Building A German Clinical Named Entity Recognition System without In-domain Training Data

Siting Liang[1], **Hans-Jürgen Profitlich**[1], **Maximilian Klass**[2], **Niko Möller-Grell**[2], **Celine-Fabienne Bergmann**[2], **Simon Heim**[2], **Christian Niklas**[2], **Daniel Sonntag**[1,*]
[1]German Research Center for Artificial Intelligence, Germany
[2]Heidelberg University Hospital, Germany
[*]University of Oldenburg, Germany
siting.liang|hans-juergen.profitlich|daniel.sonntag@dfki.de
maximilian.klass|christian.niklas@med.uni-heidelberg.de

Abstract

Clinical Named Entity Recognition (NER) is essential for extracting important medical insights from clinical narratives. Given the challenges in obtaining expert training datasets for real-world clinical applications related to data protection regulations and the lack of standardised entity types, this work represents a collaborative initiative aimed at building a German clinical NER system with a focus on addressing these obstacles effectively. In response to the challenge of training data scarcity, we propose a **Conditional Relevance Learning (CRL)** approach in low-resource transfer learning scenarios. **CRL** effectively leverages a pre-trained language model and domain-specific open resources, enabling the acquisition of a robust base model tailored for clinical NER tasks, particularly in the face of changing label sets. This flexibility empowers the implementation of a **Multilayered Semantic Annotation (MSA)** schema in our NER system, capable of organizing a diverse array of entity types, thus significantly boosting the NER system's adaptability and utility across various clinical domains. In the case study, we demonstrate how our NER system can be applied to overcome resource constraints and comply with data privacy regulations. Lacking prior training on in-domain data, feedback from expert users in respective domains is essential in identifying areas for system refinement. Future work will focus on the integration of expert feedback to improve system performance in specific clinical contexts.

1 Introduction

Clinical Named Entity Recognition (NER) plays a central role in extracting valuable information from medical texts as essential features for developing clinical decision support systems. In this work, we concentrate on the German language and its application within clinics in Germany. Clinical documents can originate from a variety of sources. Each source has its unique characteristics, making it challenging to develop a one-size-fits-all NER system (Sonntag et al., 2016; Sonntag and Profitlich, 2019; Profitlich and Sonntag, 2021; Borchert et al., 2022; Roller et al., 2022). In developing German clinical NER systems, challenges arise when strict privacy rules are applied to data sources and the complexities associated with expert annotation of training data (Kittner et al., 2021; Roller et al., 2022). Previous related research efforts in German language have tackled these challenges using a range of techniques, from rule-based approaches to transfer learning methods in low resources scenarios (Frei and Kramer, 2021; Schäfer et al., 2022; Liang et al., 2023b,a). In this paper, we describe the development and assessment of an adaptive German clinical NER system without prior training on in-domain data. This goal is motivated by the principles of Interactive Machine Learning (IML) (Fails and Olsen Jr, 2003; Dudley and Kristensson, 2018), particularly when dealing with the challenges of annotating complex medical texts.

In this work, we investigate innovative transfer learning techniques and support the evolution of dynamic annotation schemas by engaging expert users from the medical field. We aim to address the constraints posed by limited data resources. The system has been developed as part of a cooperative project involving a machine learning lab and a university institute for medical informatics. Our system is equipped with a dedicated web-based User Interface (web-UI) for correcting system-generated annotations, which is instrumental in our case study involving **cardiology**. Qualified experts who are granted access to review the specific documents sourced from the hospital's internal database in medical informatics can utilize this tool to interact with system-generated outputs via a standalone website. The main objective of the case study is to conduct a comprehensive analysis of the performance of the NER system when applied to a non-

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 70–81

distributed dataset without violating privacy regulations. This analysis helps identify areas where adjustments are needed to refine the annotation schema and offers valuable insights to guide further fine-tuning of the model's performance to maintain its relevance to domain-specific nuances, context, and entity variations. Figure 1 displays our collaborative research environment. All system-related modules and model checkpoints are deployed at the hospital endpoint to ensure strict data security.

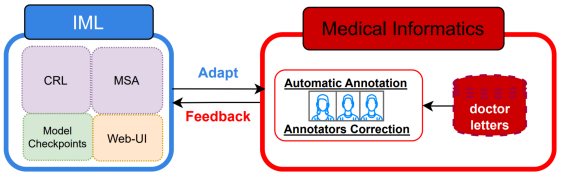

Figure 1: Overview of our collaborative research between experts in the field of interactive machine learning and medical informatics.

The main contributions of our work in the field of German clinical NER are as follows:

- Firstly, we leverage cross-domain transfer learning methods inspired by Liang et al. (2023a), using pre-trained language models and domain-relevant open-source Germain language datasets. We refer the approach to **Conditional Relevance Learning (CRL)** with an architecture that extends from a BERT-based encoder (Kenton and Toutanova, 2019) and incorporates a token-level binary classifier (see subsection 3.1). **CRL** has the potential to reduce the need of domain-specific training data compared to data-specific classifiers in low-resource scenarios. It also offers significant flexibility in adapting to changing label sets across different clinical domains.

