Hochschule Darmstadt

– Fachbereiche Mathematik und Naturwissenschaften
&
Informatik –

Exploring text mining models on case studies for rare diseases

Abschlussarbeit zur Erlangung des akademischen Grades

Master of Science (M.Sc.)
im Studiengang Data Science

vorgelegt von

Julian Bayer

Matrikelnummer: 759402

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

Ausgabedatum	:	09.12.2024
Abgabedatum	:	08.06.2025

ERKLÄRUNG

Ich versichere hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die im Literaturverzeichnis angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder noch nicht veröffentlichten Quellen entnommen sind, sind als solche kenntlich gemacht. Die Zeichnungen oder Abbildungen in dieser Arbeit sind von mir selbst erstellt worden oder mit einem entsprechenden Quellennachweis versehen. Diese Arbeit ist in gleicher oder ähnlicher Form noch bei keiner anderen Prüfungsbehörde eingereicht worden.

Während der Vorbereitung dieser Arbeit habe ich LanguageTool verwendet, um Rechtschreib-, Grammatik und Kommafehler zu prüfen und zu korrigieren. Nach der Nutzung dieses Tools/Dienstes habe ich den Inhalt nach Bedarf überprüft und bearbeitet und übernehme die volle Verantwortung für den Inhalt der Veröffentlichung.

Darmstadt, 08. Juni 2025

Julian Bayer

ABSTRACT

Rare diseases pose unique challenges in clinical research and treatment due to their low prevalence and limited data availability. This thesis investigates the application of Named Entity Recognition (NER) to extract structured information from case reports on rare diseases. Modern NER approaches, including transformer-based models and prompt-based methods, were assessed for their ability to identify entities such as diseases, symptoms, and medications.

As there are no established datasets for NER on rare disease case reports, a custom dataset was created by annotating 82 case reports from PubMed Central. In addition to the custom dataset, evaluation was performed on the MACCROBAT2020 dataset, which contains case reports for both rare and non-rare diseases.

Results show that pretrained models tailored to biomedical domains, such as `biomedical-ner-all`, outperform generalist approaches like `GliNER` and prompt-based methods like `PromptNER` in precision and recall. Challenges such as low recall, inaccurate boundaries, and entity type misclassification persist for all models, particularly in low-resource scenarios. Analysis of error cases revealed common issues such as misclassification, incorrect boundaries, and hallucinations in prompt-based methods. Prompt-based NER methods demonstrated flexibility but may need further fine-tuning and prompt engineering to improve performance in the studied scenario.

This work highlights the potential of NER systems to support rare disease research by automating information extraction from case reports. Future research should focus on expanding datasets, incorporating multilingual case reports, and exploring advanced techniques such as fine-tuning, relation extraction, and multimodal information processing. The findings contribute to advancing clinical information extraction and improving access to structured data for rare disease diagnosis and treatment.

ZUSAMMENFASSUNG

Seltene Krankheiten stellen aufgrund ihrer geringen Prävalenz und der begrenzten Datenverfügbarkeit besondere Herausforderungen für die klinische Forschung und Behandlung dar.

Diese Arbeit untersucht die Anwendung von Named Entity Recognition (NER) zur Extraktion strukturierter Informationen aus klinischen Fallberichten zu seltenen Krankheiten.

Moderne NER-Ansätze, darunter Transformer-Modelle und NER spezifisches prompten von Large Language Models, wurden auf ihre Fähigkeit zur Identifizierung von Entitäten wie Krankheiten, Symptomen und Medikamenten untersucht.

Da es keine etablierten Datensätze für NER zu Fallberichten zu seltenen Krankheiten gibt, wurde ein eigener Datensatz erstellt, in dem 82 Fallberichte aus PubMed Central annotiert wurden.

Zusätzlich zu diesem benutzerdefinierten Datensatz wurde der Datensatz MACCROBAT2020 evaluiert, der Fallberichte sowohl zu seltenen als auch zu nicht seltenen Krankheiten enthält.

Die Ergebnisse zeigen, dass vortrainierte, auf biomedizinische Domänen zugeschnittene Modelle wie `biomedical-ner-all` generalistische Ansätze wie `GliNER` und promptbasierte Methoden wie `PromptNER` in Precision und Recall übertreffen. Die Analyse von Fehlerfällen zeigte, dass Herausforderungen wie geringer Recall, ungenaue Abgrenzungen und Fehlklassifizierungen von Entitätstypen bestehen bei allen Modellen fortbestehen. Zudem können durch Halluzinationen Fehler bei promptbasierten Methoden auftreten.

Promptbasierte NER-Methoden zeigten zwar Flexibilität, benötigen aber möglicherweise weiteres Prompt-Engineering und Fine-tuning, um die Leistung im untersuchten Szenario zu verbessern.

Diese Arbeit unterstreicht das Potenzial von NER-Systemen zur Unterstützung der Forschung zu seltenen Krankheiten durch die Automatisierung der Informationsextraktion aus Fallberichten.

Zukünftige Forschung sollte sich auf die Erweiterung von Datensätzen, die Einbeziehung mehrsprachiger Fallberichte und die Erforschung fortschrittlicher Techniken wie Feinabstimmung, Relationsextraktion und multimodale Informationsverarbeitung konzentrieren.

Die Ergebnisse tragen dazu bei, die Extraktion klinischer Informationen voranzutreiben und den Zugang zu strukturierten Daten für die Diagnose und Behandlung seltener Krankheiten zu verbessern.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Part I

THESIS

1 INTRODUCTION

This chapter aims to provide an introduction and motivation for the topic of information extraction on case reports for rare diseases. The main problem presented in this work will be stated and motivation for research in this domain will be given. Scope of this work will be defined and finally the structure of the remaining thesis will be briefly presented.

1.1 MOTIVATION AND PROBLEM STATEMENT

In clinical medicine data is commonly used by medical professionals to improve decision-making. Data about specific diseases, their clinical presentations (including signs and symptoms related to the diseases) can help in development and optimization of treatment plans. For rare diseases data is often sparse and specific clinics and practitioners might only have first-hand data for a single digit number or even no patients for a given rare disease. For this reason and to help improve research in the area of rare diseases, additional data related to individual rare diseases can be beneficial. Sufficient quality must be guaranteed for the extracted data, as factually incorrect data could be of no use or even harmful in clinical contexts. Because medical data is usually considered highly sensitive and confidential, data collection from clinical texts and resources such as electronic health records can be difficult.[38] Information extraction on scientific publications could be a method of gathering additional information for this purpose. [34]

1.1.1 *Rare Diseases*

While there is no universally accepted definition for rare diseases, generally any disease that only affects a small part of the population is considered a rare disease. The "Rare Diseases Act of 2002", part of federal law in the United States, defines rare diseases as "any disease or condition that affects fewer than 200,000 people in the United States"[66] The European Commission provides the following definition: "Rare diseases, including those of genetic origin, are life-threatening or chronically debilitating diseases which are of such low prevalence that special combined efforts are needed to address them. As a guide, low prevalence is taken as prevalence of less than

5 per 10,000 in the Community".[16] Additionally the European Commission provides the following claims regarding rare diseases in the European Union:

- between 27 and 36 million people live with a rare disease
- between 6 000 and 8 000 distinct rare diseases are estimated to exist today
- while one rare disease may affect only a few patients, others may affect as many as 245 000
- Around 80% of rare diseases are of genetic origin
- 70% of rare diseases start in childhood

[51]

Organizations like the *National Organization for Rare Disorders*[43] (United States) and platforms like *Orphanet*[46] aim to provide information and guidance related to rare diseases to clinicians and patients.

Diagnosis of rare diseases can often be difficult because a clinician might lack experience and data for a specific rare disease.[53] Additionally, Rare diseases are often underrepresented in medical datasets.[39][45] This can lead to no diagnosis or even misdiagnosis which in turn could be harmful to therapeutic success of rare diseases.[68]

1.1.2 *Case Studies*

In biomedical research case studies (or case reports) are a type of research publication where clinical observations (usually those of a medical problem) of one or more patients are presented. [22] Case studies are generally considered to be anecdotal evidence. By design, case studies are methodologically limited and do not contain any form of statistical sampling. Because of this, it is generally advised to consider them last when seeking scientific / medical evidence for certain clinical interactions.[1] However they can still be useful to clinicians for example in alerting them about certain adverse drug reactions [1] or in the context of rare diseases[22].

Case reports generally briefly introduce a surveyed patient including their age and sex, their medical history and any currently present medical signs, symptoms or other findings. In the context of rare diseases, the disease is usually briefly presented to the reader including

its clinical presentation (common signs and symptoms) and existing epidemiological data. The report contains clinical findings related to the patient, as well as the results of performed diagnostic procedures. Usually large components of the reports are composed of unstructured text but may also contain structured data. In some papers results of diagnosis can be present in the form of images (e.g. slices of magnetic resonance imaging) or tables of laboratory values (e.g. blood work).[34]

Similar to other scientific literature, case reports are often available on public platforms such as PubMed[47] which is provided as part of the "National library of medicine" by the "National Center for Biotechnology Information". For a subset of the literature, the full text is available freely to the public, while for others only part of the report (usually its abstract) is freely available. PubMed lists a total of over two million case reports (as of June 2025) and over 640,000 of those are available with free full text on PubMed Central (PMC).[47] The public availability of case studies on platforms such as PubMed Central (PMC) resembles a largely untapped store of scientific and clinical information and provides an opportunity for evaluating information extraction on case studies and their full text. Unlike previous approaches using optical character recognition (OCR) to extract information from scanned documents[34], the case reports available on PubMed Central are already available in a digital format and can be used directly for information extraction.

1.1.3 *Problem Statement*

While manually extracting information from the case reports may be performed for individual cases, doing this on a large scale is considered error-prone and labor-intensive which makes it impractical. An automated information retrieval system extracting key information (e.g.: presented symptoms, related diseases, administered medication) from the texts could provide data that could be used for a variety of applications. Applications may include assisting in diagnostic and treatment of rare diseases by providing key information to clinicians or performing further research based on the extracted information. [34]

In the context of rare diseases, complex and non-standardized terminology can provide an additional challenge to information extraction systems. Fine-tuning existing named entity recognition models on this problem may be problematic due to the low number of available training samples per specific disease. Information extraction on

case studies for rare diseases can therefore be considered a *low-resource* problem.[60]

1.2 OBJECTIVES AND RESEARCH QUESTIONS

The primary goal of this study is to show how modern named entity recognition techniques can be used to effectively identify and classify diseases, symptoms, medications and treatment and diagnostic procedures as well as general patient related data (age and sex) from case reports in the context of rare diseases. For this purpose state-of-the-art named entity recognition models (e.g. Transformer based, Prompt-based) will be discussed and their performance in the stated task will be compared. Because rare-diseases can be considered a low-resource domain in named entity recognition, this work will emphasize low-resource named entity recognition and discuss model performance in this scenario. Specifically novel low-resource methods such as `GliNER` and prompt-based NER will be evaluated and compared to state-of-the-art approaches using task specific transformer-based models. Quality of extraction will be investigated using established metrics (Precision, Recall, F_1-Score) for the task of named entity recognition. In addition to usual statistic evaluation of model performance, manual error analysis will be conducted to highlight common errors in the context of medical texts and specifically rare diseases. This thesis also aims to provide a reproducible system for gathering of case report data and the extraction of information on those texts. Problems encountered in creation of this system will be discussed.

1.3 SCOPE AND CONTRIBUTIONS

Entity extraction is limited to the entity and event types defined by the ACCROBAT[11] system.

The focus is on the extraction of entities from case studies, not on the extraction of relations between them. While other subtasks of information extraction such as relationship extraction and entity normalization are briefly covered in this study they are not part of the experimental evaluation.

Modern NER methods are used, specifically transformer-based models. Legacy methods (e.g., CRF, SVM) are not considered.

Models are not fine-tuned as part of this thesis.

The models are evaluated on a small self-annotated dataset of rare disease case studies drawn from PubMed Central. The self-annotated and MACCROBAT2020 test datasets consists entirely of studies found

in the PubMed Central database. As PubMed Central only contains articles whose full-text is available for free, no case reports with restricted access will be considered. Only case reports in the English language will be considered. The dataset is not publicly available, but the annotation guidelines and the annotation tool are described in detail.

1.4 THESIS STRUCTURE

In the next part of this work foundational background for named entity recognition will be presented. Relevant approaches for named entity recognition will be shown and technical background for modern NER models (based on neural networks) will be given. Related work in the field of low-resource NER, medical NER and natural language processing in rare diseases will be presented and briefly discussed.

Following the theoretical background methodology for the evaluation performed as part of this work will be defined. This includes definition and discussion of metrics used for evaluation, as well as discussion of the datasets used for evaluation. The chapter will also present the approaches and models evaluated in this work, including any parameters and input used in the evaluation.

Practical implementation of the methodology will be presented to ensure reproducibility and highlight encountered problems that occurred in implementation. The chapter also includes information about the creation of the dataset with specifics regarding the collection and preprocessing of data for the test set. Used software and tooling will be briefly mentioned.

A chapter for discussion of experimental results will include the results of the statistical evaluation of the models with common metrics for named entity recognition. Discussion of a sample of individual errors will be discussed to highlight weaknesses of specific models. Implications of this work for information extraction on biomedical texts and texts regarding rare diseases specifically will be discussed.

A concluding chapter will provide a summary for this work and resulting findings. Limitations of the study will be acknowledgement and suggestions for future research will be given.

2

BACKGROUND AND RELATED WORK

In this chapter, named entity recognition is introduced as a task in natural language processing. Foundations of transformer models and large language models are described, as they are the basis for many modern approaches to named entity recognition. Furthermore, related work in the field of named entity recognition in medical texts is discussed.

2.1 NAMED ENTITY RECOGNITION (NER)

Named Entity Recognition (NER) is a specific task in Natural Language Processing (NLP) that involves identifying and classifying entities in unstructured texts. NER is a subtask of information extraction, which describes tasks that transform unstructured text into structured data.[31]

2.1.1 *Tokenization*

In order to perform named entity recognition (or many other NLP tasks), input text has to be tokenized.[31] Tokenization describes the process of separating words or word parts from running text, it is part of a group of tasks called text normalization which aim to convert text into a more convenient (easier to handle for machines) form.[31] Figure 2.1 shows a simple example of tokenization, in which tokens are equal to words. As English words are often separated by white spaces, tokenizing by splitting the text at white spaces would be a straight forward approach. This approach may lead to problems with terms that are treated like large single words (*New York*, *Rock 'n' Roll*) or terms like *I'm* which can be considered two words (*I* and *am*)[31]. Tokenization differs by language, an example would be German using comma and English using a period as a decimal marker.[31] Tokenization in languages like Chinese or Japanese can be considered more complex as they do not use spaces to mark potential word-boundaries and individual characters representing a single unit of meaning.[31] Tokenization may also be specific to the model used and the domain of the text at hand.

Biomedical texts often include agglutinative language, meaning that morphemes(the smallest single unit of meaning[31]) taken from Greek or Latin are combined to larger words. [4] Individual agglutinative terms may be rare in training texts and certain terms encountered after training could be out-of-vocabulary for a language model.[24] An example would be the usage of the **-itis** suffix indicating some inflammatory disease (e.g. *appendic-itis* being inflammation of the appendix).[4] Models specific to biomedical texts (e.g. PubMedBERT) may use different tokenization to generic models.[24] It should be noted however, that experiments show that domain-specific tokenization may not increase performance for biomedical-NER or only in certain low-resource scenarios. [24]

The patient was followed up with rivaroxaban.
⇒ | The | | patient | | was | | followed | | up | | with | | rivaroxaban |

Figure 2.1: Example of simple tokenization (dropping the punctuation mark entirely)

Models may also use special tokens for example the `[MASK]` token used for training in BERT based models (see Figure 2.7) or the `[SEP]` token used to separate sentences in BERT based models.[17]

2.1.1.1 *Byte-Pair-Encoding Tokenization*

Byte-pair-encoding (BPE) is a compression algorithm[21] that allows representing an open vocabulary through a fixed-size vocabulary using character sequences of variable length[59]. Originally used for byte-level encoding, the algorithm was adapted to operate on a character level.[59] Initially the vocabulary of the token learner starts out containing all individual characters (e.g. all alphabetical characters A-Z). In iterative steps the algorithm then merges the two most frequent adjacent characters into a new entry in the vocabulary. Performing this for k steps results in a vocabulary of the original size plus k new tokens.[31] Used on a corpus with the four words "ab", "bc", "bcd" and "cde" and $k = 5$ would result in the following iterations:

Iteration	**Vocabulary**
0	{a,b,c,d,e}
1	{a,b,c,d,e,bc}
2	{a,b,c,d,e,bc,cd}
3	{a,b,c,d,e,bc,cd,ab}
4	{a,b,c,d,e,bc,cd,ab,de}
5	{a,b,c,d,e,bc,cd,ab,de,bcd}

The learned vocabulary ({a,b,c,d,e,bc,cd,ab,de,bcd}) is then applied in tokenization by splitting the input text into characters and replacing character combinations existent in the vocabulary. For the example input this would result in the following tokenization:

ab	bc	bcd	cd	e

Usually the algorithm maximally operates on a word level and does not combine adjacent words into single tokens. On large input corpora BPE tokenization can iterate through thousands of merges, resulting in a tokenization where most words can be represented by a single token and only large words are represented using multiple tokens.[31]

2.1.2 *NER problem Definition*

Formally a sequence of Tokens T of length N is represented by $T = (t_1, t_2, \ldots, t_N)$. NER then generates a set of tuples (I_s, I_e, l), where s and e are integers confined to the interval $[1, N]$. I_s and I_e denote the start and end of a named entity and l indicates its category from a set of predefined categories. For example, in the sentence "Barack Obama was born in Honolulu.", NER would identify "Barack Obama" as a person's name and "Honolulu" as a location, as shown in Figure 2.2. [33]

