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

CrowNER at ROCLING 2023 MultiNER-Health Task: Enhancing NER Task with GPT Paraphrase Augmentation on Sparsely Labeled Data

Yin-Chieh Wang*
Telexpress Co., Ltd.
pony.wang@telexpress.com
Feng-Yu Kuo*
Telexpress Co., Ltd.
bruce.kuo@telexpress.com
Te-Yu Chi
Department of CSIE
National Taiwan University
d09922009@ntu.edu.tw
Sheh Chen
Telexpress Co., Ltd.
shepherd.chen@telexpress.com
Wen-Hong Wu*
Telexpress Co., Ltd.
vincent.wu@telexpress.com
Han-Chun Wu*
Telexpress Co., Ltd.
andy.wu@telexpress.com
Te-Lun Yang
Department of CSIE
National Taiwan University
d12944007@ntu.edu.tw
Jyh-Shing Roger Jang
Department of CSIE
National Taiwan University
jang@mirlab.org

Abstract

In this research, we utilized the training dataset from the ROCLING 2023 Chinese Multi-genre Named Entity Recognition in the Healthcare Domain, which comprises the Chinese HealthNER Corpus (Lee and Lu, 2021) and the ROCLING 2022 CHNER Dataset (Lee et al., 2022), along with the test set (Lee et al., 2023). The objective was to address the named entity recognition task within the Chinese healthcare domain. Our initial step involved preprocessing the training dataset. We identified instances in the training set where sentences with identical structural patterns exhibited ambiguities and errors in named entity definitions. Prioritizing data validation, we manually excluded erroneous entries. In specialized domains such as medicine, domain-specific terminologies and proprietary names are often defined within sentences as merged labels, rather than separate ones. Thus, we employed the 'Entity Relationship Construction and Merging Strategies' approach to consolidate related named entities. Subsequently, we computed the frequencies of sentence and entity occurrences. We extracted sparsely labeled data and applied two techniques for data augmentation: GPT Paraphrase and entity replacement while preserving sentence structure. These steps resulted in an augmented training set. Finally, we conducted fine-tuning experiments on various state-of-the-art BERT-based models to obtain a model suitable for the ROCLING Shared Task.

Keywords: GPT 3.5, Data augmentation, GPT paraphrase, Entity Relationship Construction and Merging Strategies

1 Introduction

Named Entity Recognition (NER) aims to identify specific meaningful entities from text, such as person names, locations, organization names, dates, and times. In specific domains like the medical field, these named entities often have unique naming conventions and characteristics. To accurately identify entities in these specialized domains, it's common to use domain-specific training data to train NER models that cater to the named entity recog-

*These authors contributed equally to this work.

nition requirements of that field. The main goal of this research is to develop and refine a Named Entity Recognition (NER) model focused on the medical field, aiming to investigate and improve its accuracy. The study involves various stages, including data preprocessing, model evaluation and selection, and experimentation with data augmentation techniques.

In the data preprocessing phase, this involves data cleaning, entity relationship construction, and merging strategies. We discovered several issues in the data, such as non-compliance with BIO tagging standards and inconsistent entity labels within the same sentence. Furthermore, by analyzing entity occurrence frequencies and sentence structures, we found many entities that should have been labeled as compound nouns were mistakenly split into separate words. Thus, we introduced the concept of Entity Relationship Groups and Merging Strategies.Initially, we developed Entity Association Groups based on the lexical structure of entities, identifying connections through shared vocabulary. Subsequently, we examined entities within sentences against these groups, merging or replacing them based on their position and association to enhance label accuracy.For example, within a sentence, entities ” 瓣膜” and ” 脫垂” might be identified separately. However, after analyzing their relationships and positions in the sentence, we merged them to form ” 瓣膜脫垂”, thereby improving the data quality.

Regarding model selection, we evaluated several pre-trained models, including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), UBERT (Lu et al., 2022), MacBERT (Cui et al., 2020), and PERT (Cui et al., 2022). Ultimately, PERT was chosen for this study. Subsequent optimization of the PERT model was carried out, and the highest F1 score was achieved by incorporating a Conditional Random Field (CRF) layer.

