BMC Medical Informatics and Decision Making

RESEARCH

Open Access

An improved data augmentation approach and its application in medical named entity recognition

Hongyu Chen[1†], Li Dan[1†], Yonghe Lu[2*], Minghong Chen[1] and Jinxia Zhang[3*]

Abstract

Performing data augmentation in medical named entity recognition (NER) is crucial due to the unique challenges posed by this field. Medical data is characterized by high acquisition costs, specialized terminology, imbalanced distributions, and limited training resources. These factors make achieving high performance in medical NER particularly difficult. Data augmentation methods help to mitigate these issues by generating additional training samples, thus balancing data distribution, enriching the training dataset, and improving model generalization. This paper proposes two data augmentation methods—Contextual Random Replacement based on Word2Vec Augmentation (CRR) and Targeted Entity Random Replacement Augmentation (TER)—aimed at addressing the scarcity and imbalance of data in the medical domain. When combined with a deep learning-based Chinese NER model, these methods can significantly enhance performance and recognition accuracy under limited resources. Experimental results demonstrate that both augmentation methods effectively improve the recognition capability of medical named entities. Specifically, the BERT-BiLSTM-CRF model achieved the highest F1 score of 83.587%, representing a 1.49% increase over the baseline model. This validates the importance and effectiveness of data augmentation in medical NER.

Keywords Data augmentation, Deep learning, Medical named entity recognition, Text features, Replacement augmentation

Introduction

Named entity recognition (NER) is a fundamental information extraction task [1]. Early NER methods relied on rule-based approaches, which required a significant amount of manual labor and expert knowledge to develop effective rules [2]. This made it difficult to apply them on a large scale. Later, statistical machine learning [3] methods emerged and significantly improved the accuracy of NER tasks. These approaches only require some annotated corpora to achieve good results, but the challenge of dealing with new words not found in the training data emerged as a major issue. In 2012, Hinton's team [4] proposed the AlexNet model based on deep learning convolutional neural networks for image recognition tasks, ushering in the era of deep learning. For NER tasks, deep learning methods can explore the deep semantic information in the text and learn it. This to some extent eases the challenge of dealing with new words and offers performance advantages in NER tasks.

[†]Hongyu Chen and Li Dan contributed equally to this work.

*Correspondence:
Yonghe Lu
luyonghe@mail.sysu.edu.cn
Jinxia Zhang
zhjinxia@foxmail.com
[1] School of Information Management, Sun Yat-Sen University, Guangzhou 510006, China
[2] School of Artificial Intelligence, Sun Yat-Sen University, Zhuhai 519082, China
[3] Department of Cardiology, General Hospital of Southern Theatre Command of PLA, Guangzhou 510010, China

Although deep learning-based NER models currently exhibit the best performance [5], their effectiveness still heavily relies on the size and accuracy of the training corpus. In the Chinese medical field, there are three main challenges: ① traditional issues in Chinese NER [6], such as the delimitation of Chinese word boundaries, ambiguity, and syntactic ambiguity; ② due to medical ethics considerations and the high level of medical knowledge specialization, annotated datasets for the Chinese medical field are relatively scarce [7]; and ③ in comparison to the general domain, most entities in the medical field are low-frequency words and include various rare characters. These challenges make the NER task in the Chinese medical field more challenging than in the general domain.

This paper focuses on the task of Chinese NER in electronic medical records. Based on deep learning models for NER, we propose two data augmentation methods, namely "Contextual Random Replacement Based on Word2Vec (CRR)" and "Targeted Entity Random Replacement Augmentation (TER)", specifically designed for Chinese medical entity recognition tasks. We conduct comparative experiments to validate the effectiveness of data augmentation methods in improving the recognition of medical entities in electronic medical records with limited annotated data. Our research contributions include:

(1) The proposal of two data augmentation methods suitable for Chinese medical NER tasks. Our experiments demonstrate that these two methods solve data scarcity in Chinese medical entity recognition tasks. The first data enhancement method, CRR, introduces the Word2Vec language model, vectorizes the entity context words in the original corpus, and compares them with the word vectors of the words in the preset lexicon to capture the synonyms with the highest cosine similarity, thus making the generated near-semantic corpus better in line with computer intuition. Better reduce the noise performance of the enhanced corpus. The second data enhancement method, TER, first analyzes the entity distribution and global performance of the data set, and then selectively selects the weak entities for enhancement, which minimizes the noise performance after enhancement while retaining the enhancement effect of the corpus, thus ensuring the effectiveness of the enhanced corpus globally.
(2) The fusion of data augmentation methods with advanced pre-trained language models to explore a solution that maximizes the performance of Chinese medical entity recognition in low-resource scenarios. Pre-trained language models such as BERT and its variant RoBERTa have greatly refreshed the performance ceiling of various tasks in natural language processing by relying on their prior knowledge and stronger language representation ability, and have better recognition ability than classical LSTM algorithms in the field of named entity recognition. In this paper, a series of comparative experiments are designed to explore and verify the representation ability of different combination models for Chinese electronic medical record text, and then the combination model is used to access the fusion corpus enhanced by the two data enhancement methods proposed in this paper, to verify the effectiveness of the enhancement methods and explore solutions to maximize the performance of Chinese medical entity recognition in the context of small samples.

Related work

NER task

The concept of NER was first proposed in the Message Understanding Conference-6 [8]. The task of NER is to identify proper nouns of pre-defined categories from selected text. NER is a subtask of information extraction and can be classified as text boundary detection or text classification [9]. Early approaches to Chinese NER primarily relied on rule-based and statistical methods. These methods involved the extraction and classification [10] of entities through techniques such as out-of-vocabulary word identification, manual annotation, and character encoding [2]. With the advent of machine learning, a common practice has been to partially annotate data in advance and then use machine learning models for feature extraction and classification [11–16]. However, the ability of machine learning models to extract features from corpora is still limited. In contrast, pre-trained deep learning models can capture more semantics and significantly enhance text representation capabilities, making them the mainstream approach for Chinese NER at present.

In the task of Chinese NER, Chinese word segmentation is a necessary preprocessing step. Since both word segmentation and NER share the commonality of boundary recognition, the idea of information sharing between these tasks lays the foundation for Chinese NER models. The Conditional Random Field (CRF) model was the first to simultaneously perform word segmentation and NER [17]. However, the continuity of Chinese text presents challenges such as text embedding methods [18], word length, and dependency relations that need further resolution.

With the development of neural networks and deep learning technologies, methods that combine Long Short-Term Memory (LSTM), Bidirectional LSTM

(BiLSTM), self-attention mechanisms, and CRF have achieved better results [19–22]. Specifically, LSTM is used to capture character features [19], while BiLSTM, being bidirectional, can account for both forward and backward sequences, providing long-distance dependencies. The self-attention mechanism can capture global dependencies [21]. Consequently, fully representing text features has become a key direction of exploration in natural language processing tasks, including NER. Models like BERT [23] and ERNIE [24] have significant advantages in this area, making the use of BERT or ERNIE as the text representation layer to extract linguistic features a fundamental method for NER models [25, 26].

Additionally, some supplementary methods considering the characteristics of Chinese also show promising optimization effects. These include features such as Chinese pronunciation [27], helpful word filtering [28], and word boundary weighting [29].

