The 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)
Taipei City, Taiwan, October 20-21, 2023. The Association for Computational Linguistics and Chinese Language Processing

YNU-ISE-ZXW at ROCLING 2023 MultiNER-Health Task: A Transformer-based Model with LoRA for Chinese Healthcare Named Entity Recognition

Xingwei Zhang, 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 subtask in the field of information extraction in natural language processing (NLP). Its main goal is to recognize named entities in text and classify them into predefined categories. In the medical field, NER technology is used to automatically identify medical-related entities, such as symptoms, examinations, diseases, and drugs, so that medical staff can better treat patients. For the named entity recognition task in the medical field proposed by ROCLING 2023, we built three models based on Transformers and used technologies such as Focal Loss and LoRA. We conducted comparative experiments on the development set and the test set, and found that the effects of the three models were not much different. Finally, our submitted DeBERTa model named RUN3 achieved a macro-f1 score of 67.79, ranking 5th.

Keywords: Chinese Named Entity Recognition, DeBERTa, Transformers, Focal Loss, LoRA

1 Introduction

Named Entity Recognition (NER) (also known as Entity Recognition, Entity Chunking, and Entity Extraction) is a subtask of Information Extraction that aims to locate and classify named entities in text into predefined categories such as people, organizations, locations etc. The shared task proposed by ROCLING 2023 is the Chinese multi-genre named entity recognition task in the medical field.

For each sentence in the data, we need to identify the type and boundary of each entity in the sentence. Table 1 details each entity type along with some examples (Lee and Lu, 2021). In this task, we adopt BIO mode. The B and I before the mark represent the start and internal tags of the entity, respectively, and the O represents that the character does not belong to any entity.

Compared with English NER, Chinese named entity recognition has the following difficulties:

- Word segmentation problem: One of the characteristics of the Chinese language is that there is no obvious word separator, so word segmentation needs to be performed first when performing NER. Wrong word segmentation will affect the results of NER, especially for some common entity words. For example, ”New York University” is incorrectly split into ”New York” and ”University”.

- Ambiguity problem: Some words may denote different entity types in different contexts. For example, ”apple” could refer to a fruit, or it could refer to a technology company. Contextual information is crucial to disambiguate this.

- Data scarcity: Compared with English, Chinese NER data resources may be relatively scarce, which makes training models more difficult. Lack of large-scale, high-quality labeled data limits model performance.

- Domain adaptability: There may be differences in named entity recognition in different domains. A general-purpose model may not perform as well in a specific domain as a model trained specifically for that domain.

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

Table 1: Named Entity Types and Detailed Descriptions.

For the task of named entity recognition, traditional methods include Hidden Markov Model (Zhou and Su, 2002), Conditional Random Field (Zheng et al., 2017), Maximum Entropy Model (Fresko et al., 2005), Support Vector Machine (Isozaki and Kazawa, 2002), etc. With the development of deep learning, some new methods continue to emerge. Such as BiLSTM+CRF (Zeng et al., 2017), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and so on. In this paper, we use pre-trained language models such as ALBERT (Lan et al., 2020), ELECTRA (Clark et al., 2020), BERT, RoBERTa and DeBERTa (He et al., 2021) to build a Chinese medical named entity recognition model. In order to solve the problem of classification imbalance in samples, we apply Focal Loss (Lin et al., 2017) to the model. In addition, we used a learning rate warm-up mechanism during training to make the neural network more stable during training. In addition to this, we also included LoRA (Hu et al., 2022) in the model, which allows us to fine-tune large models while consuming less memory, thus greatly reducing our need for video memory. Finally, we also use mechanisms such as gradient clipping to improve the performance of the model.

The rest of the paper is briefly introduced as follows. Section 2 describes the models and techniques we use. Section 3 introduces the content and results of the experiment in detail. Section 4 draws conclusions on the whole of this paper.

2 Proposed Method

This section describes the model architecture and some techniques used during training. These include BERT, RoBERTa, DeBERTa, Focal Loss, Warmup, and LoRA, among others. Figure 1 shows the overall architecture of the model.

2.1 BERT

BERT model which is based on Transformers (Vaswani et al., 2017) learns rich contextual representations through pre-training on large-scale text data, making it perform well on a variety of downstream NLP tasks. It no longer uses the traditional one-way language model or the method of shallow splicing two one-way language models for pre-training as before, but uses the new Masked Language Model (MLM), so it can generate deep two-way language characterization. After pre-training BERT, you only need to add an additional output layer for fine-tune, and you can achieve state-of-the-art performance in a variety of downstream tasks. Moreover, there is no need to modify the structure of BERT in this process. We applied the checkpoint hfl/chinese-bert-wwm-ext (Cui et al., 2021) in the model. In subsequent experiments, this model also showed good results.

