A Neural Multi-digraph Model for Chinese NER with Gazetteers

Ruixue Ding[1], Pengjun Xie[1], Xiaoyan Zhang[2], Wei Lu[3], Linlin Li[1] and **Luo Si[1]**

[1]Alibaba Group
[2]Beihang University, China
[3]Singapore University of Technology and Design
{ada.drx,chengchen.xpj,linyan.lll,luo.si}@alibaba-inc.com
xiaoyan.loic@gmail.com, luwei@sutd.edu.sg

Abstract

Gazetteers were shown to be useful resources for named entity recognition (NER) (Ratinov and Roth, 2009). Many existing approaches to incorporating gazetteers into machine learning based NER systems rely on manually defined selection strategies or handcrafted templates, which may not always lead to optimal effectiveness, especially when multiple gazetteers are involved. This is especially the case for the task of Chinese NER, where the words are not naturally tokenized, leading to additional ambiguities. To automatically learn how to incorporate multiple gazetteers into an NER system, we propose a novel approach based on graph neural networks with a multi-digraph structure that captures the information that the gazetteers offer. Experiments on various datasets show that our model is effective in incorporating rich gazetteer information while resolving ambiguities, outperforming previous approaches.

1 Introduction

Previous work (Ratinov and Roth, 2009) shows that NER is a knowledge intensive task. Background knowledge is often incorporated into an NER system in the form of named entity (NE) gazetteers (Seyler et al., 2018). Each gazetteer is typically a list containing NEs of the same type. Many earlier research efforts show that an NER model can benefit from the use of gazetteers (Li et al., 2005). On the one hand, the use of NE gazetteers alleviates the need of manually labeling the data and can handle rare and unseen cases (Wang et al., 2018). On the other hand, resources of gazetteers are abundant. Many gazetteers have been manually created by previous studies (Zamin and Oxley, 2011). Besides, gazetteers can also be easily constructed from knowledge bases (e.g., Freebase (Bollacker et al., 2008)) or com-
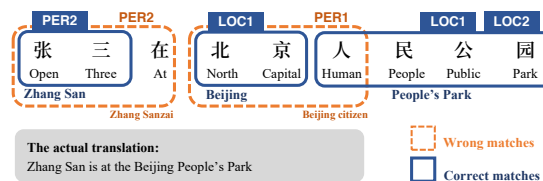

Figure 1: Example of Entity Matching

mercial data sources (e.g., product catalogues of e-commence websites).

While such background knowledge can be helpful, in practice the gazetteers may also contain irrelevant and even erroneous information which harms the system's performance (Chiu and Nichols, 2016). This is especially the case for Chinese NER, where enormous errors can be introduced due to wrongly matched entities. Chinese language is inherently ambiguous since the granularity of words is less well defined than other languages (such as English). Thus massive wrongly matched entities can be generated with the use of gazetteers. As we can see from the example shown in Figure 1, matching a simple 9-character sentence with 4 gazetteers may result in 6 matched entities, among which 2 are incorrect.

To effectively eliminate the errors, we need a way to resolve the conflicting matches. Existing methods often rely on hand-crafted templates or predefined selection strategies. For example, Qi et al. (2019) defined several n-gram templates to construct features for each character based on dictionaries and contexts. These templates are task-specific and the lengths of the matched entities are constrained by templates. Several selection strategies are proposed, such as maximizing the total number of matched tokens in a sentence (Shang et al., 2018), or maximum matching with rules (Sassano, 2014). Though general, these strategies are unable to effectively utilize the contextual information. For example, as shown in Figure 1,

maximizing the total number of matched tokens in a sentence results in wrongly matched entity 张三在 (*Zhang Sanzai*) instead of 张三 (*Zhang San*).

While such solutions either rely on manual efforts for rules, templates or heuristics, we believe it is possible to take a data-driven approach here to learn how to combine gazetteer knowledge. To this end, we propose a novel multi-digraph structure which can explicitly model the interaction of the characters and the gazetteers. Combined with an adapted Gated Graph Sequence Neural Networks (GGNN) (Li et al., 2016) and a standard bidirectional LSTM-CRF (Lample et al., 2016) (BiLSTM-CRF), our model learns a weighted combination of the information from different gazetteers and resolves matching conflicts based on contextual information.

