Exploring text mining models on case studies for rare diseases

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Julian Bayer

Referentin: Prof. Dr. Melanie Siegel Korreferentin: Prof. Dr. Antje Jahn

Introduction

Rare diseases, defined as conditions affecting fewer than 5 in 10,000 people[2], pose unique challenges in healthcare and research. Due to their low prevalence, they often receive limited attention in clinical practice and research, leading to a scarcity of data and resources.[4] Clinical case reports are a type of scientific literature that document individual patient cases, providing valuable insights into rare diseases. A wealth of information is documented in unstructured text from clinical case reports, but manually extracting this data at a large scale is impractical.

This research addresses the problem by exploring and evaluating modern Named Entity Recognition (NER) models to automatically identify and classify key information like diseases, symptoms, and medications. NER in the context of rare diseases can be considered a low-resource setting, as the amount of annotated data is limited.[5] The primary objective is to compare the performance of state-of-the-art approaches—including specialized transformer models and general-purpose large language models—in the low-resource setting characteristic of rare diseases, to find effective ways to unlock this valuable clinical data.

Key Research Questions

- Can modern NER techniques (Transformer-based, Prompt-based) effectively extract structured data from rare disease case reports?
- What are the common error patterns and challenges for this scenario?
- How can a reproducible system for data gathering, annotation, and information extraction be developed?

Methodology

To systematically assess the capabilities of modern NER techniques, a comprehensive evaluation was conducted using multiple datasets and a selection of state-of-the-art models.

Datasets: The evaluation relied on three distinct corpora to ensure robust testing. First, the **MACCROBAT2020** dataset, a public collection of 200 annotated case reports covering both rare and common diseases. Second, **MACCROBAT2020RD**, a targeted subset of MACCROBAT containing only the 18 case reports focused on rare diseases. Finally, to create a focused test set, a Custom Dataset of 82 case reports was collected from PubMed Central for five specific rare diseases and newly annotated for this study.

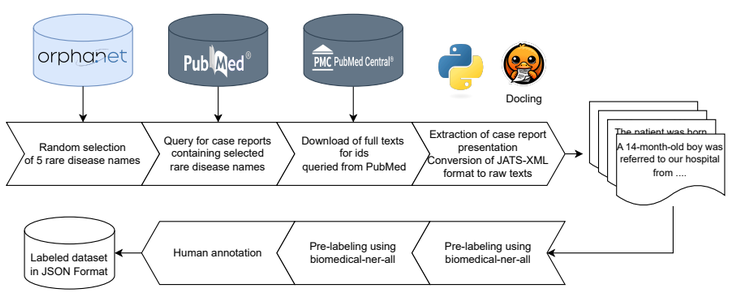

Figure 1. Process of creating the custom dataset for rare diseases.

Evaluated Models: The models were chosen to represent different modern approaches.

- **biomedical-ner-all**[1] A DistilBERT-based model specifically fine-tuned on the MACCROBAT dataset, serving as a strong, domain-specific baseline.
- **GliNER**[6] An open-type NER model designed to identify entities without being restricted to predefined types, tested in small, medium, and large variants.
- **PromptNER**[3] A zero-shot NER method that uses a structured prompt to instruct a large language model (**Gemma3 12B** in this work) to identify entities based on provided definitions.

Evaluation Framework: Model performance was measured using Precision, Recall, and F1-Score. The evaluation was performed under two criteria: Strict, requiring an exact match of both the entity's text span and its type, and Loose, which accepts a correct type prediction with any overlap in the text span.
Error Analysis: An error analysis was conducted to identify common mistakes made by the models. Four types of errors were identified:

- **Wrong Type**: an entity was successfully extracted but assigned the wrong entity type.
- **False Positive**: the model identifies some non-entity as an entity
- **Inaccurate boundary**: an entity was assigned the correct type, but start or end were incorrectly identified.
- **False Negative**: an entity was not identified at all

Results

Results show that the fine-tuned transformer model **biomedical-ner-all** outperforms all other models in both precision and recall. The results of strict evaluation on the **MACCROBAT2020** datasets are shown in Table 1.

Table 1. Strict Evaluation results on the MACCROBAT2020 dataset

Model	Precision	Recall	F_1-Score
biomedical-ner-all	**0.516**	**0.365**	**0.427**
GliNER (small)	0.322	0.086	0.136
GliNER (medium)	0.319	0.127	0.182
GliNER (large)	0.280	0.135	0.183
PromptNER (Gemma3_12b)	0.270	0.164	0.204

The results for loose evaluation, and those for the other datasets paint a similar picture, with the fine-tuned biomedical-ner-all model outperforming the others in all cases.

