedding, including character feature and the radical-level embedding (*MECT4NER*).
- Gui et al. [7] describe the *LR-CNN* model, which uses a rethinking mechanism to refine the weights of embedding words and resolves conflicts in latent words by adding a feedback layer.
- Gui et al. [8] transform the NER task to a node classification task by adopting a graph neural network to fuse the lexicon information (*LGN*).

In addition, we also compare three *Distinctive models* which were proposed recently and achieved relatively high performances.

- Schweter and Akbik [36] propose the *FLERT* model, which is based on document-level to acquire information crossing sentence boundaries for improving the NER task performance.
- Wang et al [43] propose CL-KL and CL-L_2 models, which utilize the cooperative learning method to combine the original and external information. They argue that a model that fuses the original input sentence and its external contexts can facilitate the NER task performance.

Metrics The *Exact-match* metric requires to predict the right entity boundary and entity type, which can measure the entity-level performance more rigorously [20]. We use the *Exact-match* metric in our experiments for evaluation. Specifically, we report the *Precision (P)*, *Recall (R)*, and *F1-score (F1)* metrics by *seqeval* [28].

5.2 Benchmark results

Models performance Table 8 presents the experimental results of these models. To better demonstrate the dataset characteristics, we also report the model performances on two popular Chinese NER datasets, namely Resume and Weibo. The result indicates that the CL-L_2 model can obtain the best performance on our dataset. Meanwhile, we find that the Chinese-oriented models can achieve comparable performance, such as SoftLexicon (LSTM) achieving 68.71 on F1, MECT4CNER achieving 68.31 on F1, and LEBERT achieving 68.40 on F1. The best F1 is 69.50 on

Table 8 Benchmark results on 16 NER models. '–' indicates that the corresponding metrics are not reported in original papers. We use long dash line to distinguish model categories. From top to bottom: classic deep learning model, pre-trained language model, Chinese-oriented model and distinctive models

Models	Weibo			Resume			EduNER		
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
BiLSTM+CRF	40.00	14.38	21.15	92.42	92.08	92.25	71.74	51.90	**60.23**
BERT	50.89	47.19	48.48	93.89	96.01	94.93	62.13	69.00	65.27
BERT+CRF	50.34	47.15	48.69	94.24	96.44	95.33	65.87	69.42	**67.60**
LR-CNN	57.14	66.67	59.92	95.37	94.84	95.11	64.87	60.15	62.42
TENER	–	–	57.40±0.3	–	–	95.00±0.25	63.64	68.46	65.96
LGN	55.34	64.98	60.21	95.28	95.46	95.37	71.89	61.83	66.48
FLAT+BERT	–	–	68.55	–	–	95.86	68.65	67.13	67.88
SoftLexicon(CNN)	–	–	59.65	–	–	95.02	62.34	65.32	63.80
SoftLexicon (Transformer)	–	–	61.04	–	–	94.59	67.02	63.24	65.07
SoftLexicon (LSTM)	–	–	70.50	96.08	96.13	96.11	72.05	65.67	**68.71**
MECT4CNER	–	–	70.43	–	–	95.98	73.14	64.07	68.31
SLK-NER	61.80	66.30	64.00	95.20	96.40	95.80	66.10	67.52	66.81
LEBERT	–	–	70.75	–	–	96.08	66.47	70.46	68.40
FLERT	–	–	–	–	–	–	64.90	68.85	66.82
CL-KL	–	–	–	–	–	–	68.83	68.16	68.49
CL-L_2	–	–	–	–	–	–	69.34	69.66	**69.50**

Table 9 Performance comparison (F1) of the representative models on each entity type. BiLSTM+CRF represents the classical model, BERT+CRF represents the pre-trained language model, LGN, SoftLexicon (LSTM), and LEBERT represent the Chinese-oriented models, CL-L_2 represents the distinctive NER model