We summarize our contributions as follows: 1) we propose a novel multi-digraph model to learn how to combine the gazetteer information and to resolve conflicting matches in learning with contexts. To the best of our knowledge, we are the first neural approach to NER that models the gazetteer information with a graph structure; 2) experimental results show that our model significantly outperforms previous methods of using gazetteers and the state-of-the-art Chinese NER models; 3) we release a new dataset in the e-commerce domain. Our code and data are publicly available[1].

2 Model Architecture

The overall architecture of our model is shown in Figure 2. Specifically, our model is comprised of a multi-digraph, an adapted GGNN embedding layer and a BiLSTM-CRF layer. The multi-digraph explicitly models the text together with the NE gazetteer information. The information in such a graph representation is then transformed to a feature representation space using an improved GGNN structure. The encoded feature representation is then fed to a standard BiLSTM-CRF to predict the final structured output.

Text Graph. As shown in Figure 2, given the input sentence 张三在北京人民公园 (*Zhang San is at the Beijing People's Park*) consisting of 9 Chinese characters and 4 gazetteers PER1, PER2, LOC1, LOC2 (PER1 and PER2 are gazetteers of the same type PER – "person", but are from different sources; similarly for LOC1
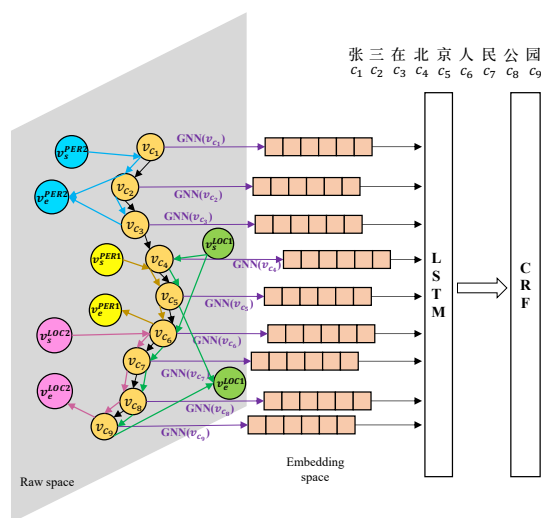
[1]https://github.com/PhantomGrapes/MultiDigraphNER

Figure 2: System architecture

and LOC2). We construct nodes as follows. We first use 9 nodes to represent the complete sentence, where each Chinese character corresponds to one node. We also use another 4 pairs of nodes (8 in total) for capturing the information from the 4 gazetteers, where each pair corresponds to the *start* and *end* of every entity matched by a specific gazetteer. Next we add directed edges between the nodes. First, for each pair of adjacent Chinese characters, we add one directed edge between them – from the left character to the right one. Next, for each matched entity from a gazetteer, edges are added from the entity start node, connecting through the character nodes composing the entity and ending with the entity end node for the corresponding gazetteer. For instance, as we have illustrated in Figure 2, with c_1c_2, or 张三 (*Zhang San*) matched by the gazetteer PER2, the following edges are constructed: $(v_s^{\text{PER2}}, v_{c_1})$, (v_{c_1}, v_{c_2}) and $(v_{c_2}, v_e^{\text{PER2}})$, where v_s^{PER2} and v_e^{PER2} are the start and end nodes for the gazetteer PER2, and each edge is associated with a label indicating its type information (PER in this case). When edges of the same label overlap, they are merged into a single edge. Such a simple process leads to a *multi-digraph* (or "directed multi-graph") representation encoding the character ordering information, the knowledge from multiple NE gazetteers, as well as their interactions.

Formally, a multi-digraph 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 and m gazetteers used in the model, the node set $V = V_c \cup V_s \cup V_e$. Here, V_c is the set of nodes representing characters. Given a gazetteer g, we introduce two special

nodes v_s^g and v_e^g to the graph which we use to denote the start and end of an entity matched with g. V_s (V_e) is a set that contains the special nodes such as v_s^g (v_e^g). Each edge in E is assigned with a label to indicate the type of the connection between nodes. We have the label set $L = \{\ell_c\} \cup \{\ell_{g_i}\}_{i=1}^{m}$. The label ℓ_c is assigned to edges connecting adjacent characters, which are used to model the natural ordering of characters in the text. The label ℓ_{g_i} is assigned to all edges that are used to indicate the presence of an text span that matches with an entity listed in the gazetteer g_i.

