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

Overview of the ROCLING 2023 Shared Task for Chinese Multi-genre Named Entity Recognition in the Healthcare Domain

Lung-Hao Lee, Tzu-Mi Lin, and Chao-Yi Chen
Department of Electrical Engineering
National Central University
lhlee@ee.ncu.edu.tw, 110521987@cc.ncu.edu.tw, 110581007@cc.ncu.edu.tw

Abstract

This paper describes the ROCLING-2023 shared task for Chinese multi-genre named entity recognition in the healthcare domain, including task description, data preparation, performance metrics, and evaluation results. Among eight registered teams, six participating teams submitted a total of 16 runs. This shared task demonstrates current NLP techniques for dealing with Chinese named entity recognition in multi-genre texts. All data sets with gold standards and evaluation scripts used in this shared task are publicly available for future research.

Keywords: named entity recognition, information extraction, health informatics, Chinese language processing

1 Introduction

Named Entity Recognition (NER) is an NLP task that involves extracting information of concern known as named entities (e.g., person, location, and organization). The NER task is a sequence labeling problem that jointly recognizes the entity boundaries and category labels. Chinese NER is difficult to process due to the lack of clear delimiters such as spaces between characters and conventional features like capitalization. Since named entity boundaries are also word boundaries for the Chinese language, incorrect word segmentation will cause error propagation. For example, "葡萄糖六磷酸鹽去氫酶" (Glucose-6-Phosphate Dehydrogenase; G6PD) is a proper name in the healthcare domain, which in a particular context may be incorrectly segmented into three words: "葡萄糖" (Glucose), "六磷酸鹽" (Hexaphosphate) and "去氫酶" (Dehydrogenase), resulting in failure to recognize it as a chemical-type named entity. Therefore, character-based methods have been found to outperform word-based approaches for breaking through this word segmentation limitation in Chinese NER (Zhang and Yang, 2018).

The ROCLING-2022 shared task (Lee et al., 2022a) focuses on Chinese healthcare NER to atomically identify healthcare named entities such as symptoms, chemicals, diseases, and treatments. In this shared task, the most frequently used system architecture is BiLSTM-CRF, which usually achieves promising results, resulting in identical findings from related studies for NER in the English language (Chiu and Nichols, 2016; Ma and Hovy, 2016).

Due to the greater challenge in the healthcare domain for Chinese NER, the ROCLING 2023 shared task features a Chinese healthcare NER task that focuses on healthcare texts written in three different genres as follows: 1) formal texts (FT) including health news and articles written by professional editors or journalists; 2) social media (SM) including texts from users in medical question/answer forums; and 3) Wikipedia articles (WA) created and edited by volunteers worldwide. Chinese healthcare named entities may be used in different word forms in different written genres. For examples, "後天免疫缺乏症候群" (Acquired Immunodeficiency Syndrome; AIDS) is commonly used as a spoken language form "愛滋病" in medical forums. "甘油三酯" (Triglyceride; TG) is a different usage referred to as "三酸甘油酯".

We organized the ROCLING 2023 shared task for Chinese **Multi**-genre **N**amed **E**ntity **R**ecognition in the **Health**care domain (denoted as MultiNER-Health), providing an evaluation platform for the development and implementation of Chinese healthcare NER systems. Given a

