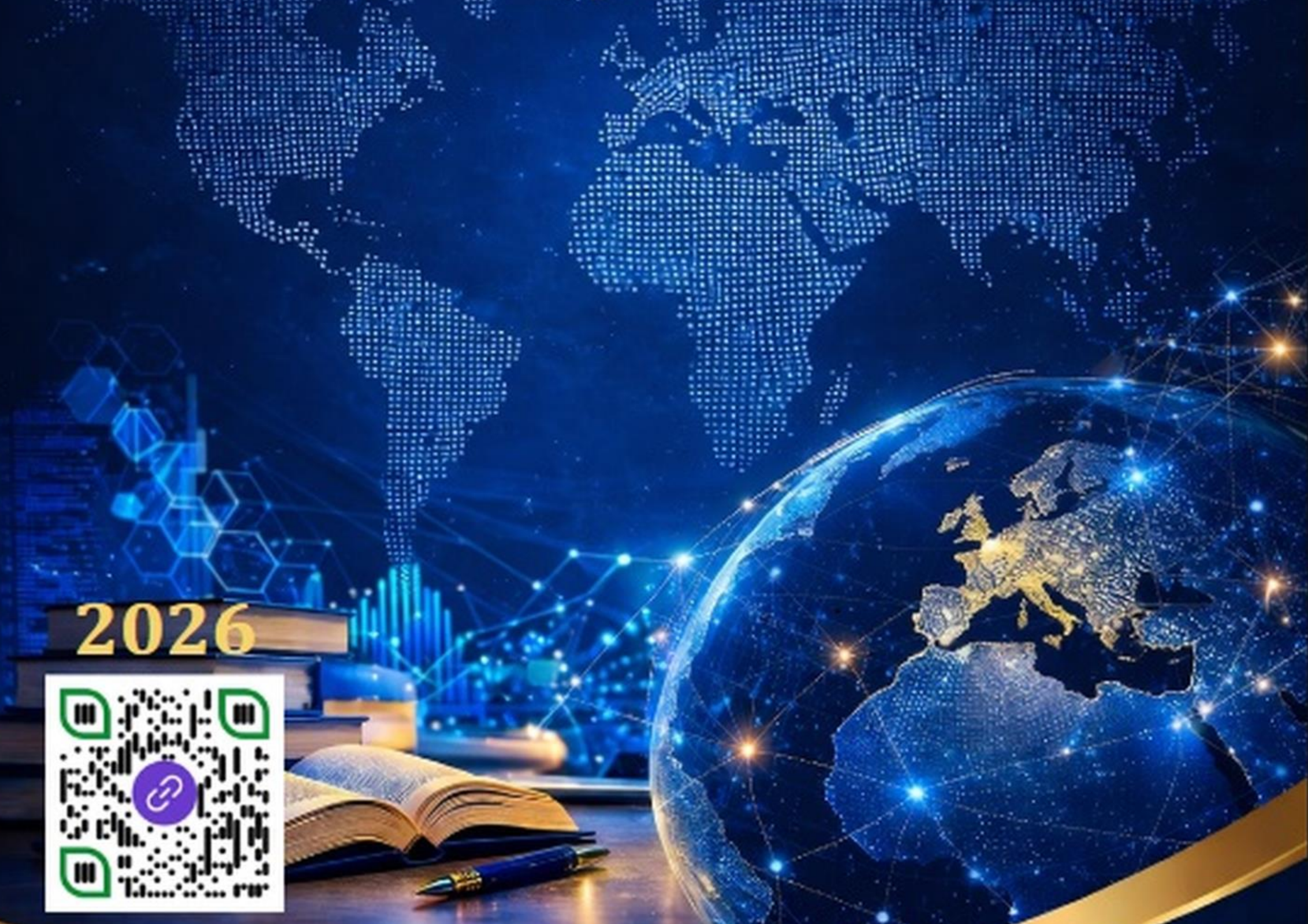
ISSN(E): 2770-8125

European International Journal of Advanced Research and Innovations

EIJARI



2026

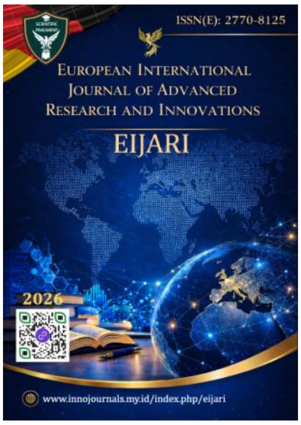
www.innojournals.my.id/index.php/eijari

EIJARI
ISSN: 2770-8125
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH AND INNOVATIONS
Volume 1, Issue 1. January - 2024

