Multiple Embeddings Enhanced Multi-Graph Neural Networks for Chinese Healthcare Named Entity Recognition

Lung-Hao Lee, *Member, IEEE*, and Yi Lu

***Abstract*—Named Entity Recognition (NER) is a natural language processing task for recognizing named entities in a given sentence. Chinese NER is difficult due to the lack of delimited spaces and conventional features for determining named entity boundaries and categories. This study proposes the ME-MGNN (Multiple Embeddings enhanced Multi-Graph Neural Networks) model for Chinese NER in the healthcare domain. We integrate multiple embeddings at different granularities from the radical, character to word levels for an extended character representation, and this is fed into multiple gated graph sequence neural networks to identify named entities and classify their types. The experimental datasets were collected from health-related news, digital health magazines and medical question/answer forums. Manual annotation was conducted for a total of 68,460 named entities across 10 entity types (body, symptom, instrument, examination, chemical, disease, drug, supplement, treatment and time) in 30,692 sentences. Experimental results indicated our ME-MGNN model achieved an F1-score result of 75.69, outperforming previous methods. In practice, a series of model analysis implied that our method is effective and efficient for Chinese healthcare NER.**

***Index Terms*—Embedding representation, graph neural networks, information extraction, named entity recognition.**

I. Introduction

NAMED Entity Recognition (NER) is a fundamental task in information extraction that locates the mentions of named entities and classifies them (e.g., person, organization and location) in unstructured texts. The NER task has traditionally been solved as a sequence labeling problem, where entity boundaries and category labels are jointly predicted. Various methods have been proposed to tackle this research problem, including Hidden Markov Models (HMM) [1], Maximum Entropy Markov Models (MEMM) [2] and Conditional Random Field (CRF) [3]. Recently, neural networks have been shown to achieve impressive results. The current state-of-the-art for English NER has been achieved by using LSTM (Long Short-Term Memory)-CRF based networks [4]–[7].

Chinese NER is more difficult to process than English NER. Chinese language is logographic and provides no conventional features like capitalization. In addition, due to a lack of delimiters between characters, Chinese NER is correlated with word segmentation, and named entity boundaries are also word boundaries. However, incorrectly segmented entity boundaries will cause error propagation in NER. For example, in a particular context, a disease entity "思覺失調症" (schizophrenia) may be incorrectly segmented into three words: "思覺" (thinking and feeling), "失調" (disorder) and "症" (disease). Hence, it has been shown that character-based methods outperform word-based approaches for Chinese NER [8]–[10].

In the digital era, healthcare information-seeking users usually search and browse web content in click-through trails to obtain healthcare-related information before making a doctor's appointment for diagnosis and treatment. Web texts are valuable sources to provide healthcare information such as health-related news, digital health magazines and medical question/answer forums. Domain-specific healthcare information includes many proper names, mainly as named entities. For example, "三酸甘油酯" (triglyceride) is a chemical found in the human body; "電腦斷層掃描" (computed tomography; CT) is medical imaging procedure that uses computer-processed combinations of X-ray measurements to produce tomographic images of specific areas of the human body, and "靜脈免疫球蛋白注射" (intravenous immunoglobulin; IVIG) is a kind of treatment for avoiding infections. Healthcare NER is an important and essential task in natural language processing to automatically identify healthcare entities such as symptoms, chemicals, diseases, and treatments for machine reading and understanding. It can be used to research topics for chemical-disease relation extraction [11], knowledge acquisition from Electronic Health Records (EHR) [12], and medical literature summarization [13].

In this paper, we propose a ME-MGNN (Multiple Embeddings enhanced Multi-Graph Neural Networks) model for Chinese healthcare NER. We derive a character representation based on multiple embeddings at different granularities from the radical, character to word levels. Multiple gated graph sequence neural networks along with standard BiLSTM-CRF are then used to identify named entities and classify their types in the healthcare domain. We manually annotated the datasets for Chinese healthcare NER because of a lack of publicly available

Manuscript received August 13, 2020; revised November 23, 2020; accepted December 25, 2020. Date of publication January 1, 2021; date of current version July 20, 2021. This work was partially supported by the Ministry of Science and Technology, Taiwan under Grant MOST 108-2218-E-008-017-MY3. *(Corresponding author: Lung-Hao Lee.)*

The authors are with the Department of Electrical Engineering, National Central University, Taoyuan 32001, Taiwan (e-mail: lhlee@ee.ncu.edu.tw; ericst91159@gmail.com).

Digital Object Identifier 10.1109/JBHI.2020.3048700

benchmark data. Experimental results show that our method significantly outperforms the previous state-of-the-art models on this constructed data. Further empirical analysis and findings confirm the model effects.

The rest of this paper is organized as follows. Section II surveys the related work on Chinese NER models. Section III describes the details of our proposed ME-MGNN model. Section IV introduces experiments for model performance evaluations. Conclusions are finally drawn in Section V.

II. Related Work

The goal of typical NER is to determine named entitiy boundaries and their categories in a given sentence by indicating the names of persons, organizations, locations, times or quantities. A hybrid Chinese NER model that integrates coarse particle part-of-speech features with fine particle word features was proposed to deal with the recognition process [14]. A lexicalized HMM-based approach was applied to assign categories of named entities identified based on a bigram model [15]. A cascaded hybrid model for Chinese NER incorporated human heuristics via Markov logic networks, producing a set of weighted first-order clauses to validate boosting NER hypotheses [16]. A hybrid model that combines Support Vector Machines (SVM) with statistical methods was adopted to improve Chinese NER performance [17]. The Ada boosting algorithm was used to combine a set of weak classifiers together for Chinese NER [18]. The CRF model and maximum entropy model were also used together to recognize Chinese name entities [19].

Recently, deep learning techniques have been widely used for Chinese NER, usually with promising results. A lattice-structured LSTM model was used to encode a sequence of characters as well as all potential words that match a lexicon [10]. Joint modeling NER and constituent parsing in the framework of neural CRF was proposed to improve NER performance [20]. A character-based LSTM-CRF model with radical-level features was proposed for Chinese NER [21]. The ME-CNER model exploited multiple embeddings based character representation to improve Chinese NER performance [22]. The dynamic meta-embedding approach is a data-specific and task-specific method used an attention mechanism to combine features of both character and word granularity in the embedding layer for Chinese NER [23]. An encoding strategy based word-character LSTM was adopted to add word information into the start or the end character of the word to alleviate the influence of word segmentation errors [24]. Chinese NER was also investigated using character-based Convolutional Neural Networks (CNN) with a local attention layer and a Gated Recurrent Unit (GRU) with a global self-attention layer to capture information from adjacent characters and sentence contexts [25]. A neural multi-graph model that captures gazetteer information was proposed to resolve ambiguities [26]. Neural NER via CNN-LSTM-CRF and joint training with word segmentation was developed to capture both local and long-distance contexts [27]. An adversarial transfer learning framework with a self-attention mechanism was used to incorporate task-shared word boundaries from a word segmentation task into a Chinese NER task [28].

