The 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)
Taipei, Taiwan, November 21-22, 2022. The Association for Computational Linguistics and Chinese Language Processing

YNU-HPCC at ROCLING 2022 Shared Task: A Transformer-based Model with Focal Loss and Regularization Dropout for Chinese Healthcare Named Entity Recognition

Xiang Luo, Jin Wang and Xuejie Zhang
School of Information Science and Engineering
Yunnan University
Kunming, China
Contact: {wangjin,xjzhang}@ynu.edu.cn

Abstract

Named Entity Recognition (NER) is a fundamental task in information extraction that locates the mentions of named entities and classifies them in unstructured texts. Previous studies typically used hidden Markov model (HMM) and conditional random fields (CRF) for NER. To learn long-distance dependencies in text, recurrent neural networks, e.g., LSTM and GRU can extract the semantic features for each token with a sequential manner. Based on Transformers, this paper describes the contribution to ROCLING-2022 Share Task. This paper adopts a transformer-based model with focal Loss and regularization dropout. The focal loss is to overcome the uneven distribution of the label. The regularization dropout (r-drop) is to address the problem of vocabulary and descriptions that are too domain-specific. The ensemble learning is to improve the performance of the model. Comparative experiments were conducted on dev set to select the model with the best performance for submission. That is, BERT model with BiLSTM-CRF, focal loss and R-Drop has achieved the best F_1-score of 0.7768 and rank the 4th place.

Keywords: Chinese Healthcare Named Entity Recognition, Sequence Labeling, Information Extraction, Transformers, Conditional Random Fields

1 Introduction

Providing computer the ability to understand the abstract meaning of real world is a fundamental task. The shared task of ROCLING-2022 is Chinese healthcare named entity recognition task. Given a sentence about Chinese healthcare, the intelligent model is required to produce the entities in this sentence.

Table 1 provides a detailed description of all target labels. For example, the input is 膽汁長期滯留就會比較容易造成膽沙和膽結石了。, the intelligence model is expected to extract three entities, including 膽汁 as BODY, and both 膽沙 and 膽結石 as DISE. By using a sequence labeling approach, the corresponding labels for all tokens should be B-BODY, I-BODY, O, O, O, O, O, O, O, O, O, O, O, O, B-DISE, I-DISE, O, B-DISE, I-DISE, I-DISE, O, O. Here, the BIO schema is adopted, where B and I respectively means the begin and inside labels, while O indicates that a token belongs to other objects.

Previous studies used probabilistic model for named entity recognition on text, such as hidden Markov model (HMM) (Zhou and Su, 2002) and conditional random field (CRF) (Zheng et al., 2017). Recent advances in deep neural networks (DNN) (Krizhevsky et al., 2012) and representation learning (Bengio et al., 2013) have considerably improved the ability of NER models. It mainly consists of an encoder to learn hidden representation for each token, as well as a classifier to assign a label for the token. For encoders, traditional models are usually used recurrent neural networks (RNN), such as long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) or gated recurrent units (GRU) to learn long-distance dependencies. Furthermore, attention mechanisms can be applied to improve the performance of RNN models to extract more task-specific features between tokens to provide meaningful information. Several effective approaches apply the pre-trained language models (PLM), such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and ALBERT (Lan et al., 2020), to provide powerful representation to boost the performance of

Entity Type	Description
Body(BODY)	The whole physical structure that forms a person or animal including biological cells, organizations, organs and systems.
Symptom(SYMP)	Any feeling of illness or physical or mental change that is caused by a particular disease.
Instrument(INST)	A tool or other device used for performing a particular medical task such as diagnosis and treatments.
Examination(EXAM)	The act of looking at or checking something carefully in order to discover possible diseases.
Chemical(CHEM)	Any basic chemical element typically found in the human body.
Disease(DISE)	An illness of people or animals caused by infection or a failure of health rather than by an accident.
Drug(DRUG)	Any natural or artificially made chemical used as a medicine
Supplement(SUPP)	Something added to something else to improve human health.
Treatment(TREAT)	A method of behavior used to treat diseases
Time(TIME)	Element of existence measured in minutes, days, years

Table 1: The detailed description of all target labels.

sequence labeling.