Genre	Examples	Input & Output
Formal Texts (FT)	Ex 1	*Input*: 早起也能預防老化，甚至降低阿茲海默症的風險 *Output*: O, O, O, O, O, O, B-SYMP, I-SYMP, O, O, O, O, O, B-DISE, I-DISE, I-DISE, I-DISE, I-DISE, O, O, O
	Ex 2	*Input*: 壓力、月經引起的痘痘患者 *Output*: B-SYMP, I-SYMP, O, B-TIME, I-TIME, O, O, O, B-DISE, I-DISE, O, O
Social Media (SM)	Ex 3	*Input*: 如何治療胃食道逆流症？ *Output*: O, O, O, O, B-DISE, I-DISE, I-DISE, I-DISE, I-DISE, I-DISE, O
	Ex 4	*Input*: 請問長期打善思達針劑是不是會變胖？ *Output*: O, O, O, O, O, B-DRUG, I-DRUG, I-DRUG, I-DRUG, I-DRUG, O, O, O, O, B-SYMP, I-SYMP, O
Wikipedia Articles (WA)	Ex 5	*Input*: 抗生素和維生素 A 酸可用於口服治療痤瘡 *Output*: B-DRUG, I-DRUG, I-DRUG, O, B-DRUG, I-DRUG, I-DRUG, I-DRUG, I-DRUG, O, O, O, O, O, O, O, B-DISE, I-DISE
	Ex 6	*Input*: 抑酸劑，又稱抗酸劑，抑制胃酸分泌，緩解燒心 *Output*: B-CHEM, I-CHEM, I-CHEM, O, O, O, B-CHEM, I-CHEM, I-CHEM, O, O, O, B-CHEM, I-CHEM, O, O, O, O, O, B-DISE, I-DISE

Table 1: Examples of the MultiNER-Health task.

Chinese specified text in a genre, the NER systems are expected to recognize healthcare entities across 10 defined types, including Body, Symptom, Instrument, Examination, Chemical, Disease, Drug, Supplement, Treatment, and Time.

The rest of this article is organized as follows. Section 2 provides a description of the Chinese MultiNER-Health shared task. Section 3 introduces the constructed data sets. Section 4 describes the evaluation metrics. Section 5 compares evaluation results from the various participating teams. Finally, we conclude this paper with findings and offer future research directions in Section 6.

2 Task Description

The goal of the MultiNER-Health shared task is to develop and evaluate the capability of Chinese NER systems for healthcare texts written in different genres. The input is a sentence indicating as one of three genres (i.e., FT, SM, and WA) that may contain named entities. The NER system should predict the boundaries and category of the named entity for each sentence.

Following the settings of the ROCLING-2022 shared task (Lee et al., 2022a), we use the common BIO format for our MultiNER-Health task. The B (Beginning)-prefix before a tag indicates that the character is the beginning of a named entity while the I (Inside)-prefix indicates that the character is inside a named entity, and O (Outside) indicates that a character belongs to no named entity.

A total of 10 entity types are used for this MultiNER-health shared task, and are defined in the Chinese HealthNER corpus (Lee and Lu, 2021) with type settings consistent with those in the ROCLING-2022 shared task (Lee et al., 2022a). The entity types and their respective tags are as follows: Body (BODY), Symptom (SYMP), Instrument (INST), Examination (EXAM), Chemical (CHEM), Disease (DISE), Drug (DRUG), Supplement (SUPP), Treatment (TREAT), and Time (TIME).

Example sentences are presented in Table 1. The input is a sentence consisting of a sequence of character-based tokens including punctuation. The NER system returns the corresponding BIO tags aligned to each token as the output. In the Example 1 from the FT genre, “老化” (aging) belongs to the Symptom (SYMP) entity type and “阿茲海默症” (Alzheimer’s disease) is a disease (DISE) type. “痤瘡” (acne) in Example 5 from the WA genre is also a kind of disease (DISE), and is a formal usage of “痘痘” in Example 2 from the SM genre. “燒心” in Example 6 from the WA genre is a spoken language form of a disease “胃食道逆流症”