Distinct from the long line of work in Chinese NER focusing on formal texts, NER for social media has attracted much attention in recent years. A joint training objective technique for different types of neural embeddings was adopted for Chinese NER in social media, based on Weibo messages [29]. The BiLSTM-CRF model was trained based on character-word mixed embeddings to improve the recognition effectiveness of Chinese NER [30]. Joint training NER and word segmentation with an LSTM-CRF model had also enhanced NER performance for Chinese social media texts [31]. A unified model which can learn out-of-domain information based on domain similarity and in-domain unannotated texts by self-training was proposed for Chinese NER in social media [32].

Clinical NER aims to recognize named entities such as symptoms, diseases and treatments from EHR. A BiLSTM-CRF model with a self-attention mechanism was proposed to integrate part-of-speech labeling information to capture the semantic features of input sentences for Chinese clinical NER [33]. A residual dilated CNN with the CRF was also presented to enhance Chinese clinical NER in terms of computational performance and training time [34].

In summary, we find that the neural computing approaches based on the BiLSTM-CRF also achieved impressive results for Chinese NER. Hence, we follow the state-of-the-art Chinese NER studies to improve character-based BiLSTM-CRF models [10], [22] [26]. Our proposed approach derives an extended character representation based on multiple embeddings at different granularities, fed into adapted multi-graph neural networks, and followed by standard BiLSTM-CRF model for Chinese healthcare NER. In addition, existing publicly available data sets for Chinese NER include the MSRA corpus [35] from formal news texts, and the Weibo corpus [29] from social networking messages. However, no Chinese NER data is available in the healthcare domain, and this study represents the first effort to produce such a data set, will be released as a language resource for further research.

III. ME-MGNN Model for Chinese NER

In this section, we introduce our proposed Multiple Embeddings enhanced Multi-Graph Neural Networks (ME-MGNN) model, for Chinese healthcare NER. The overall structure is shown in Fig. 1. Specifically, our model is comprised of three main parts: 1) an extended character representation based on multiple embeddings, 2) an adapted Gated Graph Sequence Neural Network (GGSNN), and 3) a standard BiLSTM-CRF. We represent a character via embedding techniques at different granularities, ranging from the radical, character to word levels. A directed multi-graph structure is used to explicitly model the interactions of the characters and the words in the dictionaries. Combined with an adapted GGSNN and a standard BiLSTM-CRF, our model can learn a weighted combination of the information based on sentence contexts.

A. Multiple Embeddings Based Feature Extraction

The Chinese language is naturally logographic without conventional linguistic features such as capitalization for use in

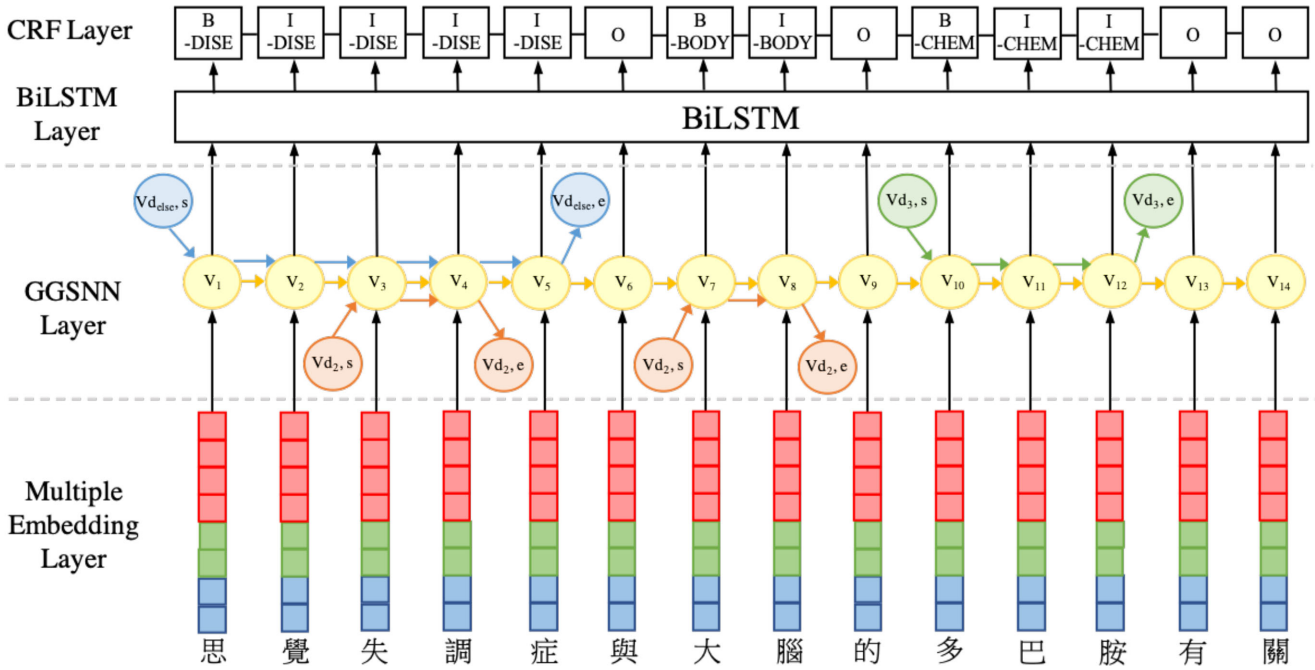

Fig. 1. ME-MGNN model is comprised of three main parts: (a) an embedding layer; (b) an adapted GGSNN layer; and 3) a standard BiLSTM-CRF.