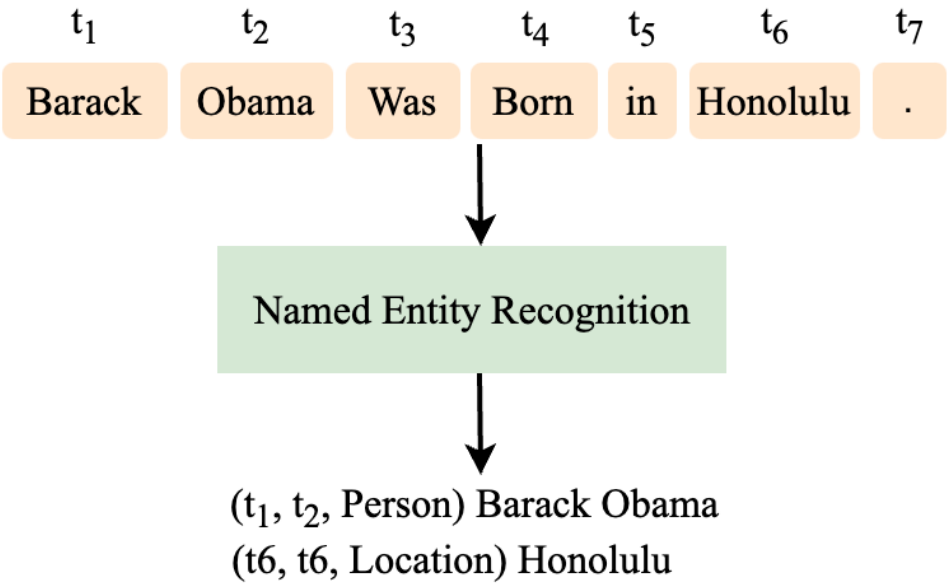
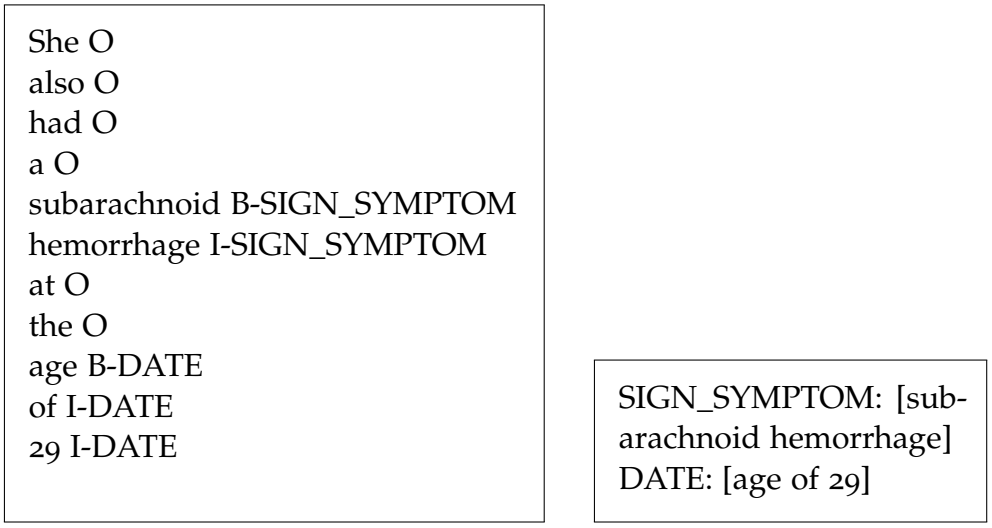


Figure 2.2: Given a sequence of tokens, NER outputs the positions of the named entities along with their associated categories[33]

NER may contain discontinuous entities, meaning that entities consist of multiple tokens that may have gaps (tokens belonging to an-

other entity or no-entity) in between.[70] In similar fashion there are also overlapping entities, where tokens can be part of multiple entities at a time. Support for discontinuous and overlapping entities is not available for all NER systems and approaches.[70]

In addition to the already described problem which can be referred to as *token-level* NER, there is also *document-level* NER. Document-level NER is a task where given a document, a system extracts all entities and their types without listing their exact spans.[29][39] A comparative example for token and document-level NER can be seen in Figure 2.3.



(a) Token-level NER (b) Document-level NER

Figure 2.3: Output of token-level NER and document-level NER systems for the given input sentence "She also had a subarachnoid hemorrhage at the age of 29.". Token level NER is shown in the BIO-Format (see Section 2.1.6)

Without token-level granularity, the accuracy of tasks such as medical NER and RE may be limited.[29] Token-level granularity may be especially important in the context of clinical applications where context can be valuable for understanding of patient-related data.[39] However some researchers argue that exact matches might not be necessary for every application and only the existence of a named entity might matter[65]

Transformation of token-level NER into document-level NER can be considered trivial, as it only requires grouping the tokens of the same entity type together and removing the span information. Document-level to token-level NER is more complex, as it requires the identification of the spans of the extracted entities in the input text. This can be

done by searching for the entity terms in the input text and determining their start and end positions, but may lead to suboptimal results when words are repeated in the text or entities are discontinuous.[2]

Research in the field of NER has been growing rapidly in the last years, especially with the introduction of transformer based NER methods as shown in Figure 2.4.

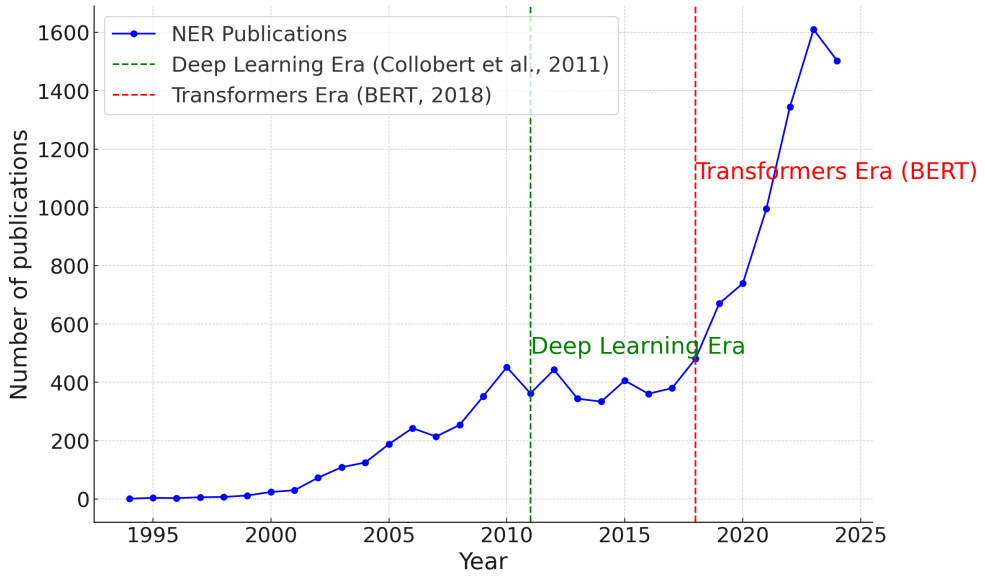


Figure 2.4: Growth of NER Publications[33]

2.1.3 *Datasets*

Numerous datasets are available for the task of named entity recognition. These datasets are often specific to a domain (e.g. biomedical, financial, news) and language (e.g. English, German, Chinese).[33] NER datasets contain texts along with their ground truth annotations for named entities. While annotation was historically done by hand, recent approaches also include automatic annotation of datasets.[6] Labeling may also be sped up by using pre-trained models to perform initial labeling of the dataset. The human annotators then only have to correct and extend the labels of the pre-labeled dataset, reducing the total workload of annotation.

2.1.4 *Approaches*

NER approaches are generally classified into knowledge- or rule-based approaches, statistical approaches or deep-learning approaches.[33] Machine learning approaches like support vector machines (SVM), conditional random fields (CRF), hidden Markov models (HMM) or

maximum entropy models were used with success in the past to perform NER, but are not discussed further in this work.[33] In recent years, prompt-based approaches (which also use deep-learning models) represent a novel approach to NER.[33] A hierarchical overview of NER approaches can be seen in Figure 2.5.

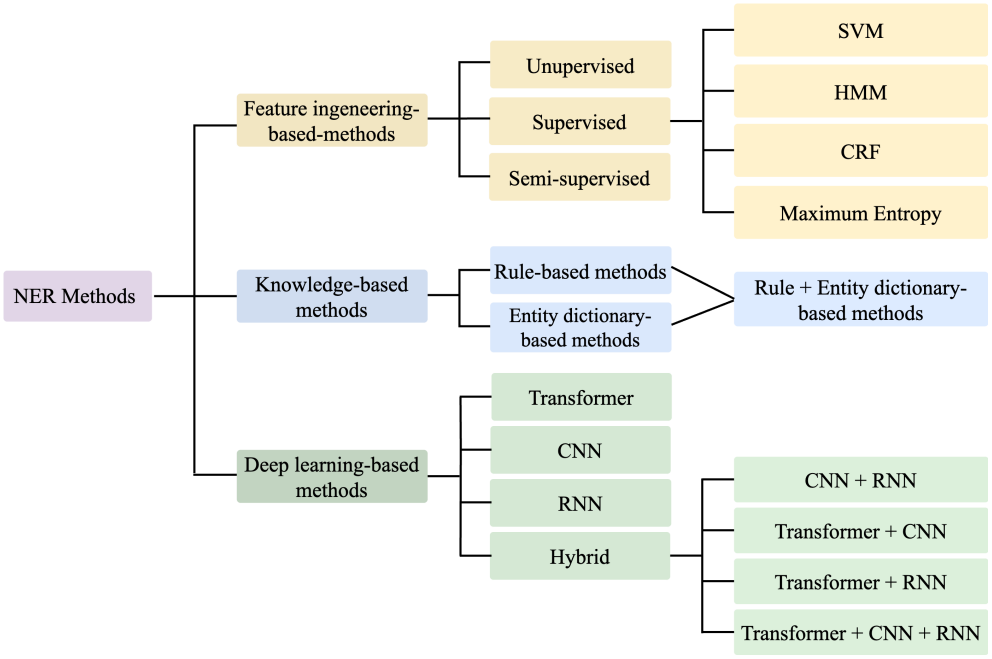


Figure 2.5: NER approaches as described by **Keraghel et al.**[33]

2.1.4.1 *Knowledge- and Rule-based NER*

Knowledge based methods originate from linguistic principles and use predefined rules and lexical resources for the task of NER.[33] The architecture of a knowledge-based NER system can be seen in Figure 2.6. A primitive knowledge-based approach could be looking up an exact match of the word in a predefined dictionary of terms for specific entity types. Downsides of this approach are numerous: the method requires the exact matching of characters where slight differences in spelling could prevent an entity from being detected. Classification of words can also be wrong for words that have different meanings in different contexts. [33] Furthermore the look-up based system provides no way of classifying out-of-dictionary words which might be especially common in specific domains such as biomedical texts.[33] Lastly adapting the technique to different entity types might require rebuilding the all dictionaries. An example might be the identification of a persons name by recognizing a preceding "Mr." or "Ms." before the last name of person.

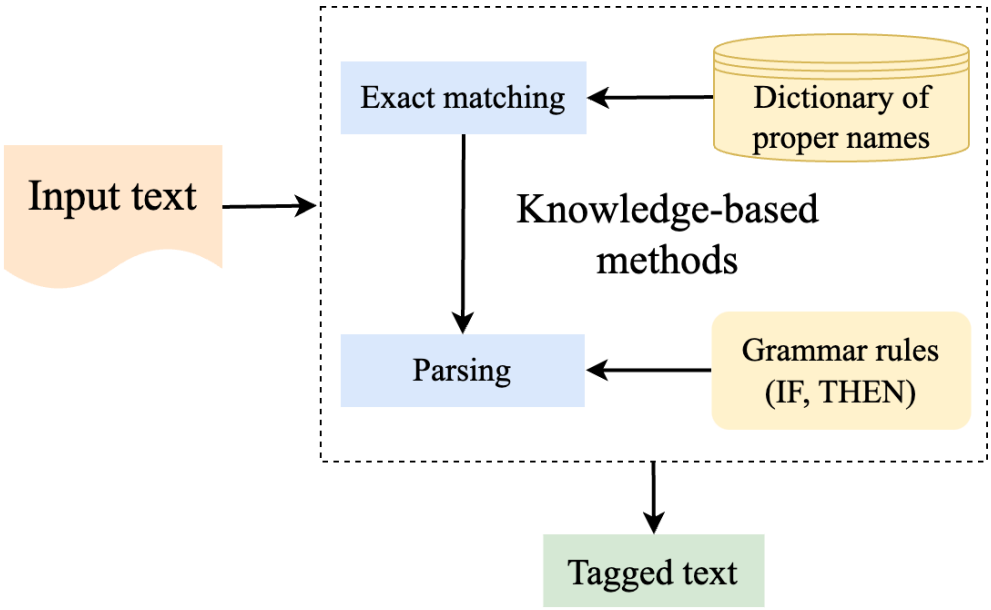


Figure 2.6: Architecture of knowledge-based NER[33]

These methods often present high precision and low recall (a definition for these metrics can be found in Section 2.1.7).

2.1.5 *Low-resource NER*

Low resource NER is a special form of the general NER task where training data for the performing model is highly limited (fewer than 50 examples [36]). A low number of available example can be due to the specific NER task at hand being performed on a language where few examples are available [14] or because the domain of the NER problem (medical, financial, etc.) provides limited available data. Rare Diseases could be considered an example for a sparse exampled domain because individual rare-diseases might only affect single-digit number of patients. Entities such as symptoms may be highly specific to a specific rare-disease limiting the available examples that could be used for training NER models for this task.[40]

Another method to improve low-resource NER can be active learning, a form of supervised or semi-supervised learning where training samples are strategically selected to maximize model performance per training iteration. This is commonly done by apply uncertainty-sampling where training data is selected based on the uncertainty shown by the model when performing inference for a specific data point.[33]

The emergence of LLMs and prompt-based NER methods such as *PromptNER* show promising results in low-resource NER scenarios.[69] However performance may be highly reliant on the prompt, requiring careful prompt engineering and / or integration of domain specific knowledge[29]

2.1.6 *Formats*

Persisting results or ground truths for named entity recognition was historically done in the *inside-outside-beginning* (IOB or BIO) format. The Conference on Computational Natural Language Learning (CoNLL) established the common CoNLL format for NER as part of its 2003 shared task.[63] In the CoNLL format there is one line for every word in the input text. The words are accompanied by a tag specifying the type of entity the word belongs to. Words that can not be associated to any entity receive the tag `O`, indicating the word is **o**utside of named entities. A tag encoding a specific entity type starts either with `B` when a word marks the **b**eginning of a new named entity or `I` when **i**nside already begun named entity. Entities are assumed to be non-recursive and non-overlapping. An example for the CoNLL format can be seen below.

```
The O
male B-SEX
patient O
stated O
pain I-SYMPTOM
during O
physical B-DIAGNOSTIC_PROCEDURE
examination I-DIAGNOSTIC_PROCEDURE
```

Other commonly used format include the BRAT-standoff format[61] used by the `brat` annotation tool[7] and JSON format used by tools like `SpaCy`[26] or `LabelStudio`[64]

2.1.7 *Evaluation*

Typically, NER systems are evaluated by comparing the extracted entities, their positions and assigned types to a gold standard. [33] Evaluation is typically differentiated into two categories: *strict* and *loose* evaluation.[58] Strict evaluation requiring exact matches of entity positions while loose evaluation only requires some overlap of predicted and gold standard entity positions. Evaluating NER systems follows the evaluation of binary classification systems.[63] When evaluating,

true positives (TP), false positives (FP) and false negatives (FN) are calculated:

- A true positive is a tuple tp where positions I_s, I_e and category l of the extracted entity match those of the gold standard.
- A false positive is a tuple fp where either one the positions or the category of the extracted entity do not match those of the gold standard.
- A false negative is a tuple fn where the positions of the extracted entity match none of the entities in the gold standard.

[37] tp, fp and fn are calculated for all extracted entities and individually summed up to their counts TP, FP and FN. The precision (sometimes called positive predictive value in other contexts) of the NER system is defined as the number of true positives TP relative to the number of total positives:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.1}$$

The best possible precision is 1, indicating no false positives, the worst precision being 0 indicating no true positives and only false positives.

Recall (sometimes called sensitivity or true positive rate) is the number of true positives TP relative to the number of true positives and false negatives.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.2}$$

The best possible recall is 1, indicating no false negative, the worst precision being 0 indicating no true positives and only false negatives.

The F_1-Score combines recall and precision using a harmonic mean[33]

$$F_1\text{-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.3}$$

[37] The highest possible value of an F-score is 1.0, indicating perfect precision and recall, and the lowest possible value is 0, if the precision or the recall is zero.[33] Evaluation in this work is further described in Section 3.3

2.1.8 *Related NLP Tasks*

2.1.8.1 *Entity Linking / Entity Normalization*

The related task of entity linking (sometimes called entity normalization) includes linking the extracted entities to specific entries in databases. Examples are finding Wikipedia articles for extracted entities.

2.1.8.2 *Relationship extraction*

The task of relationship extraction extends named entity recognition in finding relations between entities. An example in the biomedical domain could be [CHEMICAL] [causes] [Disease] or [chemical] [treats] [disease]. Datasets for NER commonly also include annotations for relations.[40][11]

2.2 NEURAL NETWORKS FOR NATURAL LANGUAGE PROCESSING

Neural networks are widely established and considered successful in natural language processing tasks.[76] In recent years, so-called *transformer* models utilizing *attention* mechanisms and *encoder-decoder* structures were established as state of the art for a variety of tasks in natural language processing and other areas of machine learning.[76]

2.2.1 *Transformer-models*

Transformer models were first described in the "Attention is all you need"[67] paper by Vaswani and colleagues. The researchers tried to create a neural network architecture that relies on the attention mechanism instead of recurrence or convolution. In machine learning, the attention mechanisms allows modeling dependencies between elements without considering their distance in a sequence.[67] For NLP this allowed higher levels of parallelization as compared to methods like Long short-term memory (LSTM) neural networks which have an inherent sequential mechanism. [67] Transformers utilize an *encoder-decoder* structure, to provide an intermediate representation of tokens inside the neural net. The encoder maps a sequence of input signals to a sequence of continuous representations. In a subsequent step the decoder generates an output sequence of tokens, one token at a time.

A depiction of the original transformer architecture can be seen in Figure 2.7.

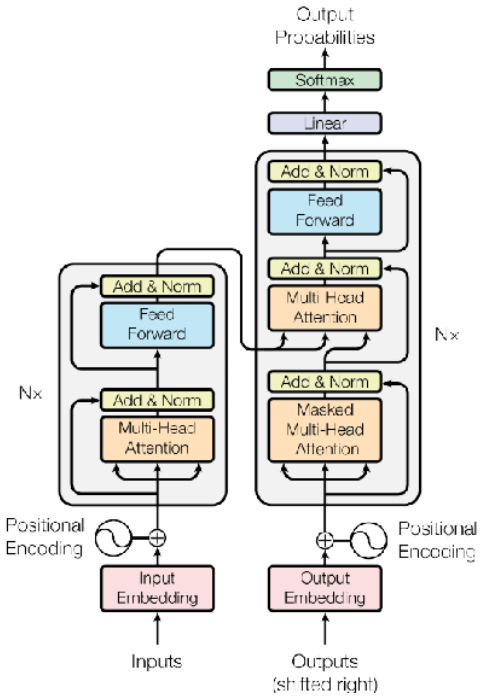


Figure 2.7: The original transformer architecture by Vaswani et al.[67]

The described architecture performed exceptionally well on machine translation tasks. Compared to other models relying on recurrence, the sole usage of attention layers allowed the for high parallelization in model training. This allowed training of larger models with this architecture. The researchers state required training costs to be an order of magnitude lower than those of models with comparable performance in text translation tasks.[67]

Numerous transformer-based language models for various tasks were created following the publishing of the general transformer architecture.[17, 49, 50] Popular transformer based models like BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers)[17] were published in the years after introduction. BERT and some of its variants like RoBERTa or DeBERTa are publically available on platforms like HuggingFace[71]. BERT is pre-trained on unlabeled text and is designed to be fine-tuned by adding a classification output layer that allows adaption to a wide range of tasks.[17] Pre-training of BERT was performed with the BooksCorpus and English Wikipedia, providing a sum of about 3.3 billion words.[17]

Some transformer based models including those described in this thesis are created using the process of *knowledge distillation*. Knowledge distillation is a process where a smaller model (the student) is trained to mimic the behavior of a larger, pre-trained model (the teacher). This process allows the smaller model to learn from the

larger model's knowledge, often resulting in a model that performs well on specific tasks while being more efficient in terms of size and computational requirements.[25][55]

Following the introduction of the transformer architecture, many variants were created. BERT (Bidirectional Encoder Representations from Transformers) was introduced by Devlin et al. in 2019 and is one of the most well-known transformer-based models.[17] BERT is pre-trained on large corpora of text and can be fine-tuned for specific tasks like NER, question answering, or text classification.[17]

2.2.2 *Large Language Models (LLMs)*

Large Language Models (LLMs) are machine learning models that used for natural language processing. LLMs are usually based on the transformer architecture described earlier. Models are considered large due to the large number of parameters they have, sizes differ but parameter counts in the order of billions are common.[23][62][9] Generally LLMs perform a language generation task and are trained to predict the most likely next tokens for a given input.

