

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 sub-task 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.