International Journal of Computational Intelligence Systems and Applications. Volume 11, Issue 1, (2024) 58-66
http://www.cayley-nielson.com/ijcisa/

Research on Entity Relationship Extraction Method Based on Negative Relationship Sampling Strategy

Lin Shi[1], **Zhigang Li**[2], **Xianming Zou**[1]

[1] School of Artificial Intelligence, North China University of Science and Technology ,Tangshan, Hebei, China

[2] Computing Center, TangShan College, Tangshan, Hebei, China

E-mail: *12389248@qq.com*

Abstract

An entity-relationship extraction model based on negative relationship sampling strategy is proposed.The model adopts a single-module one-step approach to deal with the relationship extraction task, which involves all relationships in the training process may lead to the dominance of redundant relationships, affecting the prediction results.A negative relationship sampling strategy is proposed, which improves the training effect of the model by adjusting the balance of the positive and negative relationship samples, and enhances the ability of identifying different relationship types.

Keywords: Knowledge extraction, Relationship extraction, Diabetes, Knowledge graphs.

1. Introduction

At present, the number of diabetes patients in China is growing, and the accumulation of diabetes-related data is rapid, containing a wealth of medical knowledge. These data are crucial to the diagnosis and management of healthcare professionals, which can improve the efficiency of diagnosis and treatment and the quality of life of patients. However, most of the diabetes medical data are stored in unstructured form, and there exists the problem that the knowledge is difficult to be reused efficiently. Therefore, automated extraction research is of great interest in the current diabetes healthcare field, aiming to extract useful diabetes information from unstructured data. The core tasks of knowledge extraction are named entity recognition and relation extraction.

For current relationship extraction, recent research has proposed a new approach that can identify both entities and their relationships in an end-to-end manner. Existing joint entity and relationship extraction methods typically decompose the problem into multiple modules or steps. Casrel [1] uses BERT as a backbone model and embeds entity mentions into relationships in a new labelling framework that requires multiple modules and steps. Tplinker [2] formulates the joint extraction task as a markup pair linking problem and uses a new handshake labelling paradigm, which aligns entity pair boundary labelling for each relationship type, and which requires multiple modules. These modules of the model extract entities using a table-filling paradigm that predicts start and end indices. The model acknowledges entities by predicting the position of their start and head markers. For each relation, Tplinker constructs two tables for subject and object prediction, which leads to significant redundancy and computational overload when the number increases. In addition, Tplinker's decoding phase remains in a pipelined paradigm with error propagation. To address these issues, Onerel [3] suggests that joint entity and relation extraction be addressed in a single module in one step, but all relations will be involved in the training phase. The redundant relations involved

in training will dominate the negative relation samples, making the model results more biased towards the negative relation samples and reducing the prediction results.

The above methods, the main problem is that the multi-module multi-step and multi-module one-step methods suffer from error accumulation and relationship redundancy. To address this problem, this paper proposes a one-centre relational extraction model.

2. Chinese diabetes relationship extraction model NRS-Onerel

2.1. Model Overview

In order to solve the current challenges of joint entity-relationship extraction, this paper adopts a single-module one-step extraction method and designs the relationship extraction model NRS-Onerel, which alleviates the problem of imbalance between positive and negative relationships during the single-module one-step training process and improves the accuracy of relationship extraction.

The overall framework structure of the NRS-Onerel model is shown in Fig. 1.In this paper, we use the BERT pre-trained language model for word embedding, followed by designing a negative relation sampling strategy to alleviate the defects of positive and negative relation imbalance, after which we adopt the Onerel model for entity relation extraction in the single-module one-step process.