Furthermore, some studies have tried to transform the NER task as a machine reading comprehension (MRC) (Li et al., 2020) or a candidate span extraction (Ji et al., 2020). For the former, the multi-classification problem of named entity recognition is converted into a Q&A task. The model is asked each piece of data, and then answer it through the location information of the start and end position of the entity. For the latter, the candidate span extraction is divided into two parts, The first part is candidate extraction, and this part is similar in structure to most of the previous extractive question answering models, mainly responsible for extracting candidate answers from the passage. The second part is answer selection, which is mainly responsible for selecting the most reliable answer from all the candidate answers, and considering the relationship between all the candidate answers.

By using the sequence labeling manner, the task brings two difficulties may finally impact the performance of recent NER models. One of the biggest stumbling blocks is data distribution, which often appears in conventional sequence labeling tasks and corpora. Figure 1 provided other two examples of the shared tasks. Notably, most of the labels are O. The target tokens of Chinese healthcare entities in both examples only take respective ratios of 12.9% and 20.0%. The proportion of meaningless O label is dominate. By using a cross-entropy loss function, the model may tend to assign O label for all tokens thus the model can achieve the minimal cross-entropy. However, it will be useless for the task where these minority labels, e.g., BODY, CHEM and DISE, are more important than the majority labels. That is, false negatives can have higher importance, while false positives are of course undesirable. Another important issue is that the expression is healthy-related and domain-specific, thus may limit the learning ability of the encoders which are usually pretrained on domain-independent texts.

Input	專注於每個呼吸，透過鼻子的一吸一吐的換氣，促進舒緩與放鬆身心。
Output	B-BODY, I-BODY, O, O, O, O, O, O, O, O, O, O, O, O, B-DISE, I-DISE, O, B-DISE, I-DISE, I-DISE, O, O
Ratio	12.9%

Input	平時避免酒精、咖啡和茶等利尿食品，以防低血壓加重。
Output	O, O, O, O, B-CHEM, I-CHEM, O, O, O, O, O, O, O, O, O, O, O, O, O, B-DISE, I-DISE, I-DISE, O, O, O
Ratio	20.0%

Figure 1: The imbalanced examples in labeling of Chinese HealthNER Corpus.

In this paper, we employed pretrained language models, including BERT, RoBERTa, ELECTRA (Clark et al., 2020) and ALBERT, for the Chinese healthcare named entity recognition task. To address the imbalance distribution of labels, we applied focal loss (Lin et al., 2020) on the CRF classifier. Further, a regularized dropout mechanism (Liang et al., 2021) was used to further enhance the performance of the base encoders. In addition, we tried to ensemble all base encoders as a more powerful model. Unfortunately, this did not bring any improvements on performance.

The rest of this paper is organized as follows. Section 2 describes all the models which are used in this task. Experimental results are summarized in Section 3. Conclusion is finally drawn in Section 4.

2 Model Description

This section will describe the architecture of the proposed model in details. There are several components in this section, including

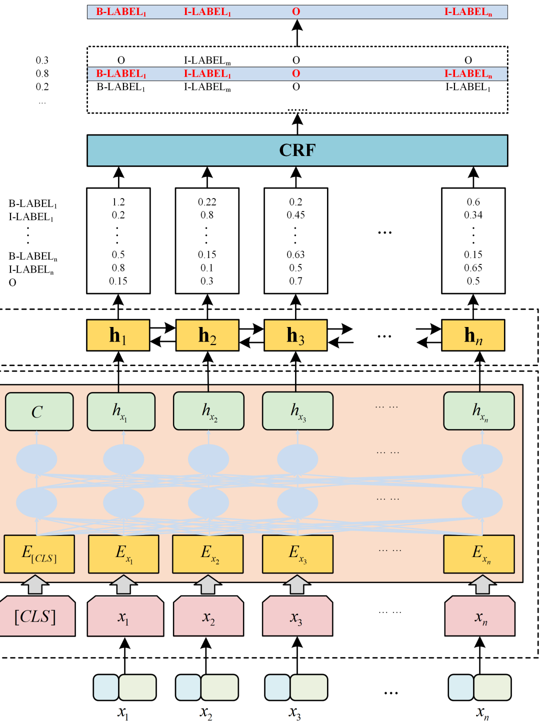

Figure 2: The overall architecture of the proposed method.