Overall, the task of Chinese NER has reached a high level of performance. However, there remains significant room for improvement in certain specialized domains, such as the Chinese medical field, which poses considerable challenges. The unique aspect of NER in the Chinese medical domain is that, while the general domain NER mainly distinguishes between person names, place names, and organizations, Chinese medical NER focuses on identifying diseases, drugs, human tissues, treatment methods, and examination items.

Moreover, the professional nature of medical texts [30] and the presence of obscure characters in the Chinese medical field result in a scarcity of available annotated data. Consequently, unsupervised algorithms, such as Semi-Supervised Support Vector Machines (SSVMs) [31]and Deep Neural Networks (DNN) [32], are initially considered for Chinese medical NER tasks. However, as technology advances and more annotated data becomes available over time, general domain NER algorithms can also be experimented with in the Chinese medical field [33–36]. Nevertheless, the recognition performance in the medical domain does not match that of the general domain, indicating that specialized methods tailored to Chinese medical corpora [37] are still worthy of further research.

Data augmentation

High-performance deep learning relies on a large amount of high-quality training corpus and computational resources, but often these two resources are difficult and costly to obtain. Therefore, in addition to continuous improvement and tuning of models, data augmentation methods, as one of the solutions to this dilemma, have been increasingly added to research by scholars in recent years [38–42].

Data augmentation is a strategy to enrich the diversity of the training set by increasing the number of training corpora without specifically collecting new training data [43]. In image, speech, and video object recognition, data augmentation is a typical processing technology, that can realize the diversity of sample features by clipping, flipping, rotating, scaling, and other methods [44–47]. In contrast, data augmentation for text has significant limitations. Unlike image data, text data cannot be augmented by methods such as clipping, scaling, or rotating to increase the sample size. Furthermore, improper text data augmentation can greatly impact the performance of downstream tasks [48]. However, in cases where the dataset is imbalanced and contains many rare words, data augmentation techniques can effectively increase text features and provide support for downstream tasks [49]. In text data augmentation, commonly used techniques include synonym replacement, random insertion, deletion, and swapping of text [50]. Among them, the most used technique is synonym replacement, which can be varied into different forms such as replacement with words of the same part of speech or category, depending on the context [51–54]. For unsupervised data, using a supervised learning-based data augmentation method showed higher performance in the text classification training set [55]. In some languages that rely on case tagging (e.g. Russian), Şahin et al. [56] proposed a crop-and-cut class of methods used in image augmentation, and they constructed a dependency tree approach that verified the effectiveness in both sequence tagging and speech tagging tasks.

The above are rule-based data augmentation methods, and there are also language model-based methods that can be used for text data augmentation, such as the DiPS method [57], G-DAUG model [58], DAGA model [59] and hybrid data augmentation method [60]. They have been validated as effective in the task of text classification and information extraction.

Although language model-based data augmentation methods theoretically can better extract contextual information and generate more "reasonable" augmented data, in practice, different training data and tasks may require different augmentation policies [61]. In cases where the dataset is imbalanced and contains many rare words, the effectiveness of language model-based data augmentation methods is still not convincing, as evidenced by the unrecognized classes with fewer instances in Wang et al. 's experiment [51]. At the same time, the results of Ding et al. 's experiment did not show absolute advantages compared to rule-based methods [59]. On the other hand, the easy data augmentation (EDA) method [50] has been widely used due to its simplicity and effectiveness, and language model-based data augmentation methods

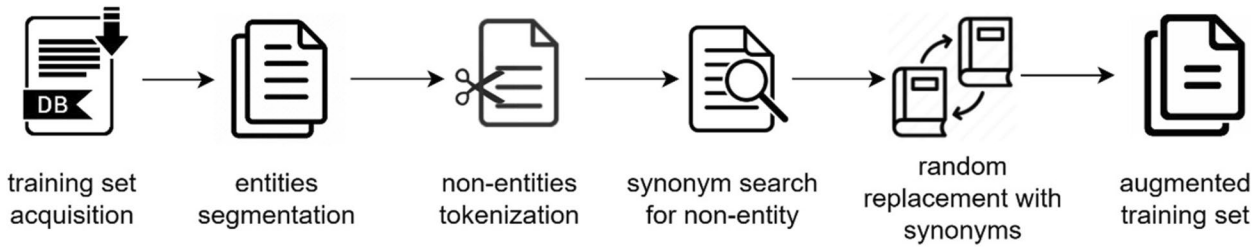

Fig. 1 The CRR method

often require additional computing resources and time since they utilize deep learning models for generation, which contradicts the low-resource and high-efficiency idea of data augmentation [62]. Therefore, we chose rule-based data augmentation methods for Chinese electronic medical records in this paper.

Methodology

Due to the specific characteristics of medical NER in Chinese electronic medical records [63, 64], not all methods can be transferred and improved for this task. NER is essentially a sequence tagging task and differs from text classification in its sensitivity to contextual features. Therefore, this paper improves upon the "random word swap" and "synonym replacement" techniques from the EDA methods proposed by Wei et al. [50], and proposes two new methods suitable for medical NER in Chinese electronic medical records: "Contextual Random Replacement Augmentation based on Word2Vec (CRR)" and "Targeted Entity Random Replacement Augmentation (TER)".

CRR

The CRR method proposed in this paper is the synonym replacement method in EDA. The aim is to minimize the semantic changes caused by data augmentation on the original corpus while consuming minimal resources. For instance, if we take the electronic medical record fragment "患者今日注射葡萄糖(The patient received a glucose injection today)" and transform it into "**病**人今日注射葡萄糖(The person received a glucose injection today)" through synonym replacement, the resulting augmented corpus can be considered as ideal because it does not alter the relationship between the context words. Such generated corpus can increase knowledge and improve the model's generalization ability. However, effective augmentation depends on two prerequisites: accurate and comprehensive synonym replacement and replacement that does not affect entity words.

To address these two prerequisites, this paper first introduced the large-scale Chinese Word2Vec word embeddings from Tencent [65], which covers almost all the vocabulary in the training corpus. Then, the original training data was preprocessed by segmenting the entity part and non-entity part to achieve replacement of non-entity context without affecting entities. The CRR method is shown in Fig. 1.

As shown in the figure, CRR for Chinese electronic medical record corpus mainly includes the following three steps:

Firstly, Read and segment the annotated corpus. To replace contextual information around entities in NER, it is necessary to first separate the entities and the non-entity parts in the sentence. The tagged corpus obtained after preprocessing is stored in pairs in the form of 'character label'. The original text is labeled by characters, but in Chinese, semantics are based on words. Therefore, characters need to be combined into sentences, which are then segmented into words for synonym replacement. The replaced text is then re-labeled by characters. For example, the tagged corpus obtained after preprocessing is: '患 O\n者 O\n**感** O\n**到** O\n疼 O\n痛 O\n, O\n**经** O\n查 O\n**明** O\n腹 B_disease and diagnosis \n腔 I_disease and diagnosis \n出 I_disease and diagnosis \n**血** I_disease and diagnosis\n'. We process it as a complete sentence denoted by '患者**感到**疼痛，**经查明**腹腔出**血**(The patient feels pain and it is found that there is intra-abdominal bleeding)'. After segmentation, the sentence is segmented as '患者/**感到**/疼痛/, /**经**/查**明**/腹腔出**血**'.