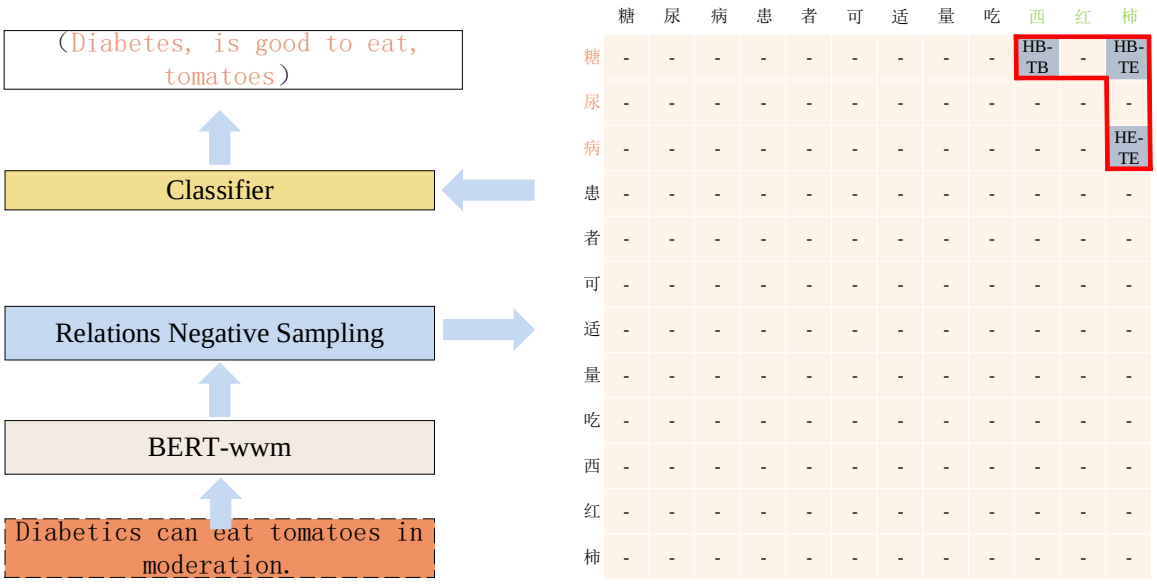

Fig. 1. Structure of the overall framework of the NRS-Onerel model

2.2. Coding layer

For the coding layer, BERT-wwm was used as the pre-training model as the encoder. Compared to the BERT_base base model used for the Onerel model, the BERT-wwm pre-training model modifies the strategy for pre-trained sample generation. Unlike BERT_base which uses [MASK] tags to replace individual words, BERT-wwm uses [MASK] tags to replace the whole vocabulary. In Chinese, a word usually consists of multiple characters, between which richer information is contained. Therefore, masking operations on the whole word can better capture the contextual information of the surrounding text.

For an input sentence, BERT-wwm is used as a sentence encoder to capture each token embedding e_i, denoted as:

$$\{e_1, e_2, \ldots, e_L\} = BERT(\{x_1, x_2, \ldots, x_L\}) \quad (1)$$

Where x_i is the input representation of each token, L indicating the length of the sentence.

2.3. Entity-Relationship Extraction Layer

In the Onerel model, a relation-specific corner tokenisation strategy is used to extract entity pairs from the table. The strategy uses the "BIE" (Beginning, Interior, End) notation to indicate the positional information of the entities. As shown in Fig. 2, for the triad of sentence expressions ("diabetes, advisable to eat, tomato"), there will be nine special labels in the relation-specific sub-matrix. Among them, (1) HB-TB labelling means that the two positions are the start labels of pairs of heads and tails, respectively, conditional on the particular relation. For example, there is a relationship between two entities "diabetes" and "tomato", "it is good to eat", so the combination ("sugar", "it is good to eat", "tomato") is assigned the classification label "HB-TB".(2) The HB-TE tag indicates that the tag corresponding to the row in the table is the beginning of the head entity and the tag corresponding to the column is the end of the tail entity. For example, "Sugar" is the start tag of "Diabetes", and "Persimmon" is the end tag of "Tomato". Therefore, the combination of ("sugar", "diabetic", and "persimmon") is assigned the tag "HB-TE".(3) The logic of the tag "HE-TE" is similar to that of "HB-TB", which indicates that the two positions are the end tags of pairs of head and tail entities, respectively, conditional on a specific relationship. For example, the combination of ("sick", eat, "persimmon") is assigned as "HE-TE".(4) The "-" symbol is used to mark all cells except the above three.

