IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 June 2023, accepted 12 July 2023, date of publication 28 July 2023, date of current version 4 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3299824

RESEARCH ARTICLE

Ensemble Transfer Learning on Augmented Domain Resources for Oncological Named Entity Recognition in Chinese Clinical Records

MEIFENG ZHOU[1], JINDIAN TAN[2], SONG YANG[2], HAIXIA WANG[1], LIN WANG[1], AND ZHIFENG XIAO[3]

[1]Department of Oncology, Hainan General Hospital (Hainan Affiliated Hospital of Hainan Medical University), Haikou, Hainan 570102, China
[2]Department of Orthopaedic Surgery, Hainan General Hospital (Hainan Affiliated Hospital of Hainan Medical University), Haikou, Hainan 570102, China
[3]School of Engineering, Penn State Erie, The Behrend College, Erie, PA 16563, USA

Corresponding authors: Lin Wang (wanglin7209@163.com) and Zhifeng Xiao (zux2@psu.edu)

This work was supported by the Scientific Research Projects of Health and Family Planning Industry in Hainan Province under Grant 20A200415.

ABSTRACT Biomedical Named Entity Recognition (NER) is a crucial task in Natural Language Processing (NLP) and can help mine knowledge from massive clinical and diagnostic records. However, the biomedical NER task often undergoes a low-resource training setting due to the high cost of human annotation, limiting the capability of traditional NER models. In this study, we propose a two-stage learning pipeline to tackle the oncological NER task in Chinese language, which is a typical task lacking training resources. In the first stage, two base models pre-trained by Word to Vector (Word2Vec) and Bidirectional Embeddings Representations from Transformer (BERT) are fine tuned to obtain domains-specific word embeddings that serve as the input for the downstream NER task. In the second stage, we feed the word embeddings into a neural network that consists of a Bidirectional Long and Short Time Memory Recurrent Neural Network (BiLSTM) and Linear-chain Conditional Random Field (CRF) for end task training. Meanwhile, we utilize a substitution-based generative model for data augmentation (DA), aiming to enhance the quantity and diversity of the training data. Experiments show that our proposed learning pipeline demonstrates superior performance compared to other model alternatives under a low-resource setting. Specifically, results show that the proposed fine-tuning strategy, when conducted on an augmented domain resource, can effectively incorporate rich domain knowledge into the final NER model, presenting a great potential in boosting a model's predictive power with limited training data.

INDEX TERMS BERT, cancer, deep learning, named entity recognition, oncology, osteosarcoma, Word2Vec.

I. INTRODUCTION

Conception and knowledge extraction is crucial for the automation of clinical research. To obtain the actionable information and discard the unstructured clinical data [2], researchers have developed a wide spectrum of learning models using natural language processing (NLP) techniques. NLP models depend heavily on supervised machine learning (ML) that requires numerous annotated data [5] for training. To train a computer to better understand a natural language, several dimensions in the real world textual resources, such as dialects, topics, languages, and genres, have been considered and exploited in the literature [31]. On the other hand, the availability of training resources play a significant role in the performance of any ML-based system, especially for a system designed for a specific domain. Depending on the amount of available domain resources, there are three scenarios for training an NLP model, including (i) High- or rich- resource settings with a large quantity of annotated data, (ii) Low-resource or Resource-poor settings

The associate editor coordinating the review of this manuscript and approving it for publication was Ines Domingues.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

with limited annotated data, and (iii) Zero-resource settings with no annotated data [21]. Despite being more challenging, scenarios (ii) and (iii) are more common in the real world, especially for a narrow domain. The lack of annotated domain resources brings difficulty for most downstream NLP tasks, including Named Entity Recognition (NER), which which aims to recognize rigid designators such as Person, Location, and Organizations that belong to predefined semantic categories. Obtaining high-quality annotated data in the biomedical field is difficult since it requires experts with domain knowledge to assign the biomedical entities appearing in the texts into different categories, making crowdsourcing [16], which is a general solution to massive annotation tasks, less likely. Therefore, it is imperative to develop novel methodology to resolve the performance degradation caused by the insufficiency of annotated medical resources.

Recent advances have witnessed the power of transfer learning that has significantly improved the performance of numerous ML tasks [29]. Transfer learning can be either supervised or unsupervised and allows pre-trained models with incorporated domain knowledge to be fine-tuned in a low-resource setting. The fine tuning process can help transfer knowledge from the source to the destination domain. The benefits of transfer learning are twofold, namely reducing the required amount of annotated data and training time. In NLP, a variety of pre-training techniques and models have been investigated. The Word to Vector (Word2Vec) model can convert words in a vocabulary to numerical representations, i.e., word embeddings, while maintaining the words' semantic meanings. Trained by Word2Vec, words with similar meanings will receive similar numeric embeddings, making it easier to calculate the distance between words. Bidirectional Encoder Representations from Transformers (BERT) [11] utilizes Transformer as a core building block and employs two self-supervised pre-training techniques, including masked language model (MLM) and next sentence prediction (NSP). The MLM is a fill-in-the-blank task that aims to predict the most probable word to fill a mask token position given the context. On the other hand, NSP is trained to predict whether a pair of sentences present a before-after relation, which is useful for tasks such as question answering that require an understanding of the semantic relation between two sentences. Combining MLM and NSP, BERT produces a robust language model that can support a wide range of downstream tasks, including NER. Transfer learning offers an path towards domain knowledge incorporation, effectively reducing the required amount of training data and the number of training epochs. Meanwhile, ensemble learning has been widely adopted at the feature level [39], [45]. However, less efforts have paid attention on the ensemble effect of two or multiple pre-trained models.

Although transfer learning saves training efforts, it does not increase the amount of annotated training data. Generative models can help enhance the size and diversity of the training set by generating synthetic samples that do not appear in the original training set. Popular models such as Generative Pre-trained Transformer 2 (GPT-2), Generative Adversarial Network (GAN), and Variational Autoencoder (VAE) have been used for DA in numerous learning tasks. However, in a domain-specific NER task under a low-resource setting, generative DA methods have not been extensively studied.

To tackle the medical NER task in a low resource setting, we propose a two-stage transfer learning framework with two performance boosters. In the first stage, we fine-tune WordVec and BERT separately on the prepared domain resources in a self-supervised manner. Meanwhile, we develop a substitution-based generative model (the first booster) for DA, which can generate synthetic and annotated data to enhance the original training set. In the second stage, we employ an ensemble of the tuned WordVec and BERT models (the second booster) to generate word embeddings, which are fed into an NER model that consists of a Bidirectional Long Short-term Memory (BiLSTM) layer and a Conditional Random Field (CRF) layer to output a prediction. In summary, this study makes the following contributions:

1) We propose a two-stage transfer learning pipeline for the biomedical NER task under a low resource setting, which is resulted from a oncological NER dataset in Chinese. The novelty of the learning pipeline mainly lies in the two performance boosters, including a substitution-based generative model for DA and an ensemble model of both Word2Vec and BERT for embedding generation. In addition, both Word2Vec and BERT are further pre-traind on domain resources to further incorporate oncological knowledge into the models. These joint efforts can benefit the downstream NER model even with a small annotated corpus in the original dataset.
2) We conduct extensive experiments a) to demonstrate the validity of using an ensemble of both Word2Vec and BERT for embedding generation since both models contribute to the performance improvement, b) to show the efficacy of the substitution-based generative DA method, verifying the importance of both quantity and diversity for an augmented training set. Results show that under the same hyper-parameter setting, our best model outperforms the state of the art (SOTA) by 4.64%, and thus can serve a new baseline for subsequent studies.

The rest of the paper is organized as follows. Section II describes the related work. Section III provides a description of the database in use. Section IV presents the core building blocks of the proposed learning pipeline. Section V reports the experimental results and offers some insights. Section VI summarizes the whole work.

II. RELATED WORK

This section reviews a collection of related studies from three aspects, including domain-specific NER, sequence model, and data augmentation in NLP.