Secondly, we used large-scale pre-trained word embeddings for paraphrasing non-entity sentences. We stored the non-entity words from a sentence and segmented them into a list variable. All non-entity words in this variable were then replaced with synonyms based on a specified probability. Each replaced word was subsequently reinserted into its corresponding position in the sentence. We utilized the gensim tool to load Tencent's extensive Chinese word embeddings. Following vector loading, we performed vector normalization to expedite similarity queries. Subsequently, for each non-entity word in a sentence, we compared it against the vocabulary in the word vector library to retrieve the top 10 nearest synonymous words based on cosine similarity with the original word. After that, we ignored punctuation,

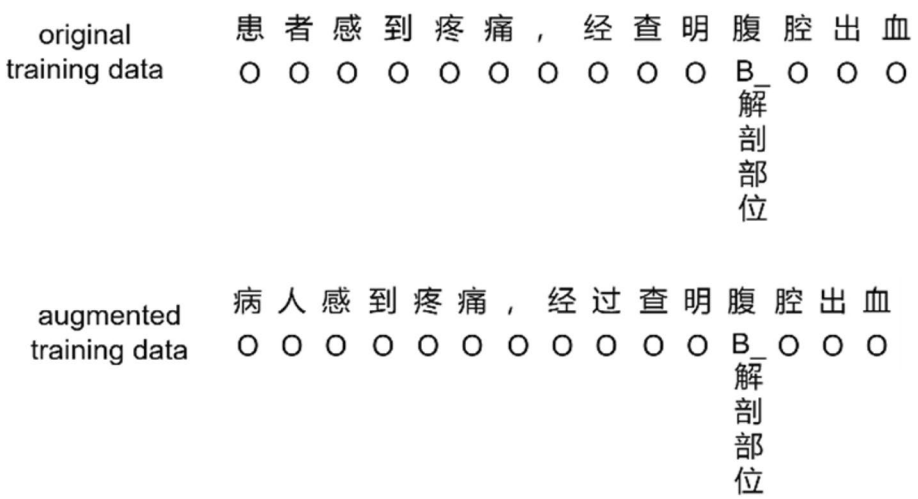

Fig. 2 The comparison between the original and the augmented training data

letters, and numbers, and replaced the original words with synonymous words with a certain probability. Finally, we obtained an augmented corpus. For example, the sentence '患者**感到**疼痛, **经**查**明**腹腔出**血** (The patient feels pain, and it is found that there is bleeding in the abdomen)' can be replaced with '**病人感到**疼痛, **经过**查**明**腹腔出**血** (The patient feels pain, and after investigation, there is bleeding in the abdomen).' It can be seen that in this sentence, '患者 (patient)' and '**经** (after)' are replaced by '**病人** (patient)' and '**经过** (after)'.

Thirdly, the replaced corpus is re-annotated and combined with the original corpus. Since the length of the text may change after replacement, the original Begin Inside Outside (BIO) tags may shift. Therefore, the replaced text needs to have all tags removed and then re-tagged according to the named entity dictionary. The augmented corpus is generated as a model-readable corpus with BIO tags and combined with the original annotated corpus to create a new and merged corpus. Specifically, the comparison between the original training data and the augmented training data is shown in Fig. 2.

It can be seen, that the augmented corpus maintains the same tagging status as the original training corpus, with the entity parts remaining unchanged. The corpus generated by this method retains the basic semantic consistency with the original corpus even after replacing some non-entity words, which meets the expected goal of the CRR data augmentation method.

TER

The TER method is inspired by the random swapping method in EDA. In text classification tasks, the reason that random word swapping can enhance performance is the overall characteristics of the sentence are not greatly altered. In Chinese electronic medical record NER tasks, if the replaced words are limited to entities of the same category, the resulting augmented sentences are semantically coherent to some extent. For example, in the Chinese electronic medical record corpus, the sentence "**术**后常规**病理**示: (直肠) **中分化**腺**癌**(Postoperative routine pathology showed that (rectum) moderately differentiated adenocarcinoma)" contains two entities, "直肠(rectum)" and "腺**癌**(adenocarcinoma)", which belong to the anatomical site and disease/diagnosis entity categories respectively. If we perform entities of the same category random replacement on this sentence, a possible augmented sentence could be "**术**后常规**病理**示: (手臂) **中分化**脑**癌**(Postoperative routine pathology showed: (arm) moderately differentiated brain tumor)". It can be seen that the two entities "直肠(rectum)" and "腺**癌**(adenocarcinoma)" are replaced with entities of the same category "手臂(arm)" and "脑**癌**(brain tumor)", generating a semantically coherent sentence without altering the contextual features.

As shown in the example above, although sentences generated by randomly replacing entities of the same category retain contextual features, they do not conform to real-world logic. In the NER task, due to the characteristics and size of the experimental data, actual NER models have a relatively lower tolerance for noise, and excessive use of such illogical augmented data may even decrease the recognition performance of the NER model. Therefore, we proposed "targeted random entity replacement" to balance the dataset and selectively replace entities in the corpus. The replacement steps are shown in Fig. 3.

Similarly, the steps of the TER method are below:

At first, extract all named entities from the original training corpus and store them in a list grouped by category.

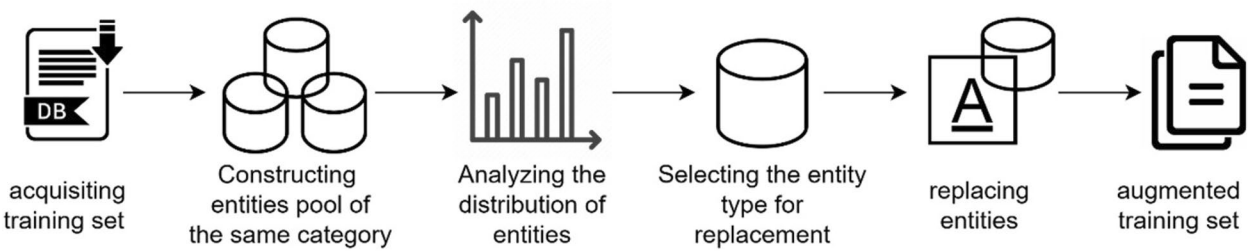

Fig. 3 The TER method

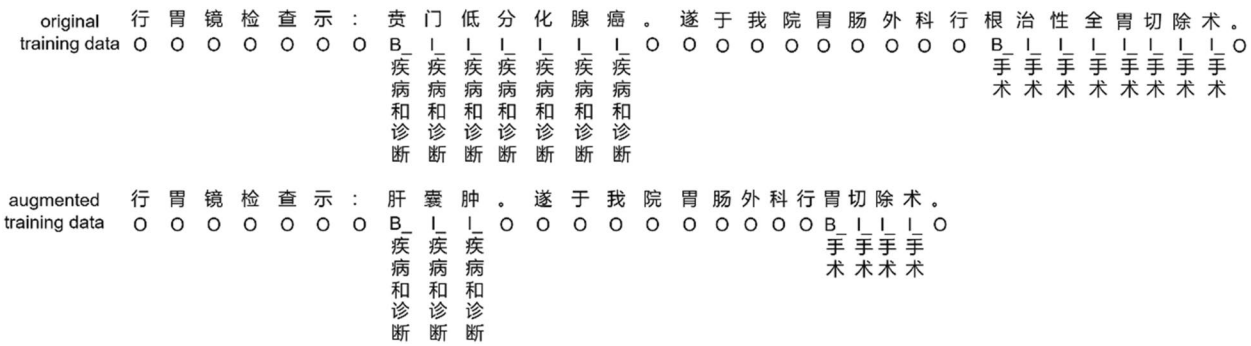

Fig. 4 The sentence of entities selected for random replacement

For example, the list storing diseases and diagnosis categories would be ["**胃**(gastric) B-disease&diagnosis, **癌**(cancer) I-disease&diagnosis", "(liver) B-disease&diagnosis, **癌**(tumor) I-disease&diagnosis", ...], and so on for other entity categories.