BERT, BiLSTM-CRF, focal loss, and R-Drop. The model architecture is shown in Figure 2.

2.1 Method

BERT. BERT was pretrained by two tasks, masked language model (MLM) and next sentence prediction (NSP), which is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. The checkpoint hfl/chinese-bert-wwm-ext (Cui et al., 2020) is used in the model, which uses 12-layer, 768-hidden, 12-heads and 110M parameters. For each layer, The attention takes its input in the form of three parameters, i.e., query, key and value. All three parameters are similar in structure, with each word in the sequence represented by a vector, denoted as,

$$Attention\left(Q,K,V\right)=softmax\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V \tag{1}$$

The attention module splits its query, key, and value parameters N-ways and passes each split independently through a separate Head. All of these similar attention calculations are then combined together to produce a final attention score as follows,

$$\begin{aligned}&MutiHead\left(Q,K,V\right)=Concat\left(head_1,...,head_h\right)W^{O}\\&where\ head_i=Attention\left(QW_i^{Q},KW_i^{K},VW_i^{V}\right)\end{aligned} \tag{2}$$

BiLSTM. The unidirectional LSTM model can only capture the information passed from head to tail. Conversely, the bidirectional LSTM can capture forward information and reverse information imultaneously, which makes the use of text information more comprehensive and the effect is better. And a linear layer is added after the final output layer of the BiLSTM network, which is used to project the output of the hidden layer generated by BiLSTM to an interval that expresses the meaning of the label features (Huang et al., 2015). The output of the BERT is used as the input of the BiLSTM as equation 3.

$$\left[H_1,H_2...H_N\right]=BiLSTM\left(\left[h_1,h_2...h_N\right]\right) \tag{3}$$

CRF. Conditional random fields is a conditional probability distribution model for solving the output sequence given the input sequence. The CRF layer can add some constraints to ensure that the final prediction result is valid. The CRF layer can learn the constraints of the sentence. These constraints

MODEL	**F_1-score**	**Submission**
BERT+RoBERTa+ELECTRA+ALBERT	0.827	
RoBERTa+ELECTRA+ALBERT	0.826	
BERT+RoBERTa+ALBERT	0.830	
BERT+RoBERTa+ELECTRA	0.830	
BERT+ELECTRA+ALBERT	0.827	
ALBERT+ELECTRA	0.822	
RoBERTa+ALBERT	0.824	
RoBERTa+ELECTRA	0.830	
BERT+ALBERT	0.829	
BERT+ELECTRA	0.829	
BERT+RoBERTa	0.831	Submission3
RoBERTa	0.832	Submission2
ELECTRA	0.822	
ALBERT	0.790	
BERT	**0.833**	**Submission1**

Table 2: F_1-score of each ensemble model in dev data.

can be learned automatically by the CRF layer when training the data. The CRF loss function is as Eq. 4:

$$\begin{aligned}
&\mathcal{L}_{CRF} = \log \frac{P_{\mathrm{Real}Path}}{P_1+P_2+\ldots+P_N} \\
&= -\log \frac{e^{S_{\mathrm{Real}Path}}}{e^{S_1}+e^{S_2}+\ldots e^{S_N}} \\
&= -\left(\log e^{S_{\mathrm{Real}Path}} - \log\left(e^{S_1}+e^{S_2}+\ldots+e^{S_N}\right)\right) \\
&= -\left(S_{\mathrm{Real}Path} - \log\left(e^{S_1}+e^{S_2}+\ldots+e^{S_N}\right)\right) \\
&= -\sum_{i=1}^{N} x_{iy_i} - \sum_{i=1}^{N-1} t_{y_i y_{i+1}} + \log\left(e^{S_1}+e^{S_2}+\ldots+e^{S_N}\right)
\end{aligned} \tag{4}$$

where the e is a constant, S is the score of the path, $x_{i,j}$ is the score at which the i-th indexed word is labeled as j. $t_{i,j}$ is the score of label i to label j.