- To enhance the adaptability and utility of our NER system across a range of medical texts (Widdows et al., 2002; Roller et al., 2022), we establish a comprehensive set of entity types categorised into six distinct semantic groups, drawing from the semantic ontology from the Unified Medical Language System (**UMLS**) Metathesaurus[1] (Bodenreider, 2004) and domain-specific annotations provided by Roller et al. (2022). This extensive annotation schema is referred to as **Multilayered Semantic Annotation (MSA)** (see subsection 3.3). This forms the basis for a dynamic and expandable semantic annotation schema as new clinical use cases emerge over time. As the NER system is deployed in specific clinical contexts, we can refine the annotation schema to align with the unique requirements of each use case it serves, ensuring the system's effectiveness and relevance across diverse clinical scenarios with minimal modification needed.

- Moreover, our **case study** plays a critical role in our broader efforts to improve the adaptability and utility of our clinical NER system. It offers essential insights into the performance of the NER system lacking in-domain training data and highlights areas where improvements are necessary.

Our collaborative effort is achieving NER system's adaptability across clinical domains and improving the robustness of the NER system's performance through the engagement of domain experts in applications. Ultimately, we aim to address the impact of strict privacy rules on data accessibility and annotation quality and contribute to research on the development of clinical information extraction systems in the German healthcare sector. Codes and demonstration are available in GitHub repository `https://github.com/sitingGZ/bert-sner-cardio`.

2 Related Work

Efforts to address the scarcity of in-domain training data in German NER training have led to the exploration of two main strategies. One strategy involves translating annotated English corpora, such as the n2c2 dataset (Henry et al., 2020) and DDI dataset (Segura-Bedmar et al., 2013), to synthesize domain-related German training datasets (Frei and Kramer, 2021; Schäfer et al., 2022). However, the accuracy and feasibility of the resulting NER models remain limited by nuances and contextual differences between languages and clinical domains, which affect their applicability to real German clinical texts. Another line of work involves the manual annotation efforts of curating generic medical datasets (Widdows et al., 2002; Borchert et al., 2022) by annotating on open-source corpora derived from medical journals. Furthermore, there are ongoing endeavours to develop domain-specific German clinical data sets that adhere to the English reference guidelines (Kittner et al., 2021; Richter-Pechanski et al.,

[1] `http://umls.nlm.nih.gov`

2023).

Task-specific datasets, like Roller et al. (2022) are rare but valuable for enhancing NER model performance by capturing domain-specific nuances. However, their adoption is hindered by the resource-intensive manual annotation workload. Llorca et al. (2023); Liang et al. (2023a) attempted to generalise various entity labels from different annotated German datasets to achieve a reasonable amount of cross-domain training data. While Llorca et al. (2023) introduced harmonised versions of German medical corpora through the BigBIO framework (Fries et al., 2022), contributing to the creation of common metadata sets for improved NER model performance across different medical text sources. However, the study found limited generalisation of NER models across different datasets. It remains challenging to effectively leverage diverse medical corpora for entity recognition in German texts. Liang et al. (2023a) augmented training data by mapping the original entity labels from source datasets to semantic types based on the ULMS ontologies and utilized a German BERT encoder with a binary token classifier to efficiently recognize medical entities by prompting with various semantic types followed by the same medical text. The novel training framework has shown effective cross-domain knowledge transfer and enhanced performance in low-resource German NER tasks.

While previous approaches offer potential solutions for the data scarcity issue, none have been able to develop a cross-domain adaptive NER system. Our work represents a step towards a more efficient and flexible long-term solution, as we combine the NER training framework from Liang et al. (2023a) with a multilayered semantic annotation schema, specifically targeting NER challenges within clinical information extraction and adapting to evolving clinical use cases.

3 Approach

3.1 Conditional Relevance Learning (CRL)

Developing in-domain training data from scratch requires substantial effort and resources for data collection and annotation, which is time-consuming and costly (Kittner et al., 2021; Roller et al., 2022; Richter-Pechanski et al., 2023). While transfer learning using pre-trained models can be beneficial, achieving acceptable performance still necessitates an effective transfer learning framework that leverages domain-specific fine-tuning techniques rather than relying solely on a large amount of labelled data in the source tasks (Llorca et al., 2023).

Liang et al. (2023a) shows that the performance of NER models firstly trained with domain-related corpus (Widdows et al., 2002) through a set of harmonized entity labels and novel training objective can be effectively generalised to clinical target tasks (Kittner et al., 2021; Roller et al., 2022) with much less fine-tuning data. While a harmonized label set proves highly advantageous in aggregating different relevant datasets to obtain a reasonable amount of training data, the limitation lies in the intricacies of the carefully designed matching process used to convert the entity labels from the target tasks to a unified label set. In this work, we only adopt the training framework from Liang et al. (2023a) which leverages pre-trained BERT-based encoder[2] and a token-level binary classifier on top of the BERT-based encoder predicts the contextual relevance score for individual tokens in a medical text input conditioned on the preceding label words. Table 1 presents two training examples of using the novel training objective from Liang et al. (2023a). The semantic type and the medical term phrases to be extracted are annotated as class 1. The remaining part of the input is marked as class 0. The training data also include negative samples, where no entity phrases can be extracted for the preceding semantic types.