Then, the number of entities of different categories in the entity pool is analyzed to select the category of augmented entities based on their quantity distribution.

Finally, random entity replacement is performed for the selected entity category. After a named entity is selected as a target for augmentation, it will be randomly replaced with another one that belongs to the same entity category in the corpus.

It should be noted that there are two possible situations for the augmented training set: ①the entire sentence is without named entities; ②the entire sentence contains named entities, but none of them are selected for random replacement, which means the sentence remains unchanged after augmentation using TER method.

For the first case, we removed all sentences consisting solely of non-entity parts from the augmented corpus to improve training speed and save training resources. For the second case, sentences containing named entities that have not been altered are retained because they help reduce the noise impact of the generated corpus on the NER model. Figure 4 shows the sentences of entities selected for random replacement.

From Fig. 4, after selecting the "Disease & Diagnosis" and "Surgery" entities as the chosen augmented entity categories, the two entities "贲门**低分化**腺**癌**(esophageal or gastric cardiac carcinoma)" and "**根**治**性**全**胃**切除**术**(radical total gastrectomy)" in the original corpus were randomly replaced with two named entities of the same category: "**肝囊肿**(liver cyst)" and "**胃**切除**术**(gastrectomy)," respectively.

NER based on data augmentation

According to the two data augmentation methods mentioned above and the relevant techniques of NER, we constructed a set of NER models based on data augmentation. The framework is designed as shown in Fig. 5.

In data preprocessing, the goal is to transform the original Chinese electronic medical record corpus into training data that can be recognized by the model. Preprocessing steps include data cleaning and correction, entity tagging, and data formatting and sorting. Data cleaning and correction mainly focus on checking the consistency of the content in the original data, handling invalid and missing values that may exist in the data, and ensuring the accuracy of the boundary of the corresponding tagged entities. Entity tagging involves annotating the entities in the text according to certain tagging rules based on the pre-defined entity information in the original dataset. Data formatting and sorting include

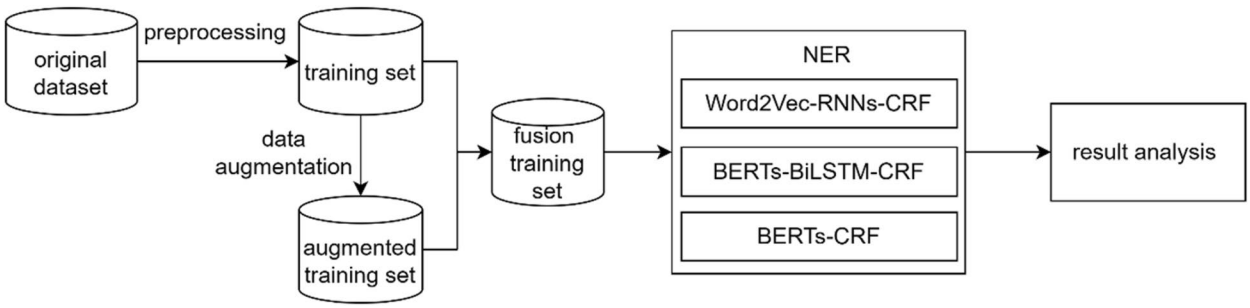

Fig. 5 NER models based on data augmentation

further formatting and aligning according to the standard requirements for reading data by the model.

In data augmentation, for the CRR augmentation method, it is necessary to segment the entities and non-entity text in the corpus, select appropriate replacement methods, and determine the replacement probability.

For the TER method, the category of augmented entities is determined based on the number and properties of each entity category. We initially selected two NER models to conduct full entities random replacement experiments for all categories of entities, without considering the entity categories and quantities. For example, the electronic medical record segment "经CT检查发现胃癌 (discovered gastric cancer through CT scan)" contains the named entities "CT检查 (CT scan)" as an imaging examination entity and "胃癌 (gastric cancer)" as a disease/diagnosis entity. After entities random replacement, this electronic medical record segment may become "经X光检查发现鼻咽癌 (discovered nasopharyngeal carcinoma through X-ray examination)", where "CT检查 (CT scan: imaging examination entity)" and "胃癌 (gastric cancer: disease/diagnosis entity)" are replaced with "X光检查(X-ray examination: imaging examination entity)" and "鼻咽癌(nasopharyngeal carcinoma: disease/diagnosis entity)" respectively. Subsequently, these two NER models were also applied to the unaugmented samples for experimentation. By comparing the results of these two experiments, we can determine how to select the augmented entity categories and replacement ratios. The results of this experiment are presented in Table 4 below.

In NER, we designed five different combination models based on three types of models to train and test on the original corpus, which served as the baseline models. Then, we conducted comparative experiments on augmented corpora using the five combination models to verify the effectiveness of the two data augmentation methods proposed in this paper. The Word2Vec model has two structures, Skip-gram and CBOW, which can better capture the relationships between words in the corpus. For example, in the corpus, the words 'patient' and 'hospital' frequently appear close to each other. Therefore, the semantic vectors generated by Word2Vec will have a closer semantic distance between these two words, achieving a better semantic representation of the text and effectively solving the problem of dimensional sparsity. Additionally, the distributed representation of Word2Vec addresses the high-dimensionality issue found in traditional semantic representation methods.

The BERT model uses bidirectional encoding Transformers and is pre-trained on a large amount of everyday corpus from sources like Wikipedia and news articles. This pre-training provides BERT with 'prior knowledge' of the target language corpus, enabling it to outperform traditional neural networks in classification and sequence labeling tasks. The RoBERTa model is a variant of the BERT model. It uses a dynamic masking mechanism instead of BERT's original static masking mechanism and removes the 'next sentence prediction' task. Compared to the BERT model, the RoBERTa model has better semantic representation and learning capabilities, resulting in improved training outcomes.

Experiment

Dataset

The dataset used in the experiment is the public electronic medical record NER dataset for the subtask "Medical Entity Extraction for Chinese Electronic Medical Records" in CCKS2019.[1] The original dataset was provided by Yidu Cloud,[2] and it contains 1000 training sets and 379 testing sets, each composed of a real medical record text. Six major categories of medical entities were manually tagged in the dataset, including "Disease and Diagnosis(疾病和诊断)", "Anatomy(解剖部位)", "Lab Test(实验室检验)", "Imaging Exam(影像检查)", "Surgery(手术)" and "Medication(药物)".