A. DOMAIN-SPECIFIC NER

Lack of sufficient annotated data is the primary challenge for domain-specific NER tasks. There is a prohibitive annotation cost in the biomedical field due to the required expertise and domain knowledge, making crowdsourcing-based annotation infeasible. To overcome these challenges, prior studies have explored various methods that are summarized as follows:

- Cross-lingual knowledge transfer aims to transfer knowledge learned from one or more source languages to a low-resource target language. Cotterell and Duh [8] trained character-level neural CRFs to predict named entities for both high-resource languages and low-resource languages jointly. Feng et al. [15] employed bilingual lexicons to bridge cross-lingual semantic mapping and design a lexicon extension strategy to alleviate the out-of-lexicon issue. Their study also considered entity type distribution as language-independent features and modeled them in the neural architecture, which became a performance booster.
- Domain-adaptive approaches allow models trained from easily obtained resources to be applied to the NER task in a target domain. Yu et al. [43] proposed a domain-adaptive approach that fine-tuned a general-domain pre-trained language model on in-domain corpora. A bootstrapping process was then started to obtain an initial NER model trained on the small fully-annotated seed data. The NER model was fine-tuned on an unannotated corpus iteratively until convergence.
- Distantly supervised learning (DSL) [36, 30, 17, 41] employs domain-specific dictionaries and knowledge bases like WikiData and YAGO to produce massive weakly annotated data, which are used to train NER models.
- Domain-specific pre-training [4], [25] leverages unsupervised pre-training on a large domain-specific corpus to improve performance on downstream NLP tasks such as NER.

In summary, there are two lines of mainstream research to address the difficulties of low-resource NER tasks. The first direction relies on transfer learning to apply knowledge learned from rich-resource domains/languages to low-resource ones. The second tries to reduce human annotation efforts through automatic annotation by utilizing existing domain-specific knowledge bases. Our work belongs to the first category. In particular, we customize a transfer learning pipeline that suits the task under investigation.

B. SEQUENCE MODEL

As an encoder-decoder RNN, the sequence-to-sequence model has similarities with the conception of end-to-end, which is extensively applied in NLP and the text labeling task. After the input of a sentence required to be translated, understood, and analyzed, the text will be divided into unitary words before being converted into word vectors via word embedding and contextual word embedding. In this stage, an encoder network is first established. Built as an RNN, it could be a GRU, or an LSTM RNN [35]. LSTM is considered an effective approach in dealing with sequence labeling problems, combining the past and future contexts together [1]. While the unidirectional LSTM is not adequate and the appearance of bidirectional LSTM (BiLSTM) can utilize the historic and the future information together to label a token [28]. One word is input to feed the network every time and after ingesting the input sequence extracted by the words in the text. A decode network is built for the output sequence, which is designed according to the wanted function. The output sequence can be constructed to cluster the words with similar meanings in a text for pre-training and training the dataset. It is difficult to correlate the words in the current label and the neighbor label, and a meaning loss and the error in NER will occur for a loss in the correlation between two seemingly unrelated labels [38]. Linear-chain Conditional Random Field (CRF) is cultivated to mitigate such a problem by controlling the structure prediction [32]. A series of potential functions is applied to approximate the conditional probability of the output label based on the input word sequence. The decoder sequence can be finished at the end word or the end sentence token. The introduction of the sequence model in our work can provide convenience to recognize the words, first, with the similar meaning according to context; second, the similarity comparison results can be obtained only in one stage algorithm realized by sequence-to-sequence model.

One problem that RNN is confronting is that the computation of the output is hard to parallel, which can result in undesired results when there are several relations between the input. If the parallelism in the neural network cannot be realized, the poor performance of learning will appear [28]. It is noted that a self-attention algorithm is introduced in the transformer, which is known as an unsupervised learning method. Without adding an algorithm related to attention, a self-attention in the transformer can be realized by three transformations for further computation, q(query) to match others, k(key) to be matched by q, and v as information to be extracted. The q of each input will undergo a computation with every k value of the input. After accomplishing the computation, processing, and another computation with the v of every input, the output b will be finally obtained. Such a transformer can be parallel to conventional RNN to get better performance, especially in text recognition when the contextual features should be a key point for NER.

Named entity recognition (NER) focuses on identifying mentions featured on rigid designators from text belongings to predefined semantic types such as person, location, organization, etc. Known as the foundation in NLP tasks such as question answering, text summarization, and machine translation, NER has achieved good performance to reduce manual work in NLP. While when coping with corpus with low available resources, new models and algorithms should be added in NER [26].

BERT is known as an advanced tool in NLP that can realize bidirectional word recognition. The Transformer mentioned before is known as a signature algorithm in BERT for better contextual information extraction. Apart from Transformer, a "masked language model" (MLM) can mask the original meaning of the vocabulary and obtain the meaning information of input words and text based on the context. Besides, different from the traditional unidirectional encoder, a bidirectional decoder can be generated with the assistance of the MLM, so it can be known that MLM is also contributive for contextual recognition in NER [11]. BERT is the encoder part of our sequence model. Correspondingly there is a decoder model called generative pre-training (GPT-2). Unsupervised learning is often applied in pre-training for clustering. Generally, rich-resource without annotation form a language model with several clusters in the first stage of GPT-2 [20]. Position embedding will be initialized randomly in this stage, and the model is expected to learn without supervision. In the second stage, the algorithm will be fine-tuning for transferring into NLP question. The objective of GPT-2 is text recognition and deal with annotated tasks, which is more suitable in NER in our work. Also, recent study has also combined sequence models with graph models to solve the NER task. For instance, Fu et al. stacked BiLSTM encoder and a GCN to model the relations of entities. The proposed model can be used to jointly extract named entities and relations [50].

C. DATA AUGMENTATION IN NLP

DA in machine refers to a collection of methods that aim to increase the quantity and diversity of training data, reducing the efforts of human annotation [13]. Most methods work by manipulating existing samples or create synthetic data to enrich the training set, acting as a regularizer to prevent overfitting. DA has been extensively used in computer vision [37] with a wide spectrum of methods developed. DA is not that straightforward when it comes to NLP due to the discrete input space and highly unstructured and dynamic semantics in texts. DA in NLP can be roughly divided into the following three categories:

- Rule-based DA refers to a class of DA methods that use predefined transforms sans model. For example, Wei and Zou [40] develop easy DA that employs a set of token-level operations such as insertion, delection, and swap, that are randomly applied to texts for perturbation. Chen et al. [6] propose to build a graph over text data. The graph can be used to infer augmented sentence pairs based on balance theory. Sahin et al. propose dependency tree morphing DA that employs sentence cropping and rotating for siblings in the tree [34].
- Example interpolation-based DA is a class of methods that works by interpolating the inputs and labels of real samples to generate synthetic samples. However, mixing sentence elements between different samples without guidance breaks the semantic meaning. Several strategies have been proposed to achieve more effective interpolation [7], [19].
- The models in model-based DA usually refer to sequence-to-sequence and language models. Yang et al. propose a DA method that uses a pre-trained transformer language model to generate synthetic examples [42]. Feng et al. [14] propose semantic text exchange, which aims to adjusting the semantics of a sentence to fit the context of a newly inserted word or phrase. Generative models such as GPT-2 have also been used. Anaby-Tavor et al. fine-tune a label-conditioned GPT-2 for DA (3). Other impactful methods include document-level paraphrasing [18], controlled paraphrasing [22], and misclassified example augmentation [12].

The proposed substitution-based generative DA is model-based and designed for NER. Prior efforts for DA in NER are sparse. One relevant work is by Dai and Adel [9], where synonym and mention replacements have been the most effective strategies in their experiments. However, their method does not utilize a generative model, while ours employs GPT-2 combined with entity mention substitution, which allows the model to generate high-quality synthetic samples.

D. PRE-TRAINING FOR NER

Pre-training has been an influential technique that leverages unsupervised or self-supervised learning to gain semantic knowledge from a large corpus at a scale [47]. A well-pre-trained model can be fine-tuned to suit various downstream tasks, including NER [48], [49], [50], [51], [52], [53]. Xue et al. proposed an NER-specific pre-training framework to inject entity knowledge into pre-trained models [48]; a similar idea was implemented by Jia et al. [50], who integrated entity information into BERT using Char-Entity-Transformer. Trewartha et al. built a domain-specific model in material science that showed superior performance in the NER task [49]. Gao et al. adopted a pipeline that consists of pre-training, fine-tuning, and self-training to boost model performance [51]. In summary, fine-tuning is an essential step that allows pre-training models to handle NER. Before fine-tuning, a second round of pre-training can be adopted, either to inject domain knowledge, like the proposed method, or to inject entity knowledge to pre-train a more entity-aware model [48], [50].