Input	**Target**
[CLS] **Clinical Drug** [SEP] **Zofran** 4mg for nausea	[0, 1, 1, 0, 1, 0, 0, 0]
[CLS] Diagnostic Procedure [SEP] Zofran 4mg for nausea	[0, 0, 0, 0, 0, 0, 0, 0]

Table 1: Training example in line with the idea of conditional relevance learning. The model learns to recognize how different tokens in the input text should be associated with specific entity types.

Depart from the approach of Liang et al. (2023a), we preserve the original labels from the training sources, rather than their transformation into a unified label set. We aim to promote a more nuanced understanding of the diverse range of entities present in medical texts, ultimately improving the adaptability and effectiveness of the NER system. Figure 2 shows the format of the training data utilized in line with the approach of **CRL**. More information about the utilized datasets can be found

[2]https://www.deepset.ai/german-bert

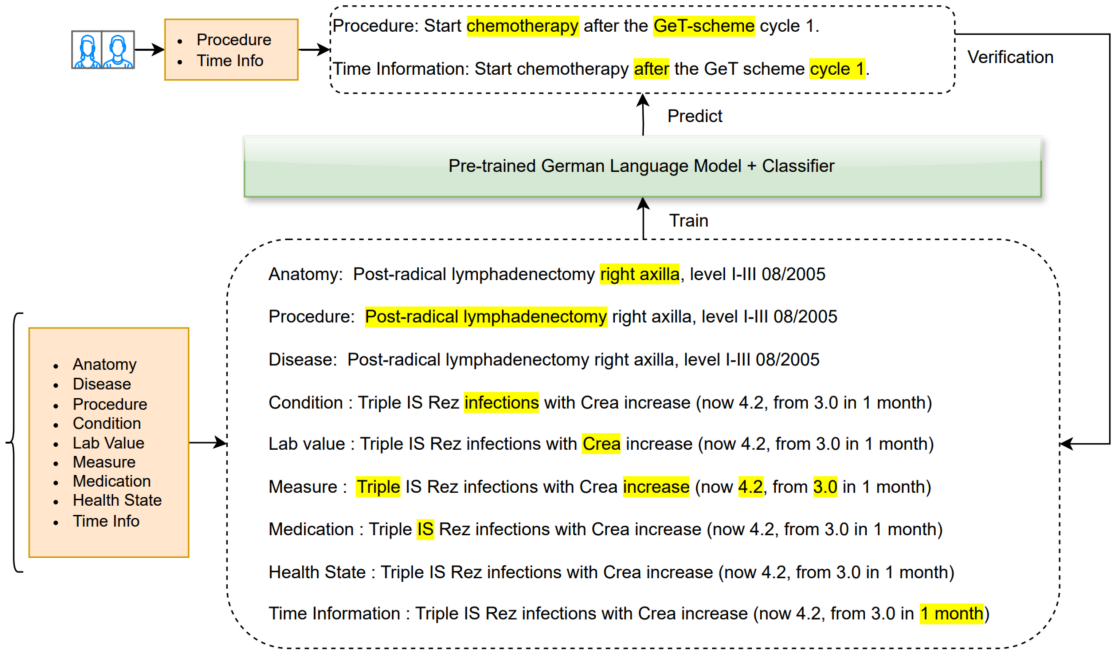

Figure 2: Our training examples for CRL approach. During inference, users can select relevant labels or introduce new ones. User feedback ensures system adaptability.

in Subsection 3.2.

CRL allows the NER system to adapt to new entity types from different label sets. During inference, users have the flexibility to choose from the entire entity label set seen during training, select only the relevant ones, or even introduce new, unseen semantic types as needed. Furthermore, user feedback regarding the applicability of the system to their specific use cases is essential. Following this adaptation, we apply the trained model on an unseen clinical dataset from the cardiovascular domain in our case study and make NER classification based on the **MSA** schema which is explained in Subsection 3.3.

3.2 Training Data

MUCHMORE[3] is a bilingual labelled corpus collecting English and German abstracts from 41 medical journals and containing the semantic annotations mapped to UMLS medical concepts (Widdows et al., 2002). **Ex4CDS**[4] is a corpus containing entity annotations related to the outcomes of kidney transplantation in nephrology clinics (Roller et al., 2022). Both are readily available open-source datasets. In this work, we train and adapt the NER system for clinical applications using these datasets with **CRL**. The entity types present in the training data from **MUCHMORE** and **Ex4CDS** are shown in Table 6 and Table 5 respectively (see Appendix A).

3.3 Multilayered Semantic Annotation (MSA)

During training, our goal is to encourage the model's comprehension of semantic diversity through a broad array of semantic types as entity labels. In practice, a medical phrase can represent an entity type of *Diagnostic Procedure* and contain the name of a *Medical Device* applied in the procedure. Multiple entity types can be predicted to the same text span, which is particularly valuable for nested and discontinuous NER tasks where entities are embedded within others (Yan et al., 2021). Hence, having a clear semantic annotation schema is beneficial for both application and performance analysis purposes.

The **UMLS** Metathesaurus integrating millions of medical concepts are widely applied knowledge sources for mining medical terms tasks (Aronson, 2001, 2006; Savova et al., 2010; Widdows et al., 2002; Borchert et al., 2022; Llorca et al., 2023). We design a **MSA** schema that aims to identify entities in multiple semantic dimensions based on **UMLS** ontology, thus the NER system can extract a broader spectrum of clinically relevant information, leading to the discovery of advanced medical knowledge. In addition to the incorporation of stan-

[3] https://muchmore.dfki.de/resources1.htm
[4] https://github.com/DFKI-NLP/Ex4CDS

dardised semantic types from **UMLS**, we add two semantic aspects, e.g. **Health State** and **Factuality** from Roller et al. (2022) in the **MSA** schema, illustrated in Table 2.