Tags	BiLSTM+CRF	BERT+CRF	LGN	SoftLexicon (LSTM)	LEBERT	CL-L_2
ALG	11.76	52.38	28.57	49.15	54.84	**56.41**
BOO	68.80	**92.86**	80.00	87.32	90.65	90.91
COF	34.62	51.28	51.35	**63.89**	43.04	62.16
CON	68.80	68.31	70.34	69.40	69.30	**71.53**
COU	76.53	**92.73**	88.21	90.09	90.58	91.90
CRN	53.75	**71.43**	64.15	65.47	68.44	69.82
DAT	86.67	87.46	**89.25**	88.24	85.71	85.81
FRM	3.77	23.53	22.78	**36.92**	21.92	27.69
JOU	70.59	87.50	72.73	**89.36**	81.63	69.57
LOC	27.27	**70.37**	61.86	55.56	66.67	69.81
ORG	50.17	75.57	71.23	75.71	76.54	**78.95**
PER	77.92	85.63	82.16	**87.16**	84.20	86.10
POL	54.05	**69.66**	69.44	63.53	69.47	69.23
TER	46.16	52.88	49.57	51.76	**54.48**	54.12
THE	68.42	74.42	78.26	76.52	75.59	**86.67**
TOO	49.56	53.35	57.83	57.37	**58.87**	55.88

EduNER (achieved by CL-L_2); meanwhile, on Weibo and Resume, the best F1 values are 70.75 achieved by LEBERT and 96.11 achieved by SoftLexicon (LSTM), respectively. Generally, there is sufficient room for improvement in the performance of NER models on EduNER.

Table 9 shows the performance of these models on each entity type. The BERT+CRF acquires the best performance on five entity types, while the CL-L_2 and SoftLexicon (LSTM) achieve the best performance on four entity types, respectively. LEBERT and LGN perform best on *TER*, *TOO*, and *DAT*. The results show that the NER models would achieve diverse performances for different entity types. It is important to emphasize that there are certain sparse entity types in our dataset, but the performance for these entity types is different. For example, *FRM*

Table 10 Performance comparison with two recent models on domain NER datasets. As the first two models are designed and revised for the medical domain, we do not reproduce these models on our dataset. '–' indicates the corresponding metrics are not reported

Models	NCBI			BC5CDR			EduNER		
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
D3NER	85.03	83.80	84.41	83.98	85.40	84.68	–	–	–
BioBERT+MRC	89.67	90.42	90.04	88.61	87.07	87.83	–	–	–
CL-KL	–	–	90.93	–	–	88.96	68.83	68.16	68.49
CL-L_2	–	–	**90.99**	–	–	**89.22**	69.34	69.66	**69.50**

and *JOU*, the number of *FRM* entities is more than *JOU*, but *FRM* falls short of *JOU* in terms of performance evaluation metrics. By retrieving and analyzing samples of these types of entities in the dataset, we summarize the following reasons may be responsible for the significant performance differences: 1) The average length of entity samples affects the performance of entity types. Generally, the probability of model misclassification tends to increase with longer entity lengths. Through comparative analysis, we found that the average length of *FRM* entities is 5.78 characters, while the average length of *JOU* entities is 5.01 characters. 2) Entities with clear, easy-learnable patterns are essential for good performance. Neural network models are more likely to learn good representation and perform well on entity types with clear patterns. The analysis results show that 97 samples of *JOU* entities are wrapped up by the symbol "≪≫," accounting for 85.08% of the samples of this type of entity. This implies that the majority of entities belonging to the *JOU* category share common characteristics for model learning. 3) Our dataset is in Chinese character-level format, while English corpora are typically structured at the token or word level. Neural network models often struggle with predicting entities across different levels of annotation, resulting in performance bottlenecks when it comes to predicting both English and mixed Chinese–English entities. The analysis results show that the percentages of English and mixed Chinese–English entities of *FRM* and *JOU* types in our dataset are 30.61% and 11.40%, respectively. This could cause the performance differences. 4) *FRM* entities are typically found in texts that exhibit highly specific domain characteristics. It is rare to come across examples of *FRM* entities in general texts. On the other hand, *JOU* entities are more prevalent and are more likely to be included in general texts used for training pre-trained language models.

