Research on Chinese Named Entity Recognition Based on RoBERTa-BIGRU-MRC Model

Huai Peng and Xianghong Tang(✉)

Key Laboratory of Public Big Data, School of Computer Science and Technology, Guizhou University, Guiyang 550000, China
xhtang@gzu.edu.cn

Abstract. Nested entity is the focus and difficulty of Chinese named entity recognition. The existing methods regard nested NER as two subtasks of Chinese word segmentation and sequence annotation. This method depends very much on the quality of input word vector, and low-quality word vector will lead to the error propagation of the model. To solve the above problems, a Chinese named entity recognition model based on RoBERTa-BiGRU-MRC is proposed. Firstly, RoBERTa is used to embed the entity type description and sentences to obtain the dynamic word vector. Secondly, BiGRU is used to extract contextual semantic features for further understanding of semantic information. Then two binary classifiers are constructed to better predict the probability value of the index at the beginning and end of the entity. Finally, the accuracy of the model is improved by constructing the loss function optimizer of predicted value and real value. Experiments were conducted on Chinese MSRA and onto4 data sets, and the accuracy, recall and F1 value were used as evaluation indexes. The experimental results show that the F1 value of the optimized model is 0.41 and 0.36 higher than that of the traditional sequence annotation model, respectively.

Keywords: Nested Entities · Entity Recognition · Word Vector · Machine Reading Comprehension

1 Introduction

Named Entity Recognition [18], also known as Entity extraction, is one of the most important research tasks in NLP research. The objective of this task is to be able to independently find, identify and classify relevant entity words with special meanings in a given unstructured text, such as people's names, place names and organization names. At the same time, NER also plays an important role in many downstream NLP tasks, such as: information retrieval [3], relationship extraction [1] and question answering system [5].

Compared with English NER, Chinese NER is more difficult. The main reason is that there are a lot of entity nesting problems in Chinese texts, so a word can be divided into multiple entity types. For example, nanjing Yangtze River Bridge can be classified as nanjing, Nanjing City, Mayor, Yangtze River, Yangtze River Bridge, bridge and other sub-entities. This leads to the difficulty of demarcating entity boundaries, which greatly

© The Author(s) 2023
N. Radojević et al. (Eds.): ICAID 2022, AHIS 7, pp. 308–319, 2023.
https://doi.org/10.2991/978-94-6463-010-7_32

affects the accuracy of NER tasks. Existing methods regard nested NER as two subtasks of Chinese word segmentation and sequence tagging. This method first trains word vector and word vector through a large corpus or uses public word vector. Then the word vector is put into the neural network model for training to get the feature vector, and finally the feature vector is input into the CRF layer to get the probability of the corresponding label. This method has the problem of error propagation and depends very much on the quality of word vector. If there is no vector representation of entity in word vector, the following model is difficult to identify the correct entity, which greatly reduces the accuracy of entity recognition. In view of the above problems, this paper proposes RoBERTa-BIGRU-MRC model by referring to the method of machine reading comprehension task. Before training, THE NER data set needs to be converted into the triplet format of machine reading comprehension tasks (Query, Answer, context), where Query represents the description of entity type, answer represents the entity in question, and context represents the sentence in the text. In this model, the description and sentence of input entity type are transformed into dynamic word embedding to obtain dynamic word vector by pre-training model RoBERTa. Then, BGRU model is used to capture semantic information from both directions and output semantic feature vectors. Then two binary classification models are constructed to predict the probability of entity starting and ending index. Finally, an optimizer is constructed for the cross entropy loss function of the predicted value and true value of the initial index, the predicted value and true value of the end index, and the predicted value and true value of the matching matrix from the start index to the end index to optimize the accuracy of the model.

The model regarded NER task as a binary task, and used two classifiers to predict the initial index and the endindex of the entity. In this way, only the word vector is required as the input of Chinese NER task, which eliminates the error propagation caused by the word fusion vector and leads to the decline in the accuracy of nested entity recognition. The model experiment in this paper shows that compared with traditional deep learning model, the proposed model has certain improvement in accuracy and recall rate on Chinese MSRA and ONTO4 data sets.