Results on a per-entity-type basis show that models perform significantly better on some entity types (patient age, medications, diseases) than others (biological structures, measurements). Per-entity investigations also show that misclassifications of entity types are more common for some entity type combinations than others, e.g., the model often misclassifies symptoms as diseases, but rarely misclassifies medications as symptoms.

Error analysis revealed common mistakes made by the models. Distribution of error types was similar across all models except **biomedical-ner-all**, which had a significantly lower (relative) amount of wrong type classifications.

Providing examples of entity types in the prompt for the **PromptNER** was tested, but did not lead to significant improvements in performance.

Investigating the raw output of the large language models used in the **PromptNER** approach reveal that they are prone to certain errors:

- **Unparsable output**: Since the raw model output needs to be parsed to extract entities, the model can produce output that is not parsable, leading to missed entities.
- **Hallucinated Entity Types**: The model sometimes assigns entity types that are not defined in the prompt, leading to incorrect classifications.
- **Modification of input text**: In some cases, spelling was modified for extracted entities, leading to incorrect matches with the original text when parsing.

Some of these errors may be mitigated by fine-tuning prompts or using more advanced parsing techniques.

Evaluation of computational resources shows wide differences in the computational time required for inference. The fine-tuned biomedical-ner-all model required an average of 67ms per case report while the **PromptNER** approach required an average of over one minute per case report. Along with the different memory requirements, which are related to the parameter count of the models, these differences should be considered when choosing a model for practical applications.

Conclusions

- Modern NER approaches considered in this work are generally capable of extracting relevant entities from case reports on rare diseases, but the general low recall and precision of extractions should be considered, especially in the context of clinical applications where data quality should be of high concern.
- Models pretrained for specific entity types, such as the **biomedical-ner-all** model perform better than entity-agnostic approaches such as **GliNER** or **PromptNER**.
- Performance on a per-entity-type basis varies significantly, which should be considered when applying NER models in practice.
- The described system for NER on case reports in the context of rare disease, is capable of providing a basis for further research or practical application in clinical contexts.

Limitations and Future Work

A primary limitation of this research is the custom dataset's reliance on a single, non-expert annotator and pre-labeling with the baseline model, which may introduce bias and affect the generalizability of the findings. The evaluation is also constrained by the lack of a fine-tuned model specifically for this task and missing specification of exact training data of the baseline **biomedical-ner-all** model, which might have been exposed to evaluation data in training. The relatively small size of the custom-annotated dataset, with 82 case reports, further limits the broad applicability of the results.

Future research should prioritize the expansion of the annotated dataset to include a wider variety of rare diseases and multilingual case reports to enhance the robustness and applicability of the models. The development and fine-tuning of specialized models on this expanded corpus could significantly improve extraction accuracy. Exploring advanced techniques such as relation extraction to identify connections between entities (e.g., symptoms and diseases) and integrating multimodal information from figures and tables in case reports presents a promising avenue for yielding more comprehensive data. Additionally, mapping extracted entities to established medical ontologies like Orphanet or UMLS could standardize the data and improve its utility for clinical research and diagnostics.

References

[1] D4data/biomedical-ner-all · Hugging Face. https://huggingface.co/d4data/biomedical-ner-all.

[2] Decision No 1295/1999/EC of the European Parliament and of the Council of 29 April 1999 adopting a programme of Community action on rare diseases within the framework for action in the field of public health (1999 to 2003), April 1999.

[3] Dhananjay Ashok and Zachary C. Lipton. PromptNER: Prompting For Named Entity Recognition, June 2023.

[4] Stéphanie Nguengang Wakap, Deborah M. Lambert, Annie Olry, Charlotte Rodwell, Charlotte Gueydan, Valérie Lanneau, Daniel Murphy, Yann Le Cam, and Ana Rath. Estimating cumulative point prevalence of rare diseases: Analysis of the Orphanet database. *European Journal of Human Genetics*, 28(2):165–173, February 2020.

[5] Cathy Shyr, Yan Hu, Lisa Bastarache, Alex Cheng, Rizwan Hamid, Paul Harris, and Hua Xu. Identifying and Extracting Rare Diseases and Their Phenotypes with Large Language Models. *Journal of Healthcare Informatics Research*, 8(2):438–461, June 2024.

[6] Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico, June 2024. Association for Computational Linguistics.