A Graph Attention Model for Dictionary-Guided Named Entity Recognition

YINXIA LOU, TAO QIAN, FEI LI, AND DONGHONG JI

Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan 430074, China
School of Computer Science and Technology, Hubei University of Science and Technology, Xianning 437000, China
Department of Computer Science, University of Massachusetts Lowell, Lowell, MA 01854 USA

Corresponding authors: Tao Qian (taoqian@whu.edu.cn) and Donghong Ji (dhji@whu.edu.cn).

This work was supported in part by the Major Project of the National Social Science Foundation of China under Grant 11&ZD189, in part by the National Key Research and Development Program of China under Grant 2017FC1200500, in part by the Major Project of Ministry of Education under Grant 18JZD015, in part by the National Natural Science Foundation of China under Grant 61772378, in part by the Natural Science Foundation Project of Hubei under Grant 2018CFB690, and in part by the Doctoral Initiation Foundation of Hubei University of Science and Technology under Grant 2016-19XB012.

ABSTRACT The lack of human annotations has been one of the main obstacles for neural named entity recognition in low-resource domains. To address this problem, there have been many efforts on automatically generating silver annotations according to domain-specific dictionaries. However, the information of domain dictionaries is usually limited, and the generated annotations may be noisy which poses significant challenges on learning effective models. In this work, we try to alleviate these issues by introducing a dictionary-guided graph attention model. First, domain-specific dictionaries are utilized to extract entity mention candidates by a graph matching algorithm, which can capture word patterns of domain entities. Furthermore, a word-mention interactive graph is leveraged to integrate the semantic and boundary information of entities into their context. We evaluated our model on the biomedical-domain datasets of recognizing chemical and disease entities, namely BC5CDR and NCBI disease corpora. The results show that our model outperforms several state-of-the-art models with different methodologies, such as feature-based models (e.g., BANNER), ensemble models (e.g., CollaboNet), multi-task learning models (e.g., MTM-CW), dictionary-based models (e.g., AutoNER). Moreover, the performance of our model is also comparable with BioBERT that owns huge parameters and needs large-scale pre-training.

INDEX TERMS Graph attention, domain-specific dictionaries, named entity recognition, text mining.

I. INTRODUCTION

Named entity recognition (NER) is a task that extracts entity mentions from texts and classifies them into predefined types, such as person and location in the general domain [1], [2], and disease and chemical in the biomedical domain [3], [4]. NER is a fundamental task in natural language processing (NLP) and bioinformatics, which is crucial for many downstream applications including relation extraction [5] and event extraction [6]. A growing number of studies have investigated this task with various approaches [2], [7]. Recently, neural networks, especially BiLSTM-CRF [2], have become popular approaches and achieved the state-of-the-art performance. However, neural networks need a large amount of annotated data. When the data are not enough (e.g., in low-resource domains), the performance degrades significantly [8].

To address this issue, domain-specific dictionaries have been used to generate silver annotations for NER in low-resource domains. A domain-specific dictionary is a manually curated list of entity names that belong to a specific entity type. The dictionary can be employed to extract entity mentions by string matching [9] or regular expression matching [10]. However, we observed two limitations in such method. The first is that the information in a dictionary has not been fully utilized. For example, considering a new disease name “ulcer of nose” that is not in the dictionary, it cannot be correctly extracted if simple string matching is used. Nevertheless, the disease dictionary contains two other disease entities: “ulcer of anus” and “cancer of nose”.

The associate editor coordinating the review of this manuscript and approving it for publication was Navanietha Krishnaraj K. Rathinam.
FIGURE 1. A word formation graph generated from two disease entities: “ulcer of anus” and “cancer of nose”. It can be used to infer two new disease entities: “ulcer of nose” and “cancer of anus”.

As shown in Figure 1, if we represent their word formation with a graph, we can infer the new disease name “ulcer of nose” via the graph. Such observation motivates us to leverage a word formation graph to extract entity mention candidates.

The second limitation is that most NER errors come from inaccurate boundary detection. As shown in Figure 2, the sentence “Provocation of postural hypotension by nitroglycerin in diabetic autonomic neuropathy” contains two disease entities, but these entities can match seven mention candidates in the dictionary. The widely-used NER model, BiLSTM-CRF [11], may suffer from sparse boundary tags [12], while ambiguous matching will make the problem worse. Liu et al. [13] observed no significant performance improvement when directly utilizing the entity mentions matched with the dictionary as additional training data.