2 Related Work

2.1 Pre-trained Language Model

Pre-training is to pre-train the model with a large corpus to obtain the word vector with contextual semantic information, and then fine-tune the word vector for the corresponding downstream task. Mikolov proposed the Word2Vec technology [11], which translates words into vectors and expresses the semantic information of words through vectors. The word vector of this technology usually only represents one meaning and has limited supporting role for downstream tasks. Peters proposed the ELMo pre-training language model [12], which is also a pre-training of a large number of texts. ELMo internally uses bidirectional LSTM to capture contextual semantics and extract deep features of texts. Radford proposed [14] the GPT model and used Transformer [16] network to replace LSTM as the language model to better capture long-distance language structures. Then the language model is used as the auxiliary task training target in supervised fine-tuning of specific tasks. The BERT [4] model proposed by Devlin uses global dynamic vectors to

fine-tune downstream tasks and adds sentence level mask mechanism in the pre-training stage to achieve the best effect in multiple typical downstream tasks. In 2019, Google released BERT's improved model ALBERT [6], which reduced the number of model parameters through the parameter sharing mechanism and had effects similar to BERT. Liu proposed the RoBERTa model and, on the basis of BERT [9], increased the number of training samples, the length of training sequence and the global mask mechanism. It has Achieved better results than BERT in the downstream task. The proposed method is based on RoBERTa pretraining language model.

2.2 Chinese NER Method

Zhang proposed a Lattice LSTM model (LSTM) for Chinese NER in 2018 [17]. Compared with character-based approach, this model can make full use of word and word order information. Compared with word-based method, word segmentation errors will not affect the recognition result. The core idea of the model is to represent words in sentences through grid LSTM, and to fuse the potential lexical information into character-based LSTM-CRF. Cao proposed a novel adjunctive transfer learning model (BiLSTM-CRF-ADV-self-attention) for Chinese NER in 2018 [2]. The author believes that Chinese Word Segmentation (CWS) is similar to Chinese NER task in many aspects, but also has many differences. Therefore, an adversity-transfer learning model is proposed to make full use of the common boundary information of Chinese word segmentation and to prevent the influence of Chinese word segmentation on NER task. At the same time, the self-attention mechanism is introduced into the model to capture the long distance dependence and grammatical information in sentences. Zhu proposed an attention-based convolutional neural network model (CAN) for Chinese NER [19]. The model consists of a character-based convolutional neural network (CNN) with a local attention layer and a gated recursive unit (GRU) with a global self-attention layer to extract information from adjacent character and sentence contexts. ZHU proposed a simple and effective Chinese NER method soft-Lexicon, which can integrate lexical information into character representation. This method reduces the complexity of sequence modeling and can introduce dictionary information into any neural network model with only a small change in character representation. In addition, BERT pre-training can be easily combined with this model. MA proposed a flat-Lattice Transformer suitable for Chinese NER [10], which was an improvement on the Lattice model of Zhang et al. In view of the low training speed of the Lattice Lstm model and the insufficient introduction of vocabulary information, the FLAT model made the following improvements: The Lattice was carried forward into a two-dimensional form, and the start and end position indexes were built for each character and word. The relative position information of each word is added to improve the position perception and direction perception of the model. FLAT model does not design or change the native coding structure, and cleverly designed position vector integrates the lexical information, which not only achieves the information lossless, but also greatly speeds up the inference speed.

2.3 Extractive Machine Reading Comprehension

Machine reading comprehension (MRC) is based on the question, using the machine to analyze and understand the context of the text, and then give the final answer. At present, MRC tasks are generally divided into five types: fill-in-the-blank, selection, extraction, generation and multi-hop inference [8]. Abstract machine reading comprehension [13] is an important type of MRC task. It mainly uses given text content and related questions to provide correct answers through analysis and understanding of text content. This task needs to predict the starting and ending position of the answer so as to select the answer segment, which is usually called span prediction or segment prediction. Questions in the extractive MRC task are generally put forward manually, and there may be a gap in length between the answers and the possibility of no answers, which is more suitable for the application scenarios in real life. Abstract machine reading comprehension representation. Suppose the abstract MRC is given triplet data (C, Q, A), where C is Context and represents the text in the data set. Q stands for Question, indicating the problem in the data set; A is Answer, indicating the Answer to the question in the text. Through data representation learning, the model can fit the relation f between input and output, which is usually expressed as formula.

$$f(C, Q) = A \tag{1}$$

A in formula (1) is usually represented as (a_{start}, a_{end}), and the part between the start position a_{start} and the end position a_{end} is the final answer to be predicted.