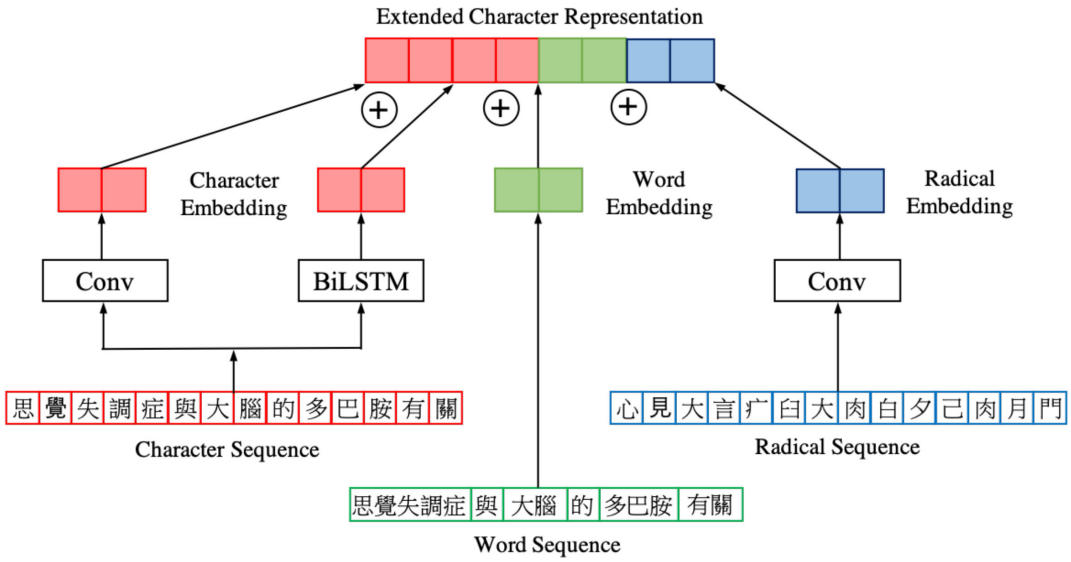

Fig. 2. The neural architecture for multiple embeddings based extended character representation.

NER. A Chinese word can be comprised a single character or multiple characters, with no separation by explicit space delimiters. Also, the meaning of a word can be inferred by its constituent characters. For example, "血漿置換術" (plasmapheresis), a kind of treatment, can be derived by "血漿" (plasma), "置換"(replacement) and "術" (surgical operation). Furthermore, Chinese characters often have hieroglyphic forms. The same radical often indicates similar semantics. For instance, "疾" (illness) and "病" (sickness) both have the same radical "疒", which is also the radical of "症" (disease). We use multiple embedding techniques based feature extraction at different granularities, ranging from the radical, character to word levels, and concatenate them for extended character representation.

Fig. 2 shows the neural architecture for multiple embeddings. Formally, an input sentence is denoted as $s = c_1, c_2, \ldots, c_n$ where c_j denotes the j^{th} character, in which r_j represents the constituent radical of c_j. Also, a sentence s can further be seen as a word sequence $w_1, w_2, \ldots, w_k$, where w_i denotes the i^{th} word in the sentence. For example, a sentence "思覺失調症與大腦的多巴胺有關" (Schizophrenia is related to dopamine in the brain.) is in terms of a character sequence "思" "覺" "失" "調" ... "關" with the corresponding radical sequence of "心" "見" "大" "言"

... “門”. This sentence can be segmented into the word sequence “思覺失調症”(schizophrenia), “與”(and), “大腦” (brain), “的” (of), “多巴胺” (dopamine) and “有關” (related).

1) Character Embedding: Characters are the smallest meaningful units in Chinese. A character can be a morpheme that has its own meaning and can compose a word itself. The meaning of a Chinese multi-character words can be inferred by considering their constituent characters. Hence, we exploit the semantics contained in the characters. First, the character embedding is fed into a convolution layer (denoted as Conv) to extract local n-gram (a consecutive of n character embeddings) features. Then, the character embedding is also fed into the sequential Bidirectional LSTM (BiLSTM) layer to capture global features among the input characters. In the BiLSTM, the backward LSTM is a reversed copy of the forward LSTM, so that we can take full advantage of the forward and backward character sequence information from the input sentence. In this way, we extract the semantic knowledge from both local context and long-term dependency together.

2) Word Embedding: A Chinese word can contain one or multiple characters. For multi-character words, we align the segmented words and the characters, and duplicate the word embedding for its constituent characters. For example, both component character “大” (big) and “腦” (brain) are aligned with a shared word embedding of “大腦” (brain). As a high-level representation, the word embedding is directly used to construct a part of the final representation based on multiple embeddings.

3) Radical Embedding: Chinese characters are naturally hieroglyphic. The same linguistic components in different characters often share the same meaning [21]. For example, characters with the radical “疒” are usually the constituent characters of named entities belonging to the disease category. Similarly, the named entities in the body part category such as “腦” (brain), “肝” (liver), “肺” (lung) and “胸” (chest) have the same radical “肉” (this radical can alternatively be written as “月”). We use a convolution operation in the CNN network to extract local context features of the radical embedding sentence. The use of radicals benefits model generalization by inferring the semantics of the characters that only appear in the testing set but not in the training set for learning.

To form the final extended character representation from these embeddings based feature extraction, individual radical/character/word vectors are initialized by looking up a pre-trained Word2vec embedding [36]. If a radical/character/word is not the vocabulary of the pre-trained radical/character/word embeddings, we initialize its embedding vector with random values. Then, the radical sequences, character sequences and word sequences of the input sentences in the training set are used to extract semantic features. (1) shows the final character representation h_i of a character c_i using a concatenation operation, in which R_i (red color notation in Fig. 2) is the output of the Conv layer and the BiLSTM layer for character embedding, followed by G_i (green color) which is the word embedding vector, and B_i (blue color) is the output of the Conv layer for radical embedding.

$$h_i = R_i \oplus G_i \oplus B_i \tag{1}$$

B. Adapted Gated Graph Sequence Neural Network

We use a directed multi-graph representation to explicitly model the input sentence together with the word information in the dictionaries. Formally, a multi-graph is defined as $G := (V, E, L)$, where V is the set of nodes, E is the set of edges, and L is the set of labels. With n Chinese characters in the input sentence, we first use n nodes to represent the complete sentence, where each Chinese character corresponds to one node. Given a named entity type in the dictionary d_i, there are two special nodes $v_{d_i,s}$ and $v_{d_i,e}$ to the graph which we use to denote the start and end of a named entity matched with a given dictionary. V_s and V_e respectively are sets that contain these special nodes such as $v_{d_i,s}$ and $v_{d_i,e}$. Hence, the entire node set $V = V_c \cup V_s \cup V_e$, where V_c is the set of nodes representing characters. For adjacent Chinese characters, we add one directed edge from the left character to the right one. Each directed edge in E is assigned a label to indicate the type of the connection between nodes. We have the label set $L = \{l_c\} \cup \{l_{d_i}\}_{i=1}^{m}$, where the label l_c is assigned to edges connecting adjacent characters, which are used to model the natural ordering of characters in the input sentence. With m dictionaries in the model, the label l_{d_i} is assigned to all edges that are used to indicate the presence of a text span that matches with a named entity listed in the dictionary d_i that contains i-character words.