Training of Large language models is performed in a semi-supervised way. Given a text, individual tokens in the text are masked and the model is trained to predict the masked token. While LLMs are pre-trained on large corpora of texts, exact specification of the data used for training is rarely provided. LLMs can be fine-tuned for specific tasks or to encode knowledge and language understanding in specific domains.

A common use for large language models are chatbots such as ChatGPT that converse with users based on an initial and potential following prompts. Techniques like reinforcement-learning are often used to prevent certain interactions (illegal or considered immoral) with users of the models or to prevent text generation about specific topics. LLMs show remarkable performance in a variety of NLP tasks such as question answering, summarization, sentiment analysis or machine translation. [73] Lots of tests are available for evaluating LLMs depending on the task in question, but the metric of perplexity is commonly used for the main task of text generation.

Large language models commonly use a temperature parameter to control the randomness of the generated text. A low temperature (e.g. 0.1) will lead to more deterministic text generation, while a higher temperature (e.g. 1.0) will lead to more diverse and random text generation.[28] The temperature parameter may be set on a per prompt basis or globally for the model.

2.3 APPROACHES TO NAMED ENTITY RECOGNITION USING NEURAL NETWORKS

Using the described neural network models, two main approaches to named entity recognition can be distinguished:

2.3.1 *Named Entity Recognition using transformer models*

NER with transformers is often performed by adding a classification layer on top of the transformer model and training the model on a specific NER task and dataset.[39][52] A depiction of the fine-tuning process for the task of NER can be seen in Figure 2.8.

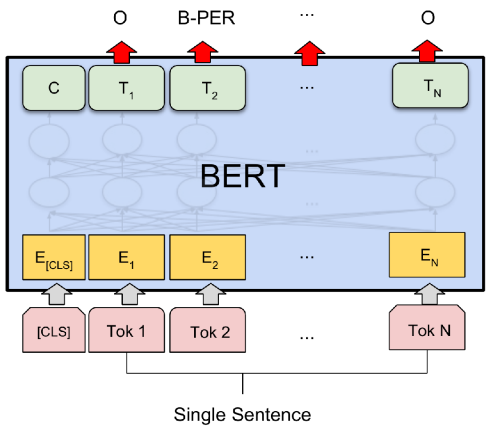


Figure 2.8: BERT fine-tuning for NER[17]

A transformer model fine-tuned in this way is specific to the entity types it was trained on. Models for various datasets and entity types are available on platforms like HuggingFace[71].

2.3.2 *Prompting LLMs for Named Entity Recognition*

Prompting Large Language Models for Named Entity Recognition does not require fine-tuning the models for specific entities. Generally, two types of prompting are described in scientific literature:

1. prompting on an per-entity-type basis, meaning that the model is prompted once for each entity type to be extracted (e.g. GPT-NER). [69]

2. prompting with a single prompt that describes the task of NER and the entity types to be extracted (e.g. PromptNER).[2]

One advantage of prompt based approaches is that no retraining or fine-tuning is needed when changing entity types because the entities to be extracted are defined in the prompt itself.[33]

Model performance can be improved by providing examples inside the prompt. This method is called few shot prompting. Examples may be positive examples or negative examples.Examples can also include a free text description explaining why the example might or might not be some specific entity.

While this provides a novel approach for low-resource NER it is not without its problems. LLMs tend to hallucinate, meaning they might generate incorrect information. In the context of NER this can include generating labels for texts that are not part of the actual input text or assigning entity labels that were not defined in the prompt.[2] Furthermore available prompt based NER systems require the output of the model to adhere to a specific format in order to be parsed and converted into a NER format such as BIO or CoNLL.As output and output format of large language models is inherently non-deterministic, failure to provide parsable output text can further reduces model performance.

In contrast to traditional NER systems, prompt based NER systems can not output classification scores for the extracted entities.[2] Another potential concern with this approach can be data contamination. For a lot of large language models the exact training data is not known, this could mean that texts subject to NER could have been "seen" in training of the model beforehand.[2] This is especially true for closed weight models, but there are also open weight models for which training data is not stated.

Estimating the exact impact of data contamination is difficult. Tests by Google on the task of question answering show that their Pathway Language Model(PaLM)[13] do not perform significantly better on previously seen texts compared to clean data.[13]

Finally the needed resources for prompting large language models compared to NER specific models should be considered, both from a performance standpoint and their environmental impact.[72]

In a scoping review **Schaefer and colleagues** describe the use of machine learning in the context of rare diseases. After investigating a total of 337 studies they found that most studies used machine learning for diagnosis and prognosis of rare disease while relatively few studies concerned treatment of rare diseases. Furthermore, they found that

the number of studies combining rare diseases and machine learning increased alongside a general upwards trend in publications concerning machine learning.[56]

Kariampuzha et al. developed an NER corpus, deep learning framework and information extraction pipeline for epidemiology data on rare diseases. Similar to this work, the authors evaluated the performance of their system on scientific literature related to specific rare diseases. Precision of different BERT-based models was compared on entity and token level. A BioBERT model was then used for extracting epidemiology, statistics, location, date, sex and ethnicity entities. The authors suggest the gathering of more training data from rare disease literature to improve the performance of their system.[32]

J. Hu and colleagues tested existing BERT models in the task of medical named entity recognition. Comparing BERT, BioBERT, ClinicalBERT, SciBERT and BlueBERT they found BioBERT to achieve the highest performance in both precision and F_1-Score when tested on the MIMIC-III dataset. The authors highlight BioBERTs adaptability and accuracy in handling complex medical terminology and claim that it can be a powerful support tool for clinical data analysis and medical decision support.[27]

Yang and colleagues show the development of two custom models `PhenoBCBERT` and `PhenoGPT` for extracting human phenotypes from clinical notes. Phenotype-driven approaches can be used in the genetic diagnosis of rare diseases. Extracted terms are linked to concepts in the UMLS database. Both presented models were fine-tuned on the public BiolarkGSC+ dataset. Prompt-learning was used for NER with the `PhenoGPT` model.[72]

In a highly related work to this task, **Bedmar et al.** compare a variety of approaches for named entity recognition on the RareDis[40] dataset (see Section 3.1 for a detailed description). The researchers apply Bidirectional Long short-term memory neural networks, BERT and other BERT-based models to the problem. They find the BioBERT to obtain the best F_1-Scores on the dataset. Prompt-based approaches with LLM were not evaluated as creation of their paper was prior to the establishment of LLM as a tool for NLP-tasks. As part of their conclusion, the researchers propose inclusion of clinical case texts of rare diseases into the dataset for improved results. In a similar fashion, this thesis aims to implement this by creation of a separate dataset for

only clinical texts taken from PubMed Central.[57]

Y. Hu and colleagues evaluated the commercially available LLMs GPT-3.5 and GPT-4 for clinical named entity recognition on clinical notes from the *MTSamples* corpus and on safety reports from the *vaccine adverse event reporting system* (VAERS). They report that while the F_1-Scores achieved with these prompt based methods trail those of BioClinicalBERT, they are very promising in light of the low amount of training samples needed. Furthermore they highlight the importance of the prompt engineering process for the task of NER. They provide and a prompt framework with different approaches based on annotation guidelines, error-analysis and annotated few-shot examples and evaluate the approaches.[28]

Lu and colleagues evaluate prompt-based NER with LLMs on EHRs. Models considered for evaluation were the general purpose LLMs GPT-3.5, GPT-4 and LLaMA-2 as well as Meditron and Llama2-MedTuned which are adapted for the medical domain. The researchers also included the UniversalNER model which is a fine-tuned version of the LLaMA-2 model for the task of NER. They find that in general LLMS struggle with the task of NER. NER was performed with the entities *Disease*, *Rare Disease*, *Skin Rare Disease*, *Sign* and *Symptom* and evaluated using the `RareDis` Dataset. The researchers were able to show that few-shot learning was able to effectively improve the performance of prompt-based NER systems. Usage of Retrieval augmented generation showed promise especially for the more specific entities *Rare Disease* and *skin rare disease* but not for the more common entities. The researcher pointed out that more research in evaluating RAG for prompt-based NER is needed. Manual error analysis was performed for 50 randomly sampled sentences from the dataset.[39]

Monajatipoor and colleagues investigated the use of LLMs in biomedical NER. They show that the selection of in-context examples can yield good improvements in performance (F_1). The researchers developed a custom approach similar to RAG that extracts medical terms using the model in a first step and then performs a lookup of these terms in the medical database UMLS. The model is then prompted for a second time with an augmented input text performing the final NER task. The researcher claim that this system termed Dictionary-Infused RAG (DiRAG) can boost F_1-Score. Established biomedical datasets I2B2, NCBI-Disease and BC2GM were used for evaluation. Performance was compared for three different NER formats (BIO, DICE and TANL).[41]

2.4 CONCLUSION

In this chapter, the task of named entity recognition (NER) was introduced, along with its relevance in natural language processing and its specific challenges in medical texts. Various approaches to NER, including rule-based, statistical, and deep learning methods, were discussed, with a focus on transformer-based and prompt-based methods. The chapter also highlighted the importance of tokenization, datasets, and related tasks such as entity linking and relationship extraction. Finally, the application of NER in medical texts, particularly in the context of rare diseases, was explored, emphasizing the potential of specialized models and prompt-based approaches. This sets the foundation for further exploration and evaluation of NER in clinical and biomedical domains in subsequent chapters.

3

SYSTEM

In this chapter the methodology of evaluating named entity recognition on case reports for rare diseases will be presented. For this purpose the creation and annotation of the used dataset will be described and compared to existing datasets in medical natural language processing. Model selection for the performed experiments will be discussed based on state-of-the-art research. This includes NER-models as well as the selection of a large language model for use with prompt-based named entity recognition. The chapter also aims to provide a reproducible system for gathering, annotation and evaluation of named entity recognition on case reports. For this reason, used tools and technologies for annotation, model usage and evaluation will be described. The evaluation procedure leading to the results in chapter Chapter 4 will be defined here.

3.1 DATASETS

In the context of medical natural language processing numerous datasets exist. Datasets such as *RareDis*[40] or *RDD*[20] focus on rare diseases, but are not comprised of case reports. This work uses three datasets for evaluation: The MACCROBAT2020[11] dataset in its entirety, MACCROBAT2020RD, a subset of the MACCROBAT2020 dataset with only rare diseases and a custom dataset created from case reports.

3.1.1 *MACCROBAT Dataset*

MACCROBAT is a dataset consisting of case reports annotated using the Open Biomedical Annotation Terms. The datasets consists of a total of 200 distinct case reports and contains reports of rare diseases cases. Case reports are annotated with a wide list of different entity types. The authors differentiated between entity and event types[1]. All event and entity types are listed and described in tables Table 3.2 and Table 3.1 respectively. In addition to the entity and event types, relationships between entities and events were also annotated. The dataset also includes discontinuous entities (total of x Entities) which were disregarded in evaluation because recognition of such entities is

1 Further in this work, Entities and events will not be differentiated.

not supported by all evaluated models. The relationships between entities are not extracted nor evaluated in any form as part of this work. The authors of the dataset to not provide quality metrics for the annotation of the dataset such as Inter-Annotator Agreement (IAA) or other metrics. Annotation was performed by one or multiple annotators with previous experience in biomedical and clinical language.[11] Annotation was performed using the BRAT annotation tool[7] and is provided in the brat standoff format.

Table 3.1: Event types used for annotation in the MACCROBAT dataset[11]

Type	Description
Activity	Patient actions and habits.
Clinical event	A clinical activity other than a medical procedure, often involving a change of Nonbiological location.
Diagnostic procedure	Any procedure done primarily in order to obtain more information.
Disease disorder	Any disease. Essentially a higher-level medical condition potentially including a collection of symptoms.
Lab value	Any result of a laboratory test or a diagnostic result, including any units or values present.
Medication	Any pharmaceutical treatment. Used with Administration and Dosage entities.
Outcome	The patient's clinical outcome.
Sign symptom	Any symptom or clinical finding.
Therapeutic procedure	Any procedure done primarily in order to address or alleviate a symptom or disease. This includes surgery, long-term therapies, and supporting procedures (e.g., intubation).
Time expressions	
Date	A time expression ending at a specific day in time.
Duration	A time expression describing a period of time, generally specifying an event has occurred continuously over the given duration.
Time	A time expression describing a specific point in time.
Other event	Any event with clinical relevance that does not fit into any of the above categories.

Table 3.2: Entity types used for annotation in the MACCROBAT dataset[11]

Type	Description
Administration	Mode of administration of a drug or other therapy.
Age	Demographics. Patient age at time of presentation.
Area	Any area. Includes value and units.
Biological structure	Any part of the body, from the cellular level to general areas.
Color	A color.
Detailed description	Any detail of an event or other entity.
Diagnostic standard	A single qualitative or qualitative standard used to make a diagnostic conclusion.
Distance	Length, width, height or other 1-dimensional attributes. Includes value and units.
Dosage	A complex numerical expression describing medication or therapy dosage. Minimally requires an amount and frequency, including values and units. May be expressed by weight.
Frequency	An expression describing how often an event occurs. For drug dosages, prefer the Dosage entity.
Gene or protein	The name or other identifier of a gene or protein.
Mass	Any measurement of physical mass. Includes value and units.
Nonbiological location	Any physical location other than those on or within a patient's body.
Occupation	Any description of a subject's daily activities.
Personal background	Demographics. Any description of subject ethnicity or national background.
Qualitative concept	A detail of an event or other entity describing it in general terms. A high-level category for which other labels may be appropriate.
Quantitative concept	A numerical value. A high-level category for which other labels may be appropriate.
Severity	Degree of a disease or symptom's severity. Sex Demographics. Patient sex at time of presentation.
Shape	A shape.
Subject	Any individual related to or of medical relevance to the patient. Does not include clinical personnel.
Texture	A texture.
Volume	Any volume, including that of bodily fluids. Includes value and units.
History expressions	
History	Any description of subject medical history.
Family history	Any description of a patient's family medical history.

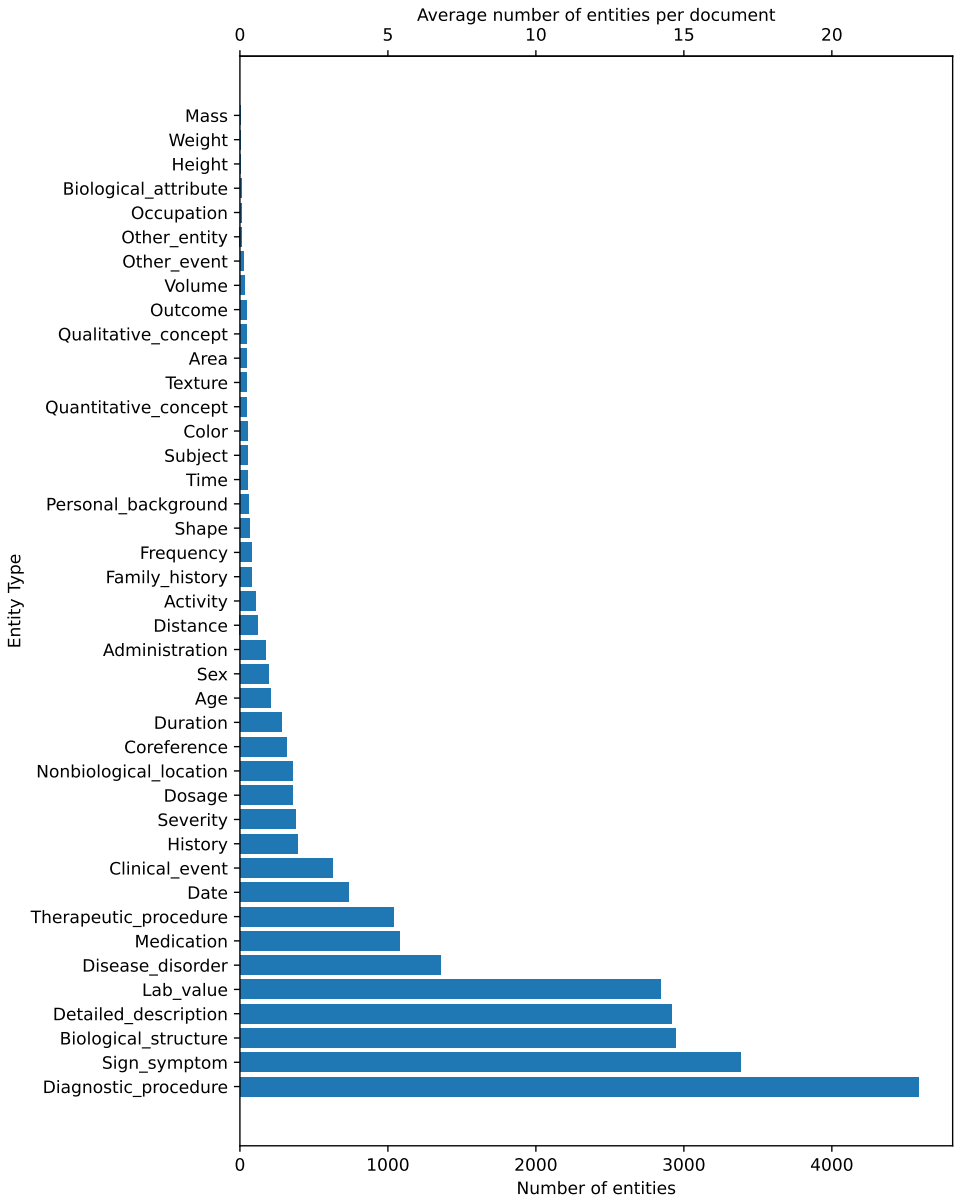


Figure 3.1: Absolute number of entities in the MACCROBAT dataset by type (total=25144)

The number of total entities by type can be seen in Figure 3.1.