In this work, we exploited the NER task in low-resource domains using domain-specific dictionaries. In particular, we studied disease and chemical entity recognition using related dictionaries in the biomedical domain. We proposed a graph attention BiLSTM-CRF model (GAT-BiLSTM-CRF), whose architecture is illustrated in Figure 2. Our model firstly extracts entity mention candidates from text using a word formation graph, which is designed to capture word formation patterns of entities and model the connections between the words in domain dictionaries. From our observation on the dataset, the words in an entity play different roles in NER. Some words are pivot words, which are usually located in the head or tail of an entity. In contrast, other words are decorative words, which are located in the middle of an entity. Hence, we label each node with using the label scheme “BIES” (i.e., beginning, inside, end and single) [14] according to its position in an entity. Furthermore, to integrate both semantic and boundary information of entity mention candidates into the neural network, we construct a word-mention interactive graph, which models the connections between mentions and words. Afterwards, a graph attention mechanism is used to alleviate the influence of false-negative candidates.

We evaluate our approach on two NER datasets, namely the BC5CDR and NCBI disease corpora [15], [16]. The experimental results show that our model significantly outperforms different kinds of baselines [8]–[10], [14], [17], [18] and even comparable with a pre-trained model, BioBERT [19]. To summarize, our main contributions include:

1) We propose a graph attention BiLSTM-CRF model (GAT-BiLSTM-CRF) using a word-mention interactive graph to integrate the semantic and boundary information of entities.

2) We propose a graph matching algorithm to improve the quality and coverage of matching mention candidates in low-resource domains.

3) Our model achieves the comparable results with BioBERT [19] and beats several strong baselines [8]–[10], [14], [17], [18].

II. RELATED WORK

Early study on NER heavily focuses on feature-based methods [20]. There are also a number of recent neural network approaches applied to NER, such as [21]–[24]. Such neural models are increasingly common in domain NER tasks [25], [26]. However, the performance of these methods degrades significantly due to low-resource in some domains such as biomedical, social media, etc.

FIGURE 2. The architecture of our NER model with the graph attention mechanism and domain-specific dictionary. The left part shows the overall network architecture, including the embedding layer, bidirectional LSTM layer, graph layer and CRF layer. The right part shows the process of extracting entity mention candidates and building word formation graphs.
A. LOW-RESOURCE NER

Many recent technologies have been proposed to tackle the low-resource domain NER, such as pre-trained embedding [19], multi-task [8], multi-models [14], transfer learning [27]. Instead of using the above technologies, we exploit the usage of a domain-specific dictionary for NER. There are attempts on the distantly supervised NER task recently. SwellShark [10] utilizes a collection of dictionaries, ontologies, and, optionally, heuristic rules to generate annotations and predict entity mentions in the biomedical domain without human annotated data. Shang et al. [28] use exact string matching to generate pseudo annotated data and apply high-quality phrases to reduce the number for false negative annotations. However, their annotation quality is limited because of low coverage of dictionaries, which leads to a relatively low recall. Compared with the distantly supervised NER, there are two main differences in our work. One is that a domain-specific dictionary is used to create a word formation graph, which can capture variants of entities and discover as new entity mentions as possible. The other is that the matched entity is used to enrich the word’s representation instead of generating pseudo annotated training data.

B. GRAPH CONVOLUTIONAL NETWORK

There are a number of recent graph convolutional network (GCN) architectures [29]–[31] for learning over graphs. Our work is closely related to the graph attention networks (GAT), introduced by [30], leveraging masked self-attention layers to assign different importance to neighboring nodes. Sui et al. [32] presented collaborative GAT to integrate self-matched lexical knowledge in Chinese NER, which aims to eliminate the ambiguity of Chinese text. Inspired by Sui’s work, we use GAT to incorporate the information of mention candidates and filter noise false negative candidates.

III. APPROACH

In this section, we first introduce the construction of word-mention interactive graph. Then we detail the architecture of GAT-BiLSTM-CRF model for Domain NER task.

A. WORD-MENTION INTERACTIVE GRAPH

As shown on the left in Figure 2, to construct a word-mention interactive graph, we need to create a word formation graph of entities, and extract mention candidates using the graph matching algorithm in advance.