Adapted GGNN. Given a graph structure, the idea of GGNN is to produce meaningful outputs or to learn node representations through neural networks with gated recurrent units (GRU) (Cho et al., 2014). While other neural architectures for graphs exist, we believe that GGNN is more suitable for the Chinese NER task for its better capability of capturing the local textual information compared to other GNNs such as GCN (Kipf and Welling, 2017).

However, the traditional GGNN (Li et al., 2016) is unable to distinguish edges with different labels. We adapt GGNN so as to learn a weighted combination of the gazetteer information suitable for our task. To cope with our multi-digraph structure, we first extend the adjacency matrix A to include edges of different labels. Next, we define a set of trainable contribution coefficients $\alpha_c, \alpha_{g_1}, \ldots, \alpha_{g_m}$ for each type of edges. These coefficients are used to define the amount of contribution from each type of structural information (the gazetteers and the character sequence) for our task.

In our model, an adapted GGNN architecture is utilized to learn the node representations. The initial state $h_v^{(0)}$ of a node v is defined as follows:

$$h_v^{(0)} = \begin{cases} W^g(v) & v \in V_s \cup V_e \\ [W^c(v)^\top, W^{bi}(v)^\top]^\top & v \in V_c \end{cases} \tag{1}$$

where W^c and W^g are lookup tables for the character or the gazetteer the node represents. In the case of character nodes, a bigram embedding table W^{bi} is used since it has been shown to be useful for the NER task (Chen et al., 2015).

The structural information of the graph is stored in the adjacency matrix A which serves to retrieve the states of neighboring nodes at each step. To adapt to the multi-digraph structure, A is extended to include edges of different labels, $A = [A_1, ..., A_{|L|}]$. The contribution coefficients are

transformed into weights of edges in A:

$$[w_c, w_{g_1}, \ldots, w_{g_m}] = \sigma([\alpha_c, \alpha_{g_1}, \ldots, \alpha_{g_m}]) \tag{2}$$

Edges of the same label share the same weight. Next, the hidden states are updated by GRU. The basic recurrence for this propagation network is:

$$H = [h_1^{(t-1)}, \ldots, h_{|V|}^{(t-1)}]^\top \tag{3}$$

$$a_v^{(t)} = [(HW_1)^\top, \ldots, (HW_{|L|})^\top]A_v^\top + b \tag{4}$$

$$z_v^{(t)} = \sigma(W^z a_v^{(t)} + U^z h_v^{(t-1)}) \tag{5}$$

$$r_v^{(t)} = \sigma(W^r a_v^{(t)} + U^r h_v^{(t-1)}) \tag{6}$$

$$\hat{h}_v^{(t)} = \tanh(W a_v^{(t)} + U(r_v^{(t)} \odot h_v^{(t-1)})) \tag{7}$$

$$h_v^{(t)} = (1 - z_v^{(t)}) \odot h_v^{(t-1)} + z_v^{(t)} \odot \hat{h}_v^{(t)} \tag{8}$$

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. Equation 3 creates the state matrix H at time step $(t - 1)$. Equation 4 shows the information to be propagated through adjacent nodes. Equations 5, 6, 7, and 8 combine the information from adjacent nodes and the current hidden state 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.

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

3 Experiments

3.1 Experimental Setup

Dataset. The three public datasets used in our experiments are OntoNotes 4.0 (Weischedel et al., 2010), MSRA (Levow, 2006), and Weibo-NER (Peng and Dredze, 2016). OntoNotes and MSRA are two datasets consisting of newswire text. Weibo-NER is in the domain of social media. We use the same split as Che et al. (2013) and Peng and Dredze (2016) on OntoNotes and on Weibo-NER. To demonstrate the effectiveness of our model in the e-commerce domain, we further constructed a new dataset by crawling and manually annotating the NEs of two types, namely PROD ("products") and BRAN ("brands"). We name our dataset as "E-commerce-NER". The NER task in the e-commerce domain is more challenging. The NEs of interest are usually the names of products