Focal Loss. Focal loss is a loss function that deals with the imbalance of sample classification. It focuses on adding weight to the loss corresponding to the sample according to the difficulty of distinguishing the sample, that is, adding a small weight to the easy-to-distinguish sample and adding a large weight to the difficult-to-distinguish sample. The expression of the focal loss is as follows.

$$\mathcal{L}_{Focal} = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{5}$$

where the α_t is a trainable parameter, the γ is a hyper-parameter and the p_t is the probability of class t.

R-Drop. Due to the existence of dropout, the same model with the same input will get two different distributions, where it can approximately be treated as two different model networks. Based on this, the different distributions produced by these two different models can be denoted as, $P_\theta(y|x)$ and $P_\theta^{'}(y|x)$. The main contribution of R-Drop is to continuously lower the KL Divergence (KL divergence) between the two distributions during the training process. Due to the asymmetry of the KL divergence itself, the globally symmetric KL divergence is indirectly used by exchanging the positions of these two distributions, which is called bidirectional KL divergence. Additionally, the model is also trained on NLL loss terms for both distributions. The final loss is as follows:

$$\begin{aligned}
\mathcal{L}_{R-drop} = &-\log P_\theta(y_i|x_i) - \log P_\theta^{'}(y_i|x_i) \\
&+\alpha[D_{KL}\left(P_\theta(y_i|x_i)\,||P_\theta^{'}(y_i|x_i)\right) \\
&+D_{KL}\left(P_\theta^{'}(y_i|x_i)\,||P_\theta(y_i|x_i)\right)]
\end{aligned} \tag{6}$$

The final objective of the used model is defined as follows:

$$\mathcal{L} = \mathcal{L}_{CRF} + \mathcal{L}_{Focal} + \mathcal{L}_{R-drop} \tag{7}$$

2.2 Ensemble Learning

In ensemble learning, multiple models are trained to solve the same problem and are combined to get better results. The most important assumption is that when weak models are combined correctly, the more accurate or robust models can be got. The stacking strategy are used as ensemble learning model. Stacking usually considers heterogeneous weak learners and stacking learning to combine base models with meta-model. Besides BERT, we tried some other models, such as RoBERTa, ELECTRA, ALBERT. The detail is as follows.

RoBERTa. RoBERTa is a robustly optimized BERT pretraining approach. It is an improved recipe for training BERT models, that can match or exceed the performance of all of the post-BERT methods. The modifications include (1) training the model longer, with bigger batches, over more data; (2) removing the next sentence prediction objective; (3) training on longer sequences; and (4) dynamically changing the masking pattern applied to the training data. The checkpoint **hfl/chinese-roberta-wwm-ext** is used in the model, which uses 12-layer, 768-hidden, 12-heads and 125M parameters.

ALBERT. ALBERT is a lite BERT for self-supervised learning of language representations which lead to models that scale much better compared to the original BERT and it uses a self-supervised loss that focuses on modeling inter-sentence coherence, and show it consistently helps downstream tasks with multi-sentence inputs. ALBERT base model with no dropout, additional training data and longer training. The checkpoint **clue/albert_chinese_tiny** is used in the model, which uses 4-layer, 312-hidden, 12-heads and 16M parameters.

ELECTRA. ELECTRA is a new method for self-supervised language representation learning. It can be used to pre-trained transformer networks using relatively little compute. ELECTRA models are trained to distinguish *real* input tokens vs. *fake* input tokens generated by another neural network, similar to the discriminator of a GAN. The checkpoint **hfl/chinese-electra-180g-small-discriminator** is used in the model, which uses 12-layer, 256-hidden, 4-heads and 12M parameters.

After comparing the meta-model, the random forest model (Breiman, 2001) are chosen to be the meta-model, and the BERT, RoBERTa, ELECTRA and ALBERT models are taken as the base models. After fine-tuning the parameters, the final result is shown in Table 2.

3 Experimental Results

In this section, comparative experiments were conducted to select the best model as the final submission. The details of the experiments are presented as follows.

3.1 Dataset

The train dataset (Lee and Lu, 2021) describes 10 entity types in total, and use the common BIO (Beginning, Inside, and Outside) format for NER tasks. The B-prefix before a tag indicates that the character is the beginning of a named entity and I-prefix before a tag indicates that the character is inside a named entity. An O tag indicates that a token belongs to no named entity.