1) WORD FORMATION GRAPH

Give a domain-specific dictionary, the vertex set is all words contained in the dictionary. We regard lemmas of words as vertexes for variants of entities in our experiment. If two words \(i\) and \(j\) are in the same entity, the \((i, j)\)-entry connects with a undirected edge instead of a directed edge in Figure 1, which can expand the coverage of the dictionary. Besides, to capture the position relationship of words in an entity, we introduce a tag set “BIES”, which refer to begin, inside, end and single respectively. For example, in the word formation graph of Figure 2, the word “hypotension”, labeled with “BIES”, may appear in the beginning, middle or end of an entity, and also may be a single entity.

2) MENTION EXTRACTION VIA GRAPH MATCHING

After building a word formation graph, we extract mention candidates using a graph matching algorithm. Give a sentence, all word sub-sequences are scanned to detect whether it is an entity mention. Specifically, by walking in the word formation graph, a word subsequence is regarded as a candidate if it can be totally matched in the graph and be constrained with node’s tags. The detail of the matching algorithm is shown in Algorithm 1.

Alg 1: Mention Candidate Detection Algorithm

Input: a word sequence \(S = (w_1, w_2, \ldots, w_m)\), a word formation graph \(G(v, e, l)\).

Output: \(S\) is a mention or not.

1: Stemming every word in \(S\).
2: if \(\exists v_m \notin G.v\) then
3: return False
4: if \(\text{len}(S) = 1\) and “S” \(\in l_{w_1}\) then
5: return True
6: if “B” \(\notin l_{w_1}\) OR “E” \(\notin l_{w_m}\) then
7: return False
8: for \((w_i, w_{i+1})\) in \(S\) do
9: if \((v_{w_i}, v_{w_{i+1}}) \notin G.e\) then
10: return False
11: return True

3) WORD-MENTION INTERACTIVE GRAPH CONSTRUCTION

After extraction all possible entity mention candidates, we need to construct a word-mention interactive graph for integrating the formation of mentions into their sentences. The vertex set of a graph is made up of the words in the input sentence and mention candidate extracted from it. For example, as shown in Figure 2, the vertex set is \(V = V_{\text{word}}\{\text{Provoction, of, neuropathy}\} \cup V_{\text{candidates}}\{\text{postural hypotension, hypotension, neuropathy}\}\). To represent the edge set, the adjacency matrix needs to be introduced. The elements of the adjacency matrix indicate whether pairs of vertices are adjacent or not in the graph. As shown in Figure 2, if a mention \(i\) contains a word \(j\), the \((i, j)\)-entry will be assigned a value of 1, and if a word \(i\) is the neighbor of word \(j\), the \((i, j)\)-entry will also assign a value of 1. Intuitively, word pairs entry can capture contextual information, and word-mention entry can capture the semantic and boundaries information of mention of the entry.

B. MODEL

The overall architecture of our proposed model is shown in Figure 2. It includes an embedding layer, a BiLSTM layer, a graph layer, and a CRF layer. The embedding layer and the
BiLSTM layer are to encode the contextual information of the sentence. The graph layer is based on GAT [30] for modeling over word-mention interactive graphs. Finally, a standard CRF model is used for decoding labels.

1) ENCODING
The input of the model is a sentence and all mention candidates that match consecutive sub-sequences of the sentence. We denote the sentence as \( s = \{w_1, w_2, \ldots, w_n\} \), where \( w_i \) is the \( i \)-th word and denote the mention candidates as \( l = \{m_1, m_2, \ldots, m_m\} \). A word is represented by concatenating its word embedding and its character representation:

\[
X_i = [e^w(w_i); x^c_i]
\]

where \( e^w \) denotes a word embedding lookup table and \( x^c_i \) denotes its character representation. Following [2], [33], we adopt Bi-LSTM for character encoding.

To capture contextual information, word representations \( X_1, X_2, \ldots, X_n \) are fed into a Bidirectional LSTM layer. Finally, the two hidden states are concatenated for a final representation.

\[
\overrightarrow{h}_i = \text{LSTM}(X_i, \overrightarrow{h}_{i-1}) \quad (2)
\]
\[
\overleftarrow{h}_i = \text{LSTM}(X_i, \overleftarrow{h}_{i-1}) \quad (3)
\]
\[
h_i = \overrightarrow{h}_i \oplus \overleftarrow{h}_i \quad (4)
\]

Each mention \( m_i \) is represented as a semantic vector, which denotes as \( w_{m_i} \). We concatenate the contextual representation and the word embeddings as the output of this layer, denoting as \( \text{Node}_i \).

\[
\text{Node}_i = [h_1, h_2, \ldots, h_n, w_{m_1}, \ldots, w_{m_m}] \quad (5)
\]

2) GRAPH ATTENTION NETWORK
We use Graph Attention Network (GAT) to model over word-mention interactive graph. In an M-layer GAT, the input of \( j \)-th layer is a set of node features, \( NF^j = f_1, f_2, \ldots, f_N \), together with an adjacency matrix \( A \), where \( N \) denotes the number of nodes and \( F \) is the dimension of features at \( j \)-th layer. The output of \( j \)-th layer is a new set of node features, \( NF^{j+1} = \{f'_1, f'_2, \ldots, f'_N\} \).