For example, as shown in Fig. 1, the input sentence “思覺失調症與大腦的多巴胺有關” (Schizophrenia is related to dopamine in the brain.) consists of 14 Chinese characters and 4 matched named entities in the dictionaries, including “思覺失調症” (schizophrenia, denoting as DISE) belonging to the disease entity category; “失調” (disorder, denoting as SYMP) is labeled as the symptom category, “大腦” (brain, denoting as BODY) is one of the body part entities; and “多巴胺” (dopamine, denoting as CHEM) is the chemical entity type. In addition to the 14 nodes to represent each Chinese character in the complete sentence, we use 8 nodes (total 4 pairs) to capture the information from dictionaries, where each pair corresponds to the start and end of every entity matched by a specific dictionary. Next, we add directed edges between the nodes. For each pair of adjacent Chinese characters, we add one directed edge from the left character to the right one. For each matched named entity from a dictionary, edges are added from the entity start node, connecting through the character nodes composing the entity and ending with the entity end node. For instance, with $c_1c_2c_3c_4c_5$, that is to say “思覺失調症” matched by the named entity type DISE in a dictionary d_{else}, the following edges are constructed: $(v_{d_{else},s}, v_{c_1})$, (v_{c_1}, v_{c_2}), (v_{c_2}, v_{c_3}), (v_{c_3}, v_{c_4}), (v_{c_4}, v_{c_5}) and $(v_{c_5}, v_{d_{else},e})$, where $v_{d_{else},s}$ and $v_{d_{else},e}$ are the start and end nodes for the named entity in the dictionary and each edge is associated with a label indicating its type information (DISE in this case). When edges of the same label overlap, they are merged into a single edge. Such an operation leads to a directed multi-graph representation encoding the character ordering information, the knowledge from multiple named entities information in the given dictionaries, as well as their interactions.

The idea of GGSNN is to generate meaningful outputs or to learn node representations through the use of neural

networks with GRU. Although other graph structure-based neural architectures exist, GGSNN has been found to be better at capturing the local textual information for the Chinese NER task [26]. However, the traditional GGSNN [36] is unable to distinguish edges with different labels. So we adapt GGSNN to learn a weighted combination of the character contexts and dictionary information. We first extend the adjacency matrix A to include edges of different labels. Consequently, we define a set of trainable contribution coefficient to reflect the contribution from each type of structural information, they are the named entities and the character sequence for our task.

In our model, the adapted GGSNN architecture is used to learn the node representation. The initial state $h_v^{(0)}$ of a node v is defined as follows.

$$h_v^{(0)} = \begin{cases} h_d(v) & v \in V_s \cup V_e \\ h_i(v) & v \in V_c \end{cases} \tag{2}$$

where $h_i(v)$ is the final output of our multiple embedding based character representation in the embedding layer for general character nodes in V_c. For special nodes in V_s or V_e, $h_d(v)$ is the randomly initialized representation of matched named entities.

The structural information of the graph is stored in the adjacency matrix A that serves to retrieve the states of neighboring nodes at each step. To adapt to directed multi-graph structure, A is extend to include edges of different labels, $A = [A_{d_1}, A_{d_2}, A_{d_3}, A_{d_4}, A_{d_{else}}]$, where d_i is the dictionary containing named entities composed of i characters, for those named entities consist of more than 4 characters that are included in the dictionary d_{else}. In our idea, a named entity containing more characters is usually highly related to the domain-specific terminology. The contribution coefficients $\alpha_c, \alpha_{d_1}, \ldots, \alpha_{d_m}$ are transformed into weights of edges in A using using the sigmoid function.

$$\begin{aligned} &[w_c, w_{d_1}, w_{d_2}, w_{d_3}, w_{d_4}, w_{d_{else}}] \\ &\quad = \sigma\left([\alpha_c, \alpha_{d_1}, \alpha_{d_2}, \alpha_{d_3}, \alpha_{d_4}, \alpha_{d_{else}}]\right) \end{aligned} \tag{3}$$

Next, the hidden states are updated by GRU. The basic recurrence for this propagation network is:

$$H = \left[h_1^{(t-1)}, \ldots\ldots, h_{|V|}^{(t-1)}\right]^T \tag{4}$$

$$a_v^{(t)} = \left[(HW_1)^T, \ldots\ldots, \left(HW_{|L|}\right)^T\right] A_v^T + b \tag{5}$$

$$z_v^{(t)} = \sigma\left(W^z a_v^{(t)} + U^z h_v^{(t-1)}\right) \tag{6}$$

$$r_v^{(t)} = \sigma\left(W^r a_v^{(t)} + U^r h_v^{(t-1)}\right) \tag{7}$$

$$\hat{h}_v^{(t)} = \tanh\left(W a_v^{(t)} + U\left(r_v^{(t)} \odot h_v^{(t-1)}\right)\right) \tag{8}$$

$$h_v^{(t)} = \left(1 - z_v^{(t)}\right) \odot h_v^{(t-1)} + z_v^{(t)} \odot \hat{h}_v^{(t)} \tag{9}$$

where $h_v^{(t)}$ is the hidden state for node v at time step t, and A_v is the row vector corresponding to node v in the adjacency matrix A. W and U are parameters to be learned. (4) creates the state matrix H at the time step $(t-1)$. (5) shows the information to be propagated through adjacent nodes. (6) to (9) combine the information from adjacent nodes and the current hidden states of the nodes to compute the new hidden state at time step t. After T steps, we have our final state $h_v^{(T)}$ for the node v.

C. Standard BiLSTM-CRF

The learned feature representations of characters in the GGSNN layer $\{h_v^{(T)} \mid v \in V_c\}$ are then fed to a standard BiLSTM-CRF [4], following the character order in the original sentence to produce the output sequence.

IV. Experiments for Performance Evaluation

We carry out an extensive set of experiments to investigate the effectiveness of models for Chinese healthcare NER.

A. Data Sets

Due to the lack of publicly available data sets, we built a corpus for NER in the healthcare domain. We firstly crawled articles from websites that provide healthcare information, online health-related news and medical question/answer forums. We then removed all HTML tags, images, videos and embedded web advertisements and split the remaining texts into several sentences. We randomly selected partial sentences to retain content diversity for manual annotation.

A total of 10 entity types are described and some examples are provided in Table I for healthcare named entity annotation. Three undergraduate students majoring in Chinese language were trained in word segmentation and the named entity tagging task. Inter-annotator agreement is 84.1%. All annotators were asked to discuss differences and seek consensus. Finally, we had 2531 testing sentences with 7305 named entities. Each sentence contains an average of 47.92 characters or 28.67 words, and an average of 2.89 named entities. When the agreement was reached, each annotator was then asked to process the remaining sentences individually. As a result, there were 28161 sentences, each with an average of 49.44 characters or 29.99 words, and 2.17 named entities.

In summary, our constructed corpus includes 30692 sentences with a total of around 1.5 million characters or 91.7 thousand words. After manual annotation, we have 68460 named entities. We then converted the annotation results using the BIO (Beginning, Inside, and Outside) format that is commonly used for NER tasks. The B-prefix before a tag indicates that the character is the beginning of a named entity and I-prefix before a tag indicates that the character is inside a named entity. An O tag indicates that a token belongs to no named entity.