We also compare our dataset with two biomedicine-oriented datasets. Table 10 shows the model performances. D3NER [3] and BioBERT-MRC [38] are models targeted for the biomedicine domain. Based on the results, we can have the following observations: Through different algorithmic optimizations, the domain-oriented model and open-domain model can both achieve consistent and satisfactory performance on biomedicine-oriented datasets. For example, by using CL-L_2, the F1 can achieve 90.99 on NCBI dataset. In contrast, the best F1 is 69.50 on EduNER (achieved by CL-L_2). It indicates that the semantics of EduNER is more complex than other domain-oriented datasets. Therefore, our dataset has more room for model improvement, which may enlighten more sophisticated and education-oriented algorithm exploration.

5.3 Results analysis

5.3.1 Performance analysis

The empirical results suggest that the existing SOTA models do not achieve promising performance results on our dataset. The exploration for possible reasons of poor performance can be helpful for future research.

Sparsity of corpus The BERT-based models achieved good results on other datasets, including Chinese-related open-domain datasets and domain-oriented datasets. But this kind of models did not achieve good performance on our dataset. A possible explanation is that the pre-trained language models are always trained on open-domain corpora (such as Wikipedia). However, EduNER contains a very large amount of domain knowledge and terminologies. Such specific knowledge is relatively sparse in open-domain corpus. So, the semantic representation ability of pre-trained language models may be weaken for domain-oriented dataset [16]. A simple fine-tuning without domain knowledge may not obtain the desired performance. This situation suggests that by combining domain-oriented knowledge, the model may improve their performance.

The long-tail problem could affect the performance of NER models [30]. Though we have adopted the two-phase annotation strategy to alleviate the long-tail problem in the dataset, the number of instances under the entity type is still uneven (see Fig. 4). Therefore, if the model is specifically designed for unbalanced data distribution, its task performance may be improved.

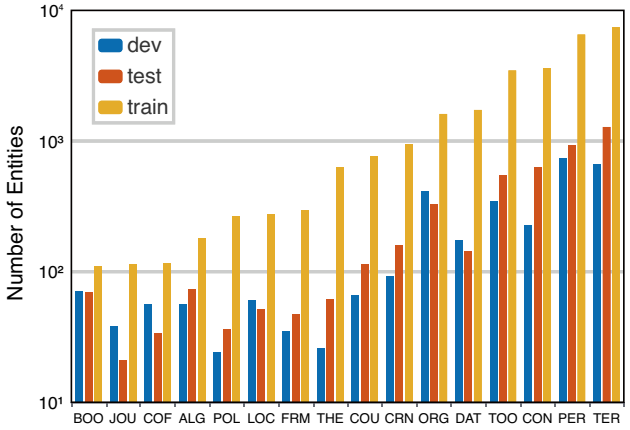

Fig. 4 An unbalanced distribution of each entity types in three subsets

The diversity of lexicon expression in Chinese may introduce entity ambiguity [23]. For example, "Learning Analytic" and "Learning Analytic technology" may be categorized into different entity types (e.g., *CRN or TOO*), but these two entities have the same meaning in certain Chinese contexts. Hence, the specific design for Chinese language may assist the algorithm in improving performance.

The OOV problem [49] may be a potential problem. We find that extensive discipline named entities are not found in the external pre-trained embedding vocabulary. These pre-trained embeddings are mentioned in our experimental settings. Specifically, the perfect match ratio is 1.95% on biword metric and 3.90% on the lexicon [26]. This situation will weaken the language-enhanced models such as SoftLexicon (LSTM). In addition, Chinese-oriented models heavily rely on external pre-trained word embeddings to enumerate potential entities, which in turn make full use of the sub-word information in the text sequences and help the models identify the boundaries of entities [49]. However, we found that the matching ratios on the EduNER dataset are very low (see Table 11). Therefore, for the EduNER dataset, these external pre-trained word embeddings provide very limited supporting capacity for NER model to enumerate potential entities and locate the boundaries of entities. The performance of such models would also be unsatisfactory.

The dataset size is a critical factor that affects models' performances [34]. Although EduNER has more than 650,000 characters, the average number of samples for each entity type is small due to the large number of entity types. As shown in Table 5, EduNER has the most entity types compared with other popular datasets. The training number of samples for each entity type will significantly affect the performance of machine learning-based NER models. As an example of a similar situation: The Weibo dataset has 1.9*k* sentences with 4 entity types, and the number of samples for each entity type is also small. Therefore, all models got poor F1 scores on this dataset as well.