A GAT operation with \( K \) independent attention head can be written as:

\[
f_i' = \oplus_{k=1}^K \sigma \left( \sum_{j \in N_i} \alpha^k_{ij} W_k f_j \right) \quad (6)
\]
\[
\alpha^k_{ij} = \frac{\exp(\text{LeakyReLU}(a^T [W_k f_i] \oplus W_k f_j))}{\sum_{k \in N_i} \exp(\text{LeakyReLU}(a^T [W_k f_i] \oplus W_k f_j))} \quad (7)
\]

where \( \oplus \) denotes concatenation operation, \( \sigma \) is a non-linear activation function, \( N_i \) is the neighborhood of node \( i \) in the graph, \( \alpha^k_{ij} \) is the attention coefficients, \( W_k \in \mathbb{R}^{F \times F} \), and \( a \in \mathbb{R}^{2F} \) is a single-layer feed-forward neural network. Note that, the dimension of the output \( f'_i \) is \( KF' \). At the last layer, averaging will be adopted, and the dimension of final output features is \( F' \).

\[
f_i^{\text{final}} = \sigma \left( \frac{1}{K} \sum_{k=1}^K \sum_{j \in N_i} \alpha^k_{ij} W_k f_j \right) \quad (8)
\]

The output of all node features is denoted as \( G \).

\[
G = \text{GAT}(\text{Node}_i, A^C) \quad (9)
\]

where \( G \in \mathbb{R}^{F' \times (n+m)} \). We keep the first \( n \) columns of these matrices and discard the last \( m \) columns, because only word representations are used to decode labels.

\[
Q = G[; 0 : n], \quad k \in \{1, 2, 3\} \quad (10)
\]

Finally, the input of CRF layer is denoted as:

\[
R = W_1 H + W_2 Q \quad (11)
\]

where \( W_1 \) and \( W_2 \) are trainable matrices. The new represent \( R \) for sentence integrate the contextual information from the BiLSTM layer and the semantic and boundary information of the entities from the GAT layer.

3) DECODING AND TRAINING
A standard CRF layer is used to capture the dependencies between successive labels. The input of the CRF layer is \( R = \{r_1, r_2, \ldots, r_n\} \), and the conditional probability of the golden tag sequence \( y = \{l_1, l_2, \ldots\} \) is

\[
p(y|s) = \frac{\exp(\sum_{i \in \text{CRF}}(W_{i,1} r_i + T_{i-1,i}^{\text{CRF}}))}{\sum_{y'} \exp(\sum_{i \in \text{CRF}}(W_{i,1} r_i + T_{i-1,i}^{\text{CRF}}))} \quad (12)
\]

Here \( y' \) is an arbitrary label sequence, \( W_{i,1}^{\text{CRF}} \) is used for modeling emission potential for the \( i \)-th word in the sentence, and \( T_{i-1,i}^{\text{CRF}} \) is the transition matrix storing the score of transferring from \( l_i \) to \( l_i \).

We use the Viterbi algorithm to find the highest scored label sequence. Given a manually annotated training data \( \{(s_1, y_1), (s_2, y_2), \ldots, (s_n, y_n)\} \), sentence-level log-likelihood loss with L2 regularization is used to train the model. The loss function is defined as:

\[
L = - \sum_{i=1}^N \log(P(y_i|s_i)) + \lambda \parallel \theta \parallel_2 \quad (13)
\]

where \( \lambda \) denotes the L2 regularization parameter and \( \theta \) is the all trainable parameters set.