Semantic Group	Entity Types
Physical Object	Anatomical Structure, Clinical Drug, Medical Device
Conceptual Entity	Clinical Attribute, Quantitative Concept, Laboratory or Test Result, Temporal Concept
Procedure	Laboratory Procedure, Diagnostic Procedure, Therapeutic or Preventive Procedure
Phenomenon or Process	Injury or Poisoning, Disease Physiologic Function, Pathologic Function
Health State	Healthy Condition, Deteriorated Condition
Factuality	Negated, Minor, Speculated

Table 2: **MSA** encompasses six semantic groups. Each semantic group contains multiple entity types to facilitate fine-grained disambiguation. Most entity types are from the **UMLS** semantic types, except for groups **Health State** and **Factuality** (Roller et al., 2022).

3.4 Automatic Entity Annotation

CRL transforms the NER task into a token-level binary classification task, it predicts a relevance score for each token based on the preceding entity labels. We employ a threshold-based approach to transform the prediction scores made by the models into entity recognition results during the inference phase. Tokens with scores above the predefined threshold are considered part of entities and are assigned the corresponding entity type. In this process, we consider the variations in prediction scores, domain shifts, entity type unbalance, and the organization of entity types into semantic groups within **MSA**. They are critical aspects of ensuring the adaptability and effectiveness of the NER system in cross-domain transfer scenarios. As a result, the key steps include: **(1)** Prediction scores are generated for each token in the new dataset. **(2)** For each specific entity type, the maximum prediction score among the tokens associated with that type is identified, which serves as the maximum confidence for the entity type in the dataset. **(3)** The prediction scores for each token associated with an entity type are normalized by dividing the maximum confidence score to ensure a common range for comparison. **(4)** A set of thresholds is applied to determine the entity type assignment for the token. Tokens with normalized prediction scores above the assigned threshold are labelled with the corresponding entity type. **(5)** In cases where entity types within the same semantic group may be assigned to the same span, we assign priorities to entity types based on their normalized prediction scores to determine the most appropriate entity type for that span. Figure 3 displays the annotation results of the selected document based on the **MSA**.

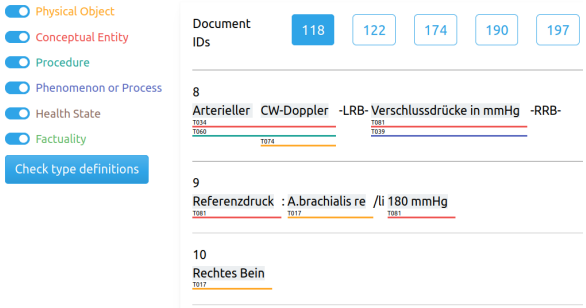

Figure 3: UI snippet of displaying MSA results for sentences of the selected document (above) and the selected semantic groups (left). The definitions of the entity types (coloured by groups) can be checked by clicking on the **[Check type definitions]** button.

CRL and **MSA** facilitate the integration of new semantic types and evolving clinical NER tasks without modification to the system architecture. However, building the NER system without in-domain training data remains challenging in understanding the domain specialities. Hence, we seek to evaluate the performance of our NER system through a detailed case study.

4 Case Study

In the absence of an annotated test dataset that closely resembles most real-world scenarios, evaluating the system's performance necessitates user evaluation. In our case study, the NER system is deployed on a German clinical corpus from the cardiovascular domain, where doctor letters undergo anonymization and time-shifting to comply with privacy regulations (Richter-Pechanski et al., 2023). Expert annotators can review and modify the system's output through the web-UI (see Figure 3 and 4), facilitating ongoing refinement based on valuable user feedback within the applied domain. This user-centric evaluation approach ensures that the system is continuously optimized to meet the specific needs and requirements of the clinical context, thereby enhancing its practical utility and effective-