In the original dataset, the category and position information of electronic medical record text and entities are

[1] Please visit: https://www.biendata.xyz/competition/ccks_2019_1/

[2] The dataset is available at http://openkg.cn/dataset/yidu-s4k

Table 1 Number of characters and sentences

	Training set	Validation set	Testing set
Number of characters	338,197	42,126	134,728
Number of sentences	6155	783	2480

recorded in a dictionary. However, such a storage format cannot be directly read by the model. Therefore, it is necessary to convert the original dataset according to the experimental requirements annotate the electronic medical record text sequence through appropriate tagging methods, and finally output it in a format that can be read by the model. In this paper, the original dataset was annotated and processed using the BIO tagging method.

After data preprocessing, we re-divided the original 1000 training sets into training sets and validation sets at an 8:2 ratio, while 379 testing set were still used as testing set. The specific information of the training set, validation set, and testing set are shown in Table 1 and Fig. 6.

Experimental environment and evaluation criteria

In this paper, we used Python 3.7.0 as the programming language for the experiments, with TensorFlow 2.4.0 and Keras 2.4.3 as the deep learning framework. All experiments were conducted on an Ubuntu 18.04.6 LTS (GNU/Linux 4.15.0–175-generic x86_64) system with an Intel(R) Xeon(R) Gold 6139 M CPU and a NVIDIA Tesla V100 GPU with 32 GB of VRAM.

In this paper, we evaluated the model's recognition performance using three metrics: precision, recall, and F1 score. For the NER task, these evaluation criteria can be divided into two types of standards in different studies: strict and relaxed standards. The strict standard requires that the predicted entity's position and category are consistent with the original entity to be considered correct, while the relaxed standard only requires that the predicted entity category is correct and has overlap with the correct entity position. In this paper, we believe that in the medical application scenario, both the tag prediction result and the boundary prediction result of the entity should be as accurate as possible. Therefore, we adopt the strict standard to evaluate the model's performance. Additionally, we should consider both the model's recognition accuracy and completeness. Therefore, we use the F1 score as the main evaluation criterion of the model, and precision and recall as auxiliary evaluation criteria.

NER model and parameter settings

Based on previous studies, we selected the traditional Word2Vec word embedding model and BERT-like pretraining model for text feature extraction [66], and CRF model for text classification [67], and considered adding BiLSTM model to improve performance [68, 69]. Five combination models were obtained, as shown in Table 2.

In the Word2Vec-BiLSTM-CRF model, the Word2Vec layer is used to generate text feature vectors that are input into the BiLSTM layer. The BiLSTM layer uses bidirectional propagation to construct an LSTM network, which helps the LSTM model effectively utilize contextual features within a specific sentence range. Considering the issue of rare characters in Chinese electronic medical records, the parameter of hs in the Word2Vec model is set to 1, and Hierarchical Softmax is used for training. The min_count parameter is set to 1, meaning all words appearing in the corpus are retained. The window parameter is set to 4, and the vector_size parameter is consistent with the 768-dimensional word vectors used in the BERT model. The hyperparameters for Word2Vec-BiLSTM-CRF include epoch setting of 10, batch_size setting of 4, learning_rate of 1e-3, and Adam [70] optimization, with a dropout rate of 0.5.

Similarly, in the BERT-CRF, BERT-BiLSTM-CRF, RoBERTa-wwm-ext-CRF, and RoBERTa-wwm-ext-BiLSTM-CRF models, all BERT model parameters use the default settings provided by the official. Both BERT and RoBERTa

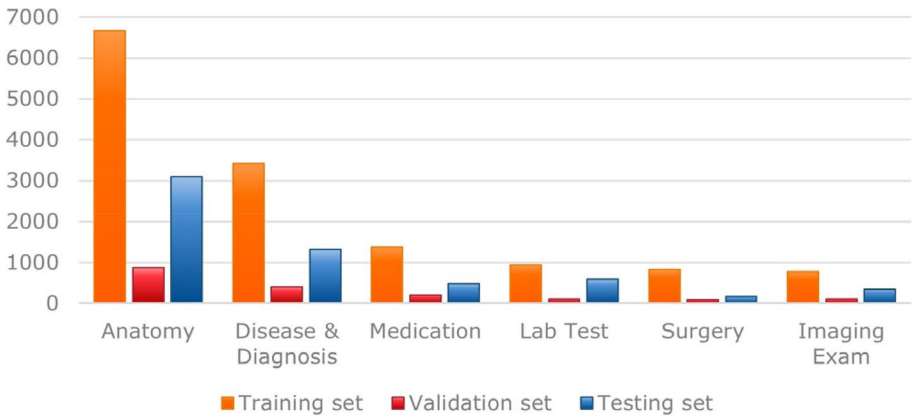

Fig. 6 Distribution of entity categories

Table 2 NER model

Models	Model Features
Word2Vec-BiLSTM-CRF	traditional word embedding + contextual features extraction + text classification
BERT-CRF	BERT pre-training + text classification
BERT-BiLSTM-CRF	BERT pre-training + contextual features extraction + text classification
RoBERTa-wwm-ext-CRF	BERT optimized pre-training + text classification
RoBERTa-wwm-ext-BiLSTM-CRF	BERT optimized pre-training + contextual features extraction + text classification

Table 3 The number of samples in CRR fusion corpus

Augmented fusion corpus type	Number
characters	691,176
sentences	9907
replaced word	18,456
non-replaced word	76,289

Table 4 The experiment results of full random entity replacement augmentation

Model	Precision(%)	Recall(%)	F1(%)
Word2Vec-BiLSTM-CRF	**80.00**	**77.91**	**78.94**
Word2Vec-BiLSTM-CRF (full entities random replacement)	79.44	77.39	78.40
BERT-CRF	**83.13**	**81.96**	**82.54**
BERT-CRF(full entities random replacement)	82.08	81.91	81.99

models use 12 encoder layers stacked, with a hidden_state dimension of 768 and 12 hidden layers. Additionally, in the BERT-BiLSTM-CRF and RoBERTa-wwm-ext-BiLSTM-CRF models, the number of unidirectional hidden layers for the RNN is 32. The hyperparameters for the four combination models, except for epoch set to 3 and optimizer selected as AdaFactor, are consistent with those of the Word2Vec-BiLSTM-CRF model.

All five combination models utilize early stopping to prevent overfitting. The early stopping monitoring rule is based on the validation loss (val_loss), and if the model does not improve within 5 iterations, training is stopped, and the best-performing model is retained.

Data augmentation and parameter settings

CRR

In this experiment, the CRR method used the Tencent word vector is open source tencent-ailab-embedding-zh-d100-v0.2.0-s.tar for pre-trained Chinese word vectors. To maintain the meaning of the augmented sentences as much as possible during synonym replacement, according to the results of reference [50], we set the replacement probability for non-entity words to 20%. The augmented corpus and the original corpus combined to form an augmented fusion corpus. The number of samples in CRR fusion corpus is in Table 3.

TER

According to the methods described earlier, we first conducted a comparative experiment between "full entities random replacement augmentation" and the unaugmented text. This experiment selected the Word2Vec-BiLSTM-CRF and BERT-CRF combination models for testing, and the experimental results are shown in Table 4.