III. DATASET AND LEARNING TASK

This section provides a description of the Yidu-S4k dataset followed by a definition of the NER task.

A. DATASET

This work utilized the Yidu-S4k dataset created for biomedical NER from the Chinese electronic biomedical records (CEMRs) dataset. The task was one of the six tasks in CCKS 2019. In particular, this task requires locating named entities and classifying them into six pre-defined categories:

- *Disease and diagnosis*: biomedically defined diseases and the judgments made by physicians in their clinical work regarding etiology, pathophysiology, and staging.

Examples of this entity type include “Endogenous chondroma of left humerus”, “Synovitis”, “Osteosarcoma”, “Ewing Sarcoma”, etc.

- *Image examination*: imaging examinations (X-ray, CT, MR, PETCT, etc.), not avoiding too many conflicts between examination operations and surgical operations. Examples include “Ultrasound of the abdomen”, “Electrocardiogram”, “Knee CT”, etc.
- *Lab testing*: physical or chemical tests performed in the laboratory, referring specifically to laboratory tests performed by the laboratory department in clinical work, excluding broad laboratory tests such as immunohistochemistry. Examples include “CER13.97NG/ML”, “83.96NG/ML”, “4.02 × 1012/L”, etc.
- *Surgery*: the main surgical treatment is excision and suturing performed locally by the doctor on the patient’s body. Examples include “Right ilium incision biopsy”, “Left fibula mass resection”, etc.
- *Medication*: specific chemical substances used for disease treatment. Examples include “Paracetamol and oxycodone”, “Tegafur”, “Sodium Aescinate”, etc.
- *Anatomical site*: the anatomical part of the body where the disease, signs and symptoms occur. Examples include “Left ilium outer plate”, “Gastric base”, “Left lateral femoral condyle”, etc.

The training set consists of a total of 2,000 diagnostic records with 1,000 annotated records and another 1,000 unannotated. Each record contains multiple named entities. The validation set has a total of 400 records. Table 1 displays the stats information of the six categories across the training and validation set.

TABLE 1. Entity quantity of the NER task in the Yidu-S4k dataset.

Entity types	Training set	Augmented set	Validation set
Disease and diagnosis	4,206	3,618	1,323
Image examination	972	725	348
Lab testing	1,195	1,263	590
Surgery	1,492	1,359	203
Medication	1,931	2,015	597
Anatomical site	8,426	7,711	3,094
total	18,222	16,691	6,155

We provide a short sample record from the dataset as follows. “*After the patient was admitted to the hospital, the relevant examinations were perfected, preoperative preparations were performed, and no surgical contraindications were found. Resection of the right middle tibial mass was scheduled for 2020- 08-31 under general anesthesia. The procedure went smoothly and returned to the ward after the operation. Prevent infection, relieve pain, strengthen bones and other symptomatic treatments. The patient is now recovering well after the operation and is required to be discharged from the hospital. The pathological results suggest: osteoid osteoma.*”. In this example, multiple named entities of different types can be identified.

B. NER LEARNING TASK

Given a sentence as an input, an NER model can locate and classify the entity mentions in the input into a set of pre-defined categories. The learning task can be formally defined as follows. Suppose s is a sentence with n word tokens, where $s = [t1, t2, \ldots, tn]$, an NER model takes s as input, and outputs a collection of entity mentions, each of which is denoted by a three-tuple $< Is, Ie, k >$, where Is and Ie are the indices of the first and last tokens of the entity mention, and k is one of the pre-defined classes. In this study, $k \in$ {“Disease and diagnosis”, “Image examination”, “Lab testing”, “Surgery”, “Medication”, “Anatomical site” }.

IV. A TWO-STAGE TRANSFER LEARNING PIPELINE

In this section, we describe the design details of the building blocks of the proposed learning pipeline.

A. SYSTEM OVERVIEW

Figure 1 shows the proposed learning pipeline. Before training an NER model, we adopt a substitutionbased generative model for data augmentation, aiming to enhance the quantity and diversity of the training data by adding a collection of synthetic samples into the original training set. The learning pipeline includes two stages. In the first stage, a pre-trained base language model is taken for fine tuning on biomedical domain resources to incorporate more domain knowledge into the model. The outcome of the first stage is a model that can generate word embeddings, which are fed into the next stage. The fine tuning step is crucial since it allows the model to capture domain information, which improves the accuracy of the downstream NER task and reduces the amount of required annotated data. In the second stage, word embeddings flow through a BiLSTM sequence model to further capture context information. At last, the output of BiLSTM is sent to a CRF layer that maximizes the log-probability of the NER sequence to encourage the network to produce a valid sequence of entity types.

B. SUBSTITUTION-BASED GENERATIVE MODEL FOR DATA AUGMENTATION

Figure 2 depicts the GPT-2-based data augmentation mechanism, aiming to alleviate the low-resource problem for the learning task. The process includes the following steps.

- First, each annotated sample in the original training set is transformed with all entity mentions substituted by special tokens representing their corresponding entity types, as shown in the “Substitution I” module in Figure 2. For example, a sentence “The patient underwent radical resection of rectal cancer.” contains two entity mentions, namely, “radical resection” and “rectal cancer”, which are replaced by the entity type tokens, resulting in a