Models	OntoNotes			MSRA		
	P	**R**	**F**	**P**	**R**	**F**
BiLSTM-CRF	72.0	75.1	73.5	92.3	92.4	92.4
(+ N-gram)	71.1	75.5	73.3	92.7	92.7	92.7
(+ PIET)	71.6	74.6	73.1	92.9	93.4	93.1
(+ PDET)	73.8	73.8	73.8	93.1	93.1	93.1
Our model (w/o gazetteers)	74.8	73.0	73.9	93.2	92.7	92.9
Our model	75.4	76.6	**76.0**	94.6	94.2	**94.4**
Zhang and Yang (2018)	76.4	71.6	73.9	93.6	92.8	93.2
Dong et al. (2016)	-	-	-	91.3	90.6	91.0
Zhang et al. (2006)	-	-	-	90.2	90.2	91.2

Table 1: Results on the newswire data

Models	Weibo-NER			E-commerce-NER		
	P	**R**	**F**	**P**	**R**	**F**
BiLSTM-CRF	60.8	52.9	56.6	71.1	76.1	73.6
(+ N-gram)	57.8	53.6	55.6	71.2	75.9	73.5
(+ PIET)	57.7	54.4	56.0	71.7	75.8	73.7
(+ PDET)	59.2	54.4	56.7	72.6	75.1	73.8
Our model (w/o gazetteers)	62.1	52.7	57.0	70.7	74.6	72.6
Our model	63.1	56.3	**59.5**	74.3	76.2	**75.2**
Zhang and Yang (2018)	-	-	58.8	-	-	-
Peng and Dredze (2016)	-	-	59.0	-	-	-

Table 2: Results on social media/e-commerce domains

or brands. In practice, the number of unique entities that can appear in such a domain can be easily tens of millions. The training data is typically far from being enough to cover even a small portion of all such NEs. Thus, the effectiveness of an NER system in the e-commerce domain relies heavily on domain-specific gazetteers.

Gazetteers. For the three public datasets, we collect gazetteers of 4 categories (PER, GPE, ORG, LOC). Each category has 3 gazetteers with different sizes, selected from multiple sources including "Sougou"[2], "HanLP"[3] and "Hankcs"[4]. We add an extra indomain gazetteer of type PER for Weibo-NER dataset since the online community has a rich set of nicknames and aliases. For our dataset in the e-commerce domain, we collect 3 product name gazetteers and 4 brand name gazetteers crawled from product catalogues from the e-commerce site Taobao[5]. To better demonstrate the problem of conflicting matches with gazetteers added as knowledge source, the entity conflict rate of each dataset with respect to the gazetteers it references is analyzed. The entity conflict rate (ECR) is defined as the ratio of non-identical overlapping entity matches to all unique entities matched with all gazetteers. The ECR of OntoNotes, MSRA, Weibo-NER and E-commerce-NER are respec-

tively 39.70%, 44.75%, 36.10% and 46.05%.

Models for Comparison. We use BiLSTM-CRF (Lample et al., 2016) with character+bigram embedding without using any gazetteer as the comparison baseline[6]. We explore the three different methods of adding gazetteer features that we compare against: N-gram features, Position-Independent Entity Type (PIET) features and Position-Dependent Entity Type (PDET) features. These feature construction processes follow the work of Wang et al. (2018). We refer the readers to their paper for further details.

To show the effect of adding gazetteer information, a trivial version of our model without using any gazetteer information is also implemented as one of our baselines (our model w/o gazetteers).

3.2 Results

From Table 1, it can be seen that our model with 12 general gazetteers of 4 entity types has an overall highest performance in the news domain. By adding domain specific gazetteers, our model is capable of improving the NER quality in both the social media and the e-commerce domains, as shown in Table 2. Previous methods of using gazetteers do improve the performance of the BiLSTM-CRF model, but the performance gains are not significant. We can observe the performance on both OntoNotes and Weibo-NER drop, when the N-gram and the PIET features were used on top of the BiLSTM-CRF model. We believe this is due to the erroneous information the model captured, especially when multiple conflicting gazetteers were used together. Compared to these methods, our model achieves a remarkably higher performance. Our model is not only able to improve recall by using the gazetteer knowledge, but is also able to offer an improved precision.