From Table 4, it can be seen that either the traditional deep learning model or the large-scale pre-trained language model, the method of full entities random replacement augmentation in the Chinese electronic medical record NER task brought negative benefits. This is because the generated augmented corpus after entities' random replacement on the electronic medical record corpus often does not conform to actual semantics and logic. Theoretically, the augmented corpus can be regarded as a "reasonable noise", which can improve the stability and robustness of the NER model if properly added. However, from the experimental results, the noise introduced by the method of full entities replacement augmentation clearly exceeds the reasonable range. Therefore, the TER method designed in this paper needs to analyze the dataset in advance and control the "noise" of the augmented corpus within a reasonable range.

According to the distribution of entity categories in the training set represented by the blue bars in Fig. 6, the number of named entities in the "Anatomy" category is much higher than in the other five categories. Therefore, we applied the TER method to randomly replace the entities of the five categories that are relatively weak in quantity (Disease and Diagnosis, Lab Test, Imaging Exam, Surgery, and Medication), and the replacement probability for each named entity remains at 1. The parameters of the augmented merged corpus are shown in Table 5.

Table 5 The number of samples in TER fusion corpus

Augmented fusion corpus type	Number
characters	625,314
sentences	9345
replaced entities	7330
deleting non-entity sentences	562

Table 6 Experimental Results

Models	Precision(%)	Recall(%)	F1(%)
①Word2Vec-BiLSTM-CRF	80.000	77.907	78.940
②Word2Vec-BiLSTM-CRF (CRR)	80.744	77.757	79.223 **(+0.28)**
③Word2Vec-BiLSTM-CRF (TER)	80.120	78.091	79.092 **(+0.15)**
④BERT-CRF	83.133	81.956	82.540
⑤BERT-CRF(CRR)	82.474	83.422	82.945 **(+0.41)**
⑥BERT-CRF (TER)	82.640	83.039	82.839 **(+0.30)**
⑦BERT-BiLSTM-CRF	80.338	83.939	82.099
⑧BERT-BiLSTM-CRF (CRR)	82.122	85.105	83.587 **(+1.49)**
⑨BERT-BiLSTM-CRF (TER)	83.328	82.606	82.965 **(+0.87)**
⑩RoBERTa-wwm-ext-CRF	82.042	83.122	82.579
⑪RoBERTa-wwm-ext-CRF (CRR)	82.323	83.256	82.787 **(+0.21)**
⑫RoBERTa-wwm-ext-CRF (TER)	81.734	83.722	82.716 **(+0.14)**
⑬RoBERTa-wwm-ext-BiLSTM-CRF	82.788	83.505	83.145
⑭RoBERTa-wwm-ext-BiLSTM-CRF (CRR)	82.756	83.639	83.195 **(+0.05)**
⑮RoBERTa-wwm-ext-BiLSTM-CRF (TER)	83.276	84.872	84.066 **(+0.92)**

Experimental results and analysis

Following the experimental design above, the experimental results are shown in Table 6. In Table 6, models without parentheses indicate that the model used raw electronic medical record data for NER. Models with the suffix "(CRR)" indicate that the model used fusion data with CRR. Models with the suffix "(TER)" indicate that the model used fusion data with TER. All results are shown as percentages and rounded to three decimal places. Based on the experimental results in the table, the following conclusions can be drawn:

(1) The two data augmentation methods proposed in this paper, "Contextual Random Replacement Based on Word2Vec (CRR)" and "Targeted Entities Random Replacement (TER)," can generally improve the ability of Chinese electronic medical record medical NER models. According to the F1 value results, the recognition performance of all five combined models improved after using fusion data. These five models include both the Word2Vec-BiLSTM-CRF model that uses traditional word embedding deep learning methods and BERT-like combined models that use pre-trained language models. This indicates that the two data augmentation methods have a promoting effect on different NER models and are generalizable.

(2) In most cases, the data augmentation method of "CRR" provides better improvement for the combined models than the "TER" method. Except for the combined model ⑭RoBERTa-wwm-ext-BiLSTM-CRF, the F1 score of the other four combined models using fusion data with CRR is better than that using TER method. This indicates that the CRR method not only enhances Chinese electronic medical record data but also preserves the original semantics better, providing better gains for the model.

(3) Adding BiLSTM layer to the pre-trained model can better improve the model's understanding of fusion data. According to the F1 value results, BERT-like models show better recognition ability for augmented data after adding the BiLSTM layer. For example, after using fusion data with CRR, the recognition performance of the combined model BERT-BiLSTM-CRF improved by up to 1.49%. This also indirectly indicates that combining BERT with BiLSTM can better capture textual features in larger-scale language data.

Table 7 Comparison of our data augmentation method with other data augmentation methods

Models	Original	EDA	MDA	CRR	TER
BERT + CRF	80.56	81.34	83.29	82.95	82.84
BiLSTM + CRF	81.00	82.04	83.45	83.39*	83.52*

*Since we do not include experiments with the BiLSTM + CRF model but use the BERT-BiLSTM-CRF and RoBERTa-wwm-ext-BiLSTM-CRF models, and since BERT and RoBERTa belong to the same family of models, the results of these two experiments are averaged to substitute for the BiLSTM + CRF model results

Based on the results from the existing literature, reference [50] compared the EDA and MDA augmentation methods using the same dataset (CCKS2019) and downstream tasks as in this paper. Therefore, the method proposed in this paper can be compared with the EDA and MDA methods. The comparison results are shown in Table 7.

As shown in Table 7, both our methods proposed outperform the EDA method and are comparable to the MDA method proposed in reference [50]. In the BERT + CRF model, the MDA method performs slightly better than our methods proposed. However, in the BiLSTM + CRF model, our methods proposed perform slightly better than the MDA method. Additionally, our methods proposed do not require additional data and computational resources, whereas the MDA method in reference [50] requires the collection of additional data resources and computational power to support the construction of a knowledge graph. Therefore, in terms of ease of use, our methods proposed have an advantage.

Conclusion

Data augmentation is essential for improving medical named entity recognition (NER) due to high data acquisition costs, specialized terminology, imbalanced data, and limited training resources. To address the lack of high-quality medical record corpora for training, we proposed two methods: Contextual Random Replacement (CRR) and Targeted Entity Random Replacement (TER). These methods balance data distribution, enrich training datasets, and improve model performance. Our methods showed significant improvements in Chinese medical NER. CRR was effective with synonym replacement, and TER highlighted the importance of selecting appropriate entity categories and augmentation data quantity.

Future work will focus on exploring additional methods like incorporating knowledge graphs, transfer learning, and graph representation learning [71, 72] to improve the prediction ability of the models. We also plan to investigate the integration of other advanced augmentation techniques and evaluate their impact on diverse medical datasets to further enhance medical NER performance.

Authors' contributions
Author Contributions: Conceptualization, Y.L. and J.Z.; methodology, L.D. and H.C.; software, H.C.; validation, J.Z., L.D., H.C. and M.C.; formal analysis, Y.L.; writing—original draft preparation, H.C.; writing—review and editing, H.C. and L.D.; visualization, L.D.; resources and supervision, Y.L.; project administration, Y.L. and J.Z.; funding acquisition, J.Z. All authors have read and agreed to the published version of the manuscript. L.D. and H.C. contribute equally to the article.