International Journal of Advanced Research and Innovations

Volume 1, Issue 1

EIJARI
ISSN: 2770-8125
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH AND INNOVATIONS
Volume 1, Issue 1. January - 2024



Information extraction from German radiological reports for general clinical text and language understanding

Michael Jantscher
Know-Center, 8010Graz, Austria
Felix Gunzer
Department of Radiology, Medical University Graz, 8036Graz, Austria

Recent advances in deep learning and natural language processing (NLP) have opened many new opportunities for automatic text understanding and text processing in the medical field. This is of great benefit as many clinical downstream tasks rely on information from unstructured clinical documents. However, for low-resource languages like German, the use of modern text processing applications that require a large amount of training data proves to be difficult, as only few data sets are available mainly due to legal restrictions. In this study, we present an information extraction framework that was initially pre-trained on real-world computed tomographic (CT) reports of head examinations, followed by domain adaptive fine-tuning on reports from different imaging examinations. We show that in the pre-training phase, the semantic and contextual meaning of one clinical reporting domain can be captured and effectively transferred to foreign clinical imaging examinations. Moreover, we introduce an active learning approach with an intrinsic strategic sampling method to generate highly informative training data with low human annotation cost. We see that the model performance can be significantly improved by an appropriate selection of the data to be annotated, without the need to train the model on a specific downstream task. With a general annotation scheme that can be used not only in the radiology field but also in a broader clinical setting, we contribute to a more consistent labeling and annotation process that also facilitates the verification and evaluation of language models in the German clinical setting.

With the wide application of artificial intelligence in medicine, there is an increasing need for the analysis of medical texts1–4 . Structured text data form the basis for many information retrieval use cases5–11 like Clinical Decision Support (CDS), diagnostic surveillance, cohort building for epidemiological studies, or query-based case retrieval. However, extracting structured and normalized information from clinical documents is a challenging task due to the lack of consistent language and standardized reports. Clinical documents, especially radiological reports, differ greatly in writing style from general medical documents such as scientific papers and articles. Due to time constraints, these

EIJARI
ISSN: 2770-8125
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH AND INNOVATIONS
Volume 1, Issue 1. January - 2024

documents/reports written by clinical staff are brief and concise and cover only important medical information (telegram style) with a subordinate focus on grammatical correctness. This leads to a pronounced divergence in semantics as well as syntax to common language.

[Editor1.1] Recently, deep language models, like bidirectional encoder representations from transformers (BERT)12, have shown an impressive performance boost for various NLP downstream tasks in the German clinical domain, such as (i) information extraction from radiological reports13–16, (ii) free-text report classification17,18 and (iii) oncology report summarization19. However, training large language models requires a significant amount of well-annotated training and testing data. Although hospitals already collect a vast amount of valuable digital free-text data (discharge reports, radiological reports, etc.) every day, they cannot be made accessible for external research due to privacy concerns and local legal restrictions. The situation is even worse for low-resource languages like German and hampers the development of modern healthcare applications in this field.

There are several initiatives for information extraction in German clinical documents derived from different medical fields. Roller et al.13 introduced a workbench for information extraction on German nephrology reports. Others focus on data from echocardiography reports 14, mental health records15 or self-generate synthetic clinical data16. However, these studies focus on a specific medical dataset, and may not allow validation of their approach in a broader clinical setting. In addition, there is a lack of a commonly agreed annotation, which makes comparison and validation with others difficult. Biased language models are one of the main resulting drawbacks

In this work, we contribute to an emerging field of research that emphasizes the role of information extraction techniques in the medical domain for low-resource languages like German. Our goal is to provide a universal German radiological language model that can be transferred to other clinical fields with minor adaptations (see Fig. 1). This allows other research teams to fine-tune the model on local datasets for specific clinical use cases without the need for expensive computing and human resources. These clinical applications cover (i) predictive clinical tasks, (ii) the generation of research cohorts, or (iii) the generation of image labels for upcoming AI-based medical imaging tasks.