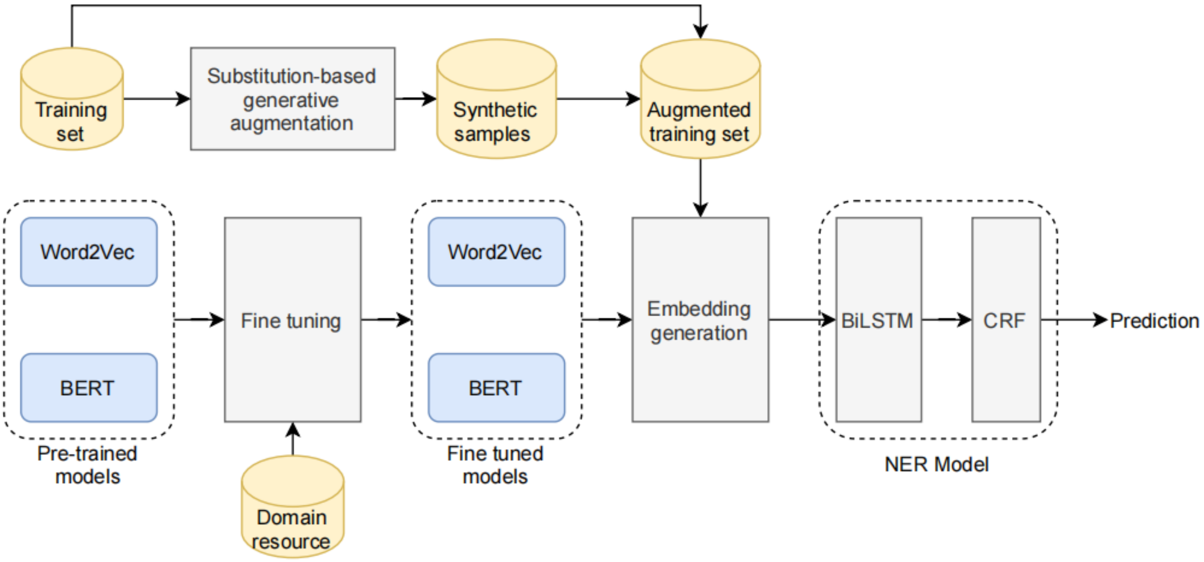

FIGURE 1. **An overview of the proposed learning framework.**

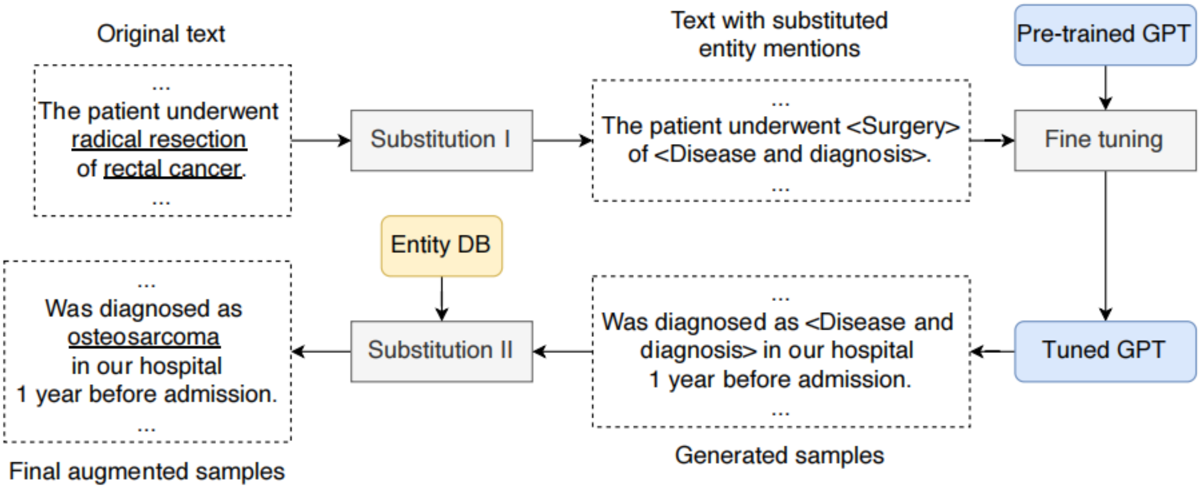
Original text
...
The patient underwent radical resection of rectal cancer.
...
Substitution I
Text with substituted entity mentions
...
The patient underwent <Surgery> of <Disease and diagnosis>.
...
Pre-trained GPT
Fine tuning
Tuned GPT
Entity DB
Substitution II
...
Was diagnosed as <Disease and diagnosis> in our hospital 1 year before admission.
...
Generated samples
...
Was diagnosed as osteosarcoma in our hospital 1 year before admission.
...
Final augmented samples

FIGURE 2. **Substitution-based generative augmentation.**

transformed sample "The patient underwent [Surgery] of [Disease and diagnosis]." The purpose of substitution is to facilitate the fine-tuning of GPT-2, which is guided to focus on the semantic meanings of entity types rather than specific entity mentions.

- After the substitution, the training samples are used to fine-tune a pre-trained GPT-2 model. Fine-tuning allows the GPT-2 model to learn the contextual knowledge in a domain-specific setting.
- The tuned GPT-2 model is then utilized to generate a collection of samples that have a similar distribution to the training set.
- The generated samples are fed into another substitution module (i.e., Substitution II in Figure 2), which randomly draw a matching entity from an entity database (DB) to substitute an type token in a generated sample. For instance, "[Disease and diagnosis]" is replaced by "bone cancer" in the sentence "Was diagnosed as [Disease and diagnosis] in our hospital 1 year before admission.", creating an augmented sample that can be automatically annotated and used to enhance the training set.
- It is noted that the entity DB specifically designed for the learning task. The DB contains a collection of six vocabularies, corresponding to the six entity types, and each vocabulary is a list of entities belonging to that category. For example, "bone cancer" is in the vocabulary of "Disease and diagnosis".

Following this procedure, a total of 1,000 augmented samples are generated. The statistical information for the entity quantity in these augmented samples are shown in Table 1.

C. STAGE ONE: PRE-TRAINING AND FINE TUNING

The first stage takes pre-trained Word2Vec and BERT language models to conduct fine tuning on the dataset in use. The fine tuning is self-supervised, which does not require annotation, and the output model can generate word embeddings in the biomedical domain.

1) PRE-TRAINING WORD2VEC ON DOMAIN RESOURCES

There are two different learning models applied in Word2Vec, including Continuous Bag-Of-Words (CBOW) and Skip-gram. CBOW learns a conditional probability of a word given its context, i.e., the surrounding words within a specified window size. Specifically, a provided sequence of words will be received as input a window of C context words by the model. Then the target word wi is predicted by minimizing the following objective:

$$E = -\frac{1}{|c|}\sum_{t=1}^{|c|} \log P(w_t \,|w_{t-c}\ , \cdots, w_{t-1}, w_{t+1}, \cdots, w_{t+c}) \tag{1}$$

$$P(w_t \,|w_{t-c,\cdots}\ , w_{t-1}, w_{t+1,\cdots}, w_{t+c}) = \frac{\exp u_t^T v_c}{\sum_i^{|V|} \exp u_t^T v_c} \tag{2}$$

where V depicts the vocabulary size, v_c is the sum of the embeddings vector of the context words $w_{t-c}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+c}$, and u is the embeddings vector of the target word.

Skip-gram, on the other hand, can predict the context words of a given word, which is regarded as the reverse of CBOW. It works by minimizing the following objective:

$$E = -\frac{1}{|c|}\sum_{t=1}^{|c|}\sum_{-c\le j\le c} \log P(w_{t+j}\,|w_t) \tag{3}$$

$$P(w_{t+j}\,|w_t) = \frac{\exp u_{t+j}^T v_t}{\sum_i^{|V|} \exp u_i^T v_t} \tag{4}$$

where u and v are the current and context word embeddings, respectively.

Either model can learn a distributed representation of words via a shallow neural network. For our task, we use a Word2Vec model pre-trained on corpora of generic Chinese text.[1] We further pre-train the Chinese Word2Vec model on two domain resources, including the training data of the Yidu-S4k dataset and a question answering dataset [44] with 623,138 medical sentences in Chinese. The adopted training algorithm is CBOW. This way, the Word2Vec model can capture the domain knowledge and produce word embeddings suitable for the NER task in this study.

2) PRE-TRAINING BERT ON DOMAIN RESOURCES

BERT is built on top of a Transformer that has an encoder to read the text input, which is a sequence of word tokens. BERT maps these tokens to vectors (i.e., token embeddings) and decorates them with some metadata by 1) adding a [CLS] token to the input at the beginning of a sentence and a [SEP] token at the end, 2) attaching a Segment embedding to each word token to indicate the sentence it belongs to, and 3) adding a positional embedding to mark its position in the sentence. These embeddings can encode rich contextual information of a word, enabling a bi-directional, or more precisely, non-directional training to capture the semantic meaning of a word within a context. Since the goal of BERT is to generate a language model, two strategies are employed: 1) to train the model using input text with masked tokens, which allows the model to learn to predict the masked tokens, and 2) to train the model with sentence pairs, say (A, B), so that the model can learn to predict whether B is the next sentence of A. In this study, we adopt the Chinese BERT wwm,[2] which has been pre-trained with the Chinese Wiki. Similarly, we apply a further pre-training to Chinese BERT wwm on the Yidu-S4k dataset and the Chinese medical QA dataset for knowledge transfer.

3) OUTPUT OF STAGE ONE

We obtain two language models of Word2Vec and BERT through fine tuning the base models. A word token passes through both models to be mapped to two word embeddings, representing non-contextual and contextual embeddings. The system then performs a concatenation of the word embeddings as the final word vector, which is sent to the next stage for NER training.

D. STAGE TWO: CHINESE BIOMEDICAL NER

In stage two, the ensemble embedding is used as an input token for the subsequent network to train a NER tagger. The downstream NER tagger consists of two layers, a BiLSTM layer and a CRF layer, which has been a popular choice for NER tasks since developed in [24].