3 RoBERTa-BIGRU-MRC Model

Existing methods regard Chinese NER task as two subtasks of Chinese word segmentation and sequence tagging, but this method has the problem of error propagation and depends very much on the quality of word vector. To solve the above problems, this paper proposes RoBERTa-BIGRU-MRC model by referring to the machine reading comprehension task. The model firstly uses RoBERTa to embed the entity type description and sentence to obtain the dynamic word vector. Then, BIGRU is used to extract contextual semantic features to further understand the semantics and obtain the semantic features between long-distance words. Then two binary classification models are constructed to predict the probability of entity starting and ending index. Finally, an optimizer is constructed for the cross entropy loss function of the predicted value and true value of the initial index, the predicted value and true value of the end index, and the predicted value and true value of the matching matrix from the start index to the end index to optimize the accuracy of the model. The structure of RoBERTa-BIGRU-MRC model is shown in Fig. 1.

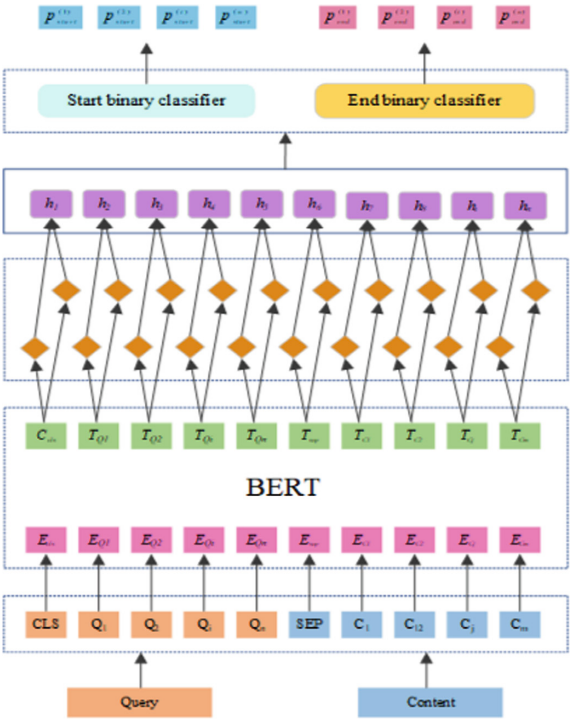

Fig. 1. Structure of RoBERTa-BIGRU-MRC mode.

3.1 RoBERTa Layer

RoBERTa evolved from BERT's model. The BERT model uses multi-layer bi-directional Transformer encoder as the main framework of the model. The model input is composed of three parts: word vector, sentence type vector and position coding vector. MLM (MaskLanguage Model) and NSP (Next Message Prediction) are used as pre-training targets. MLM model selects 15% of the tokens randomly in the input statements, and then in each training, the selected tokens are replaced with [MASK] 80% of the time, 10% of the time with random words, and the remaining 10% of the time with the original words. The purpose of this is to improve the feature representation and generalization ability of the model for sentences. NSP model is used to determine whether sentence B is A context of sentence A, so as to model the relationship between sentences. RoBERTa model mainly improved BERT model in three aspects. First, static mask was changed to dynamic mask. In BERT training data, random mask was only performed once for each sample, and the position of mask was the same during each training, while dynamic mask was dynamically generated once before each training. The model can learn more statement patterns. Secondly, NSP Loss is removed to improve the efficiency of the model. Thirdly, a larger data set is used for training and BPE is used for text data processing.

RoBERTa-BIGRU-MRC model input text sequence to RoBERTa for coding, and the dynamic word vector obtained by embedding RoBERTa model words is used as input of BGRU layer.

3.2 BIGRU Layer

Gated Recurrent Unit (GRU) is a new generation of Recurrent Neural Network (RNN), similar to Long short-term Memory (LSTM). It is used to solve the problems of gradient disappearance and gradient explosion of traditional RNN. Different from LSTM, GRU no longer uses unit state to record or transmit information, but uses hidden state to record and transmit information. Update and recirculate gate control GRU unit output. Figure 2 shows the GRU unit structure.