Table 11 Comparison of external pre-trained word embeddings

External lexicon	Vocab size	Full_match	Match_ratio	Sample	Dataset
ctb.50d.vec	704,368	1268	15%	8668	EduNER
		431	48%	897	Weibo
		6736	38%	17,826	MSRA
		4481	42%	10,594	Peopledaily
gigaword_chn_bi	3,986,686	833	10%	8668	EduNER
		352	39%	897	Weibo
		2496	14%	17,826	MSRA
		1638	15%	10,594	Peopledaily
sgns.merge.word	1,292,608	2119	24%	8668	EduNER
		545	61%	897	Weibo
		9660	54%	17,826	MSRA
		6292	59%	10,594	Peopledaily
cn_bi_fastnlp	883,344	623	7%	8668	EduNER
		342	38%	897	Weibo
		2140	12%	17,826	MSRA
		1445	14%	10,594	Peopledaily
yangjie_word_char_mix	709,995	1268	15%	8668	EduNER
		431	48%	897	Weibo
		6736	38%	17,826	MSRA
		4481	42%	10,594	Peopledaily

Models	Language	一种基于条件随机场的半监督学习方法 A Semi-supervised Learning method based on Conditional Random Fields	本研究以中国大学MOOC平台上陕西师范大学现代教育技术课程作为实验对象 This study uses the Modern Educational Technology course at Shaanxi Normal University on the Chinese University MOOCs platform as an experimental object
Gold Labels	zh	条件随机场(ALG) / 半监督学习(ALG)	中国大学MOOC平台(TOO) / 陕西师范大学(ORG) / 现代教育技术(CRN) / 实验对象(TER)
	en	Conditional Random Field(ALG) / semi-supervised learning(ALG)	Chinese University MOOCs platform(TOO) / Shaanxi Normal University(ORG) / Modern Educational Technology(CRN) / experimental subject(TER)
BiLSTM+CRF	zh	--	现代教育技术(CRN)
	en	--	Modern Educational Technology(CRN)
BERT+CRF	zh	--	中国大学MOOC平台(TOO) / 陕西师范大学(ORG) / 实验对象(TER)
	en	--	Chinese University MOOCs platform(TOO) / Shaanxi Normal University(ORG) / experimental subject(TER)
LGN	zh	--	现代教育技术(CRN)
	en	--	Modern Educational Technology(CRN)
SoftLexicon (LSTM)	zh	半监督学习(TER)	陕西师范大学(ORG) / 实验对象(TER)
	en	semi-supervised learning(TER)	Shaanxi Normal University(ORG) / experimental subject(TER)
LEBERT	zh	--	中国大学MOOC平台(TOO) / 陕西师范大学(ORG) / 实验对象(TER)
	en	--	Chinese University MOOCs platform(TOO) / Shaanxi Normal University(ORG) / experimental subject(TER)
CL-L2	zh	--	陕西师范大学(ORG) / 现代教育技术(CRN)
	en	--	Shaanxi Normal University(ORG) / Modern Educational Technology(CRN)

Fig. 5 A case study of analyzing model prediction result. Entities corresponding to the model represent the result of correct identification. '–' means that the model cannot recognize any entity

5.3.2 Case study

To take a closer view at our dataset, we present two typical hard cases. Figure 5 shows the identified entities in two sentences from a series of representative models. In the first sentence, most models suffer from a poor performance. Only SoftLexicon (LSTM) correctly recognizes the boundary of the entity, but it does not accurately predict its type (the correct type is *ALG*). For the second sentence, the model performs better. Although no models can completely predict all entities, most of them can predict a portion of the correct entities. For example, the BiLSTM+CRF, LGN, SoftLexicon (LSTM), and CL-L_2 cannot correctly locate the long entity, *"...at Shaanxi Normal University on Chinese University MOOCs platform...,"* but they can correctly predict the short ones. BERT+CRF and LEBERT can correctly recognize the *TOO* entity, *"Chinese University MOOCs platform (TOO),"* but these two models fail to recognize the boundary of the other highly ambiguous entity accurately, *"Modern Educational Technology (CRN)."* The case example indicates that the current SOTA models still face challenges to predict entities from sentences with strong ambiguity and higher entity density. Domain-oriented knowledge, the possible ambiguity of lexicon expression in Chinese, and the complex semantics and entity composition in our dataset could affect the performance of NER models. These typical characteristics can serve as important references for future research directions.