MACCROBAT is available in two different versions: MACCROBAT2018 and MACCROBAT2020, with the 2020 version including improved annotations in consistency and format.[10] The 2020 version was used in this work.

The case reports are all in English language and only single human patient. Only the text portions comprising the clinical texts are part of the dataset. Other sections like Introduction or Discussion, figures or tables and any other supplementary materials were removed from

the documents. An extract of a case report from the MACCROBAT dataset with its NER annotation is shown in Figure 3.2. The full text of the case report can be found in the appendix (Figure A.1).

A 58-year-old Age man Sex had been suffering from general fatigue Sign_symptom and severe Severity anemia Sign_symptom for several months Duration . His hemoglobin levels Diagnostic_procedure were 6.6 g/dl Lab_value (normal range: 12–16 g/dl). He had no medical history History and did not take any medicine History . Esophagogastroduodenoscopy Diagnostic_procedure and colonoscopy Diagnostic_procedure did not reveal any significant bleeding Sign_symptom . Abdominal Biological_structure computer tomography Diagnostic_procedure revealed a 2-cm Distance hypervascular Detailed_description tumor Sign_symptom in the small intestine Biological_structure (Fig.1).

Figure 3.2: Extract of a case report (PMCID: 26444414) from the MACCROBAT dataset with its NER annotation.

[11]

3.1.2 *Subset of MACCROBAT containing only rare diseases*

A subset of the MACCROBAT dataset was created containing only case reports for rare diseases. The subset was created by looking up MeSH terms on PubMed for all case reports in the MACCROBAT dataset. The MeSH terms were then used to filter the case reports for rare diseases, using the Orphanet database as a reference. For three of the case reports no MeSH terms were available. 18 case reports were found to be tagged with MeSH terms for rare diseases. No document overlap exists between these case reports and the case reports used in the custom dataset. The subset included a total of 2444 individual named entities. For easier reference, the subset was termed *MACCROBAT2020RD* (RD for Rare Disease).

3.1.3 *Custom Dataset*

In addition to the rare-disease subset of the MACCROBAT dataset, a custom dataset was created. The custom dataset aims to provide a set

of texts focused on rare diseases, while keeping structure and format (one sentence per line) similar to the MACCROBAT dataset. A figure showing the process of creating the custom dataset can be seen in Figure 3.3.

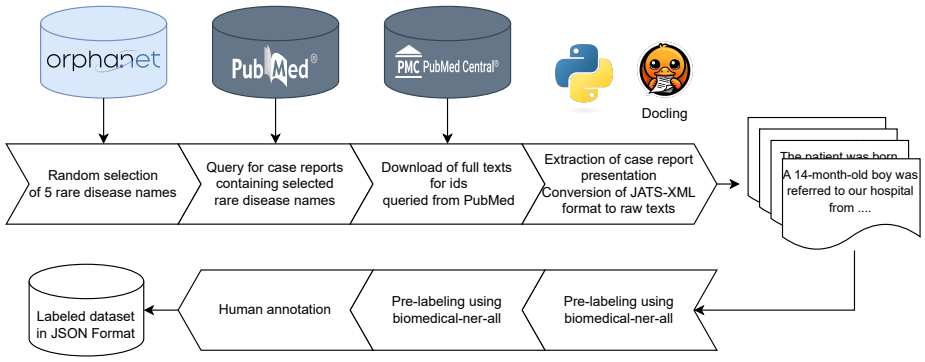


Figure 3.3: Process of creating the custom dataset for rare diseases.

3.1.3.1 *Custom dataset definition*

In order to create a custom test dataset, five rare diseases were randomly selected from the Orphanet database. The rare diseases, their identification in the Orphanet-Database (OrphaCode) and the summary in the Orphanet database can be found in Table 3.3.

Table 3.3: Selected rare diseases for the test set along with their OrphaCode and description found on Orphanet

Rare Disease	OrphaCode	Orphanet definition
Alkaptonuria	56	"A rare disorder of phenylalanine and tyrosine metabolism characterized by the accumulation of homogentisic acid (HGA) and its oxidized product, benzoquinone acetic acid (BQA), in various tissues (e.g. cartilage, connective tissue) and body fluids (urine, sweat), causing urine to darken when exposed to air as well as grey-blue coloration of the sclera and ear helix (ochronosis), and a disabling joint disease involving both the axial and peripheral joints (ochronotic arthropathy)."
Barth Syndrome	111	"Barth syndrome (BTHS) is an inborn error of phospholipid metabolism characterized by dilated cardiomyopathy (DCM), skeletal myopathy, neutropenia, growth delay and organic aciduria."
Fibrodysplasia Ossificans Progressiva	337	"Fibrodysplasia ossificans progressiva (FOP) is a severely disabling heritable disorder of connective tissue characterized by congenital malformations of the great toes and progressive heterotopic ossification that forms qualitatively normal bone in characteristic extraskeletal sites."
Leptospirosis	509	"An anthropozoonosis, rare in Europe, clinically characterized by an initial presentation of flu-like symptoms rapidly progressing into life-threatening multisystem failure (notably hepatonephritis) caused by spiral-shaped bacteria belonging to the genus Leptospira. Leptospirosis is a widespread zoonosis with a worldwide distribution and has emerged as a major public health problem in developing countries in South-East Asia and South America."
Norrie Disease	649	"A rare developmental defect during embryogenesis characterized by abnormal retinal development with congenital blindness. Common associated manifestations include sensorineural hearing loss and developmental delay, intellectual disability and/or behavioral disorders".

3.1.3.2 *Data Collection*

Similar to the process developed by **Kariampuzha et al.**[32] PubMed articles were queried based on the disease name provided by Orphanet. The National Center for Biotechnology Information provides an API (E-utilities) for querying PubMed articles and their metadata[18]. The query used for the search was: `'(<KEYWORD>"[Title/Abstract]) AND "case reports"[Publication Type] AND free full text [sb] AND (english[Filter])'` with `<KEYWORD>` being the name of the rare disease in question. After querying the articles from PubMed, full text articles were downloaded from PubMed Central using the ID (PMCID) provided in the metadata. The resulting dataset consists of a total of 82 case reports. An overview of the number of case reports per rare disease can be seen in Table 3.4.

Table 3.4: Number of collected case reports grouped by rare disease

Rare Disease	**Number of Case Reports**
Alkaptonuria	21
Barth syndrome	12
Fibrodysplasia Ossificans Progressiva	19
Leptospirosis	18
Norrie Disease	12
Total	82

3.1.3.3 *Data preprocessing*

The downloaded articles were preprocessed to isolate the case representation in the main article text. Articles available on PubMed Central adhere to the Journal Article Tag Suite (JATS)[30] format. JATS is standardized by National Information Standards Organization (NISO) as "ANSI/NISO Z39.96-2024".[42] JATS provides a common XML format for exchange of (scientific) Journal Content. The standard defines XML elements and attributes.[30]

Before conversion, front and back matter of the articles was discarded. Front matter includes (meta-)information about the article such as its author, the title, the journal the article was published, its abstract, licensing information or date of publication. Back matter always includes the list of references used in the article text and

may contain additional sections complementary to the main text such as statements about potential conflicts of interest, ethic statements or other disclaimers. To ensure compatibility to the MACCROBAT dataset, any headings, tables or figures in the article text were also removed. Finally, any references inside the text were removed (identified by the `<xref>` tag).

Conversion to a simple text (`.txt`) format was performed using the document conversion software docling[3] which supports input files in the JATS format.

3.1.3.4 *Annotation*

Annotation rules considering abbreviations, acronyms and symptoms were adopted from those of the RareDis[40] corpus. Abbreviations and acronyms which are especially common for disease and disorder names were annotated as the respective entity type.[40] If both the full name and an acronym for some term were given, both were assigned the appropriate entity label individually. In cases where entities were mentioned together with their synonyms, all synonyms were labeled the entity type in question. No relations were annotated in the custom dataset, as they were not evaluated. A variety of annotation / labeling tools were considered, an example being BRAT which was used in the creation of mentioned related datasets[40][34]. Label studio was chosen as the annotation tool for the labeling process. Label studio supports the labeling in the named entity recognition task in question as well as a variety of labeling tasks in Natural Language Processing, Computer Vision, Speech Processing and other tasks.[64] Even though strict annotation guidelines were defined, final quality of the created dataset is difficult to assess. A common metric for quality is the Inter-Annotator Agreement (IAA) which commonly uses an F_1-Score or Cohen's kappa to provide a quantitative measure for agreement of annotators when labeling the same dataset (or part of a dataset). IAA can not be calculated in case of this work because annotation was performed by a single annotator (the author of the thesis). This can be considered a methodological flaw of this work which is further discussed in Section 5.2.

Automatic pre-labeling was performed with the biomedical-ner-all model (see Section 3.2.1). Pre-labeling describes the process of assigning labels for annotated datasets which can speed up annotation in cases with limited annotation resources (single annotator in this case). Precision, Recall and F_1-Score comparing the pre-labeling to the final human annotation were 0.669, 0.486 and 0.563 respectively

3.1.4 *Exploratory Analysis and comparison of evaluated datasets*

An exploratory analysis of all three datasets was performed to compare the datasets in terms of their size, structure and entity types. Table 3.5 shows the size and structure of the evaluated datasets.

Table 3.5: Size and structure of the evaluated datasets

Dataset	**Case Reports**	**Mean Sentences**	**Mean words**
MACCROBAT2020	200	22.70	413
MACCROBAT2020RD	18	24.83	450
Custom Dataset	82	28.96	533

Table 3.6 shows entity counts for the evaluated datasets.

Table 3.6: Size and structure of the evaluated datasets

Dataset	**Entities**	**Mean Entities**
MACCROBAT2020	25 144	125
MACCROBAT2020RD	2 444	135
Custom Dataset	6 m150	75

The MACCROBAT2020 dataset contains a total of 200 case reports with 25 144 entities. MACCROBAT2020RD contains 18 case reports with 2 444 entities. The custom dataset contains 82 case reports with a total of 6 150 entities. Average number of entities is significantly lower for the custom dataset(75) compared to the MACCROBAT2020 dataset (125) and MACCROBAT2020RD (135) which may be caused by differences in annotation strategy and behavior. While generally similar, the average number of sentences per case report is slightly higher for MACCROBAT2020RD (24.83) and even higher for the custom dataset (28.96) compared to the MACCROBAT2020 dataset (22.70). This translates to the average number of words per case report also being higher for MACCROBAT2020RD (450) and custom dataset (533) compared to the MACCROBAT2020 dataset (413).

The distribution of entity types in the datasets is shown in Figure 3.4. Entity type distribution are generally similar for all datasets,

with exceptions for individual entity types. This could again be caused by differences in annotation behavior.

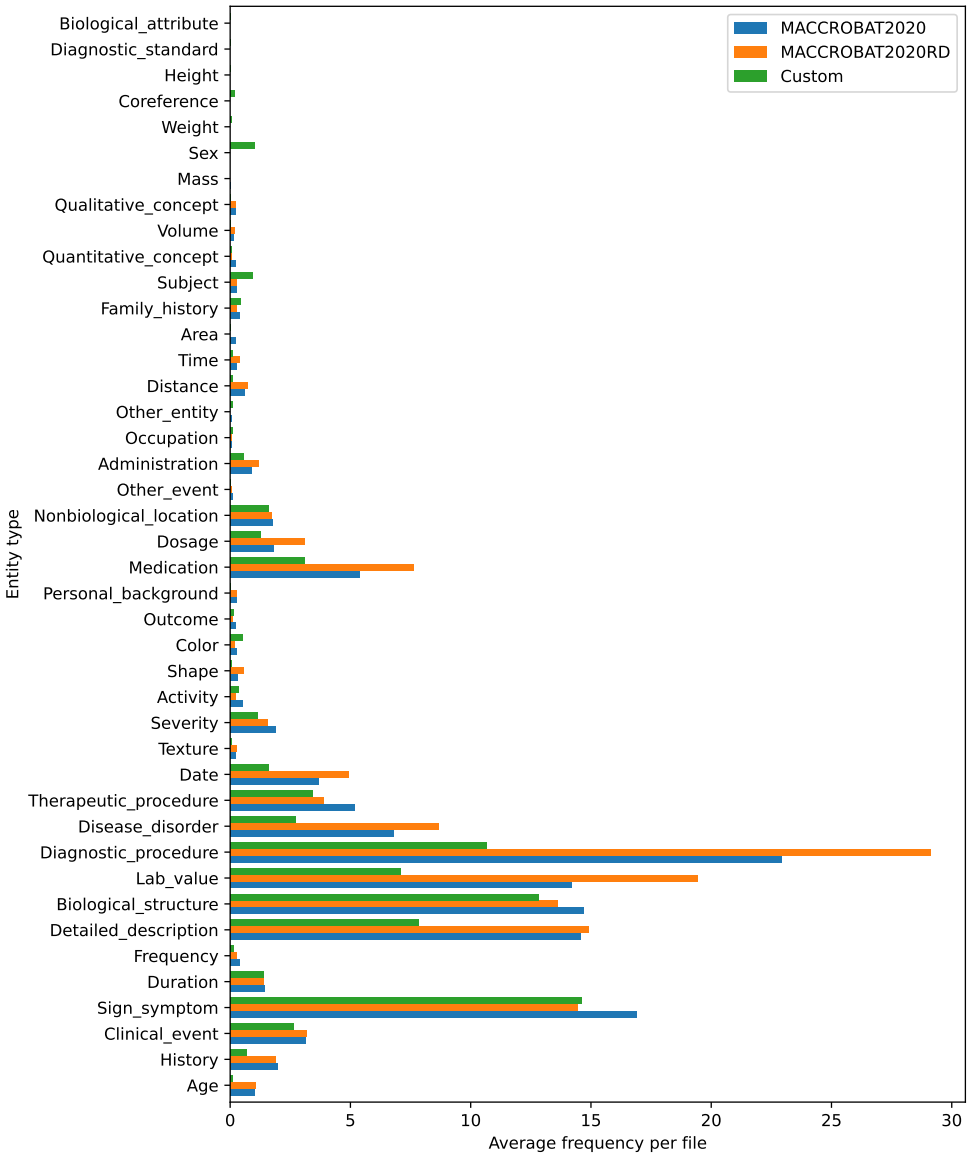


Figure 3.4: Number of entity by type relative to total number of cases in the three datasets.

3.2 SELECTED NER MODELS

For the evaluation of named entity recognition on the created dataset, a variety of models were selected. A table of all models used for evaluation can be found in Table 3.7.

Table 3.7: Model specifications for all evaluated models

Name	Variant	Architecture	Parameters
biomedical-ner-all[52]		DistilBERT with classification layer	66M
`GliNER`[75]	Small	custom architecture utilizing DeBERTa-v3	166M
"	Medium	"	209M
"	Large	"	459M
Gemma 3 [62]	12B	decoder-only transformer architecture	12B

3.2.1 *BERT-based baseline model (biomedical-ner-all)*

The model `biomedical-ner-all`[15] is a BERT-based model trained on the MACCROBAT dataset. It is based on the DistilBERT architecture which utilizes knowledge distillation to create a smaller and faster version of the original BERT model, while claiming to retain language understanding capabilities.[55] The authors modified the last layer of the DistilBERT model to a linear classification layer, specific to the entities of the MACCROBAT dataset. Classification scores are calculated for each token in the input text, making them available for evaluation. The model is fine-tuned using the MACCROBAT2020 dataset, however the exact training split used is not specified by the authors.[52] Authors of the model compared its performance on the MACCROBAT dataset (as well as two other datasets in the biomedical domain) to other models *BiLSTM-CNN-Char*, *SciBERT*, *BlueBERT*, *ClinicalBERT* and *BioBERT v1.2* and observed best results in their model. Although the MACCROBAT2020 dataset is also annotated with relationships between entities, the model does not support relationship extraction and only performs named entity recognition. The model will be used as a baseline for comparison with other models on the MACCROBAT dataset. Since the exact training split is not provided, evaluation results of the model should be considered with caution (see Section 5.2). It will also be used for pre-labeling of the custom dataset (see Section 3.1.3). The `biomedical-ner-all` model is publically available on *HuggingFace* and was used with the `transformers` library by Hugging Face.

3.2.2 *GliNER*

`GliNER`, a Generalist Model for Named Entity Recognition developed by **Zaratiana et al.** is a model for open-type NER. The authors propose this as a compact alternative to LLM based NER methods for resource-limited scenarios. `GliNER` employs a Bidirectional Transformer Language Model such as BERT or DeBERTa. Figure 3.5 depicts the architecture of the `GliNER` models. `GliNER` does not allow recognition of discontinuous entities.

The model is trained on the Pile-NER dataset (part of the UniversalNER work[77]) which incorporates texts from various domains with thousands of entity types. The dataset includes 45,889 input-output pairs with 240,725 entities of 13,020 distinct types. It should be noted that Pile-NER is a synthetic dataset created by prompting ChatGPT.[77]

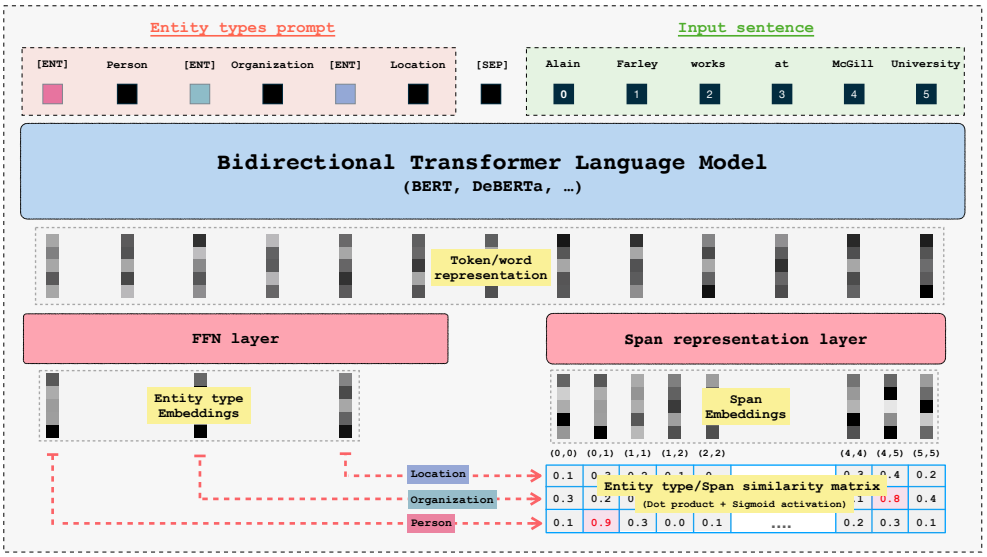


Figure 3.5: Architecture of the GliNER model[74]

For the three model variants small (166M parameters), medium(209M parameters) and large(459M parameters) were evaluated. The authors also support a multilingual model (gliner_multi, 209M parameters) variant based on mdeberta-v3-base which was not considered for this work. For all three variants version 2.1 (available on HuggingFace) were used.