The BiLSTM layer takes as input the word embeddings obtained from stage one. A BiLSTM operates on sequential data and returns another sequence that captures bi-directional context information through a forward and backward LSTM pair. The output corresponding to the input word vector is a concatenation of the left and right context representations. Specifically, a sentence is depicted by $(w_1, \cdots, w_n)$ and its contextindependent word representations is depicted by $(\tilde{v_1}, \cdots, \tilde{v_n})$. The language model can be formalized as

$$P = (w_i \,|w_1\ , \cdots, w_{i-1}) = \frac{1}{Z}\exp(w_{w_i}\vec{h}_{i-1} + b_{w_i}) \tag{5}$$

where $Z = \sum_w \exp\left(w_w \vec{h}_{i-1} + b_w\right)$ is the normalization term and $\vec{h_i}$ is the last output of the forward context representation function $\overrightarrow{LSTM}\,(\tilde{v_1}, \cdots, \tilde{v_n})$. The same definition is used for backward language modeling. A character-level CNN for the context-independent representation is also use by [33] where $\tilde{v_i} = CNN(w_i)$. The utilization of

[1] https://github.com/Embedding/Chinese-Word-Vectors

[2] https://github.com/ymcui/Chinese-BERT-wwm

a character-level CNN to represent words can help embedding from language model (ELMo) to output reasonable context-independent word embeddings for arbitrary words. Multi-layer LSTMs with skip connections are also proposed to parameterize $L\vec{ST}M$ and $L\overleftarrow{ST}M$. The iteratively multi-layer mechanism can be depicted as

$$\vec{h_i}^{(k)} = L\vec{ST}M^{()}\left(\vec{h_1}^{()}, \cdots, \vec{h_i}^{()}\right) \tag{6}$$

where the initial values of $\vec{h_i}^{(k)}$ and $\overleftarrow{h}_i^{(k)}$ are set as: $\vec{h_i}^{(0)} = \overleftarrow{h}_i^{(0)} = \tilde{V}_i$, for $k = 0$. The final contextualized embeddings are computed by weighted pooling of the activations of $L+1$ different layers as $\mathbf{ELMo}_i = \lambda \sum_{k=0}^{L} s_k \bullet \left(\vec{h_i} \bigoplus h_i\right)$. The learning objective is a summation of each word's log probabilities of two directions, given as follows:

$$\begin{aligned} E = \sum\nolimits_{i-1}^{n} & (\log p(w_i \,|w_1 , \cdots , w_{i-1}) \\ & + (\log p(w_i \,|w_{i+1} , \cdots , w_n)) \end{aligned} \tag{7}$$

BiLSTM, consisting of two hidden states, $\vec{h}$ and $\overleftarrow{h}$, can describe each sequence in the forward and reverse directions to two separate layers. The two hidden layers are then concatenated to represent the final output. Let $(x_1, x_2, \ldots, x_n)$ be an input sequence with n words, and $\vec{h_t}$ and $\overleftarrow{h}_t$ be the representations of word t given by the forward and backward layers, respectively. The final representation of a word t is yielded by BiLSTM via concatenating the outputs of both its left and right contexts, i.e., $h_t = \left[\vec{h_t}, \overleftarrow{h}_t\right]$.

The BiLSTM layer's output serves as the features of the subsequent CRF layer. During training, the CRF maximizes the log-probability of the NER sequence to encourage the network to produce a valid sequence of entity types. Specifically, assume that x= $(x_1, x_2, \ldots, x_n)$ is an input sequence, where xi is the input vector of the i-th word. Also, y = $(y_1, y_2, \ldots, y_n)$ means a sequence of predicted labels for input **x**. All y_i of **y** will range over a set $L(x)$, a possible labeling sequence for **x**. $F(y, x)$, the summation of CRF's local feature vector $F(y, x, i)$, is the global feature of CRF for input sequence **x**and label sequence **y**, and i ranges over input positions. A conditional probability $p(\mathrm{y}|\mathrm{x}, \lambda)$ is defined by the probabilistic model for the CRF covering all possible sequences of labels **y**, given **x** and weight vector λ in the following form:

$$p(y|x,\ \lambda) = \frac{1}{Z\,(\mathrm{x})} \exp\left(\mathrm{C} \bullet \ddot{\theta}(\mathrm{y}, \mathrm{x})\right) \tag{8}$$

in which $Z(\mathrm{x}) = \sum_{\mathrm{y}' \in L(x)} \exp\left(\mathrm{C} \bullet \ddot{\theta}(\mathrm{y}', \mathrm{x})\right)$ is a normalization factor.

Throughout the learning pipeline, the key is to preserve contextual information, which is featured by BERT, BiLSTM, and CRF.

V. EVALUATION

The section provides the details of the experiments, including the hardware, training configuration, evaluated models, performance metrics, generated samples used for data augmentation, embedding quality analysis, hyper-parameter tuning, and a presentation of the results.

A. EXPERIMENT CONFIGURATION

The experiments have been conducted using Python 3.6 and Pytorch 1.4 on an Ubuntu 18.04 system with an i7-10875h CPU and a Tesla V100 16G GPU. We choose BERT base, which has 12 layers of encoders with 768 hidden layers, 12 attention heads, and 110M trainable parameters.

B. MODELS

We have evaluated the following models for a comparative study.

- CRF [23]is used as a baseline to compare with other models. Using CRF alone does not preserve much contextual knowledge from the input, limiting its performance.
- W2V+BiLSTM+CRF [27] adds Word2Vec (W2V for short when used in the model name) and BiLSTM to encode semantic information, which turns out to be an effective strategy.
- BERT+BiLSTM+CRF [10] replaces Word2Vec with BERT, adopting a transformer-based deep neural network for contextual encoding. Therefore, this model is considered as the SOTA.
- W2V+BERT+BiLSTM+CRF ensembles both Word2Vec and BERT embeddings to form a more diverse representation for each word.
- W2NER [46] is a novel alternative approach for unified named entity recognition. It models the unified NER task as word-word relation classification, offering a unique perspective. By effectively capturing neighboring relations between entity words using Next-Neighboring-Word (NNW) and Tail-Head-Word-* (THW-*) relations, W2NER overcomes the kernel bottleneck of unified NER. The model utilizes a neural framework that represents the unified NER as a 2D grid of word pairs. It further incorporates multi-granularity 2D convolutions to refine the grid representations, enhancing performance. Additionally, W2NER employs a co-predictor to reason the word-word relations comprehensively. Through extensive experiments on various benchmark datasets, W2NER outperforms current top-performing baselines across flat, overlapped, and discontinuous NER tasks, pushing the boundaries of state-of-the-art NER performance.

Both Word2Vec and BERT in the above three models have been pre-trained but not fine-tuned on domain resources. To incorporate domain knowledge, we have also evaluated six additional models to form a complete ablation study:

- W2V+BiLSTM+CRF with fine tuning (F.T.)
- BERT+BiLSTM+CRF with F.T.
- W2V+BiLSTM+CRF with F.T. and data augmentation (D.A.)

- BERT+BiLSTM+CRF with F.T. and D.A.
- W2V+BERT+BiLSTM+CRF with F.T.
- W2V+BERT+BiLSTM+CRF with F.T. and D.A.

The domain resources used for fine tuning include the original Yidu-S4k dataset and an enhanced question & answer corpus [44] with 623,138 biomedical sentences in Chinese. The data augmentation strategy has been described in Section IV-B.

C. PERFORMANCE METRICS

Using accuracy as the sole performance metric is not sufficient to evaluate a mature and reliable model, especially in the case of imbalanced categories. For our dataset, the quantities of different entity instances are unbalanced, as shown in Table 1. To this end, we employ six indicators to present the experimental results for a complete comparison, including precision (Pre), recall (Rec), and the F1 score. Since our NER task is a multi-class classification problem, we take a macro-average on all six entity types to obtain a single value for each metric. Equations 9-11 show the definitions of precision, recall, and F1.

$$Pre = \frac{TP}{TP + FP} \tag{9}$$

$$Rec = \frac{TP}{TP + PN} \tag{10}$$

$$F1 = 2 \times \frac{Pre \times Rec}{Pre + Rec} \tag{11}$$

where *TP*, *FP*, and *FN* stand for true positive, false positive, and false negative, respectively.

D. GENERATING SYNTHETIC SAMPLES

Based on the augmenting strategy described in Section IV-B, we have fine-tuned a GPT-2 model using the following set of hyper-parameters: {“number of training epochs”: 5, “batch size”: 8, “evaluation steps”: 400, “warm up steps”: 300}. The tuned GPT-2 is able to generate samples similar to the ones appearing in the training set but with entity placeholders. For instance, a synthetic sample is provided: “It was due to the [Disease and Diagnosis], a [Surgery] was performed in our hospital more than a month ago. The operation went well. Postoperative pathology (pathology number: 201607041) showed [Disease and Diagnosis], which invaded the adventitia. There was no cancer infiltration at the double cut ends of the surgical specimen. One [Anatomical site] was found, and no cancer metastasis was found. …” In this example, there are two, one, and one entity mentions (placeholders for now) for the categories of Disease and Diagnosis, Surgery, and Anatomical Site, respectively. Next, the entity placeholders will be substituted with actual entities, which generates a sample that can be added into the augmented data set. It is noted that these synthetic samples are similar but different from the ones in the original training set. Also, the entity database used for substitution II can be prepared offline with more and new entities that never appear in the original dataset, which can also enhance the diversity of the augmented set, potentially benefiting the NER performance. In this study, we generated 1000 augmented samples, and the number 1000 was an empirical value that achieved a decent trade-off between set size and the overall performance.