IV. EXPERIMENTS
In this section, we carry out the experiments to investigate the effectiveness of our approach.

A. EXPERIMENTAL SETTINGS
1) DATASETS
We evaluate the performances of our models on two datasets. The BioCreative V Chemical Disease Relation (BC5CDR) dataset [16] and the NCBI disease dataset [15]. BC5CDR dataset consists of 1,500 PubMed abstracts, which has
been equally separated into training set (500), development set (500) and test set (500). The dataset contains 12,852 disease and 15,935 chemical mentions. The NCBI disease dataset consists of 793 PubMed abstracts, which has been separated into training set (593), development set (100), and test set (100). The dataset contains 6,892 disease mentions. We use two types of dictionaries to recognize disease and chemical entities, respectively. The overall statistics on the datasets are provided in Table 1.

| Datasets   | Entity type | Articles | Train | Dev | Test |
|------------|-------------|----------|-------|-----|------|
| BC5CDR     | Disease     | 1,500    | 4,182 | 4,244 | 4,424 |
| BC5CDR     | Chemical    | 1,500    | 5,203 | 5,347 | 5,385 |
| NCBI       | Disease     | 793      | 5,145 | 787  | 960  |

2) DOMAIN-SPECIFIC DICTIONARY
We build two types of dictionaries: a disease dictionary for disease mention recognition and a chemical dictionary for chemical mention recognition. We construct and process dictionaries and high-quality phases in the same way as in [9]. We combine the MeSH database¹ and the Comparative Toxicogenomic Database (CTD) Chemical and Disease vocabularies² as dictionaries. The phrases are mined from titles and abstracts of PubMed papers using the phrase mining method proposed in [9]. As suggested in [9], we apply tailored dictionaries to reduce false positive matching.

3) METRIC
Following the standard setting, we evaluate the methods using micro-averaged F1 score and also report the precision (P) and recall (R) in percentage. All the reported scores are averaged over ten different runs.

4) HYPER-PARAMETERS SETTINGS
Table 2 shows the values of hyperparameters in our models. We do not use hand-crafted features and only words and characters are considered as the inputs. We stop the training when we find the best results in the development set.

| Parameter          | Value | Parameter          | Value |
|--------------------|-------|--------------------|-------|
| Char emb size      | 30    | Word emb size      | 200   |
| Char hidden        | 50    | LSTM hidden size   | 200   |
| CNN window         | 3     | word dropout rate  | 0.6   |
| LSTM layer         | 1     | GAT layer          | 2     |
| Batch size         | 30    | GAT head           | 5     |
| Learning rate /r   | 0.001 | GAT dropout        | 0.4   |

5) PRE-TRAINED WORD EMBEDDINGS
We collected PubMed abstracts by searching the names in the original dictionaries. Then the preprocessed text was fed into FastText³ to generate 200-dimensional word embeddings on 8 billion words. Word embeddings are fine-tuned during NER training. We also use our trained word embeddings to initialize the representations of mention candidates. Character embeddings are initialized with uniform samples from $[-\sqrt{\frac{3}{dim}}, +\sqrt{\frac{3}{dim}}]$, where we set $dim = 30$.

B. BASELINES
To further demonstrate the effectiveness of our model, we compare our model with previous state-of-the-art methods for disease and chemical mention recognition. The proposed models are as follows:

- **BANNER** [17] is a CRF-based recognition model using a rich feature set.
- **TaggerOne** [18] uses a semi-Markov linear classifier while performing normalization and NER during training and prediction.
- **CollaboNet** [14] consists of multiple BiLSTM-CRF models for biomedical named entity recognition. The target model can obtain information from other collaborator models trained on different datasets in biomedical.
- **BioBERT** [19] is a domain specific language representation model pre-trained on large-scale biomedical corpora, including 18b words, which is based on the BERT architecture.
- **MTM-CW** [8] is a multi-task BiLSTM-CRF learning framework for BioNER, which shared character- and word-level information among relevant biomedical entities across differently labeled corpora.
- **SwellShark** [10] is a distantly supervised method designed for the biomedical domain, especially on the BC5CDR and NCBI-Disease datasets. It requires human efforts to customize regular expression rules and hand tune candidates.
- **AutoNER** [9] is a recent state-of-the-art distantly supervised method. After dictionary matching, it trains a BiLSTM-Softmax architecture.

C. RESULTS
We present recall, precision, and F1 scores on all datasets in Table 3. Since MTM-CW, SwellShark and AutoNER did not measure on recognizing disease and chemical entities on BC5CDR separately, we run their models on BC5CDR-Chemical and BC5CDR-Disease datasets for a fair comparison with other models. The scores are denoted with asterisks. From Table 3, one can find our model achieves the best performance on BC5CDR-Disease, and the second scores on BC5CDR-chemical and NCBI-disease, comparing with the baselines using the hot technologies, including feature engineering, multiple models, multi-task, pre-training, distantly supervision.