In the raw dataset, there are some descriptions about the sentences, such as id, genre, word, word_label, character, character_label. Because the task focuses on the character level labeling, we choose the character and character_label as the input and output.

3.2 Evaluation Metrics

The performance is evaluated by examining the difference between machine-predicted labels and human-annotated labels. We adopt standard precision, recall, and F_1-score, which are the most typical evaluation metrics of NER systems at a character level. Precision is defined as the percentage of named entities found by the NER system that are correct. The definition of Precision is as follows:

$$P = \frac{TP}{TP + FP} \tag{8}$$

Recall is the percentage of named entities present in the test set found by the NER system. The definition of Recall is as follows:

$$R = \frac{TP}{TP + FN} \tag{9}$$

F_1-score is an indicator used in statistics to measure the accuracy of a binary (or multi-class) model, which takes into account the accuracy and recall of the classification model at the same time. The definition of F_1-score is as follows:

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{10}$$

where TP is True Positive, FP is False Positive, FN is False Negative.

MODEL	LOSS	F_1-score
BERT+softmax	CrossEntropy	0.799
BERT+softmax	Focal	0.805
BERT+BiLSTM	Focal	0.812
BERT+BiLSTM+CRF	Focal	0.825
BERT+BiLSTM+CRF+R-Drop	**Focal**	**0.833**

Table 3: F_1-score of each strategy in dev data.

3.3 Implementation Details

The train data is split into train data and dev data. At first, we make pre-processing for the train data, which only obtain the characters and character labels. The tokenizer are used to convert token into vector, after that, we add the BiLSTM-CRF after the hidden output of the pre-trained model. And we find that the data is not evenly distributed in the dataset,so the focal loss is used to solve this kind of problems. It focuses on adding weight to the loss corresponding to the sample according to the difficulty of the sample discrimination.

Moreover, to strengthen the generalization of the model, the regularized dropout (R-Drop) is used. Due to the existence of dropout, the output of two models with the same parameters may also be different. In order to alleviate the inconsistency of this training process, we imposed restrictions on the output distribution, and the KL divergence loss of the data distribution metric is introduced, making the two data distributions generated by the same sample in the batch as close as possible.

Then we use dev data to select the best performing model and save it, where the evaluation metric is F_1-score. After that, the ensemble strategy is used to stack different models, and the random forest model is chosen to be the meta-model, which performs better than other classifier. There are many of combinations, we list the scores for each kind of model as well as the score for the base models in Table 2.

In addition, MRC is used in this task and MRC is quite used in NER task. When using MRC, the task is converted to a QA-type question. We need to allocate 10 queries to each sentence. Possibly due to the large amount of data, after the allocation, the whole amount of the data come to 230,000, or because the uneven distribution of the data, there are many ”O” labels, which affect the model prediction. Besides, the questioning method of query is also an aspect that affects the prediction of the model. So the MRC approach doesn't perform well.

Label embedding (Akata et al., 2015) is also another trick to enhance the understanding of the text for the model. Label embedding is to add the label of each word to the hidden representation of each word. It helps the model better understand the literal meaning of the label. But it also doesn't perform well. We guess that the insertion position may be wrong, or the embedding generated during inference is not appropriate.

3.4 Parameters Tuning

In this part, we use warm up strategy, which is an approach to optimize the learning rate. Warm up is a learning rate warm-up method mentioned in the ResNet (He et al., 2016) paper, which chooses to use a smaller learning rate at the beginning of training, and trains some epochs, and then modify it to a preset learning rate for training. Since the weights of the model are randomly initialized at the beginning of training, if a larger learning rate is selected at this time, the model may become unstable. Using the warm up method can make the learning rate smaller in several epochs at the beginning of training. Under the preheated small learning rate, the model can gradually become stable. When the model is relatively stable, the preset learning rate is selected for training, which makes the model converge faster and works better. The parameter tuning process is shown in the following Figure 3 and Figure 4.