Where, "+" indicates the add operation, "σ" indicates the Sigmoid activation function, "×" indicates the Hadamard product, and "tanh" indicates the Tanh activation function. GRU parameter update calculation formula is as follows

$$Z_t = \sigma\left(W_z \cdot \left[h_{t-1}, x_t\right]\right) \tag{2}$$

$$r_t = \sigma\left(W_r \cdot \left[h_{t-1}, x_t\right]\right) \tag{3}$$

$$\tilde{h}_t = \tanh(W_r \cdot [r * h_{t-1}, x_t]) \tag{4}$$

$$h_t = (1 - Z_t) * h_{t-1} + Z_t * \tilde{h}_t \tag{5}$$

Where, Z_t is the activation result of the update gate, which controls the information inflow in the form of gating, x_t is the input vector of the t time step, W_z is the weight matrix, h_{t-1} stores the information of the time step t−1 r_t is the activation result of reset gate, its calculation process is similar to that of update gate, and W_z is the weight matrix. h_t indicates the memory content of the current time step. h_t represents the final memory of the current time step. GRU captures information in one direction only. BIGRU is

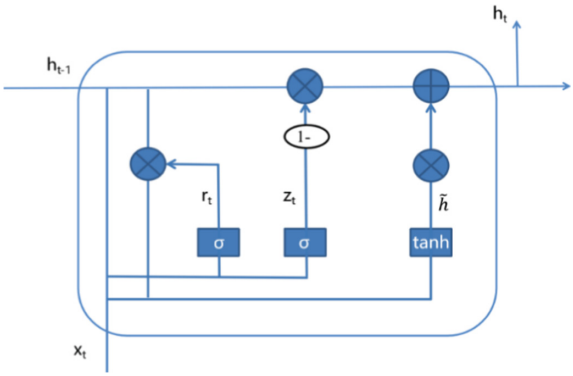

Fig. 2. GRU Unit structure

used to extract context information and is a bidirectional GRU. The forward input of the text sequence to the GRU records the memory information in its output, and the reverse input of the text sequence to the GRU obtains the memory information. The two are connected and merged to achieve the purpose of extracting context information. BIGRU layer extracts context information to obtain deeper semantic understanding, and obtains the input of emission fraction matrix to MRC.

3.3 MRC Layer

For each token in the sentence, two binary classifiers are made. The first one determines whether the token is the start word of the entity, and the second one determines whether the token is the end word of the entity. In fact, the span of an entity is found. This approach allows you to obtain multiple entities in a sentence, including nested entities, by obtaining multiple start and end indexes based on the sentence and entity type.

3.3.1 Initial Index Prediction

Assuming that the fractional matrix of the output of BIGRU is expressed as E, the model first predicts the probability of each word as the starting index, and the formula is as follows:

$$P_{start} = softmax_{eachrow}(E \cdot T_{start}) \in R^{n*2} \tag{6}$$

$T_{start} \in R^{n*2}$ is the weight to learn. The P_{start} of each line represents the probability distribution of each index as the entity start index under the premise of a given entity type.

3.3.2 End Index Prediction

End index prediction: The prediction process is exactly the same as the start index prediction, and we also get a probability distribution for the end position $P_{end} \in R^{n*2}$.

3.3.3 Probability Matrix

In a sentence context, there may be multiple entities of the same category. This means that multiple initial indexes can be predicted from the initial index model and multiple end indexes can be predicted from the end index model. The way FLAT-NER uses the start index and its most recent end index does not apply here, because entities may overlap. Therefore, we need a way to match the predicted start index with the corresponding end index. Specifically, by argmax each line P_{start} and P_{end}, we get the start and end predicted values for each position in the sentence and put them into two collections, represented by $\hat{I}_{start}$ and $\hat{I}_{end}$.

$$\hat{I}_{start} = \{i | argmax\left(P_{start}^{(i)}\right) = 1, i = 1, \ldots, n\} \tag{7}$$

$$\hat{I}_{end} = \{j | argmax\left(P_{end}^{(j)}\right) = 1, j = 1, \ldots, n\} \tag{8}$$

Where the superscript (i) represents the ith row in the matrix. Assume that any start location $I_{start} \in \hat{I}_{start}$ and end location $I_{end} \in \hat{I}_{end}$, train a binary classification model to predict the probability that they should be matched as shown in Formula (9).

$$P_{istart,jend} = sigmoid(m * concat(E_{istart}, E_{jend})) \tag{9}$$

Where E_{istart} and E_{jend} represent the start and position score matrix, and the two of them are spliced to obtain the probability matrix of start-end matching score. $m \in R^{1\times 2d}$ represents the learning weight of binary classification model.