2.2 RoBERTa

RoBERTa has improved the BERT model. The RoBERTa model includes unsupervised Pre-train and supervised Fine-tune, which improves the shortcomings of BERT training.

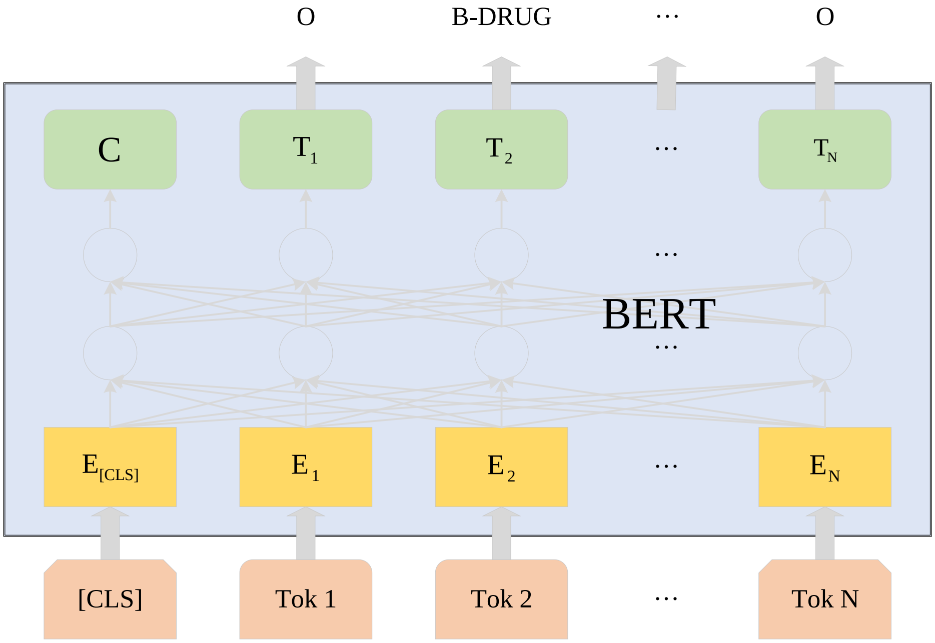

Figure 1: The overall structure of the model

Throughout the training process, larger model parameters were used, larger batch size and more training data were tried. RoBERTa builds on BERT's language masking strategy, modifying key hyperparameters in BERT. The Next Sentence Prediction (NSP) task in BERT, which has little effect on the results, is also deleted, and the model is trained with a dynamic mask. RoBERTa also received an order of magnitude more training than BERT and took longer. This enables RoBERTa to represent richer feature information than BERT, and can be better generalized to downstream tasks. We apply the checkpoint hfl/chinese-roberta-wwm-ext-large to the model. This checkpoint contains 24 layers of Transformers, 16 Attention Heads, and 1024 hidden layer units. It has achieved leading results on many Chinese datasets.

2.3 DeBERTa

The DeBERTa surpassed the performance of humans on the SuperGLUE leaderboard for the first time. The main framework of DeBERTa utilizes Transformer's Encoder. DeBERTa has mainly made two improvements on the basis of BERT, the Disentangled Attention Mechanism (DAM) and the Enhanced Mask Decoder (EMD). The principle of DAM is to represent each word with two vectors, encoding its content and relative position respectively. Then, according to the content and relative position of the word, the weight is calculated through Transformer's Self-attention mechanism, and the calculation of content to position and position to content is added. EMD introduces the absolute position information of words, and improves the timing of incorporating the absolute position information of words. It adds absolute position information before the softmax layer, avoiding the problem that the BERT model introduces absolute positions too early, which may cause the model to learn insufficient relative positions. In this experiment, we chose the checkpoint KoichiYasuoka/deberta-xlarge-chinese-erlangshen-ud-goeswith. In our experiments, this model achieved the best results.