In terms of data augmentation experiments, we first performed data correction then divided it into training set, development set, and test set. Development set and test set are filled with sparsely labeled data, consisting of challenging instances that often deviate from the patterns present in the training set. This divergence underscores the limitations of solely relying on the training set for effective predictions. This highlights the need for a more robust training approach that can better handle such intricacies and generalize well to unforeseen cases.

We formulated four data augmentation experimental setups. In this context, RUN0 was designated as the control group, representing a configuration without any data augmentation. On the other hand, we had the experimental groups including RUN1, RUN2 and RUN3. RUN1 involves incorporating the development set data into the training set. In RUN2, we leveraged ChatGPT to paraphrase the development set data, thereby enhancing the training set. Lastly, In RUN3, we incorporated entity data from the development set into the training set through substitution for data augmentation.

In conclusion, this study has effectively improved the performance of named entity recognition tasks through a comprehensive systematic process, including pre-trained model selection, data preprocessing, entity relationship construction and merging strategies, as well as data augmentation strategies. Moreover, the integration of entity relationship construction and merging strategies within the data preprocessing phase, combined with the GPT-paraphrased data for data augmentation, contributed to our team's first-place victory in the ROCLING 2023 Competition, achieving an F1 score of 69.55 (RUN2).

2 Related Work

Named Entity Recognition (NER) is the process of automatically identifying and classifying named entities in unstructured text, and then organizing them into predefined categories. There are several approaches to tackle NER task including span-based, tagging-based and generation-based. The tagging-based approach (Huang et al., 2015; Yang et al., 2017; Souza et al., 2019) involves annotating each individual word or token in the text with a specific label denoting its named entity category. The tagging-based model is often comprised of a feature extraction model such as a LSTM (Sak et al., 2014) or Transformer (Vaswani

et al., 2017) model with a conditional random field (CRF) layer that outputs the label sequence. The span-based approach (Zheng et al., 2017; Wang et al., 2020; Su et al., 2022) centers on identifying continuous sequences of words that constitute named entities, thereby marking their beginning and end positions within the text. This method is particularly adept at handling cases where named entities might comprise multiple words or where the exact boundaries are less distinct. The approach based on generation (Athiwaratkun et al., 2020; Yan et al., 2021) formulates the NER task as a problem of sequence generation using models such as BART (Lewis et al., 2020) and T5 (Raffel et al., 2020) to generate and extract the named entity tokens. By reformulating the task as a sequence generation problem, these models can directly eliminate the need for explicit boundary marking. However, since generation-based models tend to generate repetitive tokens, hallucinate information, and struggle to preserve contextual accuracy, we opted to use tagging-based and span-based approaches in our experiment. These approaches employ more structured and controlled techniques to identify and classify named entities in the text.

As an encoder of Transformer (Vaswani et al., 2017) architecture, BERT (Devlin et al., 2019) introduces deep bidirectional contextual understanding by considering both left and right context in all layers. This allows it to pretrain on unlabeled text and subsequently fine-tune with minimal architecture adjustments for various tasks. PERT (Cui et al., 2022) is an improved variant of BERT. It employs input text permutation, where the task is to predict the original token's position. PERT incorporates whole word and N-gram masking to further enhance its performance. These approaches highlight the potential for diverse pre-training tasks beyond language models. In light of PERT's higher performance compared to other BERT variants in our experiment, we opted to select PERT as the base model for further enhancement in addressing the NER task.

Data augmentation is considered a useful technique when training with limited data. Nevertheless, automatic data augmentation in NLP poses a challenge due to the complexity of language and the necessity of preserving semantic meaning. Previous approaches (Zhang et al., 2015; Yu et al., 2018; Wei and Zou, 2019) such as synonym replacement, random word insertion, word swapping, random deletion and translation from different languages may not be effective for the NER task. Since NER requires a higher level of precision in identifying and categorizing specific entities within the text. In contrast to general language understanding tasks, NER requires precise localization and classification of entities. With the rise of Large Language Models (LLMs) and in particular ChatGPT, it has the ability to generate human-like sentences. By using carefully crafted prompts, it is possible to generate sentence with similar semantic meaning as the original sentence while retaining the entity structure. Throughout this research, we will provide a comparative evaluation between human-driven and ChatGPT-powered data augmentation.