3.3.4 Loss Function

During training, each sentence has two real label Y_{start} and Y_{end} sequences of length N, representing the set of token start and end positions in the sentence. Therefore, this paper constructs two cross entropy loss functions to predict the start and end positions respectively. The loss functions are shown in Eqs. (10) to (11).

$$L_{start} = CE(P_{start}, Y_{start}) \tag{10}$$

$$L_{end} = CE(P_{end}, Y_{end}) \tag{11}$$

$Y_{start,end}$ indicates whether the start and end positions of each token really match. The actual matching value is represented by a two-dimensional label matrix. The loss function calculation formula of probability matrix is shown in Eq. (12)

$$L_{span} = CE(P_{start,end}, Y_{start,end}) \tag{12}$$

Finally, we need to add up the loss function of the three parts as the target of optimization in the training process. The calculation formula is shown in Equation.

$$L = \alpha L_{start} + \beta L_{end} + \gamma L_{span} \tag{13}$$

α, β, $\gamma \in [0, 1]$ are hyperparameters controlling the proportion of the three loss functions in the overall loss function, and joint training is carried out in an end-to-end manner. Select the start and end positions of $\hat{I}_{start}$ and $\hat{I}_{end}$ respectively. Then the position matching model is used to extract the alignment of the starting and ending positions to get the final answer.

4 Experiment

4.1 Data Set and Data Preprocessing

MSRA and Ontonotes 4.0 Chinese data set were used in this paper. MSRA is an entity recognition dataset in the news field labeled by Microsoft Research Asia and one of the datasets of SIGNAN Backoff 2006's entity recognition task. The dataset contains more than 50,000 Chinese entity recognition annotation data, which can be divided into people, places and institutions including 46,364 sentences in the training set, 4,636

sentences in the test set and 4,365 sentences in the verification set. It was launched in 2016. The Ontonotes 4.0 dataset is derived from telephone conversations, news agencies, radio news, radio conversations and blogs. Entities were classified into four categories: region, institution, person and place, including 15,724 sentences in the training set, 4,306 sentences in the test set and 4,301 sentences in the verification set.

MSRA and Ontonotes 4.0 data sets need to be converted into MRC format data in the experiment. Each row contains context (sentence), end_position (end index set), entity_label (entity type), query (description of entity type), span_position (start and end index set), and start_position (start index) Collection).

4.2 Evaluation

Precision (P), Recall (R) and F1 values were used as evaluation indexes. P represents the proportion of correctly recognized entities in all recognized entities, and R represents the proportion of correctly recognized entities in entities to be recognized. F1 is a comprehensive evaluation index combining P and R. The specific calculation process is as follows:

$$P = \frac{T_p}{T_p + F_p} \times 100\% \tag{14}$$

$$R = \frac{T_p}{T_p + F_n} \times 100\% \tag{15}$$

$$F1 = \frac{2PR}{P + R} \times 100\% \tag{16}$$

Where, T_p represents the number of correct entities recognized by the model, F_p represents the number of entities incorrectly recognized by the model, and F_n represents the number of entities not recognized by the model.

4.3 Experimental Environment

During the training, Adam optimizer was used and cross entropy loss function was used. The specific parameter Settings are shown in Table 2 (Table 1).