6 Conclusion and future work

This work presents an education-oriented NER dataset. To the best of our knowledge, EduNER is the first publicly available dataset that can be used to address the NER task in education. EduNER contains representative and diverse educational knowledge by collecting data from multiple corpus sources. The annotation followed a well-defined educational NER schema guided by domain experts. A Chinese-friendly collaborative annotation platform is built for multiple annotators collaboration. Strict data correction strategies are proposed to ensure the dataset quality. Thorough analyses and extensive benchmark experiments based on EduNER provide a good understanding of the dataset and suggest the possible problem of SOTA model. We hope the proposed EduNER dataset can address the future NER research for education domain and make contributions to the NLP community. In future, we intend to further upgrade the collaborative labeling platform. This will involve enabling an automatic annotation function by deploying highly accurate neural network models, which will then be reviewed and corrected by human experts. Additionally, we are planning to expand our dataset to include multiple languages and cover a wider range of disciplines in the education domain.

Acknowledgments This work is supported by the National Natural Science Foundation of China (72104212), the Natural Science Foundation of Zhejiang Province (LY22G030002), the Key Research

and Development Plan of Zhejiang Province (2021C03140), and the Fundamental Research Funds for the Central Universities.

Data availability The datasets generated during and/or analyzed during the current study can be foundin the GitHub repository: https://github.com/anonymous-xl/eduner.