Moreover, the grid search is used to find the optimal parameters. Finally the learning rate is set to 1e-4, the epoch is set to 25, the weight decay is set to 1e-7, and the warm up ratio is

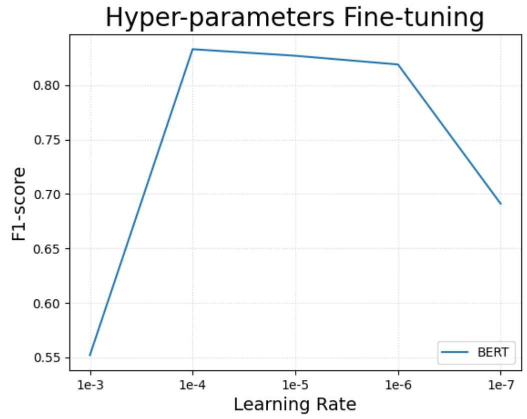

Figure 3: The performance of different learning rate on F_1-score.

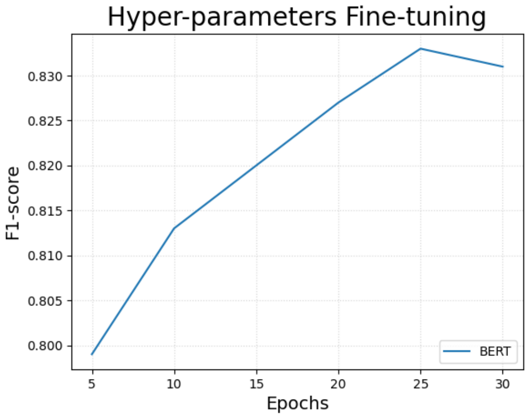

Figure 4: The performance of different epoch on F_1-score.

set to 0.1.

3.5 Comparative Results

The quantitative ablation experiments were conducted to select the best model. In the experiment, the BERT model get the highest F_1-score, which is 0.833, and the RoBERTa model get the second highest F_1-score, which is 0.832. The detailed F_1-score for each strategy is listed in the Table 3. For the final submission, we submitted three files. The results are predicted by RoBERTA, BERT+ELECTRA and BERT and their performance differences are shown in Table 2. BERT also achieved the best results in test dataset (Lee et al., 2022), which is 0.7768.

4 Conclusions

In this paper, we describe our entire experimental procedure, and finally achieve the best F_1-score of 0.7768 and rank the 4th place. For implementation, several different approaches were applied, such as MRC and label embedding. Unfortunately, they didn't perform well. We applied a BERT-BiLSTM-CRF architecture with warm up strategy and R-Drop, to get the best score.

Future works will attempt to explore more different span-based extraction methods for the Chinese healthcare NER task.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants Nos. 61966038 and 62266051. The authors would like to thank the anonymous reviewers for their constructive comments.

References

Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. 2015. Label-embedding for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(7):1425–1438.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Leo Breiman. 2001. Random Forests. *Machine Learning*, 45(1):5–32.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long* *and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA. IEEE.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. ArXiv:1508.01991 [cs].

Bin Ji, Jie Yu, Shasha Li, Jun Ma, Qingbo Wu, Yusong Tan, and Huijun Liu. 2020. Span-based joint entity and relation extraction with attention-based span-specific and contextual semantic representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 88–99.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Lung-Hao Lee, Chao-Yi Chen, Liang-Chih Yu, and Yuen-Hsien Tseng. 2022. Overview of the rocling 2022 shared task for chinese healthcare named entity recognition. In *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing*.

Lung-Hao Lee and Yi Lu. 2021. Multiple embeddings enhanced multi-graph neural networks for chinese healthcare named entity recognition. *IEEE Journal of Biomedical and Health Informatics*, 25(7):2801–2810.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online. Association for Computational Linguistics.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-Drop: Regularized Dropout for Neural Networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 10890–10905. Curran Associates, Inc.

T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. 2020. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis amp; Machine Intelligence*, 42(02):318–327.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*. ArXiv: 1907.11692.

Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1227–1236, Vancouver, Canada. Association for Computational Linguistics.

GuoDong Zhou and Jian Su. 2002. Named Entity Recognition using an HMM-based Chunk Tagger. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 473–480, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.