The entity types that were part of evaluation (see Table 3.2) were provided in lower case as recommended by the authors.[75] `GliNER` does not support definitions of entity types as input. This may be a disadvantages if entity type definitions are highly specialized or not well known to the model. GoLLIE[54] a similar approach to GliNER

supports entity type definitions, by using annotation guidelines as input to the model.[54]

The authors evaluated their model on a variety of available NER-datasets including biomedical datasets. They find their to outperform large language models in zero shot performance, but acknowledge that the evaluation was performed on exact (strict) entity matching and that partial matches may be more common in competing models.[75]

`GliNER` supports a threshold parameter to filter out low-confidence predictions. For this work the default threshold of 0.5 was used, but experiments using the large variant with different thresholds will be performed to evaluate the impact of the threshold on the performance (Precision and Recall) of the model.

The `GliNER` models were used via the `gliner` library provided by the authors[74] which is also based on the `transformers` library.

3.2.3 *Chunking of input for models with constrained input dimensions*

As both the `biomedical-ner-all` and `GliNER` models have a maximum input dimension of 512 tokens, the input text needs to be chunked into smaller parts. For this purpose the input text (which is already split into sentences) is chunked into parts with a maximum of 512 tokens. The number of tokens were counted using the tokenizer of the respective model (`DistilBERTTokenizer` for `biomedical-ner-all` and `DeBERTaTokenizer` for `GliNER` models). Chunks always contained entire sentences to avoid splitting sentences which could lead to loss of context. The chunks are then processed by the model in sequence, with the output of the model being concatenated to form the final output. Different chunking strategies like semantic chunking[48] may provide additional context to the model, but were not evaluated in this work.

3.2.4 *LLM based NER (PromptNER)*

For the evaluation of named entity recognition using large language models, the PromptNER technique[2] was selected. PromptNER specifies a prompt template for named entity recognition tasks, which is then used to query a large language model. Entity types and their definitions are provided as part of the prompt, allowing the model to extract entities based on the definitions. Additionally, examples of entities and non-entities may be provided to the model. Finally the input text is provided to the model, which then outputs a list of entities with their type and reasoning for the type assignment.[2] A template

for the prompts used in the PromptNER technique is shown in Figure 3.6.

You are an expert Named Entity Recognition (NER) system.
Your task is to accept Text as input and extract named entities.
Entities must have one of the following labels: <LIST OF ENTITY TYPES>.
If a span is not an entity label it: '==NONE=='.

Extract named entities from the text.
Below are definitions of each label to help aid you in what kinds of named entities to extract for each label.
Assume these definitions are written by an expert and follow them closely.
<ENTITY TYPE DEFINITIONS>

Q: Given the paragraph below, identify a list of entities, and for each entry explain why it is or is not an entity:

Paragraph: <EXAMPLE TEXT>
Answer:
<EXAMPLES IN OUTPUT FORMAT>

Paragraph: <INPUT TEXT>

Answer:

Figure 3.6: Prompt template used for the PromptNER technique[2], entries in angle brackets are replaced with the respective values specified in the SpaCy configuration.

3.2.4.1 *Large Language Model selection*

For experiments using the PromptNER technique, the used model was *Gemma3*[62] developed by Google DeepMind. Gemma3 utilizes a decoder-only transformer architecture and is available in four different sizes: 1B, 4B, 12B and 27B parameters. In addition to text processing, the multimodal models can also process images, however this work only uses the text processing capabilities of the models. Training data for the Gemma3 models is not specified in detail, but the authors state that the models were trained on a large multilingual dataset including 144 languages. The model has a large context window of 128k tokens, allowing for processing of long texts. The tokenizer used is

SentencePiece with split digits, preserved whitespace and byte-level encodings.[62]

3.2.4.2 *Integration with LLMs*

The PromptNER technique was used via its implementation in the `spacy-llm` library[35]. SpaCy is given a configuration including the desired entity types, their definitions and (optionally) examples or negative examples. Texts used for entity type definitions and examples were taken from the MACCROBAT dataset (see Table 3.2 and Table 3.2). Experiments will be performed with and without examples, discussing the impact of examples on the performance of the model. It then creates a prompt including the input document and the provided configurations, following the prompt template defined by PromptNER[2]. Examples in the prompt are also used to describe the output format the model needs to adhere to. After the LLM has processed the prompt and created the output, Spa parses the output for all entities extracted by the model. While LLMs are generally capable of extracting discontinuous entities, SpaCy does not support parsing of discontinuous entities. Every named entity output also includes the models reasoning for the type assignment made.

For implementation of the PromptNER method, the library `spacy-llm` was used[19]. `spacy-llm` provides integration of large language models for the widely used `spacy`[26] NLP library that focuses on the usage of large language models for various subtasks of NLP. NER in `spacy-llm` in version 3 is based on the PromptNER paper [2] and implements the chain-of-thought prompting presented in the paper.[35] `spacy-llm` supports connections to various large language models. For this evaluation a custom connector to the Ollama software was implemented. Ollama is a software designed for locally running different large language models and supports the usage of GPU resources for inference.

3.3 EVALUATION METRICS AND EXPERIMENTAL SETUP

As described in Section 2.1.7, NER systems are commonly evaluated using the metrics of *Precision*, *Recall* and F_1-Score.[58] Evaluation in this work was performed using the python library `nervaluate`[44]. `nervaluate` implements the mentioned evaluation metrics according to the definitions of the SemEval-2013 task 9.1[58]. Evaluation commonly uses two different matching strategies for named entities: strict and loose (called *type* in `nervaluate`) matching. For a strict match,

both the span and the type of the gold standard entity and the predicted entity have to match exactly. Loose matching equates to the type matching evaluation behavior defined as part of the SemEval-2013 task 9.1[58]. For a loose match the type of the entity has to match and span has to show some overlap between the gold standard and the predicted entity. For this work, evaluation was performed with both the strict and loose entity matching strategy. When comparing ground truth and predicted entities, `nervaluate` differentiates five possible outcomes:

- Correct (*COR*): both are the same
- Incorrect (*INC*): the output of a system and the golden annotation do not match
- Partial (*PAR*): system and the golden annotation are somewhat "similar" but not the same
- Missing (*MIS*): a golden annotation is not captured by a system
- Spurious (*SPU*): system produces a response which does not exist in the golden annotation

[5]

On the basis of these outcomes, two more quantitative measures can be calculated:

$$POSSIBLE(POS) = COR + INC + PAR + MIS = TP + FN \quad (3.1)$$

$$ACTUAL(ACT) = COR + INC + PAR + SPU = TP + FP \quad (3.2)$$

These two measures are then used to calculate the evaluation metrics of Precision, Recall. Definition of precision and recall are dependent on the evaluation strategy used: Strict evaluation only considers exact matches (COR):

$$\text{Precision}_{\text{Strict}} = \frac{COR}{ACT} \quad (3.3)$$

$$\text{Precision}_{\text{Strict}} = \frac{COR}{POS} \quad (3.4)$$

LOOSE EVALUATION Loose evaluation additionally considers partial matches (PAR) as correct matches. A factor of 0.5 is used to weight

the partial matches, as they are not considered as correct as exact matches:

$$\text{Precision}_{\text{Loose}} = \frac{COR + 0.5 \times PAR}{ACT} \tag{3.5}$$

$$\text{Precision}_{\text{Loose}} = \frac{COR + 0.5 \times PAR}{POS} \tag{3.6}$$

Finally, the F_1-Score is calculated as the harmonic mean of Precision and Recall:

$$F_1\text{-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.7}$$

[5]

3.3.1 *General experimental setup*

For the main experiments, all models (see Section 3.2) were evaluated on the MACCROBAT dataset, the subset of the MACCROBAT dataset containing only rare diseases and the custom dataset. All entity types defined in the MACCROBAT dataset were used for evaluation. Evaluation was also performed on a per-entity-type basis, allowing for a more detailed analysis of the performance of the models on different entity types. Both strict and loose evaluation were performed, allowing for a comparison of the performance of the models on different evaluation strategies. For PromptNER in the main experiment no examples (negative or positive) were provided in the prompt. The models were evaluated on the entire dataset, with no further splitting into training and test sets.

3.3.2 *Further experiments*

To further evaluate the PromptNER method, additional experiments were performed. This includes variations to the number of entity types used, testing the system with all (42), a subset of five and single entity types. Additionally, the impact of providing examples in the prompt was evaluated. Providing examples transforms the task from a zero-shot to a few-shot learning task, which may improve the performance of the model.[33] For the `GliNER` models, the impact of the threshold parameter was evaluated by using different thresholds for the large model.

3.4 USED HARD- AND SOFTWARE

The experiments were performed on a general purpose computer with an `AMD Ryzen 9 7950X3D` CPU (16 Cores), 64 GB of RAM and an `AMD Radeon RX 7800 XT` GPU with 16 GB of VRAM. Only the experiments using the PromptNER method were performed using the GPU, all other models were run on the CPU. A more detailed overview of the used software including their versions can be found in the appendix (Table C.1).

3.5 CONCLUSION

In this chapter, the datasets used for evaluation were described, including the MACCROBAT dataset and a custom dataset created for this work. The MACCROBAT dataset is a comprehensive dataset for named entity recognition in the biomedical domain, comprised of case reports from PubMed. The custom dataset was created to provide a set of texts focused on rare diseases, while keeping structure and format similar to the MACCROBAT dataset. The chapter also described the models used for evaluation, including a BERT-based baseline model, the `GliNER` model and the PromptNER method using large language models. The evaluation metrics used for named entity recognition were described, including Precision, Recall and F_1-Score. The experimental setup for the evaluation was described, including the datasets, models and evaluation metrics used. The next chapter will present the results of the evaluation, including the performance of the models on the MACCROBAT dataset and the custom dataset.

4

RESULTS AND DISCUSSION

In this chapter, evaluation results of the system described in the previous chapter will be discussed. In addition to the results for the common NER evaluation metrics (Precision, Recall and F_1-Score), individual errors will be discussed. Implications for medical text processing will be discussed.

4.1 EVALUATION RESULTS

4.1.1 *Evaluation on the MACCROBAT2020 dataset*

Results of the evaluation of all models on the MACCROBAT2020 dataset are shown in Table 4.1 (strict) and Table 4.2 (loose). All evaluated models generally suffer from low recall (compared to their precision). Evaluation results show that the `biomedical-ner-all` model performs the best overall, with an F_1-Score of 0.427. It also has the highest precision and recall of all models evaluated at 0.516 and 0.365 respectively. It should however be noted that the model was trained on the MACCROBAT2020 dataset and results may be highly specific to the dataset. As exact training data is not specified, the models results should be considered with caution.

The results for the three `GliNER` variants show that larger model size generally leads to better results regarding recall and the F_1-Score. However, it should be noted that precision generally decreases with larger models for this experiment.

Table 4.1: 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

As expected, results using loose evaluation are generally better than strict evaluation results, with the `biomedical-ner-all` model achieving an F_1-Score of 0.578. Changes between strict and loose evaluation are similar for all models, indicating that no model is particularly sensitive to the evaluation method.

Table 4.2: Loose Evaluation results on the MACCROBAT2020 dataset

Model	**Precision**	**Recall**	**F_1-Score**
biomedical-ner-all	**0.698**	**0.494**	**0.578**
`GliNER` (small)	0.481	0.128	0.203
`GliNER` (medium)	0.480	0.192	0.274
`GliNER` (large)	0.477	0.231	0.311
PromptNER (Gemma3_12b)	0.412	0.250	0.311

4.1.2 *Evaluation on the rare disease subset*

Table 4.3: Strict Evaluation results on MACCROBAT2020RD

Model	**Precision**	**Recall**	**F_1-Score**
biomedical-ner-all	**0.485**	**0.352**	**0.408**
`GliNER` (small)	0.347	0.097	0.151
`GliNER` (medium)	0.343	0.143	0.202
`GliNER` (large)	0.306	0.154	0.205
PromptNER (Gemma3_12b)	0.247	0.120	0.162

Loose evaluation results are shown in Table 4.4. As with the full MACCROBAT2020 dataset, results using loose evaluation are generally better than strict evaluation results, but changes show no significant differences between the models.

Table 4.4: Loose Evaluation results on MACCROBAT2020RD

Model	**Precision**	**Recall**	F_1**-Score**
biomedical-ner-all	**0.669**	**0.486**	**0.563**
`GliNER` (small)	0.498	0.139	0.217
`GliNER` (medium)	0.496	0.207	0.292
`GliNER` (large)	0.479	0.241	0.321
PromptNER (Gemma3_12b)	0.385	0.187	0.252

4.1.3 *Evaluation on the custom dataset*

Evaluation results on the custom dataset are shown in Table 4.5 and Table 4.6.

Table 4.5: Strict Evaluation results on the custom dataset

Model	**Precision**	**Recall**	F_1**-Score**
biomedical-ner-all	**0.669**	**0.486**	**0.563**
`GliNER` (small)	0.498	0.139	0.217
`GliNER` (medium)	0.496	0.207	0.292
`GliNER` (large)	0.479	0.241	0.321
PromptNER (Gemma3_12b)	0.385	0.187	0.252

Table 4.6: Loose Evaluation results on the custom dataset

Model	**Precision**	**Recall**	F_1**-Score**
biomedical-ner-all	**0.838**	**0.599**	**0.699**
`GliNER` (small)	0.457	0.101	0.165
`GliNER` (medium)	0.481	0.121	0.194
`GliNER` (large)	0.503	0.192	0.278
PromptNER (Gemma3_12b)	0.397	0.354	0.374

Results are generally similar to those on the MACCROBAT2020 dataset, with the `biomedical-ner-all` model performing best overall. It should be noted that the `biomedical-ner-all` model performed

even better on the custom dataset than on the MACCROBAT dataset. This is likely due to the fact that the custom dataset was annotated with the `biomedical-ner-all` model, which may have led to a bias in the annotations.

The prompt-based NER method `PromptNER` performed slightly better in all metrics when evaluated on the custom dataset compared to the MACCROBAT2020 dataset.

4.1.4 *Results by entity type*

Results vary widely between different entity types. Per-entity type results are shown in Figure 4.1. Tables with detailed results for each entity type can be found in the appendix (Appendix D).

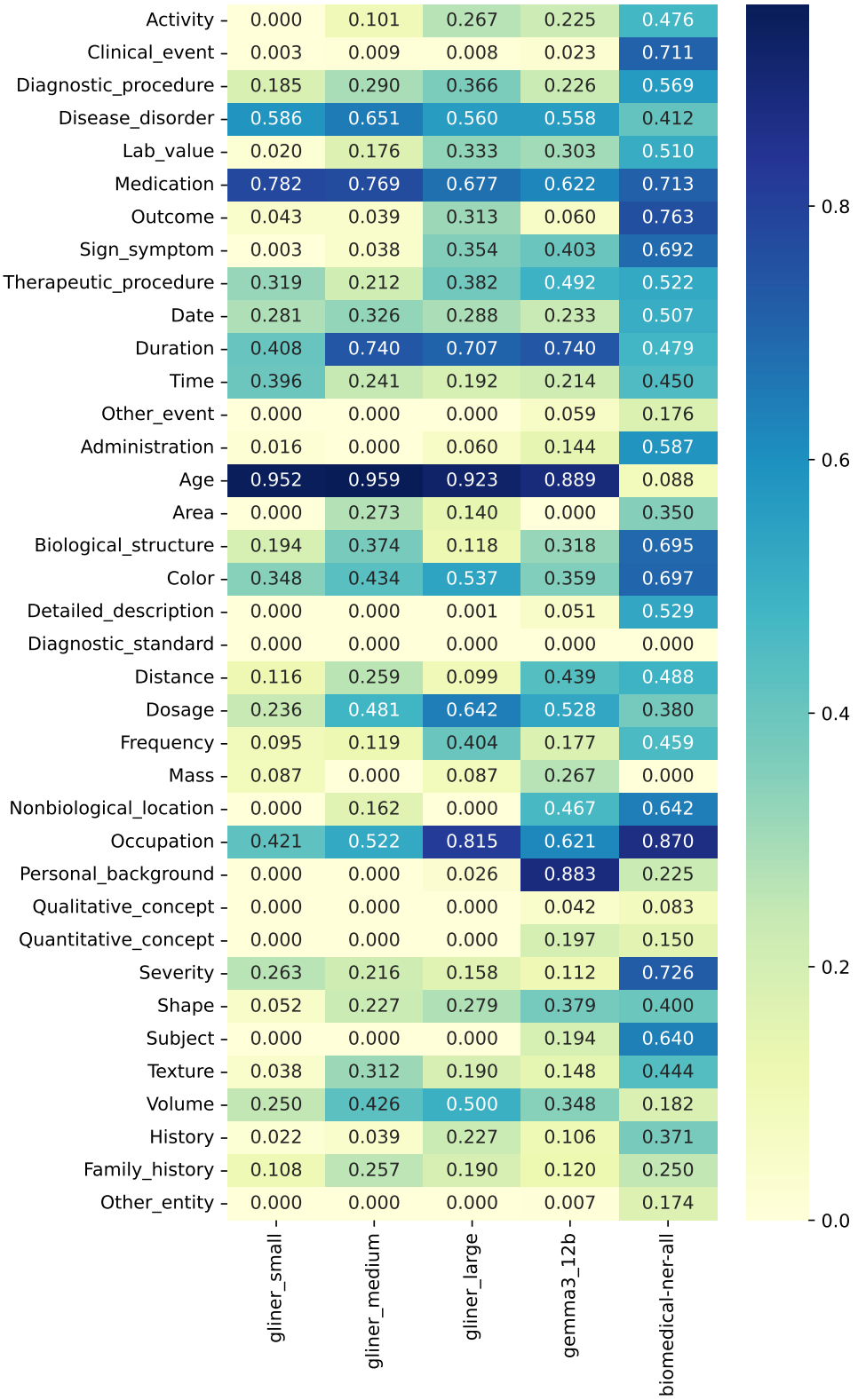


Figure 4.1: F_1-Score results (loose evaluation) for each entity type on the MACCROBAT2020 dataset.

For the entity type *Medication*, all models perform relatively well, which is likely due to the fact that medication names are often standardized and well-defined. It should also be noted that mentions of medications rarely include adjectives or other descriptive terms, which likely reduces the number of boundary errors (see Section 4.3.2) for this entity type.

4.1.5 *GliNER: Effect of threshold parameter on results*

The `GliNER` models were evaluated with different thresholds for the confidence score of the model. The results of the evaluation with different thresholds are shown in Figure 4.2.