Declarations

Conflicts of interest We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

References

1. Chen CM, Tsao HW (2021) An instant perspective comparison system to facilitate learners' discussion effectiveness in an online discussion process. Comput Educat 164(104):037. https://doi.org/10.1016/j.compedu.2020.104037
2. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Measur 20(1):37–46. https://doi.org/10.1177/001316446002000104
3. Dang TH, Le HQ, Nguyen TM et al (2018) D3NER: biomedical named entity recognition using CRF-biLSTM improved with fine-tuned embeddings of various linguistic information. Bioinformatics 34(20):3539–3546
4. Devlin J, Chang MW, Lee K, et al (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp 4171–4186, https://doi.org/10.18653/v1/N19-1423
5. Dogan RI, Leaman R, Lu Z (2014) NCBI disease corpus: a resource for disease name recognition and concept normalization. J Biomed Inf 47:1–10. https://doi.org/10.1016/j.jbi.2013.12.006
6. Figueroa A (2017) Automatically generating effective search queries directly from community question-answering questions for finding related questions. Expert Syst Appl 77:11–19. https://doi.org/10.1016/j.eswa.2017.01.041
7. Gui T, Ma R, Zhang Q, et al (2019a) Cnn-based chinese NER with lexicon rethinking. In: Kraus S (ed) Proceedings of the twenty-eighth international joint conference on artificial intelligence, IJCAI 2019, Macao, China, August 10–16, pp 4982–4988. https://doi.org/10.24963/ijcai.2019/692
8. Gui T, Zou Y, Zhang Q, et al (2019b) A lexicon-based graph neural network for Chinese NER. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp 1040–1050, https://doi.org/10.18653/v1/D19-1096
9. Hahn U, Tomanek K, Beisswanger E, et al (2010) A proposal for a configurable silver standard. In: Proceedings of the fourth linguistic annotation workshop. Association for Computational Linguistics, Uppsala, Sweden, pp 235–242
10. Hamdi A, Linhares Pontes E, Boros E, et al (2021) A multilingual dataset for named entity recognition, entity linking and stance detection in historical newspapers. In: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. Association for Computing Machinery, SIGIR '21, pp 2328–2334
11. Hu D, Wei L (2020) SLK-NER: exploiting second-order lexicon knowledge for chinese NER. In: García-Castro R (ed) The 32nd international conference on software engineering and knowledge engineering, SEKE 2020, KSIR virtual conference center, USA, July 9-19, 2020. KSI Research Inc., pp 413–417, https://doi.org/10.18293/SEKE2020-153
12. Huang Z, Xu W, Yu K (2015) Bidirectional LSTM-CRF models for sequence tagging. ArXiv preprint abs/1508.01991
13. Karlos S, Aridas C, Kanas VG et al (2021) Classification of acoustical signals by combining active learning strategies with semi-supervised learning schemes. Neural Comput Appl. https://doi.org/10.1007/s00521-021-05749-6
14. Kim JD, Ohta T, Tateisi Y et al (2003) GENIA corpus—a semantically annotated corpus for bio-textmining. Bioinformatics 19(suppl-1):i180–i182. https://doi.org/10.1093/bioinformatics/btg1023
15. Lafferty JD, McCallum A, Pereira FCN (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Brodley CE, Danyluk AP (eds) Proceedings of the eighteenth international conference on machine learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001. Morgan Kaufmann, pp 282–289
16. Lee J, Yoon W, Kim S et al (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4):1234–1240. https://doi.org/10.1093/bioinformatics/btz682
17. Levow GA (2006) The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In: Proceedings of the Fifth SIGHAN workshop on Chinese language processing. Association for Computational Linguistics, Sydney, Australia, pp 108–117
18. Li F, Wang Z, Hui SC et al (2021) A segment enhanced span-based model for nested named entity recognition. Neurocomputing 465:26–37. https://doi.org/10.1016/j.neucom.2021.08.094
19. Li J, Sun Y, Johnson R, et al (2015) Annotating chemicals, diseases, and their interactions in biomedical literature. In: Proceedings of the fifth biocreative challenge evaluation workshop. The Fifth BioCreative Organizing Committee, pp 173–182
20. Li J, Sun A, Han J et al (2022) A survey on deep learning for named entity recognition. IEEE Trans Knowl Data Eng 34(1):50–70. https://doi.org/10.1109/TKDE.2020.2981314
21. Li X, Sun X, Meng Y, et al (2020a) Dice loss for data-imbalanced NLP tasks. In: Proceedings of the 58th annual meeting of the association for computational linguistics. Association for Computational Linguistics, pp 465–476
22. Li X, Yan H, Qiu X, et al (2020b) FLAT: Chinese NER using flat-lattice transformer. In: Proceedings of the 58th annual meeting of the association for computational linguistics. Association for Computational Linguistics, pp 6836–6842. https://doi.org/10.18653/v1/2020.acl-main.611
23. Liu P, Guo Y, Wang F et al (2022) Chinese named entity recognition: the state of the art. Neurocomputing 473:37–53. https://doi.org/10.1016/j.neucom.2021.10.101
24. Liu W, Fu X, Zhang Y, et al (2021) Lexicon enhanced Chinese sequence labeling using BERT adapter. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers). Association for Computational Linguistics, pp 5847–5858. https://doi.org/10.18653/v1/2021.acl-long.454
25. Liu Z, Miao Z, Zhan X, et al (2019) Large-scale long-tailed recognition in an open world. http://arxiv.org/abs/1904.05160
26. Ma R, Peng M, Zhang Q, et al (2020) Simplify the usage of lexicon in Chinese NER. In: Proceedings of the 58th annual meeting of the association for computational linguistics.