3 Methodology and Experiments

3.1 Dataset evaluation

The evaluation process of this study employed the Precision/Recall/F1-score (P/R/F1) metrics. We utilized the data provided by the ROCLING-2023 Shared Task for our study. The training dataset comprises the Chinese Health Named Entity Recognition (NER) Corpus (Lee and Lu, 2021) as well as the ROCLING-2022 Chinese NER Dataset (Lee et al., 2022), show in Table 1. In total, it encompasses 33,897 sentences, 1,631,604 characters, and 81,829 named entities, spanning across 10 distinct entity types. The entire experimental procedure was divided into three main stages. We will sequentially conduct experiments from various perspectives, encompassing model selection and optimization, data cleaning, merging strategies as well as diverse augmentation techniques with the aim of enhancing predictive accuracy.

Genre	FT	SM	WA
Sentences	23,008	7,684	3,205
Characters	1,109,918	403,570	118,116
Named Entities	42,070	26,390	13,369
Data Sets	Chinese HealthNER Corpus		CHNER

Table 1: **Shared training sets** (FT:Formal texts, SM:Social media, WA:Wikipedia articles)

3.2 Model selection and Fine-tuning

In the first stage, the focus was on the selection of the base model, architectural design, and parameter tuning. For this phase, we utilized the "Formal Texts"subset from the Chinese Health NER Corpus (Lee and Lu, 2021) as the training set, while "Social Media" was used as the development set. We utilized the dataset to fine-tune all the pre-trained models and report their precision, recall and F1 score. We started by fine-tuning multiple pre-trained models in order to select the best base model. Including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), UBERT (Lu et al., 2022), MacBERT (Cui et al., 2020) and PERT (Cui et al., 2022).

After choosing the best base model, we enhanced it with a conditional random field (CRF) layer for the tagging-based approach and a span classification head on top for the span-based approach respectively. In addition to increasing the number of layers in the model, we utilized the focal loss function (Lin et al., 2020) to alleviate the issue of class imbalance in most of the Named Entity Recognition (NER) tasks. We applied the focal loss function to both the base model and the span-based model.

- $\text{BERT}_{\text{base}}$: 102M parameters
- $\text{RoBERTa}_{\text{base}}$: 102M parameters
- $\text{UBERT}_{\text{base}}$: 102M parameters
- $\text{UBERT}_{\text{large}}$: 325M parameters
- $\text{MacBERT}_{\text{base}}$: 102M parameters
- $\text{PERT}_{\text{base}}$: 102M parameters

3.3 Data Cleaning: Removing and fixing incorrect Data Points

During the data preprocessing phase, we initiated the analysis of all data and identified three primary types of errors: 1) Incorrect labeling formats, where certain data did not adhere to the BIO tagging standard, as illustrated in Table 2; 2) Instances of duplicated sentences with inconsistent word annotations, detailed in Table 3; 3) Cases of repeated sentences with entirely erroneous annotations, for instance, identical sentences but with entirely disparate entity labels, as demonstrated in Table 4. These errors had the potential to introduce confusion during the model training process. To mitigate such issues, we implemented programmatic checks and manually rectified sections with labeling format errors. For data instances where duplicated sentences contained incongruent entity annotations, we manually corrected overtly erroneous labels or removed erroneous data. Furthermore, duplicated sentences featuring entirely dissimilar entity labels were excluded. These rectifications contributed to an enhanced overall quality of the dataset.

3.4 Entity Relationship Construction and Merging Strategies

After the selection of the base model, we conducted data analysis and identified a significant issue wherein entities that should have been labeled as compound nouns were erroneously segmented into separate individual words. Given that the dataset under examination primarily encompasses domain-specific terminology from fields such as medicine and biochemistry, such segmentation into individual words has the potential to compromise the intended meaning and information conveyed by the entities within sentences. In light of this, we advocate that these domain-specific terms within sentence structures be defined using merged labels rather than distinct ones.

To address this issue, we conducted a two-step process. In the first step, we constructed Entity Association Groups, a concept rooted in the lexical structure of entities. Through analyzing shared vocabulary among distinct entities, we established associations between them.