2.4 Focal Loss

Focal Loss solves the imbalance of categories and differences in classification difficulty in classification problems. In the task of named entity recognition, there are much fewer entities in a sentence than non-entities, which is a severe category imbalance. So we apply Focal Loss to the model to improve performance. Focal Loss balances the weight of easy-to-classify and hard-to-classify samples by introducing an adjustment factor. Specifically, for samples that are easy to classify, the adjustment factor will reduce their weights, thereby reducing their contribution to the loss; while for samples that are difficult to classify, the adjustment factor will increase their weights, making them occupy a greater proportion in loss calculation. This mechanism helps the model to pay more attention to those misclassified samples that are difficult to classify, thereby improving the classification performance of the minority class. Focal Loss can improve the training effect of the model on the category imbalance dataset to a certain extent, making it easier for the model to learn the characteristics and distinguishing ability of a few categories. Equation 1 shows the calculation method of focal loss.

$$FL(p_t) = -\alpha_t(1-p_t)^\gamma \log(p_t) \tag{1}$$

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

2.5 LoRA

Recently, the development of pre-trained language models has promoted the research in the field of NLP to a new stage. Without manual labeling, general language representation can be learned from a massive corpus, and the performance of downstream tasks can be significantly improved. The parameters of the pre-trained language model are getting larger and larger, such as GPT-3 contains 175 billion parameters. Therefore, it is very difficult to fine-tune large-scale pre-trained language models. LoRA solves this dilemma very well. It freezes the pre-trained model weights and injects a trainable rank factorization matrix into each layer of the Transformer architecture, greatly reducing the number of trainable parameters for downstream tasks.

For a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, we constrain its update by representing the latter with a low-rank decomposition $W_0+\Delta W = W_0+BA$, where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, and the rank $r \ll \min(d,k)$. During training, W_0 is frozen and does not receive gradient updates, while A and B contain trainable parameters. Note both W_0 and $\Delta W = BA$ are multiplied with the same input, and their respective output vectors are summed coordinate-wise. For $h = W_0x$, the modified forward pass yields:

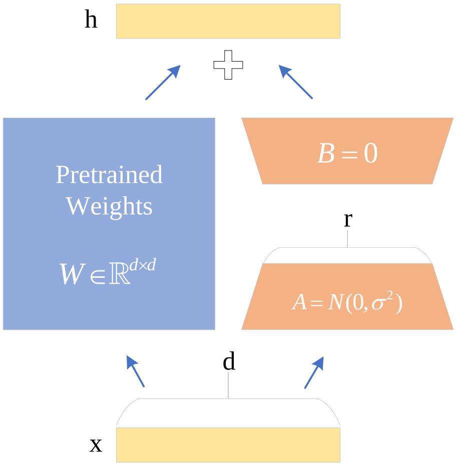

Figure 2: The reparametrization of LoRA.

$$h = W_0x + \Delta Wx = W_0x + BAx \tag{2}$$

When we apply the LoRA mechanism to our model, the number of trainable parameters in the model is greatly reduced. This makes the memory occupied by the model greatly reduced, and the training speed is also much faster. Thus, we have the opportunity to fine-tune larger models on our own devices. This helps us achieve better results.

2.6 Warm up

At the beginning of training, the weights of the model are randomly initialized. At this time, if a larger learning rate is selected, it may cause instability (oscillation) of the model. Therefore we use the warm up strategy in training. In the first few epochs of training, we choose a small learning rate, which can make the model gradually stabilize. After the model is relatively stable, select the preset learning rate for training, which can make the model converge faster and the model effect is better.

3 Experimental Results

We conducted a large number of comparative experiments using different models and parameters. Finally, we submit the results generated by the best performing model. In this section, we describe the experimental details and results.

Entity Type	Quantity	Entity Type	Quantity
BODY	52146	DISE	29072
SYMP	25513	DRUG	7230
INST	3117	SUPP	7955
EXAM	6768	TREAT	8495
CHEM	17583	TIME	3904

Table 2: Type and quantity of entities in the training set.

3.1 Datasets

The dataset used in this study is provided by ROCLING 2023 Shared Task I. In this task, data comes from three sources:

1. Formal texts (FT): this includes health news and articles written by professional editors or journalists.
2. Social media (SM): this contains texts from crowed users in medical question/answer forums.
3. Wikipedia articles (WA): this free online encyclopedia includes articles created and edited by volunteers worldwide.

Among them, FT and SM are from Chinese HealthNER Corpus (Lee and Lu, 2021), and WA is from RCOLING-2022 CHNER datasets (Lee et al., 2022a). Since the official development set is not given, we use 80% of the data in the given training set as the training set, and the remaining 20% of the data as the development set. There are a total of 10 entity types in the dataset and use the common BIO format. The B and I before the mark represent the start and internal tags of the entity, respectively, and the O represents that the character does not belong to any entity. Table 2 shows the type and quantity of each entity in the dataset.