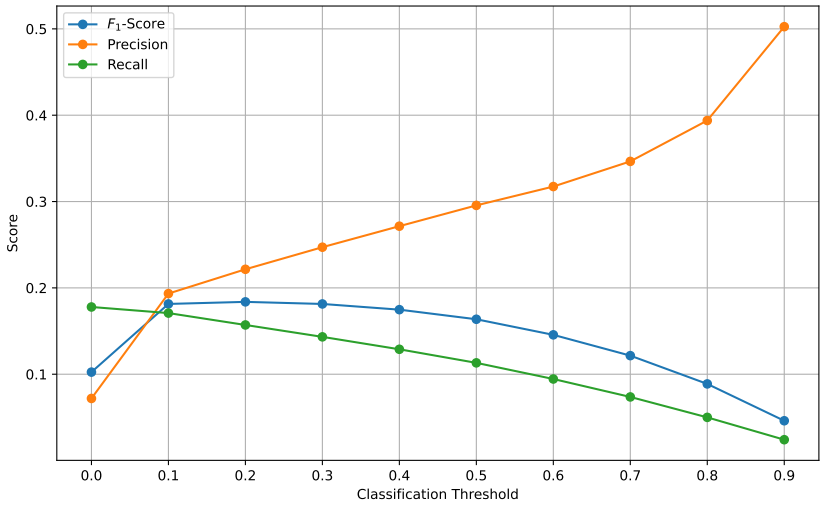
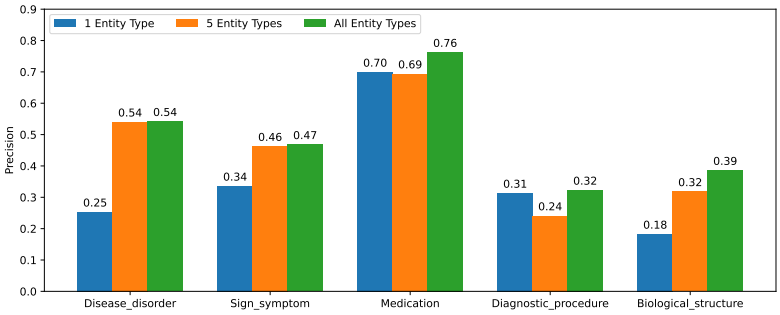


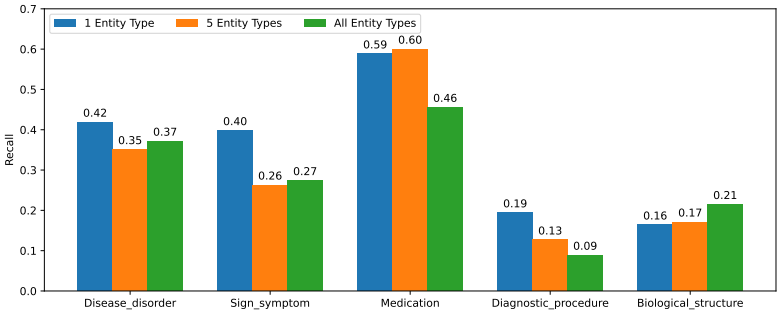
Figure 4.2: Evaluation results (F_1-Score, Precision and Recall) for the `GliNER` (large) model with different thresholds for the confidence score on the MACCROBAT2020 dataset.

As expected, the results show that the Precision increases with higher thresholds, while recall and F_1-Score decrease. Usage of `GliNER` with higher thresholds may be beneficial in applications where false positives are more problematic than false negatives.

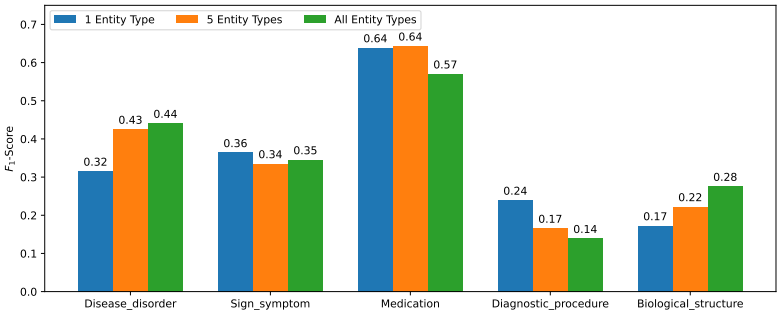
4.1.6 *Effect of number of entity types*



(a) Precision



(b) Recall



(c) F_1-Score

Figure 4.3: Evaluation results (Precision, Recall and F_1-Score) for the Prompt-NER (Gemma3_12b) model with different numbers of entity types on the MACCROBAT2020 dataset.

It was tested whether the number of entity types specified in the prompt has an effect on the results of the LLM-based NER methods. In addition to the full set of entity types(see Table 3.2 and Table 3.1), Gemma3 (12b) was also evaluated using a subset of five entity types (*Disease_disorder*, *Sign_Symptom*, *Medication*, *Therapeutic_procedure* and *Biological_structure*) and prompting for these entity types in isolation. The results of the evaluation with different numbers of entity types are shown in Figure 4.3. It should be noted that annotation of a test dataset with the used subsets of entity types was not performed, so

the results are not directly comparable to the full set of entity types. For that reason comparison is only performed on a per-entity-type basis.

These results suggest that using fewer entity types can lead to better recall and F_1-Score for the specified entity types, however precision may be reduced. Findings were not consistent across all five entity types, with some entity types showing better results with fewer entity types, while others showed better results with the full set of entity types. To reduce the negative impact on precision, it may be beneficial to provide more detailed descriptions of entity types in the prompt. It should also be noted that needed computing time was significantly reduced when using fewer entity types, most likely due to the reduced number of entities to be extracted per prompt. Using a subset of five entity types led to an average computing time of 40.64 seconds (compared to 59.24 seconds for the full set of entity types) for the `Gemma3` model. For single entity type extraction, average computing time was reduced even further to 28.43 seconds.

4.1.7 *Effect of examples in prompt*

It was expected that providing examples in the prompt would significantly improve the results of the LLM-based NER methods. In a setup with the subset of five entity types, the `Gemma3_12b` model was evaluated with and without examples in the prompt. Examples were chosen from texts outside the tested dataset to avoid biasing the model. However, the results show that providing examples in the prompt did not lead to a significant improvement in the results. While Precision improved by about 0.01 (from 0.270 to 0.281), Recall suffered (from 0.137 to 0.113) and F_1-Score sunk accordingly (from 0.198 to 0.174). As the experiment was only performed with a single set of examples, it is not possible to draw general conclusions from this result. Further experiments with different sets of examples and different entity types should be performed to determine whether providing examples in the prompt is beneficial for the LLM-based NER methods.

4.2 COMPARISON WITH BASELINE METHODS

As only the biomedical-ner-all model was previously evaluated on the MACCROBAT2020 dataset, it is not possible to compare the results of the other models to previous results. In this work the biomedical-ner-all model achieved an F_1-Score of 0.408 (using strict evaluation), which is significantly lower than the F_1-Scores of 0.919 reported by **Raza et al.**[52]. This difference is likely due to usage of different evaluation

methods, however evaluation is not described in detail by the authors which makes reproducing the results difficult.

For the custom dataset no previous results are available, so comparison is not possible.

Findings regarding prompt-based NER are consistent with those of **Liu et al.** [39], who found that prompt-based NER methods generally perform worse than fine-tuned models.

4.3 ERROR ANALYSIS

Liu et all [39] differentiate four types of errors:

- 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 (exclusively) end were incorrectly identified.
- False Negative: an entity was not identified at all

Relative parts of Wrong type, inaccurate boundary and false positives as part of total false positives are shown in Figure 4.4.

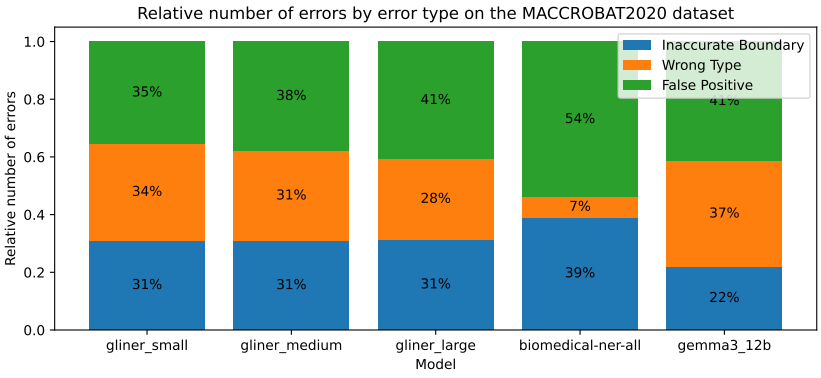


Figure 4.4: Relative number of false positive errors by error type for all evaluated models on the MACCROBAT2020 dataset

While not directly comparable because a different dataset was used, findings by **Liu et al.** showed that about one fifth to one third of errors were incorrect boundary errors when applying prompt-based NER on the RareDis dataset. For the `GliNER` models, the relative distribution of error types is similar, indicating that error type are related more to the type of model used rather than its size.

`biomedical-ner-all` shows particularly low number of type misclassification when compared to the other models. The PromptNER approach shows a high relative number of misclassifications, but fewer inaccurate boundary errors.

4.3.1 *Wrong Type*

A confusion matrix showing which entity types were misclassified as which other entity type is shown in Figure 4.5. Misclassifications are especially common for the entity types *Disease_disorder* and *Sign_Symptom*. An example of this error produced by the `GliNER` model is: *The patient was a 48-year-old man who had previously been hospitalized due to hemoptysis at 42 years of age.* Where the sentence contains the entity *hemoptysis*, for which ground-truth annotation is a symptom, but is classified as a disease by the model. [1]. Hemoptysis (the coughing up of blood) is caused by various diseases, but is usually not considered a disease itself. Apart from using more specific prompts, an additional step could be used to retrieve information about the entity type from an external knowledge base such as ICD, MeSh-Terms, Orphanet or NORD, before classifying the entity.

1 The annotation of the ground-truth aligns with the classification of hemoptysis using the ICD-11 system which classifies hemoptysis as a symptom (MD22; Symptoms, signs or clinical findings of the respiratory system: https://icd.who.int/browse/2025-01/mms/en#899998313)

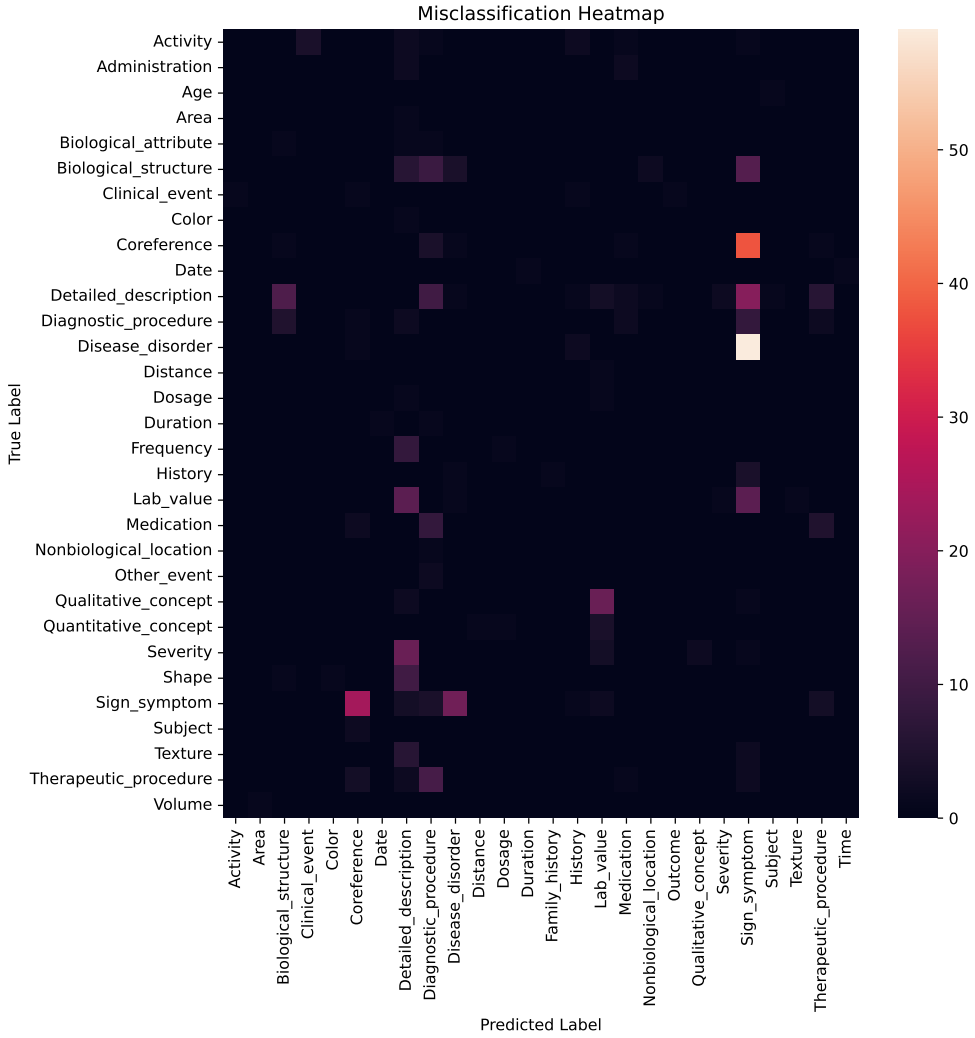


Figure 4.5: Heatmap of false entity type classifications for the `GliNER` models on the MACCROBAT2020 dataset

4.3.2 *Inaccurate boundaries*

Inaccurate boundaries are a common error type for all models. One common source for this error is the inclusion of descriptive adjectives when classifying entities. An example of this error would be the sentence: *Hysterectomy and bilateral salpingo-oophorectomy was performed for both treatment purposes and to fully categorise the malignancy to guide further adjuvant therapy.* Gemma3 (12b) correctly classifies *salpingo-oophorectomy* as a therapeutical procedure, but also includes the adjective *bilateral* in the entity, resulting in a boundary error. All evaluated models suffer from this error type. For prompt-based NER methods, providing more detailed instructions in the prompt, such as *only classify the main entity* or *do not include adjectives*, might help to reduce this error. In

general, adjusting the prompt to reflect the annotation guidelines (for evaluation) or desired output (for production use) can help to increase the quality of the output.

4.3.3 *Unparsable LLM output*

As mentioned in Section 3.2.4, the LLMs used in this work are not trained to output in a specific format. In evaluation this led to a number of errors where the output could not be parsed by SpaCy. An example for erroneous output would be:

```
February 2010 | Date | is a date
77-year-old | Age | is the patient's age
peripheral vascular disease | Disease_disorder | is a disease
```

The output does not conform to the expected format of a SpaCy NER model. The expected output format would require an additional boolean value, indicating whether the entity belongs to the stated type:

```
February 2010 | True | Date | is a date
77-year-old | True | Age | is the patient's age
peripheral vascular disease | True | Disease_disorder | is a disease
```

For the first experiment, where no few-shot examples were used the LLM produced erroneous output in a total of x cases.

A common error when performing prompt-based NER with LLMs was the generation of output in an invalid format that could not be parsed by SpaCy. This error occurred for 10 different texts or in 5% of cases for the Gemma3 (12b) model. The model was not re-prompted as part of the evaluation, so these cases were simply discarded. Methods for preventing this error are not provided directly by the authors of *PromptNER* but different prompt texts or additional components handling the output prior to parsing might prevent this error in some or all cases.

Another common error specific to LLMs was the classification of entities into types that were not specified in the prompt. SpaCy disregards these entities and only considers entities that were classified into one of the specified types. Manually parsing the output of the LLMs revealed that Gemma3 (12b) entity types were hallucinated in 23 of the 200 cases of the MACCROBAT2020 dataset. The total number of extracted named entities with hallucinated types was 55. Hallucinated entity types, included both generic entity types such as measurements(*Temperature* and *Pressure*) but also entity types specific to the biomedical domain such as *Gene*. Because invalid entity types are

not considered in the evaluation, these errors did not directly affect the evaluation results, they may however have led to lower recall for entity types similar to the hallucinated types.

Investigating the output of the LLMs revealed that Gemma3 (12b) often extracted discontinuous entities, SpaCy however does not support discontinuous entities in parsing. Modifying the underlying prompt to explicitly state that only continuous entities should be extracted could help to reduce this error. Extraction of discontinuous entities may however be useful in some cases, in which a different approach to parsing the output would be required.

Other common errors in parsing were those were the model performed a spelling correction on the entity text. For example *ataxia telengiectasia* was extracted as *ataxia telangiectasia* by Gemma3 (both spelling variations can be found in literature[12][8]). Consequently, the entity could then not be matched to the input text, leading to a missing entity in the evaluation. Such errors may also be avoided by adjusting the prompt or implementing a normalization step after extraction. Similarly, errors also occurred in some cases where the input text used special whitespace characters such as thin-spaces (Unicode `U+2009`) which were converted to regular spaces by the model. Performing additional normalization steps on the input text, such as removing special whitespace characters, could help to reduce these errors.

4.4 EVALUATION OF COMPUTING TIME AND RESOURCES

Needed computing time and resources (CPU, GPU, RAM and VRAM) vary widely between the used models. While processing time for the fastest model (biomedical-ner-all) was 60 ms, the slowest models (LLM based models) took over one minute (Average of 67 s) per case report. Processing time is highly dependent on the size of the model. For the `GliNER` models, the small variant took an average of 227 ms per case report while medium and large variants took 367 ms and 866 ms respectively. It should be noted that in the used experimental setups `GliNER` and biomedical-ner-all were run on the CPU while the LLMs were run on a GPU. Running `GliNER` and biomedical-ner-all on a GPU would likely lead to a significant speedup, but was not tested in this work. Differences in processing time should be considered for use cases where large datasets need to be processed or where processing happens on demand (e.g. in a web application). Models should be run on a GPU if available, as this can significantly reduce processing time. In the setup with Ollama, the 12B model Gemma3 used about 14 GB of the available 16 GB of VRAM. When using larger models, available

VRAM should therefore be considered. As described in Section 4.1.6 processing time for the PromptNER approach is also dependent on the number of entity types specified in the prompt.

4.5 IMPLICATIONS FOR MEDICAL TEXT PROCESSING

These results show that while NER models can be used to extract entities from medical case reports, the quality of the results is not sufficient for many applications. In the context of rare disease research, the usage of low-resource NER approaches such as the ones described in this work can help to extract relevant information from case reports, but the quality of the results is not sufficient for many applications. Quality of extractions may be improved by selecting specific entity types, such as *Medication* or *Disease_disorder*, which are more likely to be correctly identified by the models.

Prompt based NER approach allow for a large amount of flexibility in the entity types that can be extracted, but the quality of the results is highly dependent on the prompt and the model used. While using examples did not lead to significant improvements in this work, it may still be beneficial to provide examples in the prompt, especially for more complex entity types and in scenarios where precision is of high importance. When using or further developing the described system, computing resources and time should be considered, as the models used in this work require significant computing power and time to process the data. This is especially true if the system is to be applied to large datasets.

5 CONCLUSION AND FUTURE WORK

This chapter summarizes the findings of this thesis, discusses its limitations, and suggests directions for future research in the field of Named Entity Recognition (NER) applied to case reports on rare diseases. The goal is to provide a comprehensive overview of the contributions made in this thesis and to highlight areas where further research could be beneficial.