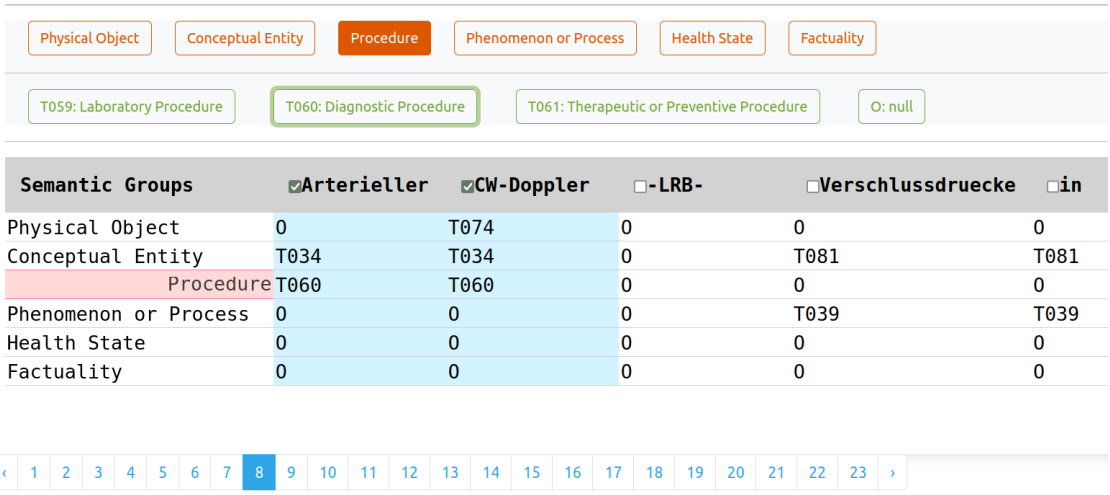

Figure 4: UI snippet of annotation revision. Annotators are requested to revise the system-generated annotations by semantic groups. Tokens from one sentence (the 8th sentence in this example) are placed in the header of the revision table, one token per column. The annotators can make changes on the entity types labelled to a single token (selecting one column) or to multiple tokens (selecting multiple columns) by clicking on the most appropriate entity type (*Diagnostic Procedure* in the second row) under the selected semantic group (*Procedure* in the example). Label *O: null* indicates that the token is not recognised as an entity according to the current annotation schema.

ness.

4.1 Evaluation Setup

In the human evaluation, we restrict the evaluation scope to the medical texts from two typical sections in the clinic routine, e.g. **Findings** and **Diagnosis**. They indicate different types of medical texts. Table 3 presents the scope of the evaluation. Table 4 shows the most frequent words in different sections of medical texts. These words represent the specialised content of each section. The comparison between these two types of medical texts is presented in each metric.

Three senior medical students, experienced in clinical annotation projects, are the expert annotators in our case study. The annotators work on the same documents. They are instructed to correct system-generated annotations through the standalone user interface shown in Figure 4. The system-generated annotations, which are utilized for revision, are generated using a threshold of 0.5. This threshold is set to avoid many false positives. Since no gold standard test dataset is established, these revisions serve as a form of ground truth to measure the NER system performance on the applied domain.

Target Data in Case Study. The most frequent words in texts from different clinic sections (**Findings** and **Diagnosis**) are listed in Table 4. They provide insights into the specialities of each

	#Docs	#Sents	#Words
Findings	8	136	1562
Diagnosis	11	155	1831

Table 3: Amount of data, of two different sections, at document-level (column 1), sentence-level (column 2) and word-level (column 3) respectively.

section. The words from **Findings** section are mostly related to the examination and lab results and those from **Diagnosis** section indicate the assessments and patient conditions.

Metrics. We measure the inter-rater reliability for each semantic group presented in Figure 5 (Cohen's Kappa[5] and Fleiss's Kappa[6]) to display the degree of agreement among multiple annotators. These scores play a crucial role in how to measure and assess the system's performance since we use the revisions of different annotators as a form of ground truth. In semantic groups with higher agreement, the evaluation metrics, such as Precision, Recall and F-scores presented in Figure 6, are more reliable indicators of the system's NER capabilities. Conversely, discrepancies in F-scores signal challenges in reaching a consensus among annotators of a given semantic group.

[5] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.

	top 20 frequent words
Findings	Kein (*no*), min, ms, QTc, QRS, Sinusrhythmus (*Sinus rhythm*), Ruhe-EKG (*ECG at rest*), Befund (*finding*) , PQ, Nachweis (*proof*), Normofrequenter (*normal frequent*), signifikanten (*significate*), R-Progression, Regelrechte (*regular*), Kammerendteilveraenderungen (*chamber end part changes*), Herzfrequenz (*heart rate*), Linkstyp (*left-side type*), rechts (*right*), regelmaessig (*regularly*), Beurteilung (*assessment*)
Diagnosis	Diagnosen (*diagnosis*), rechts (*right*), PTCA, links (*left*), Koronare (*Coronary*), Stenose (*stenosis*), RCA, Pumpfunktion (*pump function*), ED,TTE, LCX, 3-Gefaesserkrankung (*three-vessel disease*), DE-Stentimplantation (*drug eluting coronary implantation*), ohne (*without*), guter (*good*), linksventrikulaerer (*left ventricular*), Erfolgreiche (*successful*), Therapie (*therapy*), Vorhofflimmern (*Atrial fibrillation*), Rekanalisation (*Revascularization*)

Table 4: The 20 most frequent words (the *English translations*) in texts from **Findings** and **Diagnosis** sections across the applied doctor letters, excluding the stop words (articles, prepositions and numbers).