In addition to quantity, the quality of synthetic samples is essential; after all, GPT-2 may generate sentences that could be misleading or unreadable. A round of manual screening on all generated samples was conducted to ensure the quality of the augmented data used for training.

TABLE 2. Comparison results of different models on the Yidu-S4k validation set.

Model	Pre	Rec	F1
CRF [23]	0.4972	0.1788	0.263
W2V+BiLSTM+CRF [27]	0.6396	0.5425	0.5871
BERT+BiLSTM+CRF [10]	0.7725	0.7767	0.7746
W2NER [46]	0.7839	0.7813	0.7828
W2V+BERT+BiLSTM+CRF	**0.7881**	**0.7828**	**0.7853**

E. EMBEDDING QUALITY ANALYSIS

The quality of word embeddings produced by Word2Vec and BERT is an important factor of the model performance. We perform the T-Stochastic Neighbor Embedding (T-SNE) analysis on the word embeddings and plot the results for the six entity types of our task in Figure 3. It is observed that after fine-tuning on the domain resource, the entities are better clustered for both BERT and Word2Vec, meaning that a second round of pre-training using domain-specific texts allows a model to better capture contextual information; thus, entities with similar semantic meanings get closer in the embedding space, leading to a better cluster effect, as shown in Figure 3.

In addition, we select a collection of ten related words for embedding analysis. Specifically, the ten selected words are compared with the word “*root of superior mesenteric artery*” in terms of similarity,which is a score between 0 to 1, and the higher, the more similar between two words. The results are shown in Figure 4. It is observed that for eight out of ten words, BERT reports a higher similarity, which indicates the degree of semantic relatedness between each of those words and “root of superior mesenteric artery”.The only similarity Word2Vec ranks slightly higher than BERT is the word pair “small intestine”, and“root of superior mesenteric artery”. Based on our domain knowledge, BERT presents better embeddings than Word2Vec in this similarity analysis, demonstrating a superior ability to encode words’ semantic and contextual meanings.

F. OVERALL PERFORMANCE

To make the models aforementioned in Section V-B comparable, we have employed the same hyperparameter setting,

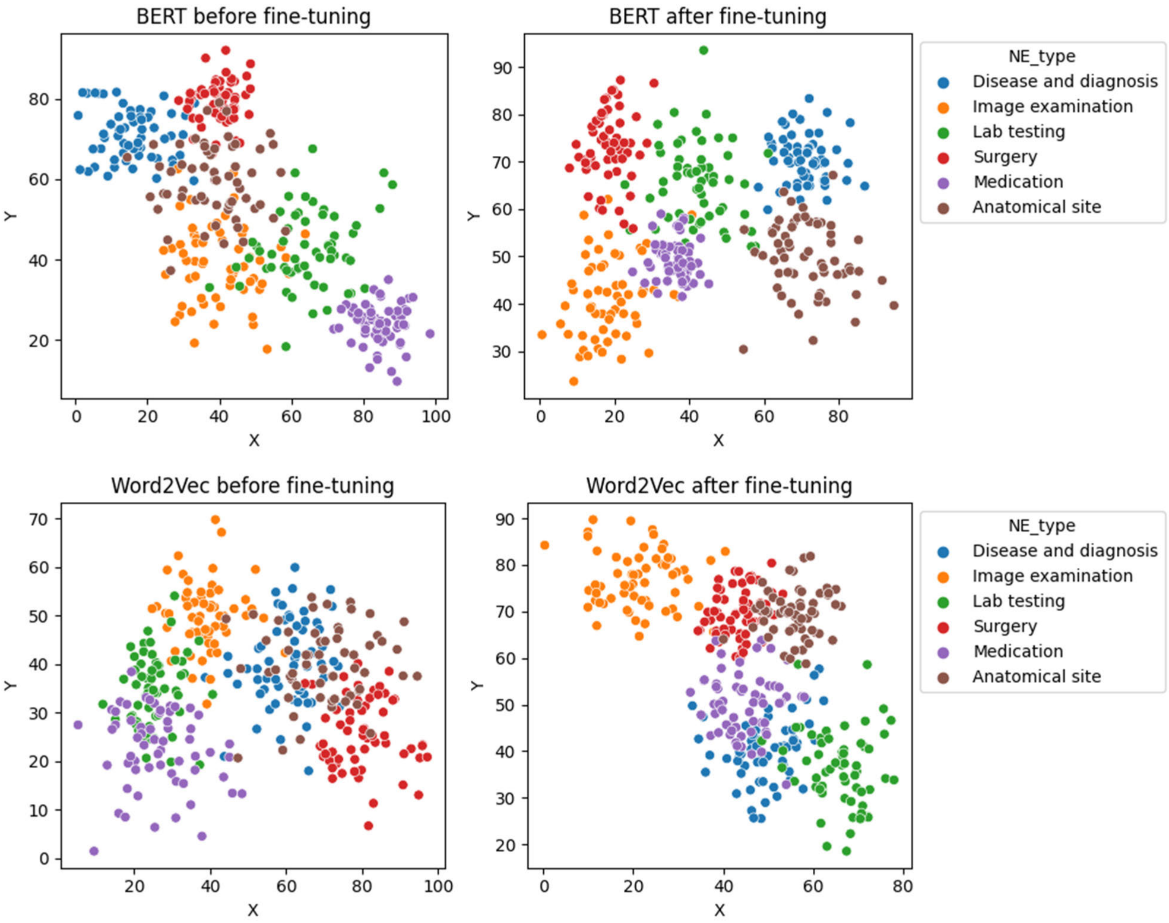

FIGURE 3. T-SNE analysis for the embeddings generated by BERT and Word2Vec.

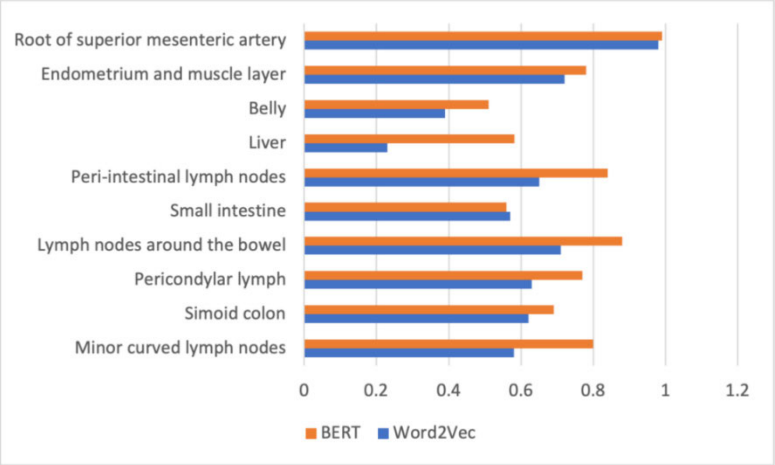

FIGURE 4. Visualization of similarities for words in BERT and Word2Vec. Ten words/phrases were selected with embeddings from both BERT and Word2Vec. A similarity score was calculated between each of these word/phrase and the phrase "root of superior mesenteric artery".

with a word vector dimension of 400, a learning rate of 1e-5, an LSTM layer number of 1, an LSTM hidden size of 200. We report the results in Table 2 and provide our observations below.

- Without any pre-training, model CRF performs the worst with an F1 score of 0.263, mainly because of the small quantity of annotated data utilized for training in our task.

TABLE 3. Ablation study.