To understand the effect of using gazetteers by different methods, we conducted some detailed experiments on OntoNotes. We first split all the sentences in the test set into 3 groups, based on if the entities also appear in the training data or not: "All" contains those sentences in which all entities can be found in the training set, "Some" contains sentences which contain some of the entities from the training set but not all, "None" contains sentences where none of the entities appear in the training set. For the last set of sentences, we con-

---

[2]A crowdsourced gazetteer used by the Chinese IME Sougou: https://pinyin.sogou.com/dict/.

[3]A gazetteer from a widely used open-source Chinese NLP toolkit: https://github.com/hankcs/HanLP.

[4]A gazetteer which consists of over ten million entries: http://www.hankcs.com/nlp/corpus.

[5]http://www.taobao.com

[6]We implemented the baseline models using the NCRFPP toolkit (Yang and Zhang, 2018).

Entities Appear In		PDET			Our model			Our model (w/o gazetteers)		
Train	*Gaze*	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
All	-	84.6	85.3	85.0	85.3	88.8	87.0	87.4	88.1	**87.7**
Some	-	78.2	73.2	75.7	79.5	76.0	**77.7**	78.0	72.0	74.9
None	-	66.7	62.9	64.7	68.5	65.0	**66.7**	66.5	59.2	62.6
None	All	69.8	64.8	67.2	74.2	67.0	**72.0**	71.4	59.9	65.1
None	Some	66.7	61.0	63.7	66.1	61.8	**63.9**	64.0	56.7	60.1
None	None	63.6	62.7	63.1	64.8	62.9	**63.8**	64.2	60.9	62.5

Table 3: Detailed results on OntoNotes (*Train*: Training data, *Gaze*: Gazetteers).

Models	P	R	F
Our model	75.4	76.6	**76.0**
(fixed coefficients)	73.8	74.5	74.2
(AI1G)	73.4	76.7	75.0
(1T1G)	78.9	73.0	75.8
(w/o gazetteers)	74.8	73.0	73.9

Table 4: Ablation study on OntoNotes

ducted additional experiments by further splitting them into three sub-groups, based on whether their entities appear in the gazetteers.

We compare three models under each setting: 1) PDET, 2) our model and 3) our model with all gazetteer nodes removed. We note that the last model can be regarded as a trivial version of both PDET and our model. As shown in Table 3, when none of the entities in a test sentence has been seen during training, with increasing gazetteer coverage our model has a more significant improvement compared to PDET. When none or some of the test entities appear in the training data, both PDET and our model perform better than the trivial model. This shows the benefit of utilizing gazetteer knowledge. Furthermore, in this case, our model still yields a relatively better F1 score, due to its better way of representing gazetteer information using multi-digraph. In the case where all the entities appear during training, both PDET and our model yield lower performance than the trivial model. We believe this is due to errors introduced by the gazetteers. Nonetheless, our model is more robust than PDET in this case.

Ablation Study. We also conducted an ablation study to explore the contributions brought by the weighted combination of gazetteers, so as to understand how our model can effectively use the gazetteer information.

As shown in Table 4, by fixing the gazetteer contribution coefficients to 1, the model's performance drops by 1.8 points in terms of F1 score. The precision is even lower than that of our model without gazetteers. This experiment shows that, without a good combination of the gazetteer information, the model fails to resolve conflicting matches. In that case, errors are introduced with the use of gazetteers. These errors harm the model's performance and have a negative effect on the precision.

We use the following ablation test to understand whether the gazetteer information can be fully utilized by our model. There are three types of infor-

mation provided by gazetteers: boundary information, entity-type information, and source information. The All in One Gazetteer (AI1G) experiment shows what role the boundary information plays in our model by merging all 12 gazetteers into one lexicon where entity type information is discarded. It outperforms the model without gazetteers by 1.1 points in terms of F1 score. The One Type One Gazetteer (1T1G) model adds the entity type information on top of the AI1G model by adding only the entity type labels (i.e., there is one gazetteer for one type, by merging all gazetteers of the same type into one). Doing so leads to a 0.8 points improvement over the AI1G model. From the experiments we can see that the entities' source information is also helpful. For example, an entity that appears in multiple PER gazetteers is more likely to be an entity of type PER than an entity appearing only in one gazetteer. Our model can effectively capture such source information and has an improvement of 0.2 points in terms of F1 compared to the 1T1G model.