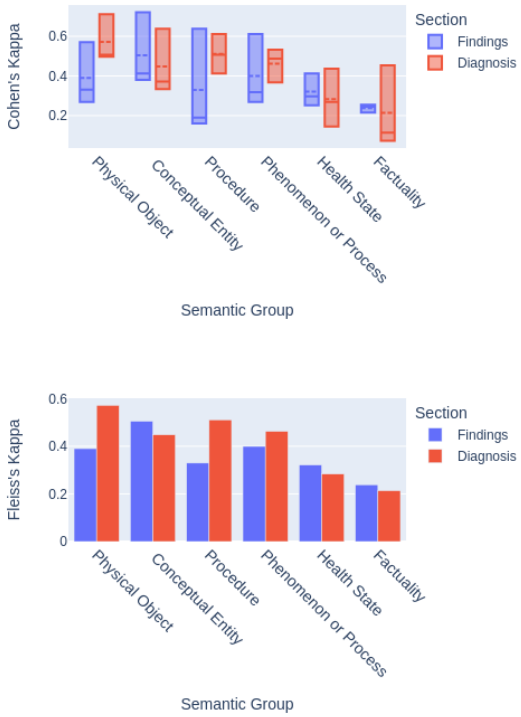

Figure 5: Inter-annotator agreement scores of pair-wise Cohen' Kappa and overall agreement of annotators based on multiple Fleiss's Kappa scores across different semantic groups and two types of medical texts.

4.2 Analysis

The metrics show that our NER system misses some entities but is relatively confident in the correctness of the labelled entities (higher precision and lower recall scores). Our NER system maintains a consistent performance, as indicated by the average F-score around 0.5 across a variety of the semantic groups with moderate agreement, e.g. **Physical Objects, Conceptual Entity, Procedure, Phenomenon or Process**. Compared to the results for section **Findings**, better system performance is observed for section **Diagnosis** based on the metrics in general. These results indicate that our NER system can provide a certain degree of reliable results in a zero-shot setting.

For the **Health State** group, higher F-scores suggest that the NER system effectively recognizes entities in this group, despite lower inter-annotator agreement scores. It suggests that annotators may agree less on which specific tokens represent the entities within this semantic group, but agree more on the broader context, leading to good performance in terms of F-scores.

The **Factuality** semantic group containing entities such as *Negated, Minor* and *Speculated*, presents a set of specific challenges. The user feedback highlights a notable ambiguity in the annotation process. While the system tends to miss many entities associated with *Minor* and *Speculated*, it is generally effective at capturing instances related to *Negated*. Furthermore, exemplified ambiguities rise in cases like *"no proof for a specific disease"* when deciding whether to annotate the terms *"no"*, *"no proof"* or the entire span with the type *Negated*. This highlights the need for further refinement and specificity in the annotation schema to ensure consistent and unambiguous annotations within the **Factuality** semantic group, as well as collecting

html#sklearn.metrics.cohen_kappa_score

[6] https://www.nltk.org/_modules/nltk/metrics/agreement.html

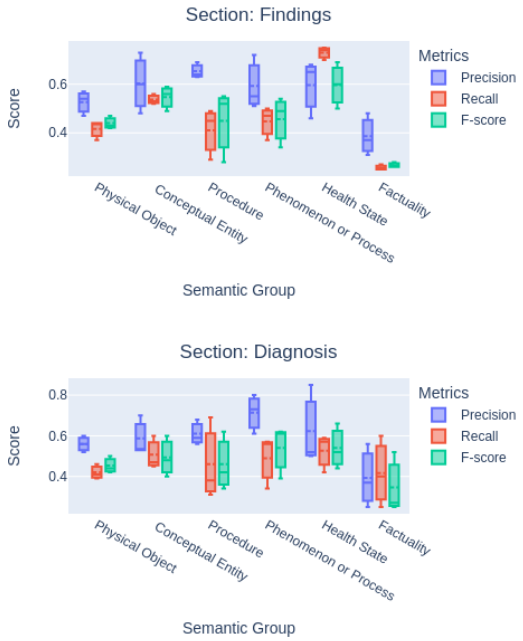

Figure 6: Precision, Recall and F-scores for NER results for medical text from **Findings** (up) and **Diagnosis** (down) section across different semantic groups.

more annotations to fine-tune the model's performance within this group.

4.3 System Usability Feedback

We administer a questionnaire to collect comprehensive feedback on the system's usability from the annotators regarding the system usability for future refinement. The answers indicate a System Usability Score (SUS) of 68, reflecting a moderately positive perception of the system's usability. Annotators generally found the annotation revision process straightforward and did not require technical support. However, a common concern was the time-consuming nature of reviewing annotations for every token within each possible semantic group. Annotators also noted that the NER system tends to overlook specific entities and there were some ambiguities related to specific entity types that require further clarifications in the annotation guidelines. These observations align with the evaluation metrics. This invaluable feedback suggests opportunities to enhance the efficiency of the NER system and improve the user experience in further use cases.

5 Conclusion

In summary, our work aims to overcome the challenges involved in developing a NER system for German clinical NLP applications without in-domain training data. We propose using advanced transfer learning methods and focusing on direct adaptability to new datasets with **CRL** and **MSA**. Our case study, which involves close collaboration with domain experts in specific clinical applications, yields invaluable insights that contribute to the overall improvement of the system. These insights are essential for tailoring the system to meet the specific information extraction requirements in target domains. Our work represents a significant advancement in clinical information extraction, alleviating limitations associated with data scarcity and cross-domain transferability. In future research, we will concentrate on incorporating expert feedback into the adaptation pipeline and models' fine-tuning, ultimately creating a continuous learning ecosystem tailored to the distinct clinical context. This effort aligns with our primary objective of advancing NER technology to effectively address challenges related to data scarcity, medical text diversity, and ever-changing label sets.

Ethical Statement

The annotators involved in the case study are co-authors of this paper and were not compensated for their research contributions.