Table II shows detailed statistics of mutually exclusive training and test sets. The entity type distribution in training and test sets is similar. The most common type was BODY (26413 cases or 38.58% of the total), followed by the SYMP (12904 cases), DISE (10079 cases) and CHEM (6834 cases). These 4 most common types of named entities thus accounted for about 82% of the total 68460 entities, with the remaining 6 types accounting for 18%.

B. Settings

We dumped Chinese Wikipedia (zh_TW version on February 3rd, 2020), automatically segmented words, obtained characters

TABLE I
NAME ENTITY TYPES WITH DESCRIPTIONS AND EXAMPLES

Entity Type	Description	Examples
Body	The whole physical structure that forms a person or animal including biological cells, organizations, organs and systems.	"細胞核" (nucleus), "神經組織" (nerve tissue), "左心房" (left atrium), "脊髓" (spinal cord), "呼吸系統" (respiratory system)
Symptom	Any feeling of illness or physical or mental change that is caused by a particular disease.	"流鼻水" (rhinorrhea), "咳嗽" (cough), "貧血" (anemia), "失眠" (insomnia), "心悸" (palpitation), "耳鳴" (tinnitus)
Instrument	A tool or other device used for performing a particular medical task such as diagnosis and treatments.	"血壓計" (blood pressure meter), "達文西手臂" (DaVinci Robots), "體脂肪計" (body fat monitor), "雷射手術刀" (laser scalpel)
Examination	The act of looking at or checking something carefully in order to discover possible diseases.	"聽力檢查" (hearing test), "腦電波圖" (electroencephalography; EEG), "核磁共振造影" (magnetic resonance imaging; MRI)
Chemical	Any basic chemical element typically found in the human body.	"去氧核糖核酸" (deoxyribonucleic acid; DNA), "糖化血色素" (glycated hemoglobin), "膽固醇" (cholesterol), "尿酸" (uric acid)
Disease	An illness of people or animals caused by infection or a failure of health rather than by an accident.	"小兒麻痺症" (poliomyelitis; polio), "帕金森氏症" (Parkinson's disease), "青光眼" (glaucoma), "肺結核" (tuberculosis)
Drug	Any natural or artificially made chemical used as a medicine	"阿斯匹靈" (aspirin), "普拿疼" (acetaminophen), "青黴素" (penicillin), "流感疫苗" (influenza vaccination)
Supplement	Something added to something else to improve human health.	"維他命" (vitamin), "膠原蛋白" (collagen), "益生菌" (probiotics), "葡萄糖胺" (glucosamine), "葉黃素" (lutein)
Treatment	A method of behavior used to treat diseases	"藥物治療" (pharmacotherapy), "胃切除術" (gastrectomy), "標靶治療" (targeted therapy), "外科手術" (surgery)
Time	Element of existence measured in minutes, days, years	"嬰兒期" (infancy), "幼兒時期" (early childhood), "青春期" (adolescence), "生理期" (on one's period), "孕期" (pregnancy)

TABLE II
DETAILED STATISTICS OF HEALTHCARE NER

Entity Type (Tag)	#Train (Ratio%)	#Test (Ratio%)
Body (BODY)	23,240 (38.01%)	3,171 (43.41%)
Symptom (SYMP)	11,423 (18.67%)	1,481 (20.27%)
Instrument (INST)	1,047 (1.71%)	42 (0.58%)
Examination (EXAM)	2,218 (3.63%)	404 (5.53%)
Chemical (CHEM)	6,090 (9.96%)	744 (10.18%)
Disease (DISE)	9,074 (14.84%)	1,005 (13.76%)
Drug (DRUG)	2,146 (3.51%)	79 (1.08%)
Supplement (SUPP)	1,403 (2.29%)	122 (1.67%)
Treatment (TREAT)	2,905 (4.75%)	203 (2.78%)
Time (TIME)	1,609 (2.63%)	54 (0.74%)
Total	61,155 (100%)	7,305 (100%)

and looked up their radicals using Master Ideographs Seeker [37]. Words occurring in Wikipedia at least 5 were retained for embedding training. We pretrained embedding vectors using Word2vec [38], obtaining distinct 863835 word vectors, 13581 character vectors and 3209 radical vectors, in which the dimension of vectors is 50. These vectors were then regarded as initial representation of the corresponding sentence sequences to feed into our multiple embedding layer for final extended character embedding training.

The dictionary items used in the GGSNN layer were collected from KingNet, National Academy for Educational Research (NAER) and Sogou that contained medical and nursing related nouns, medical instruments, medical examinations, ICD-10 disease names and so on. We have 5 dictionaries that can be distinguished according to the number of characters in each included word. A total of 351 words are included in the one-character dictionary (d_1). The two-character dictionary (d_2) contains 7978 words, while the three- (d_3) and four-character (d_4) dictionaries respectively include 19282 and 31444 words. A total of 95362 words composed of more than four characters were contained in the dictionary d_{else}.

The hold-out validation approach was used to tune the hyper-parameters. The validation set is a hold-out set from the training set, that is, 20% of training set kept aside and then used to optimize the hyper-parameters. Based on suggestions from related studies [10], [22] [26], the hyper-parameter values for our model implementation were optimized as follows: embedding size 200; batch size 32; learning rate 0.0005; dropout rate 0.5; LSTM hidden size 200; and number of training epochs 80.

The test set was finally used to compare the performance of the selected models, including the BiLSTM-CRF model [21], the ME-CNER model [22], the Gazetteers model [26], and the Lattice model [10]. We also fine-tuned the BERT transformer [39] since it can be adapted to many types of NLP tasks and usually shows promising results.

We adopted standard precision (P), recall (R) and F1-score, which are the most typical evaluation metrics of NER systems at a character level. If the predicted tag of a character in terms of BIO format was completely identical with the gold standard, that is one of the defined BIO tags, the character in the testing instance was regarded as correctly recognized. Precision is defined as the percentage of named entities found by the NER system that are correct. Recall is the percentage of named entities present in the test set found by the NER system. The F1-score is the harmonic mean of precision and recall. The McNemar's test [40], a statistical hypothesis test for comparing deep learning algorithms, was used to determine whether the performance difference among the compared methods was statically significant.