For a fair comparison with the BioBERT model [19], we also use the word representations generated by BioBERT to compute the word or mention representations of our

1. https://www.nlm.nih.gov/mesh/
2. http://ctdbase.org/downloads/
3. https://github.com/facebookresearch/fastText
TABLE 3. Main results. The results marked with the asterisks are based on our reproduction. A bold number denotes the highest value in that column.

| Methods            | Names       | BC5CDR-Disease | NCBI-Disease | BC5CDR-Chemical |
|--------------------|-------------|----------------|--------------|-----------------|
|                    |             | P   | R   | F   | P   | R   | F   | P   | R   | F   |
| Feature-based      | BANNER      | $82.00$ | $79.50$ | $80.70$ | $93.20$ | $84.00$ | $84.40$ | $82.20$ | $77.50$ | $79.80$ |
|                    | TaggerOne   | $85.20$ | $80.20$ | $82.60$ | $85.10$ | $80.80$ | $82.90$ | $94.20$ | $88.80$ | $91.40$ |
| Ensemble-based     | CollaboNet  | $85.61$ | $82.61$ | $84.08$ | $85.48$ | $87.27$ | $86.36$ | $94.26$ | $92.38$ | $93.31$ |
| Multi-task learning| MTM-CW      | *83.73* | *82.95* | *83.32* | *85.86* | *86.42* | *86.14* | *93.10* | *89.57* | *91.30* |
| Dictionary-based   | SwellShark  | *80.80* | *82.93* | *80.72* | *85.36* | *81.60* | *80.10* | *81.80* | *87.60* | *83.24* |
|                    | AutoNER     | *83.14* | *82.02* | *82.58* | *76.98* | *74.65* | *75.78* | *87.51* | *84.13* | *85.79* |
| Ours               | GAT-BiLSTM-CRF | 88.36 | 87.92 | 88.13 | 87.54 | 87.02 | 87.28 | 91.92 | 91.35 | 92.04 |

Contextualized Embeddings

| Methods | BC5CDR-Disease | NCBI-Disease | BC5CDR-Chemical |
|---------|----------------|--------------|-----------------|
|         | P   | R   | F   | P   | R   | F   | P   | R   | F   |
| BioBERT | $85.86$ | $87.27$ | $86.56$ | $89.04$ | $89.69$ | $89.36$ | $93.27$ | $93.61$ | $93.44$ |
| GAT-BiLSTM-CRF (BioBERT) | $90.28$ | $89.62$ | $89.95$ | $89.10$ | $89.72$ | $89.41$ | $93.31$ | $93.70$ | $93.50$ |

D. DISCUSSION

1) EFFECTIVENESS OF THE DICTIONARY

The dictionary is used to extract entity mention candidates from the datasets. The extraction algorithm greatly affects the quality of extracted mentions. We use precision and recall evaluating the accuracy and the coverage of mention candidates on the training set. We compare our graph matching algorithm with the exact string matching algorithm in the experiments. As shown in Table 4, using graph matching algorithm is able to boost the recall (coverage) by a large margin, increasing 18.72%, 21.63%, 14.86% on the BC5CDR-Disease, NCBI-Disease and NCBI-Disease datasets respectively, while greatly reducing the precision (which is inevitable due to the more additional noise introduced by our graph matching algorithm).

According to the Table 4, the word-mention interactive graph constructed by the graph matching algorithm contains more noises than the graph constructed by the exact string matching algorithm. We evaluate the effectiveness of two interactive graphs on the GAT-BiLSTM-CRF. Table 5 shows the performance on development sets and test sets on the three datasets. It can be seen that, despite more noise introduced by the graph matching algorithm, the graph matching algorithm helps GAT-BiLSTM-CRF achieve better recall with slightly decrease of precise. There are two possible reasons. One is that the coverage of entity mention candidates on test data play an important effect on model performance. Intuitively, the higher the coverage of matched entities on test data, the better the performance. To investigate the effect of the coverage on our model, we manually construct six dictionaries by randomly removing or adding entity names from the train and test data, with coverage between 0% and 100%. We show the F1-scores on the BC5CDR-disease, BC5CDR-chemical datasets, in Figure 3. We find that the F1-score is 79.16% when the coverage is 0% on the BC5CDR-disease, while the F1-score is 93.78% when the coverage is 100% the BC5CDR-chemical. In both solid, the higher coverage, the F1-score is larger. Moreover, we also see that the F1-scores both our model GAT-BiLSTM-CRF and BiLSTM-CRF are equal when the coverages are 57.63% and 62.17% for BC5CDR-disease and BC5CDR-chemical, respectively. This is presumably due to a higher false positive rate introduced by matching mentions. The other reason is that the graph attention module can effectively alleviate the impact of noises, which is analyzed in the next section.