	糖	尿	病	患	者	可	适	量	吃	西	红	柿
糖	-	-	-	-	-	-	-	-	-	HB-TB	HB-TI	HB-TE
尿	-	-	-	-	-	-	-	-	-	HI-TB	HI-TI	HI-TE
病	-	-	-	-	-	-	-	-	-	HE-TB	HE-TI	HE-TE
患	-	-	-	-	-	-	-	-	-	-	-	-
者	-	-	-	-	-	-	-	-	-	-	-	-
可	-	-	-	-	-	-	-	-	-	-	-	-
适	-	-	-	-	-	-	-	-	-	-	-	-
量	-	-	-	-	-	-	-	-	-	-	-	-
吃	-	-	-	-	-	-	-	-	-	-	-	-
西	-	-	-	-	-	-	-	-	-	-	-	-
红	-	-	-	-	-	-	-	-	-	-	-	-
柿	-	-	-	-	-	-	-	-	-	-	-	-

Fig. 2. Corner markers for specific relations

Simple classifiers often struggle to fully explore the interactions between entities and relationships, and are not effective in capturing the informative features of the intrinsic structure of the triad. Inspired by the idea of HoLE [4], the Onerel model defines its scoring function as:

$$f(h,t) = r^T(s * o) \tag{2}$$

where, respectively, h、t are head and tail representations. $*$ denotes the circular correlation, which is used to mine the potential dependency between two entities. s and o are subject and object entity representations.$*$ Operators are defined as nonlinear combinatorial projections:

$$s * o = ReLU(W[s; o]^T) + b \tag{3}$$

where $W \in R^{d_e \times 2d}$, d_e is the number of dimensions of the entity pair representation, b is the trainable weight and bias, [;] denotes the connection operation.

Then, using all the relational representations$R^T \in R^{d_e \times 4K}$, the correlation $(w_i, r_k, w_j)_{k=1}^{K}$ is calculated once at the same time. Where 4 is the number of categorical labels and K is the number of predefined relations.The score function of the entity relationship extraction method in this paper is:

$$v_{(w_i, r_k, w_j)} = R^T ReLU\big((W[s; o]^T) + b\big) \tag{4}$$

Ultimately, the score vectors $(w_i, r_k, w_j)$are fed into the SoftMax function to predict the corresponding labels:

$$P\left(y_{(w_i, r_k, w_j)} \middle| S\right) = SoftMax\left(v_{(w_i, r_k, w_j)}\right) \tag{5}$$

2.4. Negative Relational Sampling Strategy

Negative relation sampling strategy is a key technique in the field of relation extraction for solving the data imbalance problem. In relation extraction tasks, it is common to have an imbalance of positive and negative samples due to the fact that the number of real relations (positive examples) is much less than the number of pairs of entities without relations (negative examples).For example, in relation extraction in the healthcare domain, relations between diseases and therapeutic drugs may be a minority of samples, whereas entity pairs of diseases and other drugs that do not have relations may make up the majority. This situation leads to a greater tendency for the model to predict samples as negative examples, which reduces the ability to recognise true relationships and affects model training and performance. The core of negative relationship sampling is to select a portion of the large number of negative samples that are both as representative and diverse as the positive samples, without introducing too much noise, in order to balance the proportion of positive and negative samples. This can effectively avoid excessive preference for negative examples during model training, thus improving the model's ability to recognise real relationships.

In the experiments of this paper, for a predefined set of relations$R = \{r_1, r_2, \ldots r_K\}$, the set of sampled relations is obtained by negatively sampling the relations to alleviate the problem of imbalance between positive and negative relations:

$$\tilde{R} = \{\widetilde{r_1}, \widetilde{r_2}, \ldots, \widetilde{r_{NS}}\} = NegativeSample(R, NS) \tag{6}$$

Where $NegativeSample(R, NS)$ denotes the retention of all positive relationships in the set of relationships R, while negative relationships are randomly sampled and the total number of positive and negative relationships is ensured as NS.This is done to ensure a balanced sample of positive and negative relations during training.

3. Experimental results and analysis

3.1. Experimental parameter settings

This experiment uses a unified runtime environment with Pytorch version 1.8.1, Python version 3.7, and the operating system Ubuntu 18.04. All experiments were conducted on a computer equipped with a

12th Generation Intel® Core™ i5-12400 CPU, 16GB RAM, and an NVIDIA RTX A5000 with 24GB of memory.Conducted.The specific experimental environment is shown in Table 1.