5.1 SUMMARY OF FINDINGS

Results show that the modern NER approaches considered in this work are generally capable of extracting relevant entities from case reports on rare diseases. However, 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`.

Experiments with different number of entity types show that the number of entity types can have a significant impact on the quality of extractions. Using a small number or single entity types can improve recall, but may lead to a drop in precision.

The described system for gathering, annotating and evaluating NER on case reports, especially in the context of rare disease, is capable of providing a basis for further research or practical application in clinical contexts.

5.2 LIMITATIONS

The main limitations of this work are related to the annotation of the custom test set that was used for evaluation. One limitation being the small size of the dataset, which was limited to 82 case reports. This small size limits the generalizability of the results and may not be representative of the broader population of case reports on rare diseases. Another limitation is that the annotation was not performed by a medical expert or rare disease expert, which may have led to inconsistencies or inaccuracies in the annotations. Additionally, the annotation was performed by a single annotator, which may have in-

troduced bias or subjectivity into the annotations. Since pre-labeling was done with the baseline model `biomedical-ner-all`, annotations may be biased towards the entities that the model is capable of extracting.

The lack of a custom fine-tuned approach (e.g. based on a BERT model) for the specific task of NER on case reports prevents comparison to a highly specialized model.

A major limitation in evaluation of the NER methods is the baseline model `biomedical-ner-all` used for comparison. Since exact training data is not specified and might contain data used in evaluation, the results of the model on the MACCROBAT2020 dataset should be considered with caution.

5.3 SUGGESTIONS FOR FUTURE RESEARCH

Expanding the annotated corpus with case reports for a variety of rare diseases could allow for evaluation with higher quality a dataset. Using the system described in this work could also be used, adapted and fine-tuned to specific rare diseases.

While English is the most common language for case reports, incorporating case reports in other languages could expand extracted data further. Since not all models used in this thesis are multilingual, usage of either language specific or multilingual models should be considered.

As part of the annotation process it deemed useful to expand the list of considered entity types. Genetic information (Genes, Phenotypical Abnormalities) often plays a part in causation and diagnosis of rare disease. Similarly, chemicals too can play a significant part in both causation and diagnosis and are not directly extracted as part of this work (only if the chemical in question would be considered medication).

Development of tooling utilizing the extracted data could provide useful to clinicians and researchers alike. Tooling could directly display extracted entities related to specific rare diseases or perform additional processing steps.

Mapping extracted entities to existing ontologies provided by NORD or OrphaData could help in normalizing entities to a common set of terms. This step could either be used after extraction, to provide a normalized set of entities, or as an error correction step to improve

the quality of the extracted entities. Other thesauruses and ontologies like Medical Subject Heading (MeSH) or the unified medical language system (UMLS) could be used for entities not covered by NORD and Orphanet which focus entirely on Diseases and Disorders.

In case reports, part of the information is transported in images such as imaging from diagnostic procedures like MRI or medical ultrasound. Extracting information from these images by means of specialized computer vision models or more generalized multi-modal models could provide additional useful information for research and clinical application.

Fine-tuning custom models on the corpora used in this work could improve the quality of the extractions and should be considered in future research.

Comparison and evaluation of LLM / prompt-based NER approaches can be extended by including different LLMs used for prompt-based approaches in the evaluation. This could include other freely or commercially available LLMs, such as OpenAI's GPT-4, Google's Gemini, or other models that are capable of performing NER tasks.

Implementing Relation Extraction could highlight relevant connections between entities, examples might be "[Symptom] is a manifestation of [Disease]" or "[Medication] treats [Disease]". The RareDis Corpus contains similar relation information, but additional data could be extracted with case reports as a basis of data.

Similarly, other methods of information answering especially question answering could provide additional use especially in the context of automated medical consultation for patients. Usage of full text of case reports for retrieval augmented generation might provide additional information to language models performing question answering.

5.4 CONCLUDING REMARKS

In conclusion, this thesis has explored the application of Named Entity Recognition (NER) methods to case reports on rare diseases, providing insights into their capabilities and limitations. The findings demonstrate the potential of modern NER approaches to extract relevant entities, while also highlighting areas for improvement, partic-

ularly in recall and precision. The limitations of this work, including the small size of the annotated dataset and the lack of expert annotation, underscore the need for further research to enhance the reliability and generalizability of the results. Expanding the dataset, incorporating additional entity types, and developing specialized models are promising directions for future work. By addressing these limitations and exploring the suggested avenues for research, the methodologies and systems described in this thesis could contribute to advancing the field of clinical information extraction, ultimately benefiting both researchers and clinicians working with rare diseases.

Part II

APPENDIX

A

APPENDIX: EXAMPLE OF ANNOTATED CASE REPORT

A 58-year-old [Age] man [Sex] had been suffering from general fatigue [Sign_symptom] and severe [Severity] anemia [Sign_symptom] for several months [Duration]. His hemoglobin levels [Diagnostic_procedure] were 6.6 g/dl [Lab_value] (normal range: 12–16 g/dl). He had no medical history [History] and did not take any medicine [History]. Esophagogastroduodenoscopy [Diagnostic_procedure] and colonoscopy [Diagnostic_procedure] did not reveal any significant bleeding [Sign_symptom]. Abdominal [Biological_structure] computer tomography [Diagnostic_procedure] revealed a 2-cm [Distance] hypervascular [Detailed_description] tumor [Sign_symptom] in the small intestine [Biological_structure] (Fig.1). Oral DBE [Diagnostic_procedure] detected a 2-cm-diameter [Distance] reddish [Color], submucosal [Biological_structure] tumor-like [Detailed_description] lesion [Sign_symptom] with surface ulceration [Detailed_description] in the jejunum [Biological_structure], approximately 20 cm away from the Treitz ligament [Biological_structure] (Fig.2). We did not perform biopsy [Diagnostic_procedure] because it can be difficult to stop bleeding in the case of hypervascular lesions. Under the diagnosis of a small bowel tumor [Disease_disorder], gastrointestinal stromal tumor [Disease_disorder] ([GIST [Disease_disorder]]), malignant lymphoma [Disease_disorder], or cancer [Disease_disorder], we performed laparoscopic-assisted [Detailed_description] segmental [Detailed_description] resection [Therapeutic_procedure] of the jejunum [Biological_structure] with the dissection of lymph nodes [Detailed_description]. Examination [Diagnostic_procedure] of the resected [Detailed_description] tumor [Sign_symptom] showed that it measured 19 × 16 mm [Area] in diameter (Fig.3). Histology [Diagnostic_procedure] revealed the proliferation of blood capillaries [Sign_symptom] and granulation tissue [Sign_symptom], which was consistent with PG [Disease_disorder] (Fig.4). The patient was discharged [Clinical_event] on postoperative day 9 [Date] without complication [Sign_symptom] and his anemia [Sign_symptom] improved [Lab_value] gradually without the need for oral iron [Medication] after surgery.

Figure A.1: Extraction of a case report (PMCID: 26444414) from the MACCROBAT dataset with its NER annotation.

B

APPENDIX: PROMPTNER EXAMPLE PROMPT

You are an expert Named Entity Recognition (NER) system.
Your task is to accept Text as input and extract named entities.
Entities must have one of the following labels: Activity, Administration, Age, Area, Biological_attribute, Biological_structure, Clinical_event, Color, Coreference, Date, Detailed_description, Diagnostic_procedure, Diagnostic_standard, Disease_disorder, Distance, Dosage, Duration, Family_history, Frequency, Height, History, Lab_value, Mass, Medication, Nonbiological_location, Occupation, Other_entity, Other_event, Outcome, Personal_background, Qualitative_concept, Quantitative_concept, Severity, Sex, Shape, Sign_symptom, Subject, Texture, Therapeutic_procedure, Time, Volume, Weight.
If a span is not an entity label it: '==NONE=='.

Extract named entities from the text.
Below are definitions of each label to help aid you in what kinds of named entities to extract for each label.
Assume these definitions are written by an expert and follow them closely.
Activity: Patient actions and habits.
Clinical_event: A clinical activity other than a medical procedure, often involving a change of Nonbiological location.
Diagnostic_procedure: Any procedure done primarily in order to obtain more information.
Disease_disorder: Any disease. Essentially a higher-level medical condition potentially including a collection of symptoms.
Lab_value: Any result of a laboratory test or a diagnostic result, including any units or values present.
Medication: Any pharmaceutical treatment. Used with Administration and Dosage entities.
Outcome: The patient's clinical outcome.
Sign_symptom: Any symptom or clinical finding.
Therapeutic_procedure: Any procedure done primarily in order to address or alleviate a symptom or disease. This includes surgery, long-term therapies, and supporting procedures (e.g., intubation).
Date: A time expression ending at a specific day in time.
Duration: A time expression describing a period of time, generally specifying an event has occurred continuously over the given duration.
Time: A time expression describing a specific point in time.
Other_event: Any event with clinical relevance that does not fit into any of the above categories.
Administration: Mode of administration of a drug or other therapy.
Age: Demographics. Patient age at time of presentation.
Area: Any area. Includes value and units.
Biological_structure: Any part of the body, from the cellular level to general areas.
Color: A color.
Detailed_description: Any detail of an event or other entity.
Diagnostic_standard: A single qualitative or qualitative standard used to make a diagnostic conclusion.
Distance: Length, width, height or other 1-dimensional attributes. Includes value and units.
Dosage: A complex numerical expression describing medication or therapy dosage. Minimally requires an amount and frequency, including values and units. May be expressed by weight.
Frequency: An expression describing how often an event occurs. For drug dosages, prefer the Dosage entity.
Mass: Any measurement of physical mass. Includes value and units.
Nonbiological_location: Any physical location other than those on or within a patient's body.
Occupation: Any description of a subject's daily activities.
Personal_background: Demographics. Any description of subject ethnicity or national background.
Qualitative_concept: A detail of an event or other entity describing it in general terms. A high-level category for which other labels may be appropriate.
Quantitative_concept: A numerical value. A high-level category for which other labels may be appropriate.
Severity: Degree of a disease or symptom's severity. Sex Demographics. Patient sex at time of presentation.
Shape: A shape.
Subject: Any individual related to or of medical relevance to the patient. Does not include clinical personnel.
Texture: A texture.
Volume: Any volume, including that of bodily fluids. Includes value and units.
History: Any description of subject medical history.
Family_history: Any description of a patient's family medical history.
Other_entity: Any event with clinical relevance that does not fit into any other type.
Biological_attribute: Any biological attribute
Coreference: Coreferences label words or phrases referring to previously defined events or entities
Height: Demographics. Patient height at time of presentation
Sex: Demographics. Patient sex at time of presentation
Weight: Demographics. Patient weight at time of presentation

Q: Given the paragraph below, identify a list of entities, and for each entry explain why it is or is not an entity:

Paragraph: The chest X-ray showed an enlarged heart.
Answer:
1. chest | True | Biological_structure | is a part of the human body
2. X-ray | True | Diagnostic_procedure | is a medical imaging technique
3. enlarged heart | True | Sign_symptom | is a condition that indicates potential heart disease or other health issues

Paragraph: A 58-year-old man had been suffering from general fatigue and severe anemia for several months.
His hemoglobin levels were 6.6 g/dl (normal range: 12–16 g/dl).
He had no medical history and did not take any medicine.
Esophagogastroduodenoscopy and colonoscopy did not reveal any significant bleeding.
Abdominal computer tomography revealed a 2-cm hypervascular tumor in the small intestine (Fig.1).
Oral DBE detected a 2-cm-diameter reddish, submucosal tumor-like lesion with surface ulceration in the jejunum, approximately 20 cm away from the Treitz ligament (Fig.2).
We did not perform biopsy because it can be difficult to stop bleeding in the case of hypervascular lesions.
Under the diagnosis of a small bowel tumor, gastrointestinal stromal tumor (GIST), malignant lymphoma, or cancer, we performed laparoscopic-assisted segmental resection of the jejunum with the dissection of lymph nodes.
Examination of the resected tumor showed that it measured 192009×200916 mm in diameter (Fig.3).
Histology revealed the proliferation of blood capillaries and granulation tissue, which was consistent with PG (Fig.4).
The patient was discharged on postoperative day 9 without complication and his anemia improved gradually without the need for oral iron after surgery.

Answer:

Figure B.1: Prompt used by the PromptNER model. The input text example (PMCID: 26444144) is taken from the MACCROBAT2020 dataset.

C

APPENDIX: USED SOFTWARE COMPONENTS

Table C.1: Software used for data processing, implementation and evaluation

Software	Version	Usage
python	3.12.9	Programming language used for implementation and evaluation
biopython	1.85	Querying PubMed and PubMed Central for case reports
docling	2.28.0	Document conversion from JATS to Markdown
spacy-llm	0.7.3	NER with LLMs using the PromptNER method
spacy	3.8.5	NER with LLMs using the PromptNER method
label-studio	1.17.0	Annotation of the custom dataset
ollama	0.7.0	Running LLMs locally
transformers	4.49.0	NER with Hugging Face models
`GliNER`	0.2.16	NER with the `GliNER` model
nervaluate	0.2.0	Evaluation of NER results

D

APPENDIX: EVALUATION RESULTS BY ENTITY TYPE

Table D.1: Precision by entity type for all evaluated models on the MACCROBAT2020 dataset.

Entity Type	GliNER (small)	GliNER (medium)	GliNER(large)	Gemma3 (12b)	biomedical-ner-all
Activity	0.000	0.500	0.643	0.282	0.587
Clinical_event	0.032	0.064	0.027	0.101	0.820
Diagnostic_procedure	0.522	0.510	0.506	0.308	0.742
Disease_disorder	0.798	0.800	0.741	0.691	0.536
Lab_value	0.180	0.566	0.607	0.451	0.669
Medication	0.853	0.815	0.747	0.798	0.806
Outcome	0.200	0.111	0.278	0.080	0.853
Sign_symptom	0.011	0.104	0.566	0.532	0.767
Therapeutic_procedure	0.643	0.359	0.617	0.696	0.633
Date	0.392	0.427	0.349	0.328	0.651
Duration	0.689	0.891	0.784	0.853	0.553
Time	0.486	0.345	0.200	0.200	0.720
Other_event	0.000	0.000	0.000	0.039	0.300
Administration	0.025	0.000	0.087	0.194	0.771
Age	0.935	0.952	0.940	0.898	0.151
Area	0.000	0.391	0.286	0.000	0.378
Biological_structure	0.502	0.582	0.237	0.411	0.766
Color	0.706	0.581	0.733	0.538	0.717
Detailed_description	0.000	0.000	0.002	0.073	0.653
Diagnostic_standard	0.000	0.000	0.000	0.000	0.000
Distance	0.500	0.760	0.368	0.485	0.584
Dosage	0.634	0.764	0.745	0.692	0.464
Frequency	0.172	0.240	0.333	0.183	0.544
Mass	0.048	0.000	0.048	0.154	0.000
Nonbiological_location	0.000	0.647	0.000	0.613	0.778
Occupation	0.667	0.600	0.786	0.529	1.000
Personal_background	0.000	0.000	0.053	0.925	0.615
Qualitative_concept	0.000	0.000	0.000	0.030	0.103
Quantitative_concept	0.000	0.000	0.000	0.143	0.188
Severity	0.577	0.575	0.259	0.203	0.804
Shape	0.143	0.400	0.522	0.545	0.442
Subject	0.000	0.000	0.000	0.144	0.711
Texture	0.143	0.556	0.353	0.171	0.455
Volume	0.714	0.714	0.684	0.333	0.364
History	0.038	0.066	0.287	0.129	0.420
Family_history	0.143	0.305	0.267	0.154	0.271
Other_entity	0.000	0.000	0.000	0.004	0.222

Table D.2: Recall by entity type for all evaluated models on the MACCROBAT2020 dataset.

Recall	GliNER (small)	GliNER (medium)	GliNER(large)	Gemma3 (12b)	biomedical-ner-all
Activity	0.000	0.056	0.168	0.187	0.400
Clinical_event	0.002	0.005	0.005	0.013	0.628
Diagnostic_procedure	0.112	0.202	0.286	0.178	0.462
Disease_disorder	0.464	0.548	0.450	0.468	0.335
Lab_value	0.011	0.104	0.230	0.228	0.412
Medication	0.722	0.728	0.619	0.510	0.639
Outcome	0.024	0.024	0.357	0.048	0.690
Sign_symptom	0.002	0.024	0.257	0.324	0.631
Therapeutic_procedure	0.212	0.150	0.277	0.380	0.445
Date	0.219	0.264	0.245	0.180	0.416
Duration	0.290	0.633	0.643	0.654	0.423
Time	0.333	0.185	0.185	0.229	0.327
Other_event	0.000	0.000	0.000	0.120	0.125
Administration	0.011	0.000	0.046	0.114	0.475
Age	0.971	0.966	0.908	0.880	0.062
Area	0.000	0.209	0.093	0.000	0.326
Biological_structure	0.121	0.275	0.079	0.260	0.635
Color	0.231	0.346	0.423	0.269	0.679
Detailed_description	0.000	0.000	0.001	0.039	0.444
Diagnostic_standard	0.000	0.000	0.000	0.000	0.000
Distance	0.066	0.156	0.057	0.402	0.419
Dosage	0.145	0.351	0.564	0.427	0.321
Frequency	0.066	0.079	0.513	0.171	0.397
Mass	0.500	0.000	0.500	1.000	0.000
Nonbiological_location	0.000	0.093	0.000	0.378	0.546
Occupation	0.308	0.462	0.846	0.750	0.769
Personal_background	0.000	0.000	0.018	0.845	0.138
Qualitative_concept	0.000	0.000	0.000	0.070	0.070
Quantitative_concept	0.000	0.000	0.000	0.319	0.125
Severity	0.170	0.133	0.114	0.077	0.661
Shape	0.032	0.159	0.190	0.290	0.365
Subject	0.000	0.000	0.000	0.296	0.582
Texture	0.022	0.217	0.130	0.130	0.435
Volume	0.152	0.303	0.394	0.364	0.121
History	0.015	0.028	0.188	0.090	0.333
Family_history	0.086	0.222	0.148	0.099	0.232
Other_entity	0.000	0.000	0.000	0.077	0.143

Table D.3: *F*1-Score by entity type for all evaluated models on the MACCROBAT2020 dataset.