Datasets		**Training Sets**			**Test Sets**		
Source		Chinese HealthNER Corpus		ROCLING 2022 CHNER Dataset	ROCLING 2023 MultiNER-Health Datasets		
Genre		FT	SM	WA	FT	SM	WA
#Sentence		23,008	7,648	3,205	2,035	2,208	2,381
#Character		1,109,918	403,570	118,116	149,276	98,317	92,498
#Named Entity		42,070	26,390	13,369	10,845	8,292	9,761
Entity Type	Body	17,639	8,772	5,315	2,461	2,572	3,843
	Symptom	6,432	6,472	1,944	2,635	2,280	1,890
	Instrument	743	346	250	190	41	149
	Examination	444	2,178	207	223	511	180
	Chemical	5,716	1,118	1,718	1,124	321	748
	Disease	5,865	4,214	2,609	2,300	1,322	1,970
	Drug	1,165	1,060	481	932	746	451
	Supplement	1,338	187	183	47	92	56
	Treatment	2,031	1,077	468	512	363	308
	Time	697	966	194	421	44	166

Table 2: Detailed data statistics.

(gastroesophageal reflux disease) in Example 3 from the SM genre.

3 Data Preparation

The training sets for this MultiNER-health task consist of two parts: the Chinese HealthNER corpus (Lee and Lu, 2021) was used for both the FT and SM genres and the ROCLING-2022 CHNER dataset (Lee et al., 2022a) was designed for the WA genre. For the FT genre, we have 23,008 sentences with a total of 1,109,918 characters, sourced from web-based health-related articles. The SM genre collected from medical question/answer forums includes 7,648 sentences with a total of 403,570 characters. The quantity in the FT genre about 3 times than that in the SM genre in the Chinese HealthNER corpus. After manual annotation, this corpus consists of 68,460 named entities across 10 defined entity types, of which 42,070 entities (about 61%) came from the FT genre and the remaining 26,390 entities belong to the SM genre. The training instances for the WA genre originate from the ROCLING 2022 CHNER dataset, which includes 3,205 sentences with a total of 118,116 characters and 13,369 named entities.

We use the existing named entities in the Chinese HealthNER corpus as the query terms to identify corresponding texts written in different genres. Four undergraduate students majoring in Chinese language were trained in the named entity tagging task, producing a Fleiss' Kappa value of inter-annotator agreement of 82.17%. All annotators were asked to discuss differences and seek consensus. When agreement was reached, each annotator was then asked to process sentences individually. As a result, our constructed test set includes 2,035/2,208/2,381 sentences respectively for the FT/SM/WA genres, resulting in a total of 340,091 characters and 28,898 named entities.

Table 2 presents detailed statistics for the mutually exclusive training and test sets, showing similar entity type distributions. The most frequently occurring type was Body, followed by Symptom, Disease and Chemical regardless of genre. In the training sets, these 4 types collectively accounted for about 82.9% of all named entity instances, with the remaining 6 types accounting for 17.1%. In the test sets, these 4 types accounted for 81.2% of the total, with the other 6 types accounting for the remaining 18.8%.

In the training set, sentences used for the FT and SM genres may or may not contain named entities, but sentences belonging to the WA genre contain at least one named entity. Each sentence had an average of 48.19 characters and 2.42 named

Team	Run#	F1-score (%)				Rank
		Formal Texts	Social Media	Wikipedia Articles	Macro-averaging	
CrowNER	Run 2	**65.49**	69.54	**73.63**	**69.55**	1
YNU-HPCC	Run 2	61.96	71.11	72.13	68.40	2
ISLab	Run 1	62.52	**71.42**	71.19	68.38	3
SCU-MESCLab	Run 1	62.51	71.33	70.57	68.14	4
YNU-ISE-ZXW	Run 3	62.79	70.22	70.37	67.79	5
LingX	Run 2	51.23	59.28	60.54	57.02	6
Baseline (BiLSTM-CRF)	Word2vec	60.99	67.16	67.91	65.35	-
	BERT	61.08	70.77	72.54	68.13	-

Table 3: Testing results of the MultiNER-Health task.

entities. For system performance evaluation, at least 2,000 sentences per genre were tested, each with an average of 51.34 characters and 4.36 named entities. The average sentence length in the test set was slightly longer and the named entity density was relatively higher than those in the training set.