Table 1. Configuration table of experimental environment

Name	deployment
CPU	Intel®Core™i5-12400 CPU
GPU	NVIDIA RTX A5000 24GB
Operating System	Ubuntu 18.04
Memory	24GB RAM
Deep Learning Framework	Pytorch 1.8.1
Programming Languages	Python 3.7

3.2. Introduction of dataset

In the entity-relationship extraction task, the dataset used in this paper's experiment is the self-constructed diabetes domain dataset and the Baidu DuIE 2.0 dataset.

The Baidu DuIE 2.0 dataset contains text data from multiple domains, such as news, Baidu encyclopaedia, microblogs, and so on. Each sample consists of an entity, a relationship and a context, and its dataset is the largest schema-based Chinese information extraction dataset in the industry, containing more than 210,000 Chinese sentences and 48 defined relationship types.

3.3. Existing model comparison experiment results and analysis

The Chinese diabetes relationship extraction model NRS-Onerel model proposed in this paper is compared with the existing entity relationship extraction model in the two datasets used for the comparison experiment, the models of entity relationship extraction methods include two kinds of Pipeline pipeline extraction methods and joint extraction. The comparison models selected for this experiment are as follows:

Pipeline: in this paper, a pipeline model is designed for named entity recognition using the BERT-BiLSTM-CRF model, from which entity pairs are obtained. Then, the long and short-term memory neural network model approach based on the attention mechanism by Zhou et al [5] is used to identify the relationship between entities in a sentence.

BERT: a pre-trained model introduced by Google, which is used as the baseline model for the experiments in this paper.

BERT-Gather: integrates contextual information of sentence head entities using the pre-trained BERT model.

WDec: an approach proposed by Nayak et al [6] to jointly extract entities and relations using an encoder-decoder architecture.

CasRel: an approach that uses interaction information and contextual features between entities to infer relationships between entities by constructing sentence-level graph networks.

TPLinker: an approach that employs the representation of entities and inter-entity relationships in a table to capture semantic associations between entities using graph neural network structure.

MGRSA: a framework for multilevel gated recurrent mechanism proposed by Zhong [7] to merge word

granularity information into character granularity information for entity relationship extraction.

Seq2UMTree: a method proposed by Zhang et al [8] that uses sequence-to-tree conversion technique to convert text sequences into tree structures and performs relationship extraction on the tree structure.

Onerel: joint extraction task transformed into a fine-grained ternary classification problem.

Each model is experimented on the three datasets mentioned above to compare the existing entity-relationship extraction models with the NRS-Onerel model proposed in this paper in terms of the metrics of precision rate, recall rate and F1 value, Table 2 shows the experimental results on the diabetes domain dataset, and Table 3 shows the experimental results on the Baidu DuIE 2.0 dataset.

Table 2. Results of comparative experiments on the diabetes domain dataset

Model	P(%)	R(%)	F1(%)
Pipeline	76.4	42.7	54.8
BERT	82.5	73.6	77.8
BERT-Gather	81.2	79.5	80.4
WDec	62.8	69.0	65.7
CasRel	85.5	81.6	83.5
TPLinker	82.0	80.9	81.5
MGRSA	79.4	83.0	81.2
Seq2UMTree	78.3	80.7	79.5
Onerel	85.6	84.5	85.0
NRS-Onerel	**88.8**	**85.1**	**86.9**

Table 3. Comparative experimental results on Baidu DuIE 2.0 dataset

Model	P(%)	R(%)	F1(%)
Pipeline	40.3	49.0	44.3
BERT	72.9	65.7	69.1
BERT-Gather	75.4	70.6	72.9
WDec	64.1	54.2	58.7
CasRel	72.4	71.0	71.7
TPLinker	73.4	62.4	67.4
MGRSA	73.5	65.1	69.2
Seq2UMTree	70.5	65.1	68.7
Onerel	77.0	68.9	72.8
NRS-Onerel	**77.5**	**71.8**	**74.5**

The experimental results show that on the diabetes domain dataset, the NRS-Onerel model proposed in this paper achieves an F1 value of 86.9%, which is an improvement of 9.1% relative to the baseline model BERT.And on the Baidu DuIE 2.0 dataset, the model achieves an F1 value of 74.5%, which is improved by

7.1% and 2.8% relative to the F1 values of the current mainstream entity-relationship extraction models TPLinker and CasRel, respectively.These results fully validate the effectiveness of the entity-relationship extraction model proposed in this paper.