Funding
This research was funded by Guangzhou Science and Technology Planning Project, grant number 202002020036. The funding body played no role in the design of the study and collection, analysis, interpretation of data, and in writing the manuscript.

Availability of data and materials
The data that support the findings of this study are available at http://openkg.cn/dataset/yidu-s4k, named yidu-s4k.zip. The dataset was uploaded by Yidu Cloud and is stated to be used for scientific research on NLP.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 26 May 2023 Accepted: 29 July 2024
Published online: 05 August 2024

References

1. Chieu HL, Ng HT. Named Entity Recognition: a maximum entropy approach using global information. In: Proceedings of the 19th international conference on computational linguistics. Morristown: Association for Computational Linguistics; 2002.
2. Levow GA. The third international chinese language processing bakeoff: word segmentation and named entity recognition. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. Sydney: Association for Computational Linguistics; 2006. p. 108–17.
3. Makridakis S, Spiliotis E, Assimakopoulos V. Statistical and machine learning forecasting methods: concerns and ways forward. PLoS One. 2018;13(3):e0194889.
4. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. Commun ACM. 2017;60(6):84–90.
5. Yadav V, Bethard S. A survey on recent advances in named entity recognition from deep learning models. arXiv [cs.CL]. 2019. 1910.11470.
6. Jia C, Shi Y, Yang Q, Zhang Y. Entity Enhanced BERT Pre-Training for Chinese NER. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: Association for Computational Linguistics; 2020.
7. Ji B, Li S, Yu J, Ma J, Tang J, Wu Q, Tan Y, Liu H, Ji Y. Research on Chinese Medical Named Entity Recognition Based on Collaborative Cooperation of Multiple Neural Network Models. J Biomed Inform. 2020;104:103395.
8. Grishman R, Sundheim B. Message Understanding Conference-6: a brief history. In: Proceedings of the 16th conference on Computational linguistics. Morristown: Association for Computational Linguistics; 1996.
9. Goyal A, Gupta V, Kumar M. Recent named entity recognition and classification techniques: a systematic review. Comput Sci Rev. 2018;29:21–43.
10. Parlak B, Uysal AK. On classification of abstracts obtained from medical journals. J Inf Sci. 2019;46(5):648–63.
11. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural Language Processing (Almost) from Scratch. J Mach Learn Res. 2011;12:2493–537.
12. Liu Z, Zhu C, Zhao T. Chinese Named Entity Recognition with a Sequence Labeling Approach: Based on Characters, or Based on Words? In: Proceedings of the Advanced intelligent computing theories and applications, and 6th international conference on Intelligent computing. Berlin, Heidelberg: Association for Springer-Verlag; 2010. p. 634–40.
13. Li H, Hagiwara M, Li Q, Ji H. Comparison of the Impact of Word Segmentation on Name Tagging for Chinese and Japanese. In: Proceedings of International Conference on Language Resources and Evaluation. Linguistics: Association for Computer Science; 2014. p. 2532–6.
14. Parlak B, Uysal AK. On feature weighting and selection for medical document classification. In: Studies in computational intelligence. 2017. p. 269–82.
15. Parlak B. A novel feature ranking algorithm for text classification: Brilliant probabilistic feature selector (BPFS). Comput Intell. 2023;39(5):900–26.
16. Parlak B, Uysal AK. A novel filter feature selection method for text classification: Extensive Feature Selector. J Inf Sci. 2021;49(1):59–78.
17. He J, Wang H. Chinese Named Entity Recognition and Word Segmentation Based on Character. In: Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing. 2008.