Association for Computational Linguistics, Online, pp 5951–5960. https://doi.org/10.18653/v1/2020.acl-main.528
27. Meifeng L, Jinjiao L, Cui K (2010) Educational technology in China. Br J Edu Technol 41(4):541–548. https://doi.org/10.1111/j.1467-8535.2010.01094.x
28. Nakayama H (2018) seqeval: A python framework for sequence labeling evaluation. Software. https://github.com/chakki-works/seqeval
29. Nasar Z, Jaffry SW, Malik MK (2021) Named entity recognition and relation extraction: state-of-the-art. ACM Comput Surv 54(1):1–39. https://doi.org/10.1145/3445965
30. Nguyen T, Nguyen D, Rao P (2020) Adaptive name entity recognition under highly unbalanced data. arXiv:2003.10296 [cs, stat]. https://arxiv.org/abs/arXiv:2003.10296 [cs, stat]
31. Peng N, Dredze M (2015) Named entity recognition for chinese social media with jointly trained embeddings. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics, Lisbon, Portugal, pp 548–554. https://doi.org/10.18653/v1/D15-1064
32. Peng N, Dredze M (2016) Improving named entity recognition for Chinese social media with word segmentation representation learning. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: short papers). Association for Computational Linguistics, Berlin, Germany, pp 149–155. https://doi.org/10.18653/v1/P16-2025
33. Poibeau T, Kosseim L (2001) Proper name extraction from non-journalistic texts. In: Computational Linguistics in the Netherlands 2000. Brill, pp 144–157
34. Qian H, Li X, Zhong H, et al (2021) Pchatbot: a large-scale dataset for personalized chatbot. In: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. Association for Computing Machinery, SIGIR '21, pp 2470–2477
35. Salinas Alvarado JC, Verspoor K, Baldwin T (2015) Domain adaption of named entity recognition to support credit risk assessment. In: Proceedings of the Australasian language technology association workshop 2015, Parramatta, Australia, pp 84–90
36. Schweter S, Akbik A (2021) Flert: Document-level features for named entity recognition. http://arxiv.org/abs/2011.06993
37. Sui D, Tian Z, Chen Y, et al (2021) A Large-Scale Chinese Multimodal NER Dataset with Speech Clues. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (Volume 1: long papers). Association for Computational Linguistics, pp 2807–2818
38. Sun C, Yang Z, Wang L et al (2021) Biomedical named entity recognition using BERT in the machine reading comprehension framework. J Biomed Inf 118(103):799
39. Tanabe L, Xie N, Thom LH et al (2005) GENETAG: a tagged corpus for gene/protein named entity recognition. BMC Bioinf 6(Suppl 1):S3
40. Tjong KSEF (2002) Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In: COLING-02: The 6th conference on natural language learning 2002 (CoNLL-2002), pp 142–147
41. Truong TH, Dao MH, Nguyen DQ (2021) COVID-19 named entity recognition for Vietnamese. In: Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies. Association for Computational Linguistics, Online, pp 2146–2153. https://doi.org/10.18653/v1/2021.naacl-main.173
42. Viera AJ, Garrett JM (2005) Understanding interobserver agreement: the kappa statistic. Fam Med 37(5):360–363
43. Wang X, Jiang Y, Bach N, et al (2021) Improving named entity recognition by external context retrieving and cooperative learning. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (Volume 1: Long Papers). Association for Computational Linguistics, Online, pp 1800–1812. https://doi.org/10.18653/v1/2021.acl-long.142
44. Wu S, Song X, Feng Z (2021) MECT: Multi-metadata embedding based cross-transformer for Chinese named entity recognition. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (Volume 1: Long Papers). Association for Computational Linguistics, Online, pp 1529–1539. https://doi.org/10.18653/v1/2021.acl-long.121
45. Yan H, Deng B, Li X, et al (2019) Tener: adapting transformer encoder for named entity recognition. ArXiv preprint
46. Yang J, Zhang Y, Dong F (2017) Neural word segmentation with rich pretraining. In: Proceedings of the 55th annual meeting of the association for computational linguistics (Volume 1: long papers). Association for Computational Linguistics, Vancouver, Canada, pp 839–849. https://doi.org/10.18653/v1/P17-1078
47. Zhang J (2016) Modern educational technology, 4th edn. Higher Education Press, Beijing
48. Zhang S, Jafari O, Nagarkar P (2021) A survey on machine learning techniques for auto labeling of video, audio, and text data. https://doi.org/10.48550/arXiv.2109.03784
49. Zhang Y, Yang J (2018) Chinese NER using lattice LSTM. In: Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: long papers). Association for computational linguistics, Melbourne, Australia, pp 1554–1564. https://doi.org/10.18653/v1/P18-1144
50. Zheng K, Sun L, Wang X et al (2021) Named entity recognition in electric power metering domain based on attention mechanism. IEEE Access 9:152,564-152,573. https://doi.org/10.1109/ACCESS.2021.3123154
51. Zupanc K, Bosnić Z (2017) Automated essay evaluation with semantic analysis. Knowl-Based Syst 120:118–132