The NRS-Onerel model in this paper introduces a negative relationship sampling strategy, a strategy that aims to ensure that the model can fully learn the features and relationships between positive and negative examples during the training process by balancing the number of positive and negative relationship samples in the dataset.In this way, the model is able to better distinguish between different relationship types and improve the training effect.The use of negative relation sampling helps to prevent the model from being overly biased towards negative samples during the learning process, thus improving the model's ability to recognise positive examples.

3.4. Ablation experiment results and analysis

In order to assess the impact of the negative relation sampling strategy proposed in this paper on the effectiveness of the model, ablation experiments were conducted to verify the impact of different numbers of negative relation sampling on the model performance under the two selected datasets.Fig. 3 demonstrates the results of the experiments under the diabetes domain dataset, while Fig. 4 shows the results of the experiments under the Baidu DuIE 2.0 dataset.As can be observed from Fig. 3, under the diabetes domain dataset, the NRS-Onerel model achieves the highest F1 value when NS=10 is used, which is due to the fact that the negative relation sampling used makes the number of positive and negative relation samples close to each other, which reduces the model bias caused by data imbalance, and thus obtains the highest F1 value score.In addition, with the gradual increase in the number of NS, the F1 value shows a decreasing trend.And as seen in Fig. 4, under the Baidu DuIE 2.0 dataset, the NRS-Onerel model has the highest F1 value when NS=50.When NS is less than 50 or greater than 60, the F1 value shows a decreasing trend.These results demonstrate the impact of negative relational sampling on model performance and provide a basis for selecting the appropriate number of negative samples to be sampled under a particular dataset.

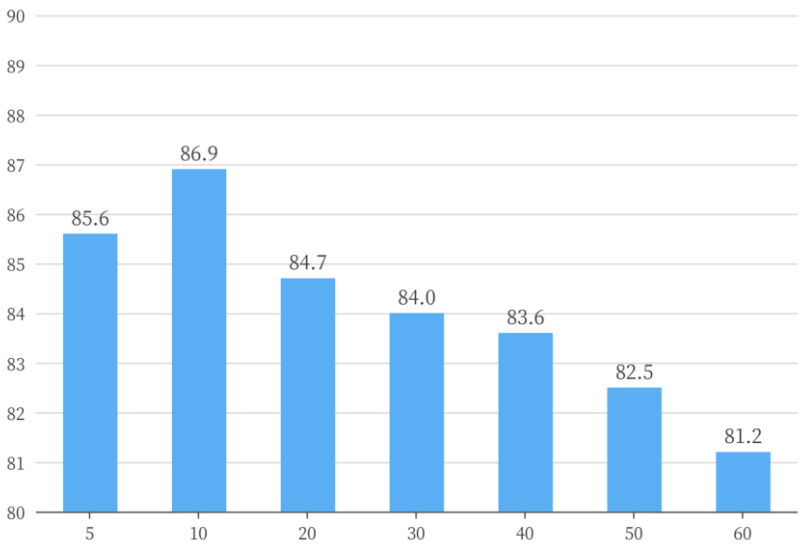

Fig. 3. Experimental results for different negative relationship sampling numbers under the diabetes domain

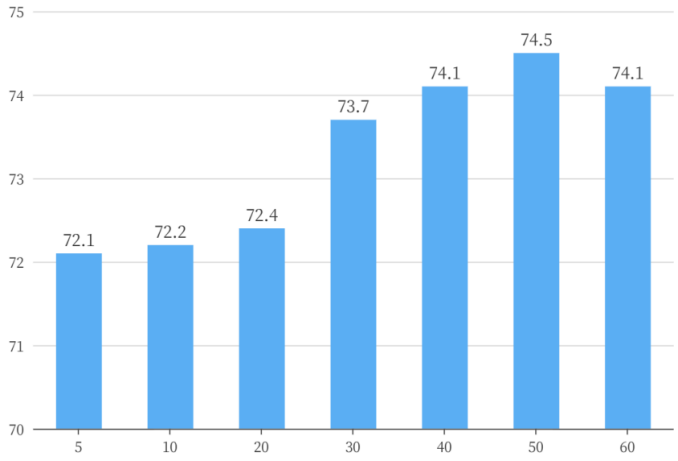

Fig. 4. Experimental results of different negative relationship sampling numbers under Baidu DuIE 2.0 dataset

It can be seen that as the number of NS gradually approaches the total number of relations in the dataset, the more significant the effect of the model is improved. This phenomenon verifies that when the number of NS is larger than the number of relations in the dataset, the positive and negative relation samples in the training data will present an imbalanced state. In this case, the training of the model is more inclined to learn negative relations, and therefore the final prediction results will be more biased towards negative relations, which leads to a decrease in model performance. On the contrary, when the number of NS is smaller than the number of relations in the dataset, it will lead to insufficient model training, which in turn affects the model performance.