C. Results

Table III compares the results for various NER models on our constructed healthcare NER datasets. The F1-score obtained by our proposed ME-MGNN model showed significant differences with all compared models (p-value < 0.01). The BiLSTM-CRF [21] architecture (F1-score 71.56) was regarded as the baseline

TABLE III
EXPERIMENTAL RESULTS ON HEALTHCARE NER DATASETS

Model	P	R	F1
BiLSTM-CRF (Dong et al. 2016)	70.38	72.77	71.56
BERT (Devlin et al. 2018)	71.45	**76.36**	73.82
ME-CNER (Xu et al. 2019)	73.68	74.62	74.15
Gazetteers (Ding et al. 2019)	73.00	75.56	74.26
Lattice (Zhang and Yang 2018)	74.69	75.76	75.22
ME-MGNN (ours)	**75.46**	75.76	**75.69**

TABLE IV
ABLATION STUDY FOR EMBEDDING EFFECTS

ME-MGNN Model	P	R	F1
All (character + word + radical)	**75.46**	75.76	**75.69**
- radical	73.50	**76.73**	75.08
- word	73.48	75.10	74.28
- radical - word	73.46	74.54	74.00
- radical - word, Conv ⊕ BiLSTM → Conv	72.75	72.35	72.55
- radical – word, Conv ⊕ BiLSTM → BiLSTM	73.74	72.68	73.20

result. The BERT transformer [39] improved F1-score by 2.26. The ME-CNER model [22] used multiple embeddings for character representation and the BiGRU-CRF network as NER tagger. Compared against the baseline BiLSTM-CRF, ME-CNER enhanced F1-score by 2.59, revealing the effectiveness of multiple embeddings. The Gazetteers method [26] incorporated rich word information to resolve the ambiguities in the multi-digraph architecture. It slightly outperformed the ME-CNER model, indicating that the gazetteers are useful resources for the NER task. The Lattice model [10] used a lattice-structured BiLSTM-CRF network, the state-of-the-art approach achieving the second-best F1-score of 75.22.

Our proposed ME-MGNN model extracted features using multiple embeddings to train multi-graph neural networks for Chinese healthcare NER, providing a best F1-score of 75.69, slightly outperforming the Lattice model with a statistically significant difference (p-value is 0.00146). We also compared the training time spent with the Lattice model. On a Nvidia DGX-1 sever using a V100 GPU with the same system settings, the Lattice model spent about 6 days and 4 hours for training the NER model, which took about 6.43 times longer than ours (about 23 hours). In summary, we can find that our ME-MGNN model is an effective and efficient solution for the Chinese healthcare NER task.

D. In-Depth Analysis

We further discuss the findings of proposed ME-MGNN model in the following aspects.

1) Embedding effects: We conducted an ablation study for ME-MGNN by removing radical and word embeddings and replacing the Conv-BiLSTM network with either a Conv or a BiLSTM network.

TABLE V
PERFORMANCE COMPARISONS FOR ENTITIES APPEAR IN THE TRAINING SET

Entities appear in train	Without Dictionaries			With Dictionaries		
	P	R	F1	P	R	F1
All	79.18	80.47	79.82	79.56	82.00	**80.76**
Partial	72.72	70.04	71.36	74.43	72.67	**73.56**
None	69.32	64.21	66.67	72.04	70.53	**70.73**

TABLE VI
PERFORMANCE COMPARISONS FOR ENTITIES APPEAR IN THE DICTIONARIES

Entities appear in dict.	Without Dictionaries			With Dictionaries		
	P	R	F1	P	R	F1
All	71.11	80.00	75.29	76.74	82.50	**79.52**
Partial	60.00	39.13	47.37	76.47	56.52	**65.00**
None	71.43	62.50	**66.67**	63.64	65.62	64.62

Table IV shows a performance comparison. We first removed radical embedding (marked as "-radical"). It's clear that our ME-MGNN experienced a degree of performance degradation without radical embeddings, confirming that radicals can provide insight into the meanings of similar characters. Removing word embeddings (marked as "-word") produces larger performance loss than removing radical embeddings, indicating that word information is an essential clue to determining named entity boundaries. Only using character embeddings (marked as "-radical -word") reduced F1-score by 1.69. Another observation is that either Conv (marked as *Conv* ⊕ *BiLSTM* → *Conv*) or BiLSTM (marked as *Conv* ⊕ *BiLSTM* → *BiLSTM*) significantly reduced performance. This suggests the Conv-BiLSTM network is indeed more effective in deriving semantic knowledge from local and global features.

2) Entities appear in the training set: To understand the effect of using dictionaries, we conducted some detailed experiments on the test set divided into three groups based on whether or not the entities in a testing sentence appear in the training set. "All" contains those sentences in which all entities can be found in the training set; "Partial" includes sentences that have some of the entities from the training set but not all; "None" means sentences where none of the entities appear in the training set. We compare the performance difference with and without the dictionaries in the GGSNN layer of our ME-MGNN model, with results summarized in Table V. It's obvious that increasing the coverage of entities in the training set can benefit the effectiveness regardless of whether dictionaries are used or not. In addition, incorporating dictionary information can improve performance on all groups, with even the "None" group showing an F1-score improvement of 4.06.

3) Entities appear in the dictionaries: Testing sentences where no entities appear in the training set are further split into three sub-groups: "All", "Partial," and "None", based on whether their entities appear in the dictionaries, with results summarized in Table VI. We can find a significant F1-score improvement of 17.63 for the "Partial" sub-group, followed by a 4.23 improvement for the "All" sub-group with the help

TABLE VII
PERFORMANCE COMPARISONS FOR GENRES OF WRITTEN TEXTS

Genres of written texts	Formal texts			Social media		
	P	**R**	**F1**	**P**	**R**	**F1**
Formal texts	72.49	71.21	**71.84**	71.06	65.31	68.06
Social media	63.01	66.00	64.47	78.45	80.08	**79.26**

of dictionary information. These observations confirm that our proposed model had effectively included word boundaries in the dictionary and correctly identified entity types based on multiple embeddings and multi-graph neural networks. However, for "None" sentences in which entities do not appear in the training set or the dictionaries, including other word information in the dictionaries may be regarded as noises in our model training, slightly decreasing the performance.

4) Genres of written texts: We also analyze the effects of different written genres, so that the training and test set were respectively divided into two groups based on text sources. "Formal texts" contains sentences originating from articles with healthcare information and online health-related news; "Social media" includes sentences collecting from medical question/answer forums. Table VII shows the performance comparisons. Clearly, the same genre of written texts in both training and test set had better F1-scores than those obtained from different written genres. This also confirms the importance of using text types for training models that are identical to testing instances.

E. Error Analysis

Although our ME-MGNN model outperforms previous work, results in the Chinese healthcare NER are still much lower than formal texts in the news domain. For example, the state-of-the-art results of NER on the MSRA dataset (three entity types: persons, locations and organizations in news texts) from the Gazetteers model [26] has an F1-score of 94.4 and the Lattice model [10] has an F1-score of 93.18. Hence, we did more analysis to help understanding the error causes. Table VIII shows a total of five error types at entity-level and their corresponding examples.