Model	F.T.	D.A.	Pre	Rec	F1
W2V+BiLSTM+CRF	☒	☒	0.6396	0.5425	0.5871
	☑	☒	0.6515	0.5831	0.6137
	☑	☑	0.6773	0.6094	0.6383
BERT+BILSTM+CRF	☒	☒	0.7725	0.7767	0.7746
	☑	☒	0.7879	0.7921	0.7902
	☑	☑	0.8194	0.7981	0.8079
W2V+BERT+BiLSTM+CRF	☒	☒	0.7881	0.7828	0.7853
	☑	☒	0.8342	0.7841	0.8084
	☑	☑	**0.8397**	**0.8123**	**0.821**

- Pre-training has significantly improved the performance. In particular, the Word2Vec base model (i.e., W2V + BiLSTM + CRF) achieves an F1 of 0.5871, while the BERT base model (i.e., BERT + BiLSTM + CRF) posts an F1 of 0.7746. It also shows that the ensemble of Word2Vec and BERT embeddings boosts the F1 score to 0.7853, which demonstrates the usefulness of introducing diversity in word representation. In addition, W2NER performed well with an F1 score of 0.7828, showing its powerful NER capability via the strategy of word-word relation classification.
- Fine tuning the Word2Vec and BERT base models on the domain resources have shown effectiveness. Specifically, with fine tuning, the W2V+BiLSTM+CRF and BERT+BiLSTM+CRF models have improved the F1 scores by 2.66% and 1.56%, respectively. It is observed that although Word2Vec presents a higher percentage improvement than BERT (2.66% vs. 1.56%), BERT shows dominant performance in F1 (0.7902 vs. 0.6137). Similar effect has been observed in the ensemble model W2V+BERT+BiLSTM+CR, which presents an F1 of 0.8084 with a gain of 2.53%. The results show that the evaluated models have achieved consistent performance gains when fine-tuned on domain resources, which allow oncological knowledge to be injected into the models during training.
- When fine-tuned on the augmented domain resources, the performance of the evaluated models have been further elevated. Specifically, the Word2Vec, BERT, and the ensemble model have posted a gain of 2.46%, 1.77%, and 1.24%, respectively. The demonstrated improvement shows that our data augmentation strategy has been effective in generating synthetic samples that have enhanced the quality and diversity of the dataset used for training, allowing the models to learn more domain knowledge even in a low-resource setting.

The results in Table 3 serves a complete ablation study, demonstrating the necessity of each performance boosting strategy utilized in our learning framework, including embedding ensemble, fine tuning, and data augmentation. It is noted that each strategy has achieved consistent performance gains on each evaluated model, and a combination of the three boosters has led to the best F1 (0.821), a gain of 4.6% compared to the SOTA.

G. HYPER-PARAMETER TUNING

We perform hyper-parameter tuning on the best ensemble model obtained from the previous section. The tuned three hyper-parameters include the learning rate, the number of LSTM layers, and the number of LSTM hidden units. For the learning rate, we consider three values including 1.0E-03, 1.0E-04, and 1.0E-05. For the number of LSTM layers, we have evaluated 1 layer and 2 layers. For the number of LSTM hidden units, values in the set {64, 256, 384} have been considered. We have conducted a grid search that explores a total of 18 experiments in the search space and obtained the optimal setting, namely, a learning rate of 1.0E-03, the number of LSTM layers being 1, and the number of LSTM hidden units being 64. This optimal setting has yielded a model with a Pre of 0.8365, a Rec of 0.8281, and an F1 of 0.8317.

VI. DISCUSSION

In this work, we proposed and verified a two-stage strategy to cope with the low-resource NER in the Chinese biomedical domain. Specifically, we fine-tuned two base language models, Word2Vec and BERT, on domain resources to generate word embeddings, which were supplied to the downstream NER task realized by BiLSTM and CRF. Experiment results demonstrated that the proposed strategy can effectively reduce the amount of annotated data required for training while achieve superior performance. The best performing model, W2V+BERT+BiLSTM+CRF with F.T. and D.A., is promising in solving other low-resource NER problems.

This study has the following limitations, which also point out future directions we would pursue. First, the current study only considers two word embedding models, namely, Word2Vec and BERT, while there are other options to be explored. Second, BERT in this work only serves as an embedding model, while its capability in building a predictive model for the NER task has not been fully released. A promising direction is to fine-tune BERT on a domain-specific NER task without using BiLSTM and CRF. Lastly, recent advances have witnessed the prosperity of large language models (LLMs) and their powerfulness in NLP and human-computer interaction. LLMs can be fine-tuned with domain knowledge without losing the base language capability so as to suit a specific learning task such as NER. Several studies have been conducted to verify the ability of an LLM for knowledge extraction. It is appealing to pursue the direction and examine how well LLM can handle the NER task.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Conceptualization and methodology, Meifeng Zhou, Jindian Tan, Song Yang, Haixia Wang, Lin Wang, and Zhifeng Xiao; software, validation, and original draft preparation, Meifeng Zhou, Jindian Tan, and Zhifeng Xiao; funding acquisition, Meifeng Zhou, Jindian Tan, and Lin Wang; review and editing, Haixia Wang and Zhifeng Xiao. All authors have read and agreed to the published version of the manuscript.

SUPPLEMENTAL DATA

Not applicable.

DATA AVAILABILITY STATEMENT

The dataset supporting the conclusions of this article is available at http://openkg.cn/dataset/ yidu-s4k (created on June 2nd 2021).

ACKNOWLEDGMENT

(Meifeng Zhou and Jindian Tan are co-first authors.)

REFERENCES

[1] M. Abdel-Nasser and K. Mahmoud, "Accurate photovoltaic power forecasting models using deep LSTM-RNN," *Neural Comput. Appl.*, vol. 31, no. 7, pp. 2727–2740, Jul. 2019.

[2] A. Abdo, B. Chen, C. Mueller, N. Salim, and P. Willett, "Ligand-based virtual screening using Bayesian networks," *J. Chem. Inf. Model.*, vol. 50, no. 6, pp. 1012–1020, Jun. 2010.

[3] A. Anaby-Tavor, B. Carmeli, E. Goldbraich, A. Kantor, G. Kour, and S. Shlomov, "Do not have enough data? Deep learning to the rescue!" in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 7383–7390.

[4] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," 2019, *arXiv:1903.10676*.

[5] C. L. Bruce, J. L. Melville, S. D. Pickett, and J. D. Hirst, "Contemporary QSAR classifiers compared," *J. Chem. Inf. Model.*, vol. 47, no. 1, pp. 219–227, Jan. 2007.

[6] H. Chen, Y. Ji, and D. Evans, "Finding friends and flipping frenemies: Automatic paraphrase dataset augmentation using graph theory," 2020, *arXiv:2011.01856*.

[7] J. Chen, Z. Yang, and D. Yang, "MixText: Linguistically-informed interpolation of hidden space for semi-supervised text classification," 2020, *arXiv:2004.12239*.

[8] R. Cotterell and K. Duh, "Low-resource named entity recognition with cross-lingual, character level neural conditional random fields," in *Proc. 8th Int. Joint Conf. Natural Lang. Process.*, 2017, pp. 91–96.

[9] X. Dai and H. Adel, "An analysis of simple data augmentation for named entity recognition," 2020, *arXiv:2010.11683*.

[10] Z. Dai, X. Wang, P. Ni, Y. Li, G. Li, and X. Bai, "Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records," in *Proc. 12th Int. Congr. Image Signal Process., Biomed. Eng. Informat.*, Oct. 2019, pp. 1–5.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[12] T. Dreossi, S. Ghosh, X. Yue, K. Keutzer, A. Sangiovanni-Vincentelli, and S. A. Seshia, "Counterexample-guided data augmentation," 2018, *arXiv:1805.06962*.

[13] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, "A survey of data augmentation approaches for NLP," 2021, *arXiv:2105.03075*.

[14] S. Y. Feng, A. W. Li, and J. Hoey, "Keep calm and switch on! Preserving sentiment and fluency in semantic text exchange," 2019, *arXiv:1909.00088*.

[15] X. Feng, X. Feng, B. Qin, Z. Feng, and T. Liu, "Improving low resource named entity recognition using cross-lingual knowledge transfer," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 4071–4077.

[16] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze, "Annotating named entities in Twitter data with crowdsourcing," in *Proc. NAACL HLT Workshop Creating Speech Lang. Data Amazon's Mech. Turk*. 2010, pp. 80–88.

[17] J. Fries, S. Wu, A. Ratner, and C. Ré, "SwellShark: A generative model for biomedical named entity recognition without labeled data," 2017, *arXiv:1704.06360*.

[18] V. Gangal, S. Y. Feng, M. Alikhani, T. Mitamura, and E. Hovy, "NAREOR: The narrative reordering problem," 2021, *arXiv:2104.06669*.

[19] D. Guo, Y. Kim, and A. M. Rush, "Sequence-level mixed sample data augmentation," 2020, *arXiv:2011.09039*.