4. Conclusion

This paper firstly introduces the ideas of multi-module multi-step, multi-module one-step and single-module one-step extraction processes in entity-relationship joint extraction. Aiming at the existing problems of single-module one-step relationship extraction, the relationship extraction model NRS-Onerel is designed to alleviate the problem of positive and negative relationship imbalance in the single-module one-step training process in order to improve the accuracy of the relationship extraction. The NRS-Onerel model uses the BERT pre-training language model to generate word embeddings, and adopts a negative relationship sampling strategy in order to solve the problem of the imbalance of positive and negative relationships. And the extraction of entity relations is achieved by a single-module one-step process of the Onerel model. The experimental results show that the NRS-Onerel model proposed in this paper outperforms the models of the pipelined approach and the joint entity-relationship extraction approach for entity-relationship extraction under both datasets used, exhibiting better performance and confirming the effectiveness of the entity-relationship extraction method based on the negative relationship sampling strategy proposed in this paper.

References

[1] WEI Z, SU J, WANG Y, et al. A novel cascade binary tagging framework for relational triple extraction, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020: 1476-1488.

[2] WANG Y, YU B, ZHANG Y, et al. TPLinker: Single-stage joint extraction of entities and relations through token pair linking, Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain: International Committee on Computational Linguistics, 2020: 1572-1582.

[3] SHANG Y-M, HUANG H, MAO X. Onerel: Joint entity and relation extraction with one module in one step. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(10): 11285-11293.

[4] NICKEL M, ROSASCO L, POGGIO T. Holographic embeddings of knowledge graphs. Proceedings of the AAAI conference on artificial intelligence, 2016: 30(1).

[5] ZHOU P, SHI W, TIAN J, et al. Attention-based bidirectional long short-term memory networks for relation classification. Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers), Berlin, Germany: Association for Computational Linguistics, 2016: 207-212.

[6] NAYAK T, NG H T. Effective modeling of encoder-decoder architecture for joint entity and relation extraction. Proceedings of the AAAI conference on artificial intelligence, 2020, 34(5): 8528-8535.

[7] ZHONG Z. Chinese entity relation extraction based on multi-level gated recurrent mechanism and self-attention. 2021 2nd International Conference on Artificial Intelligence and Information Systems, 2021, 306: 1-7.

[8] ZHANG R H, LIU Q, FAN A X, et al. Minimize exposure bias of seq2seq models in joint entity and relation extraction. arXiv preprint arXiv:2009.07503, 2020.

[9] MILLER J J. Graph database applications and concepts with Neo4j, Proceedings of the southern association for information systems conference, Atlanta, GA, USA. 2013, 2324(36): 141-147.

[10] FRANCIS N, GREEN A, GUAGLIARDO P, et al. Cypher: An evolving query language for property graphs, Proceedings of the 2018 international conference on management of data. Houston, United States, 2018: 1433-1445.

[11] HINTON G E. Learning distributed representations of concepts. Proceedings of the eighth annual conference of the cognitive science society, 1986: 12.

[12] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[13] PETERS M E, NEUMANN M, ZETTLEMOYER L, et al. Dissecting contextual word embeddings: Architecture and representation. arXiv preprint arXiv:1808.08949, 2018.

[14] CUI Y, CHE W, LIU T, et al. Pre-training with whole word masking for Chinese BERT. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.

[15] CUI Y, CHE W, LIU T, et al. Revisiting pre-trained models for Chinese natural language processing. arXiv preprint arXiv:2004.13922, 2020.

[16] XU L, DONG Q, LIAO Y, et al. CLUENER2020: fine-grained named entity recognition dataset and benchmark for Chinese. arXiv preprint arXiv:2001.04351, 2020.