In addition to the provided datasets, participating systems are allowed to use other publicly available data for this shared task, but such usage should be specified in their system description paper.

4 Performance Metrics

Performance evaluation is mainly conducted by examining the difference between the machine-predicted and human-annotated BIO tags. The most typical evaluation metrics of NER systems at a character level are precision, recall and F1-score. If the predicted BIO tag of a character in the testing instances was completely identical to the gold standard, then it was regarded as correctly recognized. Precision is the percentage of correct named entities found by the NER system. Recall is the percentage of named entities present in the test set found by the NER system. Given the tradeoff between precision and recall, the F1-score further combines precision and recall using their harmonic mean to provide an overall performance judgement.

Each team was allowed to provide at most three submissions during the evaluation period. At each submission, different genres were evaluated independently. The macro-averaging F1-score among three genres is used as the ranking metric in the leaderboard.

5 Evaluation Results

Among eight registered teams, six submitted their testing results, providing a total of 16 submissions, from which the submission with the best macro-averaging F1-score of each team was kept for official performance ranking. The baseline systems were mainly based on the BiLSTM-CRF neural architecture. We used two representation embeddings: word2Vec (Mikolov et al., 2013) and BERT (Devlin et al., 2019). The BERT-BiLSTM-CRF (Lee et al., 2022b) baseline system was also used to recognize Chinese complex named entities at the SemEval-2022 MultiCoNER task (Malmasi et al., 2022), ranking 7th out of 21 participating teams. YUN-HPCC (Pang et al., 2023) used RoBERTa-large (Liu et al., 2019) representation combined with a BiLSTM-CRF model to build an NER model. The YUN-ISE-ZXW (Zhang et al., 2023) team applied the focal loss (Lin et al., 2017) and RoLA (Hu et al., 2022) to fine-tune a pre-trained DeBERTa transformer (He et al., 2021). SCU-MESCLab (Su et al., 2023) presented three transformer-based models with the ensemble mechanism to determine the boundaries and categories of named entities. ISLab (Wu et al., 2023) proposed a three-stage NER system using a

label semantics model (Ma et al., 2022) based on the RoBERTa transformer (Liu et al., 2019) to predict the labels of named entities, followed by a label correction model and heuristic rules to process abnormal labels. The LingX (Wang and Yang, 2023) team designed extraction-style prompts to explore the potential of ChatGLM2-6B (Du et al., 2022) to recognize named entities. The CrowNER (Wang et al., 2023) team used the PERT (Cui et al., 2022) representation followed by CRF to recognize named entities, and investigated the impacts of entity replacement and sentence paraphrase using ChatGPT (OpenAI, 2023).

Table 3 summarizes the task testing results. The overall best results came from the CrowNER team (Wang et al., 2023), achieving the best macro-averaging F1 score of 69.55, followed by YUN-HPCC (Pang et al., 2023) and ISLab (Wu et al., 2023). CrowNER also obtained the best results for the FT and WA genres, while the system designed by ISLab performed best for the SM genre. In summary, in addition to combining transformers embeddings with a whole or partial BiLSTM-CRF architecture as the mainstream solution, incorporating large language models like ChatCPT and ChatGLM presents a new direction for the NER task.

6 Conclusions and Future Work

This paper provides an overview of the ROCLING-2023 MultiNER-Health task for Chinese multi-genre named entity recognition in the healthcare domain, including task descriptions, data preparation, performance metrics and evaluation results. We received a total of 16 test submissions from six participating teams. Regardless of actual performance, all submissions contribute to the development of an effective healthcare NER solution, and each system description paper for this shared task also provides useful insights for further research.