[20] Z. Hu, Y. Dong, K. Wang, K.-W. Chang, and Y. Sun, "GPT-GNN: Generative pre-training of graph neural networks," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 1857–1867.

[21] K. Kann, K. Cho, and S. R. Bowman, "Towards realistic practices in low-resource natural language processing: The development set," 2019, *arXiv:1909.01522*.

[22] A. Kumar, K. Ahuja, R. Vadapalli, and P. Talukdar, "Syntax-guided controlled generation of paraphrases," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 330–345, Dec. 2020.

[23] C. Sutton, K. Rohanimanesh, and A. McCallum, "Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004, pp. 282–289.

[24] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," 2016, *arXiv:1603.01360*.

[25] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.

[26] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 1, pp. 50–70, Jan. 2022, doi: 10.1109/TKDE.2020.2981314.

[27] M. Li, Y. Zhang, M. Huang, J. Chen, and W. Feng, "Named entity recognition in Chinese electronic medical record using attention mechanism," in *Proc. Int. Conf. Internet Things (iThings) IEEE Green Comput. Commun. (GreenCom) IEEE Cyber, Phys. Social Comput. (CPSCom) IEEE Smart Data (SmartData)*, Jul. 2019, pp. 649–654.

[28] P. H. Li, T. J. Fu, and W. Y. Ma, "Why attention? Analyze BiLSTM deficiency and its remedies in the case of NER," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 8236–8244.

[29] Y. Li, Z. Wang, L. Yin, Z. Zhu, G. Qi, and Y. Liu, "X-Net: A dual encoding–decoding method in medical image segmentation," *Vis. Comput.*, vol. 39, pp. 1–11, Jan. 2021.

[30] A. Liu, J. Du, and V. Stoyanov, "Knowledge-augmented language model and its application to unsupervised named-entity recognition," 2019, *arXiv:1904.04458*.

[31] M. N. Melissourgou and K. T. Frantzi, "Genre identification based on SFL principles: The representation of text types and genres in English language teaching material," *Corpus Pragmatics*, vol. 1, no. 4, pp. 373–392, Dec. 2017.

[32] Y. Ming, "A conditional random field (CRF) based machine learning framework for product review mining," Ph.D. thesis, Dept. Statist., North Dakota State Univ., Fargo, ND, USA, 2019.

[33] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018, *arXiv:1802.05365*.

[34] G. G. Şahin and M. Steedman, "Data augmentation via dependency tree morphing for low-resource languages," 2019, *arXiv:1903.09460*.

[35] F. Shahid, A. Zameer, and M. Muneeb, "Predictions for COVID-19 with deep learning models of LSTM, GRU and bi-LSTM," *Chaos, Solitons Fractals*, vol. 140, Nov. 2020, Art. no. 110212.

[36] J. Shang, L. Liu, X. Ren, X. Gu, T. Ren, and J. Han, "Learning named entity tagger using domain-specific dictionary," 2018, *arXiv:1809.03599*.

[37] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019.

[38] B. Wang, A. Wang, F. Chen, Y. Wang, and C.-C.-J. Kuo, "Evaluating word embedding models: Methods and experimental results," *APSIPA Trans. Signal Inf. Process.*, vol. 8, no. 1, pp. 1–12, 2019.

[39] K. Wang, M. Zheng, H. Wei, G. Qi, and Y. Li, "Multi-modality medical image fusion using convolutional neural network and contrast pyramid," *Sensors*, vol. 20, no. 8, p. 2169, Apr. 2020.

[40] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," 2019, *arXiv:1901.11196*.

[41] Z. Xiao, "Towards a two-phase unsupervised system for cybersecurity concepts extraction," in *Proc. 13th Int. Conf. Natural Comput., Fuzzy Syst. Knowl. Discovery (ICNC-FSKD)*, Jul. 2017, pp. 2161–2168.

[42] Y. Yang, C. Malaviya, J. Fernandez, S. Swayamdipta, R. Le Bras, J.-P. Wang, C. Bhagavatula, Y. Choi, and D. Downey, "Generative data augmentation for commonsense reasoning," 2020, *arXiv:2004.11546*.

[43] H. Yu, X.-L. Mao, Z. Chi, W. Wei, and H. Huang, "A robust and domain-adaptive approach for low-resource named entity recognition," in *Proc. IEEE Int. Conf. Knowl. Graph (ICKG)*, Aug. 2020, pp. 297–304.

[44] S. Zhang, X. Zhang, H. Wang, L. Guo, and S. Liu, "Multi-scale attentive interaction networks for Chinese medical question answer selection," *IEEE Access*, vol. 6, pp. 74061–74071, 2018.

[45] Z. Zhu, M. Zheng, G. Qi, D. Wang, and Y. Xiang, "A phase congruency and local Laplacian energy based multi-modality medical image fusion method in NSCT domain," *IEEE Access*, vol. 7, pp. 20811–20824, 2019.

[46] J. Li, H. Fei, J. Liu, S. Wu, M. Zhang, C. Teng, D. Ji, and F. Li, "Unified named entity recognition as word-word relation classification," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 10, pp. 10965–10973.

[47] H. H. Mao, "A survey on self-supervised pre-training for sequential transfer learning in neural networks," 2020, *arXiv:2007.00800*.

[48] M. Xue, B. Yu, Z. Zhang, T. Liu, Y. Zhang, and B. Wang, "Coarse-to-fine pre-training for named entity recognition," 2020, *arXiv:2010.08210*.

[49] A. Trewartha, N. Walker, H. Huo, S. Lee, K. Cruse, J. Dagdelen, A. Dunn, K. A. Persson, G. Ceder, and A. Jain, "Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science," *Patterns*, vol. 3, no. 4, Apr. 2022, Art. no. 100488.

[50] C. Jia, Y. Shi, Q. Yang, and Y. Zhang, "Entity enhanced BERT pre-training for Chinese NER," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 6384–6396.

[51] S. Gao, O. Kotevska, A. Sorokine, and J. B. Christian, "A pre-training and self-training approach for biomedical named entity recognition," *PLoS ONE*, vol. 16, no. 2, Feb. 2021, Art. no. e0246310.

[52] Y. Wang, Y. Sun, Z. Ma, L. Gao, Y. Xu, and T. Sun, "Application of pre-training models in named entity recognition," in *Proc. 12th Int. Conf. Intell. Hum.-Mach. Syst. Cybern. (IHMSC)*, vol. 1, Aug. 2020, pp. 23–26.

[53] S. Chen, Y. Pei, Z. Ke, and W. Silamu, "Low-resource named entity recognition via the pre-training model," *Symmetry*, vol. 13, no. 5, p. 786, May 2021.

[54] Z. Li, D. Qu, C. Xie, W. Zhang, and Y. Li, "Language model pre-training method in machine translation based on named entity recognition," *Int. J. Artif. Intell. Tools*, vol. 29, Dec. 2020, Art. no. 2040021.

[55] T.-J. Fu, P.-H. Li, and W.-Y. Ma, "GraphRel: Modeling text as relational graphs for joint entity and relation extraction," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1409–1418.

MEIFENG ZHOU received the Ph.D. degree in internal medicine from Zhongshan University. She is currently with the Department of Oncology, Hainan General Hospital.

JINDIAN TAN received the master's degree from Guangdong Medical University. He is currently with the Department of Orthopaedic Surgery, Hainan General Hospital.

SONG YANG received the master's degree in orthopedics from Dalian Medical University. He is currently with the Department of Orthopaedic Surgery, Hainan General Hospital.

HAIXIA WANG received the master's degree in internal medicine from Hainan Medical University. She is currently with the Department of Oncology, Hainan General Hospital.

LIN WANG received the Ph.D. degree in oncology from Southern Medical University. He is currently with the Department of Oncology, Hainan General Hospital.

ZHIFENG XIAO received the B.S. degree in computer science from Shandong University, China, in 2008, and the Ph.D. degree in computer science from the University of Alabama, in 2013. He was an Assistant Professor with Penn State Erie, The Behrend College, from 2013 to 2019, where he is currently an Associate Professor with the Department Computer Science and Software Engineering. His research interests include interdisciplinary AI and cybersecurity, with a particular focus on the areas of AI-powered decision science, accountable systems, and bioinformatics.