ID	Character	Original Tags / Corrected Tags
297	經	'O-SYMP' 'B-SYMP'
1241	'通', '道', '蛋', '白'	'B-CHEM', 'I-CHEM', 'i-CHEM', 'I-CHEM' 'B-CHEM', 'I-CHEM', 'I-CHEM', 'I-CHEM'
1241	'異', '相', '睡', '眠'	'T-TIME', 'I-TIME', 'I-TIME', 'I-TIME' 'B-TIME', 'I-TIME', 'I-TIME', 'I-TIME'

Table 2: Instances of incorrect labeling formats and non-adherence to BIO tagging standard using CHNER

Sentence	Word	Original	Corrected
中暑了，一旦發現有人核心體溫高過攝氏４０度且意識混亂或昏迷，要趕緊打１１９送急診。 〔中暑了，一旦發現有人核心體溫高過攝氏４０度且意識混亂或昏迷，要趕緊打１１９送急診。	中暑 中暑	(DISE) (O)	– (DISE)
然而，點滴輸液過與不及都會出問題，水分不足導致休克，過多卻可能引起體內積水，如肺積水、腹腔積水等。 然而，點滴輸液過與不及都會出問題，水分不足導致休克，過多卻可能引起體內積水，如肺積水、腹腔積水等。	點滴 點滴	(INST) (O)	– (INST)
血液中高量的維生素Ｂ可以永續地降低肺癌的風險。 血液中高量的維生素Ｂ可以永續地降低肺癌的風險。	肺癌 肺癌	(DISE) (O)	– (DISE)

Table 3: Annotation Table (Part 1)

For example, the term " 瓣膜" (valve) shares a subword relationship within entities like " 人工瓣膜" and " 二尖瓣膜脫垂", as shown in Figure 1. Utilizing graph analysis techniques, we created an Entity Association Graph as depicted in Figure 2.The Entity Association Groups were constructed based on annotated datasets from the Chinese HealthNER Corpus (Lee and Lu, 2021) and the ROCLING-2022 Chinese NER Dataset (Lee et al., 2022)

a.

Entity	Entity Association Groups
瓣膜	**瓣膜**
	人工**瓣膜**
	二尖**瓣膜**脫垂
	瓣膜疾病
	瓣膜性心臟
	心臟**瓣膜**
	人工機械**瓣膜**
	瓣膜脫垂
	瓣膜性心臟病
	二尖**瓣膜**
	心臟**瓣膜**疾病
	心**瓣膜**
	瓣膜閉鎖不全

b.

Entity	Entity Association Groups
脫垂	**脫垂**
	2尖辦膜**脫垂**
	二尖瓣膜**脫垂**
	中度**脫垂**
	二尖瓣**脫垂**
	二尖瓣三尖瓣**脫垂**
	輕度**脫垂**
	瓣膜**脫垂**
	二間辦**脫垂**
	直腸**脫垂**

Figure 1: Subword relationships of the Entities

The second step focuses on the Merging Entities or Terms. This process involves examining the entities within sentences and their corresponding Entity Association Groups. The goal is to determine whether there are associated entities from these groups present in the sentence and, based on their positions within the sentence, decide whether they meet the criteria for merging to correct the labels.

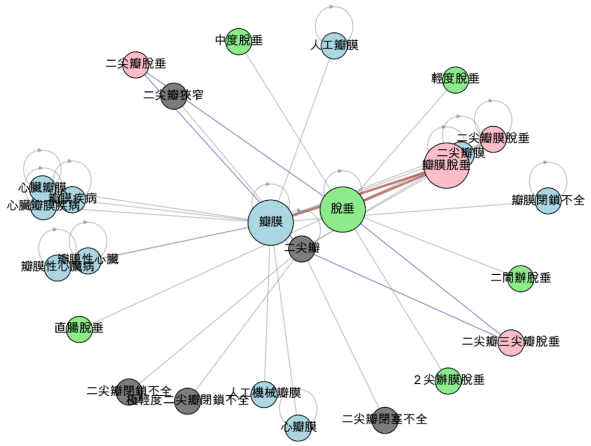

Figure 2: Entity Association Graph Generated using Graph Analysis Techniques

For example, a sentence has both annotations for the entities " 瓣膜" (valve) and " 脫