2) EFFECTIVENESS OF THE MODEL

The graph attention network(GAT) is the core module in the GAT-BiLSTM-CRF model, which performs the functions that integrate extracted mentions information and filter those
TABLE 5. The results of GAT-BiLSTM-CRF using different matching methods.

| Methods      | BC5CDR-disease | NCBI-disease | BC5CDR-chemical |
|--------------|----------------|--------------|-----------------|
|              | P   | R   | F   | P   | R   | F   | P   | R   | F   |
| Exact matching | Dev set | 90.15 | 83.79 | 86.15 | 88.52 | 81.81 | 85.03 | 92.97 | 87.78 | 90.30 |
|              | Test set | 90.32 | 84.23 | 87.17 | 88.84 | 85.74 | 87.26 | 93.61 | 88.57 | 91.02 |
| Graph matching | Dev set | 90.14 | 87.61 | 88.85 | 88.49 | 86.91 | 87.69 | 92.71 | 93.91 | 93.31 |
|              | Test set | 90.21 | 89.58 | 89.89 | 88.77 | 88.12 | 88.45 | 93.01 | 93.65 | 93.33 |

TABLE 6. Ablation analysis for our model. GCT-BiLSTM-CRF uses the graph convolutional network (GCT) [34] instead of the graph attention network.

| Model                  | BC5CDR-disease | NCBI-disease | BC5CDR-chemical |
|------------------------|----------------|--------------|-----------------|
|                        | P   | R   | F   | P   | R   | F   | P   | R   | F   |
| GAT-BiLSTM-CRF         | 90.21 | 89.58 | 89.89 | 88.77 | 88.12 | 88.45 | 93.01 | 93.65 | 93.33 |
| BiLSTM-CRF             | 86.56 | 82.25 | 84.35 | 86.12 | 84.37 | 85.24 | 91.82 | 88.12 | 89.91 |
| GCT-BiLSTM-CRF        | 88.16 | 88.91 | 88.33 | 86.54 | 87.51 | 87.02 | 91.87 | 92.61 | 92.24 |

TABLE 7. Comparison of the ability to predict new mentions between BiLSTM-CRF and GAT-BiLSTM-CRF.

3) CASE STUDY

In order to verify the ability of our model to discover new mentions, we compared our model GAT-BiLSTM-CRF with BiLSTM-CRF in the disease recognition task. We randomly choose four examples from BC5CDR-disease test set as shown in Table 7. Column 1 represents sentences containing new mentions, which are neither in BC5CDR-disease training set nor in the disease dictionary. Column 2 represents gold mentions in the sentences. Column 3-4 represent the predicted results by BiLSTM-CRF and GAT-BiLSTM-CRF, respectively. From Table 7, we can see that both models can correctly predict the mentions in example 1 and 2, in which mentions have obvious indicator features “s” and “of”. But BiLSTM-CRF can only detect the part of gold mentions in the example 3 and 4, since “glomerular” and “duct destruction” have appeared in the training set but the gold mentions do not. While our model can correctly discover the mentions because our model can extract the gold mentions from sentences using the graph matching algorithm.

V. CONCLUSION

In this paper, we propose a graph attention-based BiLSTM-CRF model for low-resource domain NER with a domain-specific dictionary. A domain-specific dictionary is used to extract entity mention candidates by the graph matching-based entity mention extraction algorithm. Using the graph matching algorithm is able to boost the coverage by a large margin. Although it produces more noises, the graph
attention module can effectively alleviate the impact of noises. The experiments on disease and chemical datasets in the biomedical domain show that our model has complementary strengths to the art-of-art models, and both word formation graph and word-mention interactive graph is effective. In the future, we plan to further extract and incorporate dictionary knowledge from other learning principles and empower more sequence labeling tasks on low-resource domains.