Table 1. Experimental environment parameter setting

Soft Hardware	version
OS	Ubuntu 16.04
CPU	Intel(R) Xeon(R) Gold 5220 CPU @ 2.20 GHz
GPU	A100-SXM4-40 GB
memory	240 GB
python	3.8.0
torch	1.7.1

Table 2. Experimental environment parameter setting

Parameter names	parameter value
max_seq_len	128
batch_size	8
gru_nums	1
hidden_size	768
optimizer	Adam
learning	1e−05

Table 3. Experimental results of MSRA dataset model

Model	P	R	F1
Lattice-LSTM	93.57	92.79	93.18
LGN	94.19	92.73	93.46
CGN	94.22	92.71	93.47
FLAT	95.12	93.63	94.35
BERT-Tagger	94.97	94.62	94.80
Glyce-BERT	95.27	95.51	95.75
RoBERTa-BIGRU-MRC	96..54	96.79	96.15

Table 4. Experimental results of ontonotes dataset model

Model	P	R	F1
Lattice-LSTM	76.35	71.56	93.18
LGN	76.13	73.68	93.46
CGN	76.23	73.55	93.47
FLAT	78.46	75.12	94.35
BERT-Tagger	78.01	80.35	94.80
Glyce-BERT	81.87	81.40	95.75
RoBERTa-BIGRU-MRC	82.31	81.68	96.15

4.4 Experimental Results and Analysis

Experimental results of RoBERTa-BIGRU-MRC model and other models on MSRA and Ontonotes data sets are shown in Tables 3 and 4.

From the above experimental results, RoBERTa-BIGRU-MRC model has better accuracy, recall and F1 value than other models on MSRA and Ontonotes data sets.

We believe that there are two main reasons that lead to the better effect of the model than the traditional sequence annotation model: (1) the input of the model in this paper only uses the input word vector table without considering the representation of word vector, so as to avoid the absence of some word vectors, resulting in the error propagation of subsequent models and the low accuracy of some entity recognition; (2) In the classification task, when the amount of data is the same, the accuracy of multi classification task is lower than that of two classification task. In this paper, entity type prediction is regarded as a binary classification task, so the entity type prediction of the model is better than the classification effect of sequence annotation task.

5 Conclusions

This paper aims at the defects of Chinese named entity recognition task based on sequence annotation. Referring to machine reading comprehension task, RoBERTa-BIGRU-MRC model is proposed, and entity recognition is regarded as two binary classification tasks. To solve the problem of model error propagation, the sequence annotation method depends very much on the quality of word fusion vector. On MSRA and Ontonotes data sets, the accuracy, recall and F1 value of RoBERTa-BIGRU-MRC model are better than those of traditional sequence annotation model.

References

1. Bunescu R (2005) A shortest path dependency kernel for relation extraction. In: Proceedings of human language technology conference and conference on empirical methods in natural language processing
2. Cao YK (2018) Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism. In: Proceedings of the 2018 conference on empirical methods in natural language processing, pp 182–192
3. Chen LK (2015) Event extraction via dynamic multi-pooling convolutional neural networks. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing, pp 167–176
4. Devlin WK (2019) BERT: pre-training of deep bidirectional transformers for language
5. Diefenbach D, Lopez V, Singh K, Maret P (2017) Core techniques of question answering systems over knowledge bases: a survey. Knowl Inf Syst 55(3):529–569
6. Lan MS (2019) ALBERT: a lite BERT for self-supervised learning of language
7. Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. 724, 731
8. Li HX (2020) FLAT: Chinese NER using flat-lattice transformer. arXiv Preprint arXiv:2004.11795
9. Liu MN (2019) RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692
10. Ma MQ (2020) Simplify the usage of lexicon in Chinese NER. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 5951–5960
11. Mikolov IK (2013) Distributed representations of words and phrases and their compositionality. arXiv preprint arXiv:1310.4546
12. Peters M (2018) Deep contextualized word representations. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 2227–2237

13. Qiu XJ (2019) A survey on neural machine reading comprehension. arXiv arXiv:1906.03824
14. Radford KT (2018) Improving language understanding by generative pre-training. J. Representations. J. CoRR. 1909.11942
15. Understanding. C. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 2019, 1: 4171–4186
16. Vaswani N (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
17. Zhang J (2018) Chinese NER using lattice LSTM. arXiv Preprint arXiv:1805.02023
18. Zhao ZH (2019) Adversarial training based lattice LSTM for Chinese clinical named entity recognition. J Biomed Inf 99:103290
19. Zhu G (2019) CANCER: Convolutional attention network for Chinese named entity recognition. arXiv Preprint arXiv:1904.02141

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.