Issue	Example
Same Sentence	富含 胡蘿蔔素的食物可以預防肺癌
Same Word	’富含’, ’ 胡蘿蔔素’, ’的’, ’食物’, ’可以’, ’預防’, ’肺癌
Same Character	‘富’, ’含’, ’’, ’胡’, ’蘿’, ’蔔’, ’素’, ’的’, ’食’, ’物’, ’可’, ’以’, ’預’, ’防’, ’肺’, ’癌’
Character label(22759)	‘O’, ’O’, ’B-SUPP’, ’I-SUPP’, ’I-SUPP’, ’I-SUPP’, ’I-SUPP’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’B-DISE’, ’I-DISE’
Character label (00859)	’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’

Table 4: Annotation Table (Part 2)

垂” (prolapse). However, through analysis of the Entity Association Group for ” 瓣膜” and ” 脫垂”, we identify a more comprehensive entity annotation, ” 瓣膜脫垂”, which serves as a subword for both entities. As a result, if the positions of both entities align for merging, we combine these two entities into ” 瓣膜脫垂” and subsequently retrieve its named entity type from the Entity Association Groups, as illustrated in Table 5. These corrections contribute to enhanced semantic precision and strengthen the model’s expressive capacity.

In the stage of the experiment, an experimental design was conducted to validate the feasibility of the merging strategy proposed in this study. The Health NER and CHNER datasets were initially merged. Following essential data correction, the dataset were partitioned into a training set consisting of 29,411 samples, a test set consisting of 2,533 samples, and a development set consisting of 2,000 samples. It’s worth highlighting that that in the entire dataset, sentences containing entities that appeared only once were classified as sparse labeled data. Both the test and development sets originated from this sparse labeled dataset categorization. In the experiment, the training set was divided into experimental and control groups. In the experimental group, data were subjected to merging corrections based on the second phase method, while the control group remained in its original state. Both groups were trained using the final model from the first phase and evaluated on the unmodified test set.

3.5 Data Augmentation Strategies

In order to enhance the performance of the model, we propose a data augmentation strategy through entity replacement to expand the training set. This approach employs development sets partitioned from dataset with sparse annotations.

” 腎功能失調” is a unique entity in the development set, labeled as ’DISE.’ We then found another piece of data in the training set with the same entity type (DISE) and containing only a single entity. New data is generated By using a substitution approach. Figure 3 provides an example of this replacement.

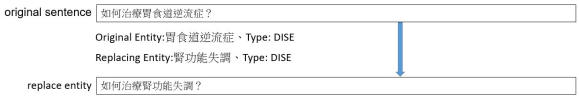

Figure 3: The method of entity replacement.

However, this method may encounter challenges in maintaining semantic coherence, as the generated sentences may not consistently preserve semantic meaning. In the following paragraph, we propose the method that use Chat GPT to paraphrase sentences to mitigate the issue of semantic inconsistency. To resolve the potential issue of semantic inconsistency in the previous method, we attempted data augmentation using GPT. This approach allows us to maximize semantic coherence while paraphrasing sentences.

During the development of the augmentation process, we observed that GPT also tends to rewrite entities within sentences. To ensure that entities are not rewrited, we first replaced the entities within the sentences with placeholders such as entity1, entity2, and store these entities’ information, such as original word and entity type, in a list called ner-list. Then, we used GPT to paraphrase the sentences with these placeholders, and finally, we putted the corresponding entities back into the sentences according to the ner-list. To label generated sentences, we first create a character label list which its length equals to the generated sentence with all ”O”. Then according to the ner-list, find index of each entity and replace the element at the index to the corresponding entity type. This approach guar-

Example	Original Words	Original Label	Corrected Words	Corrected Label
如果發現瓣膜脫垂嚴重導致血液逆流	瓣膜, 脫垂	(BODY),(SYMP)	瓣膜脫垂	(DISE)
但應不會乳房皮膚紅腫熱痛	紅腫, 熱, 痛	(SYMP),(SYMP),(SYMP)	紅腫熱痛	(SYMP)
須留意的是泌尿感染或是骨盆腔發炎的問題	骨盆腔, 發炎	(BODY),(SYMP)	骨盆腔發炎	(DISE)
和另一種類胡蘿蔔素玉米黃素(Zeaxanthin)	類, 胡蘿蔔素	(O),(SUPP)	類胡蘿蔔素	(SUPP)

Table 5: Example of Merging Entities or Terms

antees that the sentences are rewritten while still preserving the original entities. Due to the time limitation, we only ensure that generated sentences are different from original sentences and keep all entities in original sentences. Figure 4 provides an example prompt template used for this GPT-based rewriting, while Figure 5 demonstrates the procedural representation of the generation process.