In the data set, each piece of training data has 7 parameters, including id, genre, sentence, word, word_label, character, character_label. Since the NER task only needs to use character-level information, we only extract the character and character_label information in each piece of data during training.

3.2 Evaluation Metrics

We adopt standard precision, recall, and F1-score, which are the most typical evaluation metrics of NER systems at a character level. If the predicted tag of a character in terms of BIO format was completely identical with the gold standard, that is one of the defined BIO tags, the character in the testing instance was regarded as correctly recognized. Precision is defined as the percentage of named entities found by the NER system that are correct. Recall is the percentage of named entities present in the test set found by the NER system. Precision and recall are defined as equation 3 and equation 4, where TP, FP, and FN represent True Positive, False Positive, and False Negative, respectively.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

The official test set for testing scores is divided into three types, namely Formal Texts, Social Media and Wikipedia Articles. Different types of test sets will be evaluated independently. The Macro-averaging F1 score among three genres will be used for final ranking in the leaderboard. The definition of F1-score is as follows:

$$F_1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

3.3 Implementation Details

First, we split the officially provided training set into a training set and a development set. Second, we preprocess the training data and only extract the characters and character labels. Then use the tokenizer to convert the token into a vector. Finally, the vector is sent to the pre-training model to get the output. During the training process, we found errors in some labels in the training set, such as ”I-CHEM” was mislabeled as ”i-CHEM”, ”B-TIME” was mislabeled as ”T-TIME”, etc. We corrected all wrong labels. After a detailed analysis of the dataset, we found that the data is not evenly distributed in the sample. Therefore, we use Focal Loss to apply different weights to different samples to solve this problem. We also applied LoRA in the model. This can help us reduce the amount of parameters during training, and at the same time

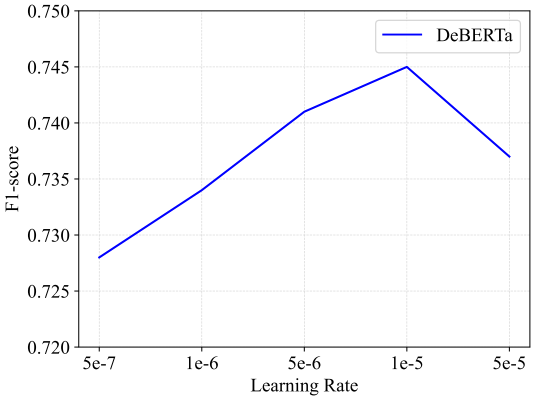

Figure 3: F1-score at different learning rates.

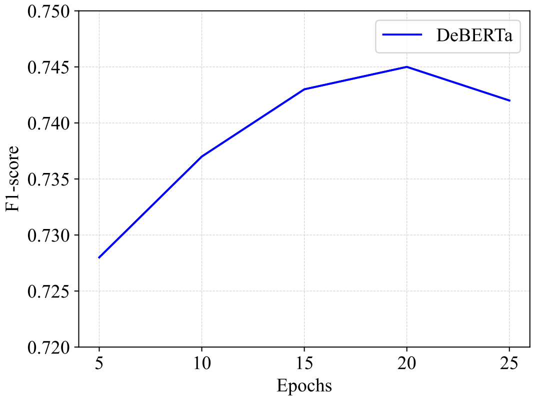

Figure 4: F1-score at different epochs.

fine-tune large-scale pre-trained language models on devices with small memory. Finally, we save the three models with the highest F1-score on the development set and submit the results of using these three models to predict the test set.

3.4 Parameters Fine-tuning

During training, we found that when the learning rate is large, there will be a phenomenon of gradient explosion. So we use the warm up strategy. The initial weights of the model are initialized randomly. Warm up will choose a smaller learning rate at the beginning of training, which can make the model gradually stabilize. Wait for the model to be relatively stable before training with the preset learning rate.

We also use a gradient clipping strategy. Gradient clipping is a method of changing or clipping the error derivative to a threshold during network backpropagation, and using the clipped gradient to update the weights. By rescaling the error derivative, updates to the weights will also be rescaled, significantly reducing the chance of overflow or underflow. In the training parameter part of the model, the learning rate is 1e-5, the batch size is 1, and the epochs is 20. Also, set max_grad_norm to 5 to prevent exploding gradients. Moreover, r is set to 16 in LoRA.