Entity Type	GliNER (small)	GliNER (medium)	GliNER(large)	Gemma3 (12b)	biomedical-ner-all
Activity	0.000	0.101	0.267	0.225	0.476
Clinical_event	0.003	0.009	0.008	0.023	0.711
Diagnostic_procedure	0.185	0.290	0.366	0.226	0.569
Disease_disorder	0.586	0.651	0.560	0.558	0.412
Lab_value	0.020	0.176	0.333	0.303	0.510
Medication	0.782	0.769	0.677	0.622	0.713
Outcome	0.043	0.039	0.313	0.060	0.763
Sign_symptom	0.003	0.038	0.354	0.403	0.692
Therapeutic_procedure	0.319	0.212	0.382	0.492	0.522
Date	0.281	0.326	0.288	0.233	0.507
Duration	0.408	0.740	0.707	0.740	0.479
Time	0.396	0.241	0.192	0.214	0.450
Other_event	0.000	0.000	0.000	0.059	0.176
Administration	0.016	0.000	0.060	0.144	0.587
Age	0.952	0.959	0.923	0.889	0.088
Area	0.000	0.273	0.140	0.000	0.350
Biological_structure	0.194	0.374	0.118	0.318	0.695
Color	0.348	0.434	0.537	0.359	0.697
Detailed_description	0.000	0.000	0.001	0.051	0.529
Diagnostic_standard	0.000	0.000	0.000	0.000	0.000
Distance	0.116	0.259	0.099	0.439	0.488
Dosage	0.236	0.481	0.642	0.528	0.380
Frequency	0.095	0.119	0.404	0.177	0.459
Mass	0.087	0.000	0.087	0.267	0.000
Nonbiological_location	0.000	0.162	0.000	0.467	0.642
Occupation	0.421	0.522	0.815	0.621	0.870
Personal_background	0.000	0.000	0.026	0.883	0.225
Qualitative_concept	0.000	0.000	0.000	0.042	0.083
Quantitative_concept	0.000	0.000	0.000	0.197	0.150
Severity	0.263	0.216	0.158	0.112	0.726
Shape	0.052	0.227	0.279	0.379	0.400
Subject	0.000	0.000	0.000	0.194	0.640
Texture	0.038	0.312	0.190	0.148	0.444
Volume	0.250	0.426	0.500	0.348	0.182
History	0.022	0.039	0.227	0.106	0.371
Family_history	0.108	0.257	0.190	0.120	0.250
Other_entity	0.000	0.000	0.000	0.007	0.174

BIBLIOGRAPHY

[1] Jeffrey K Aronson. "Anecdotes as Evidence." In: *BMJ : British Medical Journal* 326.7403 (June 2003), p. 1346. ISSN: 0959-8138. DOI: 10.1136/bmj.326.7403.1346.

[2] Dhananjay Ashok and Zachary C. Lipton. *PromptNER: Prompting For Named Entity Recognition*. June 2023. DOI: 10.48550/arXiv.2305.15444. arXiv: 2305.15444 [cs].

[3] Christoph Auer et al. *Docling Technical Report*. Dec. 2024. DOI: 10.48550/arXiv.2408.09869. arXiv: 2408.09869 [cs].

[4] George L. Banay. "An Introduction to Medical Terminology I. Greek and Latin Derivations *." In: *Bulletin of the Medical Library Association* 36.1 (Jan. 1948), pp. 1–27. ISSN: 0025-7338.

[5] David S. Batista. *Named-Entity Evaluation Metrics Based on Entity-Level*. 2018. URL: https://davidsbatista.net/blog/2018/05/09/Named_Entity_Evaluation/.

[6] Sergei Bogdanov, Alexandre Constantin, Timothée Bernard, Benoit Crabbé, and Etienne Bernard. *NuNER: Entity Recognition Encoder Pre-training via LLM-Annotated Data*. Feb. 2024. DOI: 10.48550/arXiv.2402.15343. arXiv: 2402.15343 [cs].

[7] *Brat Rapid Annotation Tool*. https://brat.nlplab.org/index.html.

[8] Andrea L. Bredemeyer, Ching-Yu Huang, Laura M. Walker, Craig H. Bassing, and Barry P. Sleckman. "Aberrant V(D)J Joining in ATM-deficient Lymphocytes Is Dependent on Non-Homologous DNA End Joining." In: *Journal of immunology (Baltimore, Md. : 1950)* 181.4 (Aug. 2008), pp. 2620–2625. ISSN: 0022-1767. DOI: 10.4049/jimmunol.181.4.2620.

[9] Tom B. Brown et al. *Language Models Are Few-Shot Learners*. July 2020. DOI: 10.48550/arXiv.2005.14165. arXiv: 2005.14165 [cs].

[10] J. Harry Caufield. *MACCROBAT*. Jan. 2020. DOI: 10.6084/m9.figshare.9764942.v2. URL: https://figshare.com/articles/dataset/MACCROBAT2018/9764942.

[11] J. Harry Caufield, Yichao Zhou, Yunsheng Bai, David A. Liem, Anders O. Garlid, Kai-Wei Chang, Yizhou Sun, Peipei Ping, and Wei Wang. *A Comprehensive Typing System for Information Extraction from Clinical Narratives*. Oct. 2019. DOI: 10.1101/19009118.

[12] Luciana Chessa et al. "Intra-Erythrocyte Infusion of Dexamethasone Reduces Neurological Symptoms in Ataxia Teleangiectasia Patients: Results of a Phase 2 Trial." In: *Orphanet Journal of Rare Diseases* 9 (Jan. 2014), p. 5. ISSN: 1750-1172. DOI: 10.1186/1750-1172-9-5.

[13] Aakanksha Chowdhery et al. *PaLM: Scaling Language Modeling with Pathways*. Oct. 2022. DOI: 10.48550/arXiv.2204.02311. arXiv: 2204.02311 [cs].

[14] Ryan Cotterell and Kevin Duh. "Low-Resource Named Entity Recognition with Cross-lingual, Character-Level Neural Conditional Random Fields." In: ().

[15] *D4data/Biomedical-Ner-All · Hugging Face*. https://huggingface.co/d4data/biomedical-ner-all.

[16] *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)*. Apr. 1999.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. May 2019. DOI: 10.48550/arXiv.1810.04805. arXiv: 1810.04805 [cs].

[18] *Entrez Programming Utilities Help*. National Center for Biotechnology Information (US), 2010.

[19] *Explosion/Spacy-Llm*. Explosion. May 2025.

[20] Hermenegildo Fabregat, Lourdes Araujo, and Juan Martinez-Romo. "Deep Neural Models for Extracting Entities and Relationships in the New RDD Corpus Relating Disabilities and Rare Diseases." In: *Computer Methods and Programs in Biomedicine* 164 (Oct. 2018), pp. 121–129. ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2018.07.007.

[21] Phil Gage. *FEB94 A New Algorithm for Data Compression*. http://www.pennelynn.com/Documents/CUJ/HTML/94HTML/19940045.HT 1994.

[22] Joel J Gagnier, David Riley, Douglas G Altman, David Moher, Harold Sox, and Gunver Kienle. "The CARE Guidelines." In: *Deutsches Ärzteblatt International* 110.37 (Sept. 2013), pp. 603–608. ISSN: 1866-0452. DOI: 10.3238/arztebl.2013.0603.

[23] Aaron Grattafiori et al. *The Llama 3 Herd of Models*. Nov. 2024. DOI: 10.48550/arXiv.2407.21783. arXiv: 2407.21783 [cs].

[24] Bernal Jiménez Gutiérrez, Huan Sun, and Yu Su. *Biomedical Language Models Are Robust to Sub-optimal Tokenization*. July 2023. DOI: 10.48550/arXiv.2306.17649. arXiv: 2306.17649 [cs].

[25] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. *Distilling the Knowledge in a Neural Network*. Mar. 2015. DOI: 10.48550/arXiv.1503.02531. arXiv: 1503.02531 [stat].

[26] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. *spaCy: Industrial-strength Natural Language Processing in Python*. 2020. DOI: 10.5281/zenodo.1212303.

[27] Jiacheng Hu, Runyuan Bao, Yang Lin, Hanchao Zhang, and Yanlin Xiang. *Accurate Medical Named Entity Recognition Through Specialized NLP Models*. Dec. 2024. DOI: 10.48550/arXiv.2412.08255. arXiv: 2412.08255 [cs].

[28] Yan Hu et al. "Improving Large Language Models for Clinical Named Entity Recognition via Prompt Engineering." In: *Journal of the American Medical Informatics Association* 31.9 (Sept. 2024), pp. 1812–1820. ISSN: 1527-974X. DOI: 10.1093/jamia/ocad259.

[29] Myeong Jin, Choi Sang-Min, and Kim Gun-Woo. "COMCARE: A Collaborative Ensemble Framework for Context-Aware Medical Named Entity Recognition and Relation Extraction." In: *Electronics* 14.2 (2025), p. 328. DOI: 10.3390/electronics14020328.

[30] *Journal Article Tag Suite*. https://jats.nlm.nih.gov/.

[31] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd. 2025.

[32] William Z. Kariampuzha, Gioconda Alyea, Sue Qu, Jaleal Sanjak, Ewy Mathé, Eric Sid, Haley Chatelaine, Arjun Yadaw, Yanji Xu, and Qian Zhu. "Precision Information Extraction for Rare Disease Epidemiology at Scale." In: *Journal of Translational Medicine* 21 (Feb. 2023), p. 157. ISSN: 1479-5876. DOI: 10.1186/s12967-023-04011-y.

[33] Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. *Recent Advances in Named Entity Recognition: A Comprehensive Survey and Comparative Study*. Dec. 2024. DOI: 10.48550/arXiv.2401.10825. arXiv: 2401.10825 [cs].

[34] Srinivasa Rao Kundeti, J Vijayananda, Srikanth Mujjiga, and M Kalyan. "Clinical Named Entity Recognition: Challenges and Opportunities." In: *2016 IEEE International Conference on Big Data (Big Data)*. Dec. 2016, pp. 1937–1945. DOI: 10.1109/BigData.2016.7840814.

[35] *Large Language Models · spaCy API Documentation*. https://spacy.io/api/large-language-models#ner-v3.

[36] Dong-Ho Lee, Akshen Kadakia, Kangmin Tan, Mahak Agarwal, Xinyu Feng, Takashi Shibuya, Ryosuke Mitani, Toshiyuki Sekiya, Jay Pujara, and Xiang Ren. *Good Examples Make A Faster Learner: Simple Demonstration-based Learning for Low-resource NER*. Mar. 2022. DOI: 10.48550/arXiv.2110.08454. arXiv: 2110.08454 [cs].

[37] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. "A Survey on Deep Learning for Named Entity Recognition." In: *IEEE Transactions on Knowledge and Data Engineering* 34.1 (Jan. 2022), pp. 50–70. ISSN: 1558-2191. DOI: 10.1109/TKDE.2020.2981314.

[38] Rumeng Li, Xun Wang, and Hong Yu. "Two Directions for Clinical Data Generation with Large Language Models: Data-to-Label and Label-to-Data." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing* 2023 (Dec. 2023), pp. 7129–7143. DOI: 10.18653/v1/2023.findings-emnlp.474.

[39] Qiuhao Lu, Rui Li, Andrew Wen, Jinlian Wang, Liwei Wang, and Hongfang Liu. *Large Language Models Struggle in Token-Level Clinical Named Entity Recognition*. Aug. 2024. DOI: 10.48550/arXiv.2407.00731. arXiv: 2407.00731 [cs].

[40] Claudia Martínez-deMiguel, Isabel Segura-Bedmar, Esteban Chacón-Solano, and Sara Guerrero-Aspizua. "The RareDis corpus: A corpus annotated with rare diseases, their signs and symptoms." eng. In: *Journal of Biomedical Informatics* 125 (Jan. 2022), p. 103961. ISSN: 1532-0480. DOI: 10.1016/j.jbi.2021.103961.

[41] Masoud Monajatipoor, Jiaxin Yang, Joel Stremmel, Melika Emami, Fazlolah Mohaghegh, Mozhdeh Rouhsedaghat, and Kai-Wei Chang. *LLMs in Biomedicine: A Study on Clinical Named Entity Recognition*. July 2024. DOI: 10.48550/arXiv.2404.07376. arXiv: 2404.07376 [cs].

[42] National Information Standards Organization JATS Standing Committee. *ANSI/NISO Z39.96-2024, JATS: Journal Article Tag Suite, Version 1.4*. DOI: 10.3789/ansi.niso.z39.96-2024.

[43] *National Organization for Rare Disorders | NORD*. Mar. 2022.

[44] *Nervaluate: NER Evaluation Considering Partial Match Scoring*.

[45] 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." In: *European Journal of Human Genetics* 28.2 (Feb. 2020), pp. 165–173. ISSN: 1476-5438. DOI: 10.1038/s41431-019-0508-0.

[46] *Orphanet: Sarcoidosis*. URL: https://www.orpha.net/en/disease/detail/797?name=Sarkoidose&mode=name.

[47] *PubMed*. https://pubmed.ncbi.nlm.nih.gov/.

[48] Renyi Qu, Ruixuan Tu, and Forrest Bao. *Is Semantic Chunking Worth the Computational Cost?* Oct. 2024. DOI: 10.48550/arXiv.2410.13070. arXiv: 2410.13070 [cs].

[49] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. *Improving Language Understanding by Generative Pre-Training*. Tech. rep. OpenAI, June 2018.

[50] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. Sept. 2023. DOI: 10.48550/arXiv.1910.10683. arXiv: 1910.10683 [cs].

[51] *Rare Diseases - European Commission*. https://health.ec.europa.eu/rare-diseases-and-european-reference-networks/rare-diseases_en. Feb. 2025.

[52] Shaina Raza, Deepak John Reji, Femi Shajan, and Syed Raza Bashir. "Large-Scale Application of Named Entity Recognition to Biomedicine and Epidemiology." In: *PLOS Digital Health* 1.12 (Dec. 2022), e0000152. ISSN: 2767-3170. DOI: 10.1371/journal.pdig.0000152.

[53] S. Christy Rohani-Montez, Jennifer Bomberger, Cong Zhang, Jacob Cohen, Lucy McKay, and William R.H. Evans. "Educational Needs in Diagnosing Rare Diseases: A Multinational, Multispecialty Clinician Survey." In: *Genetics in Medicine Open* 1.1 (Apr. 2023), p. 100808. ISSN: 2949-7744. DOI: 10.1016/j.gimo.2023.100808.

[54] Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. *GoLLIE: Annotation Guidelines Improve Zero-Shot Information-Extraction*. Mar. 2024. DOI: 10.48550/arXiv.2310.03668. arXiv: 2310.03668 [cs].

[55] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. *DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter*. Mar. 2020. DOI: 10.48550/arXiv.1910.01108. arXiv: 1910.01108 [cs].

[56] Julia Schaefer, Moritz Lehne, Josef Schepers, Fabian Prasser, and Sylvia Thun. "The Use of Machine Learning in Rare Diseases: A Scoping Review." In: *Orphanet Journal of Rare Diseases* 15.1 (June 2020), p. 145. ISSN: 1750-1172. DOI: 10.1186/s13023-020-01424-6.

[57] Isabel Segura-Bedmar, David Camino-Perdonas, and Sara Guerrero-Aspizua. *Exploring Deep Learning Methods for Recognizing Rare Diseases and Their Clinical Manifestations from Texts*. Nov. 2021. DOI: 10.48550/arXiv.2109.00343. arXiv: 2109.00343 [cs].

[58] Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. "SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013)." In: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Ed. by Suresh Manandhar and Deniz Yuret. Atlanta, Georgia, USA: Association for Computational Linguistics, June 2013, pp. 341–350.

[59] Rico Sennrich, Barry Haddow, and Alexandra Birch. "Neural Machine Translation of Rare Words with Subword Units." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Katrin Erk and Noah A. Smith. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. DOI: 10.18653/v1/P16-1162.

[60] 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." en. In: *Journal of Healthcare Informatics Research* 8.2 (June 2024), pp. 438–461. ISSN: 2509-498X. DOI: 10.1007/s41666-023-00155-0. URL: https://doi.org/10.1007/s41666-023-00155-0.

[61] *Standoff Format - Brat Rapid Annotation Tool*. https://brat.nlplab.org/standoff.html.

[62] Gemma Team et al. *Gemma 3 Technical Report*. Mar. 2025. DOI: 10.48550/arXiv.2503.19786. arXiv: 2503.19786 [cs].

[63] Erik F. Tjong Kim Sang and Fien De Meulder. "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition." In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. 2003, pp. 142–147.

[64] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. *Label Studio: Data labeling software*. Comp. software. Open source software available from https://github.com/HumanSignal/label-studio. 2020-2025. URL: https://github.com/HumanSignal/label-studio.

[65] Richard Tzong-Han Tsai, Shih-Hung Wu, Wen-Chi Chou, Yu-Chun Lin, Ding He, Jieh Hsiang, Ting-Yi Sung, and Wen-Lian Hsu. "Various Criteria in the Evaluation of Biomedical Named Entity Recognition." In: *BMC Bioinformatics* 7 (Feb. 2006), p. 92. ISSN: 1471-2105. DOI: 10.1186/1471-2105-7-92.

[66] United States: National Archives and Records Administration: Office of the Federal Register. *An Act to Amend the Public Health Service Act to Establish an Office of Rare Diseases at the National Institutes of Health, and for Other Purposes.* Nov. 2002.

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. *Attention Is All You Need*. Aug. 2023. DOI: 10.48550/arXiv.1706.03762. arXiv: 1706.03762 [cs].

[68] Anna Visibelli, Bianca Roncaglia, Ottavia Spiga, and Annalisa Santucci. "The Impact of Artificial Intelligence in the Odyssey of Rare Diseases." In: *Biomedicines* 11.3 (Mar. 2023), p. 887. ISSN: 2227-9059. DOI: 10.3390/biomedicines11030887.

[69] Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. *GPT-NER: Named Entity Recognition via Large Language Models*. arXiv:2304.10428. Oct. 2023. DOI: 10.48550/arXiv.2304.10428. URL: http://arxiv.org/abs/2304.10428.

[70] Yucheng Wang, Bowen Yu, Hongsong Zhu, Tingwen Liu, Nan Yu, and Limin Sun. "Discontinuous Named Entity Recognition as Maximal Clique Discovery." 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). Online: Association for Computational Linguistics, 2021, pp. 764–774. DOI: 10.18653/v1/2021.acl-long.63.

[71] Thomas Wolf et al. *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. July 2020. DOI: 10.48550/arXiv.1910.03771. arXiv: 1910.03771 [cs].

[72] Jingye Yang, Cong Liu, Wendy Deng, Da Wu, Chunhua Weng, Yunyun Zhou, and Kai Wang. "Enhancing Phenotype Recognition in Clinical Notes Using Large Language Models: PhenoBCBERT and PhenoGPT." In: *Patterns* 5.1 (Jan. 2024), p. 100887. ISSN: 2666-3899. DOI: 10.1016/j.patter.2023.100887.

[73] Junjie Ye et al. "A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models." In: ().

[74] Urchade Zaratiana. *Urchade/GLiNER*. May 2025.

[75] Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. "GLiNER: Generalist Model for Named Entity Recognition Using Bidirectional Transformer." In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 5364–5376. DOI: 10.18653/v1/2024.naacl-long.300.

[76] Ming Zhou, Nan Duan, Shujie Liu, and Heung-Yeung Shum. "Progress in Neural NLP: Modeling, Learning, and Reasoning." In: *Engineering* 6.3 (Mar. 2020), pp. 275–290. ISSN: 2095-8099. DOI: 10.1016/j.eng.2019.12.014.

[77] Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. *UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition*. Jan. 2024. DOI: 10.48550/arXiv.2308.03279. arXiv: 2308.03279 [cs].