1) Contain: The gold entity contains one system recognition result. For example, the disease entity "德國麻疹" (Rubella) contains our NER result "麻疹" (measles). The type is correct, but with incorrect boundaries with different meaning.
2) Be-Contained: The gold entity is contained by the system recognition result. For example, a body part entity "横隔膜" (diaphragm) is incorrectly contained by a non-existent word "指横隔膜".
3) Split: There are gaps in the system recognition results. For example, a kind of symptom "喉嚨痛" (sore throat) is split as two entities: "喉嚨" (throat) as body type and "痛" (pain) as symptom type.
4) Cross: The gold entities partially cross two system recognition results. For example, two adjacent body entities "陽溪穴" (Yang Ravine) and "左右手" (left and right hands) are incorrectly recognized as "穴左" (left side of acupoint) and "右手" (right hand).
5) No-Cross: the gold entity is missing in the system recognition result. There are no common characters between gold one and the recognition. For example, "阻塞" (obstruction) a kind of symptom that is missing in our NER result.

The most common error type "No-Cross" accounted for 72.23% of all errors (1584 out of 2193), with the remainder comprising 4 other error types. In our observations, some healthcare entities are domain-specific such as "血清胺基丙酮酸轉化酶" (Serum Glutamate Pyruvate Transaminase; SGPT), "攝護腺肥大症候群" (Benign Prostatic Hyperplasia; BPH), and "胞漿精子注射" (Intracytoplasmic Sperm Injection; ICSI), which can not be successfully recognized by our model, and are also not included in the dictionaries. Through such error analysis, we confirmed the importance of dictionary usage and word coverage in our model.

F. Discussion

The most obvious difference between traditional and simplified Chinese is the way that the characters look. Simplifying a traditional Chinese character to have simpler and fewer strokes may change the form of its radical. For example, the traditional radical of "髮" (hair) is "髟", while the simplified character "发" (hair) has the radical "又". In addition, a single simplified character may occasionally correspond to two traditional characters. For instance, the simplified character "脏" (dirty or organ) corresponds to the two traditional characters "髒" (dirty) and "臟" (organ). In addition, a word with same meaning may have different common usages, such as a "laser scalpel", which is respectively depicted as "雷射刀" and "激光刀" in traditional and simplified Chinese. In summary, if our constructed data in traditional Chinese can be converted into simplified Chinese carefully, our proposed model can be directly applied to simplified Chinese.

EHR datasets are usually not publicly available due to privacy concerns, and unstructured Chinese texts in EHRs are also relatively rare. For example, doctors usually write summarized medical records in English. Based on our in-depth analysis, the proposed model can be applied to different text types with possible performance losses.

The used dictionaries play an important role in our proposed ME-MGNN model for Chinese healthcare NER. Our error analysis shows that including domain terminologies in the dictionaries produces better results. Without a dictionary, our model will degenerate to a single graph neural network, with relatively poor performance.

Our healthcare NER model can be regarded as the fundamental task for health informatics. For example, it can be applied to de-identify concerned entities such as confidential information in medical records. After successfully recognizing healthcare entities, we can further extract relations among entities to find entity-relationship triples for knowledge base construction and visualization.

TABLE VIII
ERROR TYPES AND THEIR CORRESPONDING EXAMPLES

Contain	Sentence	國際間 德國麻疹$_{\text{DISE}}$ 仍有疫情發生，所以有出國計畫要預先做好安排。 (Rubella is still endemic in many places around the world, so make arrangements in advance if you plan to go abroad.)
	NER	國際間 德國麻疹$_{\text{DISE}}$ 仍有疫情發生，所以有出國計畫要預先做好安排。
Be-Contained	Sentence	肺主脈指 橫膈模$_{\text{BODY}}$ 銜接心臟的部分 (The pulmonary artery is a connected by the diaphragm to the heart.)
	NER	肺主脈 指橫膈模$_{\text{BODY}}$ 銜接心臟的部分
Split	Sentence	喉嚨痛$_{\text{SYMP}}$ 主要是扁桃腺發炎 (Sore throat is mainly due to inflammation of the tonsils.)
	NER	喉嚨$_{\text{BODY}}$ 痛$_{\text{SYMP}}$ 主要是扁桃腺發炎
Cross	Sentence	指壓 陽溪穴$_{\text{BODY}}$ 左右手$_{\text{BODY}}$ 各10分鐘 (Apply pressure to the *yangxi* pressure points on the left and right hands for 10 minutes.)
	NER	指壓 陽溪 穴左$_{\text{BODY}}$ 右手$_{\text{BODY}}$ 各10分鐘
No-Cross	Sentence	鉀離子量若攝取充足，可降低腦血管 阻塞$_{\text{SYMP}}$ 風險。 (Adequate potassium intake can reduce the risk of cerebrovascular obstruction.)
	NER	鉀離子量若攝取充足，可降低腦血管 阻塞 風險。

V. CONCLUSION

We propose a Multiple Embeddings enhanced Multi-Graph Neural Networks (ME-MGNN) model, for Chinese healthcare NER, making the the following contributions:

1) We propose a multiple-embedding-driven multi-graph neural network architecture. Various semantic information in different granularities containing radical, character and word are well exploited. Multi-graph neural networks are adapted to the healthcare NER task. Our method achieved an F1-score of 75.69 that significantly outperforms the previous models on Chinse healthcare datasets.
2) To our best knowledge, our constructed data is the first Chinese NER corpus in the healthcare domain. It includes 30692 sentences with a total around 1.5 million characters or 91.7 thousand words. After manual annotation, we have 68460 named entities across 10 entity types: body, symptom, instrument, examination, chemical, disease, drug, supplement, treatment, and time. We plan to release it as a language resource for further research.

This pilot study is our first exploration in the Chinese healthcare NER task. In future, we will exploit other semantic features and develop advanced models to further improve performance.

ACKNOWLEDGMENT

We sincerely thank all the annotators for their efforts in the named entity tagging task. We also thank the anonymous reviewers for their insightful comments.