We hope the data sets collected and annotated for this shared task can facilitate and expedite future development of Chinese NER in the healthcare domain. Therefore, the gold standard test set and evaluation scripts are made publicly available in GitHub repositories as follows:

- Chinese HealthNER Corpus (FT/SM genres) https://github.com/NCUEE-NLPLab/Chinese-HealthNER-Corpus
- ROCLING-2022 CHNER Task (WA genre) https://github.com/NCUEE-NLPLab/ROCLING-2022-ST-CHNER
- ROCLING-2023 MultiNER-Health Task (FT/SM/WA genres) https://github.com/NCUEE-NLPLab/ROCLING-2023-ST-MultiNERHealth

Future directions will focus on the development of Chinese healthcare entity-relationship extraction. We plan to build new language resources to develop techniques for the future enrichment of this research topic, especially for open information extraction.

Acknowledgments

We thank all the participants for taking part in our shared task. We appreciate annotators for their efforts in data annotations. This work is partially supported by the National Science and Technology Council, Taiwan, under the grant MOST 111-2628-E-008-005-MY3.

References

Jason P. C. Chiu, and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4: 357-370.

Yiming Cui, Ziqing Yang, and Ting Liu. 2022. PERT: pre-training BERT with permuted language model. *arXiv:2203.06906*. https://doi.org/10.48550/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)*. Association for Computational Linguistics, pages 4171–4186.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: general language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 320-335.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: decoding-enhanced BERT with disentangled attention. *In*

Proceedings of the 9th International Conference on Learning Representations.

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 *Proceedings of the 10th International Conference on Learning Representations*.

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.

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*. The Association for Computational Linguistics and Chinese Language Processing, pages 363-368.

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*. Association for Computational Linguistics, pages 1597-1602.

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

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 pre-training approach. arXiv: 1907.11692v1. https://doi.org/10.48550/arXiv.1907.11692

Jie Ma, Miguel Ballesteros, Srikanth Doss, Rishita Anubhai, Sunil Mallya, Yaser Al-Onaizan, and Dan Roth. 2022. Label semantics for few shot named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, pages 1956-1971.

Xuezhe Ma, and Eduard Hovy. 2016. End-to-end sequence labeling via Bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1064-1074.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. SemEval-2022 Task 11: multilingual complex named entity recognition (MultiCoNER). In *Proceedings of the 16th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, pages 1412-1437.

Tomas Mikolov, llya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th Conference on Neural and Information Processing Systems*. NeurIPS Foundation, pages 3111-3119.

Chonglin Pang, You Zhang, and Xiaobing Zhou. YUN-HPCC at ROCLING 2023 MultiNER-Health Task: a transformer-based approach for Chinese healthcare NER. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing*.

Tzu-En Su, Ruei-Cyuan Su, Ming-Hsiang Su, and Tsung-Hsien Yang. 2023. SCU-MESCLab at ROCLING 2023 MultiNER-Health Task: named entity recognition using multiple classifier model. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing*.

Yin-Chieh Wang, Wen-Hong Wu, Feng-Yu Kuo, Han-Chun Wu, Te-Yu Chi, Te-Lun Yang, Sheh Chen, and Jyh-Shing Roger Jang. 2023. CrowNER at ROCLING 2023 MultiNER-Health Task: enhancing NER task with GPT paraphrase augmentation on sparsely labeled data. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing*.

Xuelin Wang and Qihao Yang. 2023. LingX at ROCLING 2023 MultiNER-Health Task: intelligent capture of Chinese medical named entities by LLMs. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing*.

Jun-Jie Wu, Tao-Hsing Chang, and Fu-Yuan Hsu. 2023. ISLab at ROCLING 2023 MultiNER-Health Task: a three-stage NER model combining textual content and label semantics. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing*.

Xingwei Zhang, Jin Wang, and Xuejie Zhang. 2023. YUN-ISE-ZXW at ROCLING 2023 MultiNER-Health Task: a transformer-based model with LoRA for Chinese healthcare named entity recognition. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing*.

Yue Zhang, and Jie Yang. 2018. Chinese NER using lattice LSTM. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1554–1564.