Limitations

Importantly, our NER system's adaptability to new datasets to address data scarcity limitations is a key achievement of our research. However, the absence of interactive annotation rounds for resolving disagreements among annotators has prevented the creation of a more refined standard test dataset. In future endeavours, we plan to overcome these limitations by focusing on long-term data collection initiatives aimed at fine-tuning the system and adjusting model weights to better suit specific domains. Due to resource constraints, our ability to conduct comprehensive evaluations across a wider spectrum of clinical domains is restricted. Additionally, we recognize the necessity of exploring additional use cases to expand our understanding of medical text diversity, introduce new entity labels, and enhance the overall robustness of our cross-domain NER system. By addressing these limitations and pursuing a more extensive and diverse set of clinical data, we aim to further elevate the adaptability and utility of our NER system.

Acknowledgements

This research has been supported by the **pAItient** project (BMG, 2520DAT0P2), the **Ophthalmo-AI** project (BMBF, 16SV8639) and the Endowed Chair of Applied Artificial Intelligence, Oldenburg University. This work is further supported by **Accenture Labs** (research grant).

References

Alan R Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. page 17. American Medical Informatics Association.

Alan R Aronson. 2006. Metamap: Mapping text to the umls metathesaurus. *Bethesda, MD: NLM, NIH, DHHS*, 1:26.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Florian Borchert, Christina Lohr, Luise Modersohn, Jonas Witt, Thomas Langer, Markus Follmann, Matthias Gietzelt, Bert Arnrich, Udo Hahn, and Matthieu-P. Schapranow. 2022. GGPONC 2.0 - the German clinical guideline corpus for oncology: Curation workflow, annotation policy, baseline NER taggers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3650–3660, Marseille, France. European Language Resources Association.

John J Dudley and Per Ola Kristensson. 2018. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–37.

Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 39–45.

Johann Frei and Frank Kramer. 2021. Gernermed – an open german medical ner model.

Jason Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Sunny Kang, Rosaline Su, Wojciech Kusa, Samuel Cahyawijaya, et al. 2022. Bigbio: A framework for data-centric biomedical natural language processing. *Advances in Neural Information Processing Systems*, 35:25792–25806.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association: JAMIA*, 27(1):3.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Madeleine Kittner, Mario Lamping, Damian T Rieke, Julian Götze, Bariya Bajwa, Ivan Jelas, Gina Rüter, Hanjo Hautow, Mario Sänger, Maryam Habibi, et al. 2021. Annotation and initial evaluation of a large annotated german oncological corpus. *JAMIA open*, 4(2):ooab025.

Siting Liang, Mareike Hartmann, and Daniel Sonntag. 2023a. Cross-domain German medical named entity recognition using a pre-trained language model and unified medical semantic types. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 259–271, Toronto, Canada. Association for Computational Linguistics.

Siting Liang, Mareike Hartmann, and Daniel Sonntag. 2023b. Cross-lingual German Biomedical Information Extraction: from Zero-shot to Human-in-the-Loop. *arXiv e-prints*, pages arXiv–2301.

Ignacio Llorca, Florian Borchert, and Matthieu-P. Schapranow. 2023. A meta-dataset of German medical corpora: Harmonization of annotations and cross-corpus NER evaluation. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 171–181, Toronto, Canada. Association for Computational Linguistics.

Hans-Jürgen Profitlich and Daniel Sonntag. 2021. A case study on pros and cons of regular expression detection and dependency parsing for negation extraction from german medical documents. technical report. *CoRR*, abs/2105.09702.

Phillip Richter-Pechanski, Philipp Wiesenbach, Dominic M Schwab, Christina Kiriakou, Mingyang He, Michael M Allers, Anna S Tiefenbacher, Nicola Kunz, Anna Martynova, Noemie Spiller, et al. 2023. A distributable german clinical corpus containing cardiovascular clinical routine doctor's letters. *Scientific Data*, 10(1):207.

Roland Roller, Aljoscha Burchardt, Nils Feldhus, Laura Seiffe, Klemens Budde, Simon Ronicke, and Bilgin Osmanodja. 2022. An annotated corpus of textual explanations for clinical decision support. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2317–2326, Marseille, France. European Language Resources Association.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Henning Schäfer, Ahmad Idrissi-Yaghir, Peter Horn, and Christoph Friedrich. 2022. Cross-language transfer of high-quality annotations: Combining neural

machine translation with cross-linguistic span alignment to apply NER to clinical texts in a low-resource language. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 53–62, Seattle, WA. Association for Computational Linguistics.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. 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)*, pages 341–350.

Daniel Sonntag and Hans-Jürgen Profitlich. 2019. An architecture of open-source tools to combine textual information extraction, faceted search and information visualisation. *Artificial intelligence in medicine*, 93:13–28.

Daniel Sonntag, Volker Tresp, Sonja Zillner, Alexander Cavallaro, Matthias Hammon, André Reis, Peter A Fasching, Martin Sedlmayr, Thomas Ganslandt, Hans-Ulrich Prokosch, et al. 2016. The clinical data intelligence project. *Informatik-Spektrum*, 39(4):290–300.

Dominic Widdows, Beate Dorow, and Chiu-Ki Chan. 2002. Using parallel corpora to enrich multilingual lexical resources. pages 240–245.