REFERENCES

[1] E. F. Sang and F. De Meulder, “Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition,” 2003, arXiv:cs.cs/0306050. [Online]. Available: https://arxiv.org/abs/cs.cs/0306050

[2] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, “Neural architectures for named entity recognition,” 2016, arXiv:1603.01360. [Online]. Available: http://arxiv.org/abs/1603.01360

[3] L. Luo, Z. Yang, P. Yang, X. Zhang, L. Wang, H. Lin, and J. Wang, “An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition,” Bioinformatics, vol. 34, no. 8, pp. 1381–1388, Apr. 2018.

[4] Y. Lou, Y. Zhang, T. Qian, F. Li, S. Xiong, and D. Ji, “A transition-based joint model for disease named entity recognition and normalization,” Bioinformatics, vol. 33, no. 15, pp. 2363–2371, Aug. 2017.

[5] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, “Distant supervision for relation extraction without labeled data,” in Proc. Joint Conf. 47th Annua. Meeting ACL, 4th Int. Joint Conf. Natural Lang. Process. (AFNLP ACL-IJCNLP), 2009, pp. 1003–1011.

[6] T. H. Nguyen and R. Grishman, “Graph convolutional networks with argument-aware pooling for event detection,” in Proc. 32nd AAAI, 2018, pp. 1003–1011.

[7] D. Nadeau and S. Sekine, “A survey of named entity recognition and classification,” Lingvisticae Investigationes, vol. 30, no. 1, pp. 3–26, 2007.

[8] X. Wang, Y. Zhang, X. Ren, Y. Zhang, M. Zitnik, J. Shang, C. Langlotz, and J. Han, “Cross-type biomedical named entity recognition with deep multi-task learning,” Bioinformatics, vol. 35, no. 10, pp. 1745–1752, May 2019.

[9] 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. [Online]. Available: http://arxiv.org/abs/1809.03599

[10] 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. [Online]. Available: http://arxiv.org/abs/1704.06360

[11] X. Ma and E. Hovy, “End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF,” 2016, arXiv:1603.01354. [Online]. Available: http://arxiv.org/abs/1603.01354

[12] J. Li, D. Ye, and S. Shang, “Adversarial transfer for named entity boundary detection with pointer networks,” in Proc. Int. Joint Conf. Artif. Intell. AAAI Press, Aug. 2019, pp. 5053–5059.

[13] S. Liu, Y. Sun, B. Li, W. Wang, and X. Zhao, “HAMNER: Headword amplified multi-span distantly supervised method for domain specific named entity recognition,” 2019, arXiv:1912.01731. [Online]. Available: http://arxiv.org/abs/1912.01731

[14] W. Yoon, C. H. So, J. Lee, and J. Kang, “CollaboNet: Collaboration of deep neural networks for biomedical named entity recognition,” BMC Bioinf., vol. 20, no. S10, p. 249, May 2019.

[15] R. I. Dogan, R. Leaman, and Z. Lu, “NCBI disease corpus: A resource for disease name recognition and concept normalization,” J. Biomed. Informat., vol. 47, pp. 1–10, Feb. 2014.

[16] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, and Z. Lu, “BioCreative V CDR task corpus: A resource for chemical disease relation extraction,” Database, vol. 2016, Jan. 2016, Art. no. bav068.

[17] R. Leaman and G. Gonzalez, “BANNER: An executable survey of advances in biomedical named entity recognition,” in Proc. Biocomputing, Dec. 2007, pp. 652–663.

[18] R. Leaman and Z. Lu, “TaggerOne: Joint named entity recognition and normalization with semi-Markov models,” Bioinformatics, vol. 32, no. 18, pp. 2839–2846, Sep. 2016.

4 All codes and resources are released at the website https://github.com/yx100/GAT-BiLSTM-CRF.
FEI LI received the B.S. degree in computer science and technology from Jilin University, Changchun, in 2005, the M.S. degree in computer application technology from Ningbo University, Ningbo, in 2008, and the Ph.D. degree in computer software and theory from Wuhan University, Wuhan, in 2017. He is currently a Postdoctoral Research Associate with the University of Massachusetts. His research interests include natural language processing, and deep learning and their application in the biomedical domain.

DONGHONG JI received the B.S., M.S., and Ph.D. degrees from the Computer School, Wuhan University, China, in 1988, 1991, and 1995, respectively. He was a Postdoctoral Research Fellow with Tsinghua University, from 1995 to 1998. From 1998 to 2008, he was a Research Scientist with the Institute for Infocomm Research of Singapore. He is currently a Professor and a Ph.D. Supervisor with the School of Cyber Science and Engineering, Wuhan University. His research interests include natural language processing, machine learning, and data mining.