Model	Precision	Recall	F1-score
ALBERT	0.649	0.757	0.699
ELECTRA	0.666	0.781	0.719
BERT	0.687	0.797	0.738
RoBERTa	0.712	0.771	0.740
DeBERTa	0.706	0.788	0.745

Table 3: F1-score of different strategies.

3.5 Comparative Results

We also used ALBERT, ELECTRA and other models for experiments. Table 3 details Precision, Recall and F1-score for each strategy. In our experiments, DeBERTa achieved the best performance with an F1-score of 0.745. RoBERTa's F1-score is next at 0.740. In the official test results (Lee et al., 2023), the officially provided BERT-BiLSTM-CRF model (Lee et al., 2022b) achieved a macro-averaging F1 score of 0.6813. Our DeBERTa model achieved a macro-averaging F1 score of 0.6779, slightly lower than the official BERT-BiLSTM-CRF model.

3.6 Ablation experiment

To verify the effectiveness of LoRA, we conduct a series of experiments. We trained the model without LoRA and the model with LoRA separately. Table 4 shows the number of parameters for both and the time required for each epoch. From the table 4, we can see that after using LoRA, the parameter amount and training time of each model are greatly reduced. This shows that LoRA is beneficial to the training of the model. Thanks to LoRA, we can try more pre-trained models to get better results.

The distribution of samples in the training data is not balanced, so we use Focal Loss. After using Focal Loss, the performance of each model has been improved. Table 5 shows the detailed data of different strategies.

Model	Trainable params	Epoch
BERT	102299178	30min
RoBERTa	326088746	67min
DeBERTa	714992682	105min
BERT+LoRA	622122	25min
RoBERTa+LoRA	1615914	53min
DeBERTa+LoRA	2423850	82min

Table 4: The impact of LoRA on the model.

Model	Precision	Recall	F1-score
BERT	0.674	0.778	0.722
RoBERTa	0.683	0.774	0.726
DeBERTa	0.685	0.783	0.731
BERT+Focal Loss	0.687	0.797	0.738
RoBERTa+Focal Loss	0.712	0.771	0.740
DeBERTa+Focal Loss	0.706	0.788	0.745

Table 5: The impact of Focal Loss on the model.

4 Conclusions

This paper build three models to solve the named entity recognition task in the medical field proposed by ROCLING 2023. We describe the experiments in various details and select the predictions of the best performing models as the final submission. In the end, the best Macro-averaging F1 we obtained was 0.6779, ranking fifth. The sources of data in the data set provided by this task are divided into Formal Texts, Social Media, and Wikipedia Articles. Correspondingly, the final test set is also divided into these three types. In future work, we will try to split the data set into three parts based on the data source, and train three different models respectively. Different models are then used to predict labels for the test set. This might lead to better results.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61966038 and 62266051.

References

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, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese BERT. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:3504–3514.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Moshe Fresko, Binyamin Rosenfeld, and Ronen Feldman. 2005. A hybrid approach to NER by MEMM and manual rules. In *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management, Bremen, Germany, October 31 - November 5, 2005*, pages 361–362. ACM.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Hideki Isozaki and Hideto Kazawa. 2002. Efficient support vector classifiers for named entity recognition. In *19th International Conference on Computational Linguistics, COLING 2002, Howard International House and Academia Sinica, Taipei, Taiwan, August 24 - September 1, 2002*.

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 *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Lung-Hao Lee, Chao-Yi Chen, Liang-Chih Yu, and Yuen-Hsien Tseng. 2022a. 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, ROCLING 2022, Taipei, Taiwan, November 21-22, 2022*, pages 363–368. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Lung-Hao Lee, Tzu-Mi Lin, and Chao-Yi Chen. 2023. Overview of the rocling 2023 shared task for chinese multi-genre named entity recognition in the healthcare domain. *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing*.

Lung-Hao Lee, Chien-Huan Lu, and Tzu-Mi Lin. 2022b. NCUEE-NLP at semeval-2022 task 11: Chinese named entity recognition using the bert-bilstm-crf model. In *Proceedings of the 16th International Workshop on Semantic Evaluation, SemEval@NAACL 2022, Seattle, Washington, United States, July 14-15, 2022*, pages 1597–1602. Association for Computational Linguistics.

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

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007. IEEE Computer Society.

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. *CoRR*, abs/1907.11692.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Donghuo Zeng, Chengjie Sun, Lei Lin, and Bingquan Liu. 2017. LSTM-CRF for drug-named entity recognition. *Entropy*, 19(6):283.

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, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1227–1236. 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, July 6-12, 2002, Philadelphia, PA, USA*, pages 473–480. ACL.