4 Conclusion and Future Work

We present a novel neural multi-digraph model for performing Chinese named entity recognition with gazetteers. Based on the proposed multi-digraph structure, we show that our model is better at resolving entity-matching conflicts. Through extensive experiments, we have demonstrated that our approach outperforms the state-of-the-art models and previous methods for incorporating gazetteers into a Chinese NER system. The ablation study confirms that a suitable combination of gazetteers is essential and our model is able to make good use of the gazetteer information. Although we specifically investigated the NER task for Chinese in this work, we believe the proposed model can be extended and applied to other languages, for which we leave as future work.

Acknowledgments

References

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proc. of SIGMOD*.

Wanxiang Che, Mengqiu Wang, Christopher D Manning, and Ting Liu. 2013. Named entity recognition with bilingual constraints. In *Proc. of NAACL-HLT*.

Xinchi Chen, Xipeng Qiu, Chenxi Zhu, Pengfei Liu, and Xuanjing Huang. 2015. Long short-term memory neural networks for chinese word segmentation. In *Proc. of EMNLP*.

Jason P C Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. In *Proc. of TACL*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proc. of EMNLP*.

Chuanhai Dong, Jiajun Zhang, Chengqing Zong, Masanori Hattori, and Hui Di. 2016. Character-based lstm-crf with radical-level features for chinese named entity recognition. In *Proc. of ICCPOL*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proc. of ICLR*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proc. of NAACL-HLT*.

Gina-Anne Levow. 2006. The third international chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proc. of the Fifth SIGHAN Workshop on Chinese Language Processing*.

Yaoyong Li, Kalina Bontcheva, and Hamish Cunningham. 2005. Svm based learning system for information extraction. *Deterministic and statistical methods in machine learning*.

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. 2016. Gated graph sequence neural networks. In *Proc. of ICLR*.

Nanyun Peng and Mark Dredze. 2016. Improving named entity recognition for chinese social media with word segmentation representation learning. In *Proc. of ACL*.

Zhang Qi, Liu Xiaoyu, and Fu Jinlan. 2019. Neural Networks Incorporating Dictionaries for Chinese Word Segmentation. In *Proc. of AAAI*.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proc. of CoNLL*.

Manabu Sassano. 2014. Deterministic Word Segmentation Using Maximum Matching with Fully Lexicalized Rules. In *Proc. of EACL*.

Dominic Seyler, Tatiana Dembelova, Luciano Del Corro, Johannes Hoffart, and Gerhard Weikum. 2018. A study of the importance of external knowledge in the named entity recognition task. In *Proc. of ACL*.

Jingbo Shang, Liyuan Liu, Xiang Ren, Xiaotao Gu, Teng Ren, and Jiawei Han. 2018. Learning named entity tagger using domain-specific dictionary. In *Proc. of EMNLP*.

Qi Wang, Yuhang Xia, Yangming Zhou, Tong Ruan, Daqi Gao, and Ping He. 2018. Incorporating dictionaries into deep neural networks for the chinese clinical named entity recognition. *arXiv preprint*.

Ralph Weischedel, Martha Palmer, and Mitchell P Marcus. 2010. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*.

Jie Yang and Yue Zhang. 2018. Ncrf++: An open-source neural sequence labeling toolkit. In *Proc. of ACL (System Demonstrations)*.

Norshuhani Zamin and Alan Oxley. 2011. Building a corpus-derived gazetteer for named entity recognition. In *Software Engineering and Computer Systems*.

Suxiang Zhang, Ying Qin, Juan Wen, and Xiaojie Wang. 2006. Word Segmentation and Named Entity Recognition for SIGHAN Bakeoff3. In *Proc. of the Fifth SIGHAN Workshop on Chinese Language Processing*.

Yue Zhang and Jie Yang. 2018. Chinese NER Using Lattice LSTM. In *Proc. of ACL*.