Figure 4: Example Prompt Template used for GPT-based rewriting

Figure 5: Procedural representation of the generation process.

The stage aims to investigate the impact of various augmentation strategies on model performance. We employed the training set (29,411 samples), test set (2,533 samples), and development set (2,000 samples) partitioned as described in the previous data cleaning **Section** 3.4.

We performed data augmentation on the training set using the development set (2,000 samples) in various ways and evaluated model performance using the test set. The experimental design consisted of four experimental groups: RUN0, the control group, which did not undergo any data augmentation; RUN1, where development set data was added to the training set; RUN2, involving the rephrasing of development set data using ChatGPT for training set augmentation; and RUN3, entailing the incorporation of entity data from the development set into the training set using entity substitution.

4 Experiment results and discussion

4.1 Model Selection Results: Enhancing Model Performance

We use the HealthNER corpus (Lee and Lu, 2021) to fine-tune all the pre-trained models. We select the AdamW optimizer with learning rate of 5e-5, batch size of 28 as the hyperparameters and train with 50 epochs. We evaluate the model per 100 steps during training and select the best one by the F1 score. The result is shown in Table 6. $\text{PERT}_{\text{base}}$ performed better than other models in terms of F1 score. Therefore, we selected $\text{PERT}_{\text{base}}$ as the base model for the subsequent experiment.

Model	P	R	F1
$\text{BERT}_{\text{base}}$	74.82	75.77	74.88
$\text{RoBERTa}_{\text{base}}$	74.01	75.93	74.96
$\text{UBERT}_{\text{base}}$	75.61	74.96	75.29
$\text{UBERT}_{\text{large}}$	69.06	75.33	72.06
$\text{MacBERT}_{\text{base}}$	74.75	76.66	75.69
$\text{PERT}_{\text{base}}$	75.31	76.74	76.02

Table 6: Comparison of Models (P:Precision, R:Recall, F1:F1 score)

We continue to improve PERT by incorporating a conditional random field (CRF) layer for the tagging-based approach and add a start and end classification head for the span-based approach. In addition to increasing the number of layers in the model, we utilize the focal loss function (Lin et al., 2020) to alleviate the issue of class imbalance in most of

the Named Entity Recognition (NER) tasks. We apply the focal loss function to both the $\text{PERT}_{\text{base}}$ model and the $\text{PERT}_{\text{Span}}$ model. We use the same hyperparamters as mentioned and PERT_{CRF} achieved the highest F1 score compared to other methods as shown in Table 7.

Model	P	R	F1
PERT	75.31	76.74	76.02
$\text{PERT}_{\text{focal}}$	74.92	76.56	75.74
PERT_{CRF}	76.90	76.84	76.87
$\text{PERT}_{\text{Span}}$	74.25	77.57	75.88
$\text{PERT}_{\text{Span with focal}}$	76.95	74.89	75.91

Table 7: Comparative evaluation of different architecture and loss function. (P:Precision, R:Recall, F1:F1 score)

4.2 Entity Relationship Construction and Merging Strategies: Impact on Model Enhancement

To validate the feasibility of the proposed merging strategy in this study, we conducted an experimental design. The experimental outcomes revealed that the performance of the model was enhanced through the implementation of the merging strategy, as illustrated in the Table 8.

Methods	P	R	F1
PERT_{CRF} [a]	76.10	77.64	75.57
PERT_{CRF} [b]	78.12	80.35	76.02

Table 8: Comparative evaluation of training set with and without fixed. (P:Precision, R:Recall, F1:F1 score)

[a] Training set without fixed using method 3.3

[b] Training set fixed using method 3.3

4.3 Data Augmentation Strategies: Evaluating Techniques for Performance Enhancement

As shown in Table 9, PERT_{CRF} with data augmentation (RUN1, RUN2, RUN3) outperformed PERT_{CRF} without data augmentation (RUN0). The augmentation method using the replacement approach (RUN3) showed less enhancement compared to the other two methods. This might be attributed to the fact that employing only replacement-based data augmentation can not ensure semantic coherence, thereby affecting the model's performance. In the experiments of RUN1 and RUN2, incorporating GPT-paraphrased development set into the training set resulted in similar performance compared to directly adding development set to the training set, with both F1 values approximately around 79.8. This result demonstrates that the GPT-paraphrased sentences retained their semantic meaning and therefore did not significantly affect the performance, in comparison to RUN1.