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

A Entity Types in Training Data

Entity types derived from **Ex4CDS** are presented in Table 5. 134 semantic types from **UMLS** semantic network ontology in 2001 are annotated in **MUCHMORE** corpora. However, the number of annotations of each semantic type is extremely imbalanced ranging from less than 10 terms to at most 8202, see Table 6.

Entity Type	Description	Corresponding Type Names
Condition	A pathological medical condition of a patient can describe for instance a symptom or a disease.	Sign or Symptom; Disease or Syndrome; Finding
DiagLab	Particular diagnostic procedures have been carried out.	Laboratory Procedure; Diagnostic Procedure
LabValues	Mentions of lab values.	Clinical Attribute
Measure	Mostly numeric values, often in the context of medications or lab values, but can also be a description if a value changes, e.g. raises.	Quantitative Concept
Medication	A medication.	Pharmacologic Substance
Process	Describes particular process, such as blood pressure, or heart rate, often related to vital parameters.	Physiologic Function
TimeInfo	Describes temporal information, such as 2 weeks ago or January.	Temporal Concept
Health State*	A positive condition of the patient.	Healthy Condition
Factuality*	Factuality regarding symtoms and diseases (present or not, present but in a lower amount, kind of speculation).	Negated, Minor, Speculated

Table 5: Entity types, descriptions in Ex4CDS and the corresponding type names (* Type names are matched to the UMLS semantic types except for *HealthState, Factuality*, where no proper semantic type is found and retained the natural words of the entity types).

ID	Semantic Type	Description	Amount
T101	Patient or Disabled Group	An individual or individuals classified according to a disability, disease, condition or treatment.	8202
T047	Disease or Syndrome	A condition which alters or interferes with a normal process, state, or activity of an organism. It is usually characterized by the abnormal functioning of one or more of the host's systems, parts, or organs. Included here is a complex of symptoms descriptive of a disorder.	7636
T023	Body Part, Organ, or Organ Component	A collection of cells and tissues which are localized to a specific area or combine and carry out one or more specialized functions of an organism. This ranges from gross structures to small components of complex organs. These structures are relatively localized in comparison to tissues.	7070
T169	Functional Concept	A concept which is of interest because it pertains to the carrying out of a process or activity.	5569
T061	Therapeutic or Preventive Procedure	A procedure, method, or technique designed to prevent a disease or a disorder, or to improve physical function, or used in the process of treating a disease or injury.	5542
T046	Pathologic Function	A disordered process, activity, or state of the organism as a whole, of a body system or systems, or of multiple organs or tissues. Included here are normal responses to a negative stimulus as well as patholologic conditions or states that are less specific than a disease. Pathologic functions frequently have systemic effects.	3974
T191	Neoplastic Process	A new and abnormal growth of tissue in which the growth is uncontrolled and progressive. The growths may be malignant or benign.	3806
T170	Intellectual Product	A conceptual entity resulting from human endeavor. Concepts assigned to this type generally refer to information created by humans for some purpose.	3266
T081	Quantitative Concept	A concept which involves the dimensions, quantity or capacity of something using some unit of measure, or which involves the quantitative comparison of entities.	3049
T033	Finding	That which is discovered by direct observation or measurement of an organism attribute or condition, including the clinical history of the patient. The history of the presence of a disease is a 'Finding' and is distinguished from the disease itself.	2621
T060	Diagnostic Procedure	A procedure, method, or technique used to determine the nature or identity of a disease or disorder. This excludes procedures which are primarily carried out on specimens in a laboratory.	2621
T184	Sign or Symptom	An observable manifestation of a disease or condition based on clinical judgment, or a manifestation of a disease or condition which is experienced by the patient and reported as a subjective observation.	2547
T024	Tissue	An aggregation of similarly specialized cells and the associated intercellular substance. Tissues are relatively non-localized in comparison to body parts, organs or organ components.	2533
T121	Pharmacologic Substance	A substance used in the treatment or prevention of pathologic disorders. This includes substances that occur naturally in the body and are administered therapeutically.	2403
T037	Injury or Poisoning	A traumatic wound, injury, or poisoning caused by an external agent or force.	2080
T029	Body Location or Region	An area, subdivision, or region of the body demarcated for the purpose of topographical description.	1865
T040	Organism Function	A physiologic function of the organism as a whole, of multiple organ systems, or of multiple organs or tissues.	1540
T041	Mental Process	A physiologic function involving the mind or cognitive processing.	1429
T078	Idea or Concept	An abstract concept, such as a social, religious or philosophical concept.	1309
T032	Organism Attribute	A property of the organism or its major parts.	1281
T073	Manufactured Object	A physical object made by human beings.	1226
T091	Biomedical Occupation or Discipline	A vocation, academic discipline, or field of study related to biomedicine.	1213
T123	Biologically Active Substance	A generally endogenous substance produced or required by an organism, of primary interest because of its role in the biologic functioning of the organism that produces it.	1187
T100	Age Group	An individual or individuals classified according to their age.	1149
T062	Research Activity	An activity carried out as part of research or experimentation.	1148
T079	Temporal Concept	A concept which pertains to time or duration.	1124

Table 6: Most frequent **UMLS** semantic types annotated in the MUCHMORE corpus. The numbers in the third column are the amount of annotated terms appear in the training data.