REFERENCES

[1] N. Ponomareva, F. Pla, A. Molina, and P. Rosso, "Biomedical named entity recognition: A poor knowledge HMM-based approach," in *Proc. Int. Conf. Appl. Natural Lang. Inf. Syst.*, 2007, pp. 382–387.
[2] H. L. Chieu and H. T. Ng, "Named entity recognition with a maximum approach," in *Proc. 7th Conf. Natural Lang. Learn.*, 2003, pp. 160–163.
[3] C.-H. Wei, R. Leaman, and Z. Lu, "SimConcept: A hybrid approach for simplifying composite named entities in biomedical text," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 4, pp. 1385–1391, Jul. 2015.
[4] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol.*, 2016, pp. 260–270.
[5] X. Ma and E. Hovy, "End-to-end sequence labeling via Bi-directional LSTM-CNNs-CRF," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguist.*, 2016, pp. 1064–1074.
[6] J. Chiu and E. Nichols, "Named entity recognition with bidirectional LSTM-CNNs," *Trans. Assoc. Comput. Linguist.*, vol. 4, pp. 357–370, Jul. 2016.
[7] L. Liu *et al.*, "Empower sequence labeling with task-aware neural language model," in *Proc. 32nd Conf. Artif. Intell.*, 2018, pp. 5253–5260.
[8] J. He and H. Wang, "Chinese named entity recognition and word segmentation based on character," in *Proc. 6th SIGHAN Workshop Chin. Lang. Process.*, 2008, pp. 128–132.
[9] H. Li, M. Hagiwara, Q. Li, and H. Ji, "Comparison of the impact of word segmentation on name tagging for Chinese and Japanese," in *Proc. 9th Int. Conf. Lang. Resour. Eval.*, 2014, pp. 2532–2536.
[10] Y. Zhang and J. Yang, "Chinese NER using lattice LSTM," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguist.*, 2018, pp. 1554–1564.
[11] H. Zhou *et al.*, "Combining context and knowledge representations for chemical-disease relation extraction," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 6, pp. 1879–11889, Nov./Dec. 2019.
[12] S. Perera, C. Henson, K. Thirunarayan, A. Sheth, and S. Nair, "Semantics driven approach for knowledge acquisitions from EHRs," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 2, pp. 515–524, Mar. 2014.
[13] X. Zhang, P. Geng, T. Zhang, Q. Lu, P. Gao, and J. Mei, "Aceso: PICO-guided evidence summarization on medical literature," *IEEE J. Biomed. Health Inform.*, to be published, doi: 10.1109/JBHI.2020.2984704.
[14] Y. Wu, J. Zhao, B. Xu, and H. Yu, "Chinese named entity recognition based on multiple features," in *Proc. Hum. Lang. Technol. Conf. Empirical Methods Natural Lang. Process.*, 2005, pp. 427–434.
[15] G. Fu and K.-K. Luke, "Chinese named entity recognition using lexicalized HMMs," *ACM SIGKDD Explorations Newslett.*, vol. 7, no. 1, pp. 19–25, Jun. 2005.
[16] X. Yu, "Chinese named entity recognition with cascaded hybrid model," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol., Short Papers*, 2007, pp. 197–200.
[17] L. Li, T. Mao, D. Huang, and Y. Yang, "Hybrid models for chinese named entity recognition," in *Proc. 5th Workshop Chin. Lang. Process.*, 2006, pp. 72–78.
[18] X. Yu, M. Carpuat, and D. Wu, "Boosting for chinese named entity recognition," in *Proc. 5th Workshop Chin. Lang. Process.*, 2006, pp. 150–153.

[19] A. Chen, F. Peng, R. Shan, and G. Sun, "Chinese named entity recognition with conditional probabilistic models," in *Proc. 5th Workshop Chin. Lang. Process.*, 2006, pp. 173–176.
[20] R. Wang, X. Xin, W. Chang, K. Ming, B. Li, and X. Fan, "Chinese NER with height-limited constituent parsing," in *Proc. 33rd Conf. Artif. Intell.*, 2019, pp. 7160–7167.
[21] C. Dong, J. Zhang, C. Zong, M. Hattori, and H. Di, "Character-based LSTM-CRF with radical-level features for chinese named entity recognition," in *Proc. Int. Conf. Comput. Process. Oriental Languages*, 2016, pp. 239–250.
[22] C. Xu, F. Wang, J. Han, and C. Li, "Exploiting multiple embeddings for chinese named entity recognition," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, 2019, pp. 2269–2272.
[23] N. Zhang, F. Li, G. Xu, W. Zhang, and H. Yu, "Chinese NER using dynamic meta-embeddings," *IEEE Access*, vol. 7, pp. 64450–64459, May 2019.
[24] W. Liu, T. Xu, Q. Xu, J. Song, and Y. Zu, "An encoding strategy based word-character LSTM for chinese NER," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol.*, 2019, pp. 2379–2389.
[25] Y. Zhu and G. Wang, "CAN-NER: Convolutional attention network for chinese named entity recognition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol.*, 2019, pp. 3384–3393.
[26] R. Ding, P. Xie, X. Zhang, W. Lu, L. Li, and L. Si, "A neural multi-diagraph model for chinese NER with gazetteers," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguist.*, 2019, pp. 1462–1467.
[27] F. Wu, J. Liu, C. Wu, Y. Huang, and X. Xie, "Neural chinese named entity recognition via CNN-LSTM-CRF and joint training with word segmentation," in *Proc. Web Conf.*, 2019, pp. 3342–3348.
[28] P. Cao, Y. Chen, K. Liu, J. Zhao, and S. Liu, "Adversarial transfer learning for chinese named entity recognition with self-attention mechanism," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 182–192.
[29] N. Peng and M. Dredze, "Named entity recognition for chinese social media with jointly trained embeddings," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 548–554.
[30] S. E and Y. Xiang, "Chinese named entity recognition with character-word mixed embedding," in *Proc. 26th ACM Int. Conf. Inf. Knowl. Manage.*, 2017, pp. 2055–2058.
[31] N. Peng and M. Dredze, "Improving named entity recognition for chinese social media with word segmentation representation learning," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguist.*, 2016, pp. 149–155.
[32] J. Xu, H. He, X. Ren, S. Li, and X. Sun, "Cross-domain and semisupervised named entity recognition in chinese social media: A unified model," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 11, pp. 2142–2152, Jul. 2018.
[33] G. Wu, G. Tang, Z. Wang, Z. Zhang, and Z. Wang, "An attention-based BiLSTM-CRF model for chinese clinic named entity recognition," *IEEE Access*, vol. 7, pp. 113942–113949, Aug. 2019.
[34] J. Qin, Y. Zhou, Q. Wang, T. Ruan, and J. Gao, "Chinese clinical named entity recognition using residual dilated convolutional neural network with conditional random field," *IEEE Trans. Nanobiosci.*, vol. 18, no. 3, pp. 306–315, Jul. 2019.
[35] G.-A. Levow, "The third international chinese language processing bakeoff: Word segmentation and named entities recognition," in *Proc. 5th Workshop Chin. Lang. Process.*, 2006, pp. 108–117.
[36] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," in *Proc. 4th Int. Conf. Learn. Representation*, 2016, *arXiv:1511.05493*. [Online]. Available: https://arxiv.org/abs/1511.05493
[37] Master Ideographs Seeker for CNS 11643 Chinese Standard Interchange Code. [Online]. Available: https://www.cns11643.gov.tw/index.jsp?ID=0&SN=&lang=en
[38] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. 27th Conf. Neural Inf. Process. Syst.*, 2013, pp. 1–9.
[39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv: 1810.04805v1*. [Online]. Available: https://arxiv.org/abs/1810.04805
[40] T. G. Dietterich, "Approximate statistical tests for comparing supervised learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998.