18. Peng N, Dredze M. Named entity recognition for Chinese social media with jointly trained embeddings. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: Association for Computational Linguistics; 2015. p. 548–54.
19. Peng N, Dredze M. 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). Berlin: Association for Computational Linguistics; 2016. p. 149–55.
20. Zhang Y, Yang J. Chinese NER Using Lattice LSTM. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne: Association for Computational Linguistics; 2018. p. 1554–64.
21. Cao P, Chen Y, Liu K, Zhao J, Liu S. Adversarial Transfer learning for Chinese named entity recognition with self-attention mechanism. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics; 2018,. p. 182–92.
22. Jin Y, Xie J, Guo W, Luo C, Wu D, Wang R. LSTM-CRF Neural Network with Gated Self Attention for Chinese NER. IEEE Access. 2019;7:136694–703.
23. Devlin J, Chang MW, Lee K, Toutanova K. 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. Minneapolis: Association for Computational Linguistics; 2019. p. 4171–86.
24. 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. ERNIE 3.0: Large-Scale Knowledge Enhanced Pre-Training for Language Understanding and Generation. CoRR. arXiv [cs.CL], 2021;2107:02137. https://arxiv.org/abs/2107.02137.
25. Wang Y, Sun Y, Ma Z, Gao L, Xu Y. An ERNIE-based joint model for Chinese named entity recognition. Appl Sci. 2020;10(16):5711.
26. Wang Y, Lu L, Yang W, Chen Y. Local or global? A novel transformer for Chinese named entity recognition based on multi-view and sliding attention. Int J Mach Learn Cybern. 2024;15:2199–208.
27. Mai C, Liu J, Qiu M, Luo K, Peng Z, Yuan C, Huang Y. Pronounce Differently, Mean Differently: a multi-tagging-scheme learning method for Chinese NER integrated with lexicon and phonetic features. Inf Process Manage. 2022;59(5):103041.
28. Tian X, Bu X, He L. Multi-task learning with helpful word selection for lexicon-enhanced Chinese NER. Appl Intell. 2023;53(16):19028–43.
29. Guo Y, Feng S, Liu F, Lin W, Liu H, Wang X, Su J, Gao Q. Enhanced Chinese domain named entity recognition: an approach with lexicon boundary and frequency weight features. Appl Sci. 2023;14(1):354.
30. Hu L, Zhang M, Hu P, Zhang J, Niu C, Lu X, Jiang X, Ma Y. Dual-channel hypergraph convolutional network for predicting herb–disease associations. Brief Bioinform. 2024;25(2):bbae067.
31 Tang B, Cao H, Wu Y, Jiang M, Xu H. Recognizing clinical entities in hospital discharge summaries using structural support vector machines with word representation features. BMC Med Inform Decis Making. 2013;13(S1):S1.
32. Wu Y, Jiang M, Lei J, Xu H. Named entity recognition in chinese clinical text using deep neural network. PubMed. 2015;216:624–8.
33. Chalapathy R, Borzeshi ZE, Piccardi M. Bidirectional LSTM-CRF for Clinical Concept Extraction. In: Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP). Osaka: Association for The COLING 2016 Organizing Committee; 2016. p. 7–12.
34. Ravikumar J, Kumar PR. Machine learning model for clinical named entity recognition. Int J Power Electron Drive Syst Int J Electric Comput Eng. 2021;11(2):1689.
35. Xu G, Wang C, He X. Improving Clinical Named Entity Recognition with Global Neural Attention. In: Proceedings of APWeb-WAIM 2018. Macau: Association for Lecture Notes in Computer Science; 2018. p.264–279.
36. Liu K, Hu Q, Liu J, Xing C. Named Entity Recognition in Chinese Electronic Medical Records Based on CRF. In: Proceedings of 2017 14th Web Information Systems and Applications Conference (WISA). Liuzhou: Association for IEEE; 2017. p. 105–10.
37. Zhao B, He Y, Su X, Yang Y, Li G, Huang Y, Hu P, You Z, Hu L. Motif-aware miRNA-disease association prediction via hierarchical attention network. IEEE J Biomed Health Inform. 2024;28(7):4281–94.
38. Croce D, Filice S, Castellucci G, Basili R. Learning to Generate Examples for Semantic Processing Tasks. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle: Association for Computational Linguistics; 2022. p. 4587–601.
39. Kashefi O, Hwa R. Quantifying the Evaluation of Heuristic Methods for Textual Data Augmentation. In Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020); Online: Association for Computational Linguistics; 2020. p. 200–208.
40. Guo B, Han S, Huang H. Selective Text Augmentation with Word Roles for Low-Resource Text Classification. arXiv [cs.CL], 2022;2209:01560. https://arxiv.org/abs/2209.01560.
41. Li Y, Li X, Yang Y, Dong R. A diverse data augmentation strategy for Low-Resource neural machine translation. Information. 2020;11(5):255.
42. Fadaee M, Bisazza A, Monz C. Data augmentation for Low-Resource Neural Machine Translation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vancouver: Association for Computational Linguistics; 2017. p. 567–73.
43. Feng SY, Gangal V, Wei J, Chandar S, Vosoughi S, Mitamura T, Hovy E. A Survey of Data Augmentation Approaches for NLP. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Online: Association for Computational Linguistics; 2021. p. 968–988.
44. Kumar T, Mileo A, Brennan R, Bendechache M. Image data augmentation approaches: a comprehensive survey and future directions. arXiv [cs.CV]. 2023; 2301:02830. https://arxiv.org/abs/2301.02830.
45. Yoo J, Kang S. Class-adaptive data augmentation for image classification. IEEE Access. 2023;11:26393–402.
46. He K, Liu C, Lin S, Hopcroft JE. Local Magnification for Data and Feature Augmentation. arXiv [cs.CV]. 2022;2211:07859. https://arxiv.org/abs/2211.07859.
47. Atmaja BT, Sasou A. Effects of data augmentations on speech emotion recognition. Sensors (Basel). 2022;22(16):5941.
48 Shorten C, Khoshgoftaar TM, Furht B. Text data augmentation for deep learning. J Big Data. 2021;8(1):101.
49. Du J, Grave E, Gunel B, Chaudhary V, Celebi O, Auli M, Stoyanov V, Conneau A. Self-Training Improves Pre-Training for Natural Language Understanding. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics; 2021.
50. Wei J, Zou K. Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. 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). Stroudsburg: Association for Computational Linguistics; 2019.
51. Wang A, Li L, Wu X, Zhu J, Yu S, Chen X, Li J, Zhu H. Entity relation extraction in the medical domain: based on data augmentation. Ann Transl Med. 2022;10(19):1061.
52. Shi B, Zhang L, Huang J, Zheng H, Wan J, Zhang L. MDA: an intelligent medical data augmentation scheme based on medical knowledge graph for chinese medical tasks. Appl Sci (Basel). 2022;12(20):10655.
53. Coulombe C. Text Data Augmentation Made Simple by Leveraging NLP Cloud APIs. arXiv [cs.CL]. 2018;1812:04718. https://arxiv.org/abs/1812.04718.
54. Kobayashi S. Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies , Volume 2 (Short Papers). New Orleans: Association for Computational Linguistics; 2018. p. 452–7.
55. Xie Q, Dai Z, Hovy E, Luong MT, Le QV. Unsupervised Data Augmentation for Consistency Training. arXiv [cs.LG]. 2020;1904:12848. https://arxiv.org/abs/1904.12848.
56. Şahin GG, Steedman M. Data Augmentation via Dependency Tree Morphing for Low-Resource Languages. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics; 2018.
57. Kumar A, Bhattamishra S, Bhandari M, Talukdar P. Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation. In: Proceedings of the 2019 Conference of the North. Stroudsburg: Association for Computational Linguistics; 2019.

58. Yang Y, Malaviya C, Fernandez J, Swayamdipta S, Le Bras R, Wang JP, Bhagavatula C, Choi Y, Downey D. Generative Data Augmentation for Commonsense Reasoning. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg: Association for Computational Linguistics; 2020.
59. Ding B, Liu L, Bing L, Kruengkrai C, Nguyen TH, Joty S, Si L, Miao C. DAGA: Data Augmentation with a Generation Approach for Low-Resource Tagging Tasks. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: Association for Computational Linguistics; 2020.
60. Zhang R, Yu Y, Zhang C. SeqMix: Augmenting Active Sequence Labeling via Sequence Mixup. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: Association for Computational Linguistics; 2020.
61. Li S, Ao X, Pan F, He Q. Learning policy scheduling for text augmentation. Neural Netw. 2022;145:121–7.
62. Wang Z, Wu Y, Liu F, Liu D, Hou L, Yu H, Li J, Ji H. Augmentation with Projection: Towards an Effective and Efficient Data Augmentation Paradigm for Distillation. arXiv [cs.CL]. 2023;2210:11768. https://arxiv.org/abs/2210.11768.
63. Wu Y, Jiang M, Xu J, Zhi D, Xu H. Clinical Named Entity Recognition Using Deep Learning Models. AMIA Annu Symp Proc. 2017;2017:1812-19.
64. Liu P, Guo Y, Wang F, Li G. Chinese named entity recognition: the state of the Art. Neurocomputing. 2022;473:37–53.
65. Song Y, Shi S, Li J, Zhang H. Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). Stroudsburg: Association for Computational Linguistics; 2018.
66. Parlak B, Uysal AK. The impact of feature selection on medical document classification. In: Proceedings of 2016 11th Iberian Conference on Information Systems and Technologies (CISTI). Gran Canaria: IEEE; 2016.
67. Song S, Zhang N, Huang H. Named entity recognition based on conditional random fields. Cluster Comput. 2019;22(S3):5195–206.
68. Chiu JPC, Nichols E. Named entity recognition with bidirectional LSTM-CNNs. Trans Assoc Comput Linguist. 2016;4:357–70.
69. Li L, Jiang Y. Integrating language model and reading control gate in BLSTM-CRF for biomedical named entity recognition. IEEE/ACM Trans Comput Biol Bioinform. 2020;17(3):841–6.
70. Kingma DP, Ba JL. Adam: A method for stochastic optimization. arXiv[cs.LG]. 2017;1412:6980. https://arxiv.org/abs/1412.6980.
71 Zhao B, Su X, Hu P, Huang Y, You Z, Hu L. iGRLDTI: An Improved Graph Representation Learning Method for Predicting Drug-Target Interactions over Heterogeneous Biological Information Network. Bioinformatics. 2023;39(8):btad451.
72. Zhao B, Su X, Hu P, Ma Y, Zhou X, Hu L. A geometric deep learning framework for drug repositioning over heterogeneous information networks. Brief Bioinform. 2022;23(6):bbac384.

Publisher's Note