Experiment	P	R	F1
$\text{RUN0}_{[a]}$	80.29	76.38	78.28
$\text{RUN1}_{[b]}$	81.45	78.36	79.88
$\text{RUN2}_{[c]}$	81.47	78.22	79.81
$\text{RUN3}_{[d]}$	81.18	76.74	78.90

Table 9: Comparison between different data augmentation methods. (P:Precision, R:Recall, F1:F1 score)

[a] PERT_{CRF} without data augmentation.

[b] PERT_{CRF} augmented with human written data(development set).

[c] PERT_{CRF} augmented with GPT-paraphrased development set.

[d] PERT_{CRF} with low frequency entities augmentation.

5 Conclusions

In this study, we conducted a series of experiments and explorations for named entity recognition (NER) task. Initially, we selected PERT as the baseline model since it outperformed other pre-trained models on Health-NER corpus. Subsequently, we further improved PERT model by incorporating Conditional Random Fields (CRF) layer, achieving the highest F1 scores among other architectures and loss function. Furthermore, our proposed strategies involving the construction of Entity Association Groups and the merging of entities were validated to enhance model performance.

Additionally, we investigated the impact of various data augmentation strategies on model

performance. Through methods such as entity replacement and sentence paraphrasing using GPT, we observed improvements in F1 scores. However, employing GPT for sentence paraphrasing requires further adjustments to achieve more pronounced effects.

The study presents a comprehensive and systematic approach encompassing pre-trained model selection, data point correction, entity relationship construction, merging strategies, and data augmentation techniques. These efforts contributed to our team's first-place achievement in the ROCLING 2023 competition, attaining an F1 score of 69.55 (RUN2). The outcomes of the three submissions and official baseline result are presented in Table 10. The official baseline used BERT-BiLSTM-CRF as their model. The main difference between our model and the baseline is that we did not add the BiLSTM layer in the middle of our embedding model and CRF layer since the self-attention mechanism in the transformer-like architecture already considered the relationship between each word in the sentence.

	P	R	F1
RUN1	71.14	67.64	69.28
RUN2	72.35	67.08	69.55
RUN3	72.55	66.27	69.22
Official Baseline	-	-	68.13

Table 10: Evaluation scores for the three experimental results in the ROCLING 2023 competition. (P:Precision, R:Recall, F1:F1 score)

References

Ben Athiwaratkun, Cicero Nogueira dos Santos, Jason Krone, and Bing Xiang. 2020. Augmented natural language for generative sequence labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 375–385, Online. Association for Computational Linguistics.

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.

Yiming Cui, Ziqing Yang, and Ting Liu. 2022. Pert: pre-training bert with permuted language model. *arXiv preprint arXiv:2203.06906*.

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.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

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 (ROCLING 2022)*, pages 363–368.

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. in proceedings of the 35th 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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):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*, abs/1907.11692.

Junyu Lu, Ping Yang, Ruyi Gan, Jing Yang, and Jiaxing Zhang. 2022. Unified bert for few-shot natural language understanding. *arXiv preprint arXiv:2206.12094*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Hasim Sak, Andrew W. Senior, and Françoise Beaufays. 2014. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *INTERSPEECH*, pages 338–342.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2019. Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*.

Jianlin Su, Ahmed Murtadha, Shengfeng Pan, Jing Hou, Jun Sun, Wanwei Huang, Bo Wen, and Yunfeng Liu. 2022. Global pointer: Novel efficient span-based approach for named entity recognition. *arXiv preprint arXiv:2208.03054*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. TPLinker: Single-stage joint extraction of entities and relations through token pair linking. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1572–1582, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A unified generative framework for various NER subtasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5808–5822, Online. Association for Computational Linguistics.

Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. *arXiv preprint arXiv:1703.06345*.

Adams Wei Yu, David Dohan, Thang Luong, Rui Zhao, Kai Chen, and Quoc Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

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.