In this paper, we pay attention to radiological reports due to their lower syntactic complexity compared to other clinical documents. Unlike clinical documents such as discharge letters or surgery reports, imaging reports focus on specific anatomical regions and consequently specific pathologies and observations. This reduces the amount of potential clinical and medical information. Therefore, radiological reports are a good starting point for the

EIJARI
ISSN: 2770-8125
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH AND INNOVATIONS
Volume 1, Issue 1. January - 2024

development of a German clinical language model. In order to train this language model, we introduce active learning along with a strategic sampling method to generate highly informative labeled training data (Fig. 1). This process of data labeling is particularly challenging in the clinical setting, as the complexity of clinical texts requires the involvement of medical experts in the labeling process. As a final step, we present a general annotation scheme for named entity recognition (NER) and relation extraction (RE) that not only considers the radiological domain but a broader clinical setting. We make the following contributions to the interdisciplinary field of natural language processing and clinical research:

- We provide a transformer-based language model that has been trained on German radiological reports for the tasks of (i) NER and (ii) RE. We emphasize, that our language model provides a significant boost in performance for medical predictive studies. To the best of our knowledge, this work is the first German radiological language model which can be used as a starting point for many clinical downstream tasks.
- By implementing active learning along with strategic sampling, we present an efficient method to generate consistently labeled training data with little annotation effort. We show that this method, in combination with a pre-trained language model, has great potential for general knowledge acquisition across different imaging domains
- We present a general annotation scheme that includes supervision from radiologists, incorporation of medical ontologies (RadLex, MesH) and previous work22,23 that can be used not only in the radiological field but also in a broader clinical setting. We therefore contribute to a more consistent labeling and annotation process, which is also beneficial in verifying and comparing modeling approaches in the German clinical domain.

Results. Experimental setup. For this study, we obtained radiological reports from three different imaging modalities from our institution. Table 1 gives an overview of the different training and test sets. Sentence count states the number of unique sentences per dataset. Based on the schemata from Tables 4, 5, clinical entities (Named Entity) and relationships (Relationship+ and Relationship−) between those entities were extracted from the reports, using an active learning approach. Relationship+ refers to actual relations of entities while entities that have no relationship with each other within a sentence (Relationship−) are automatically labeled as no relation. A pre-trained BERT language model is fine-tuned for (i) NER and (ii) RE. Following the masked language modeling method, 15% of the words in the CT Head (CTH) dataset were randomly masked and used for self-supervised pre-training. The pre-trained model serves as a basis for further fine-

EIJARI
ISSN: 2770-8125
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH AND INNOVATIONS
Volume 1, Issue 1. January - 2024

tuning in the following experiments. A more detailed description of the active learning procedure can be found in the "Material and Methods" section.

Results for CT Head reports (source domain). In this first experiment, the language model is fine-tuned on labeled data derived from the CTH dataset. Initially, 268 randomly selected sentences were annotated by two clinical experts. During the first seven learning cycles, 100 additional randomly selected samples were chosen from the pool of unlabeled sentences per iteration and proposed to the clinical experts for annotation.[Rev1.1] During the random sampling iterations, a subset of 20% of the sentences per iteration was annotated by both clinical experts. In addition, 10% of all sentences were annotated by both clinical experts during the strategic sampling iterations. Based on these expert reviews, an interannotator agreement was estimated separately for both the random sampling phase and the strategic sampling phase. The Cohen's kappa values of 0.74 (random sampling) and 0.83 (strategic sampling) show a substantial as well as a near perfect agreement between the annotators. Inspired by Ramponi and Plank24 as well as Salhofer et al.25 we intervene at iteration 8 with a strategic sampling approach that favors longer sentences/samples with a higher perplexity score (Eq. 1). In order to ignore outliers but still select highly informative training samples, we select samples (sentences) in the upper 75% to 90% percentile of the perplexity scores and sentence lengths. Examples of sentences with different lengths and perplexity scores are stated in Table 2. It can be observed that sentences containing medical entities such as diseases that are more common in reports (e.g., regular sinus ventilation) lead to low perplexity values. Sentences containing rare diseases (or medical conditions), like "calcified bulbus oculi", are much more surprising to the model and lead to higher perplexity scores. We conclude that drawing longer sentences with a higher perplexity score leads to samples with a higher number of named entities and, implicitly, more relationships compared to the random sampling strategy. The values in Table 3 support this conclusion. The second advantage of this strategic sampling approach is that entity and relationship classes are more uniformly distributed in the training set. Figure 2 shows that the distributions are much more skewed toward the minor classes in the strategic sampling method.

As the vast majority of relationships correspond to the None relations (Relationship−), the positive relations (Relationship+) were oversampled four times to avoid strong class imbalance during training. Since the model is to be evaluated based on the actual distribution of relationship classes, no oversampling is performed for the test set. To confirm the performance of the language model in terms of NER and RE within (i) the same imaging modality and (ii) the same anatomical region, the evaluation is based on the CTH test set. Since we assume that each entity and relationship class is equally important, the macro F1-score is

EIJARI
ISSN: 2770-8125
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH AND INNOVATIONS
Volume 1, Issue 1. January - 2024

used as an evaluation measurement. In Fig. 3, the main results are reported. The score slightly increases during the iterations where random sampling is applied. At iteration 8, when we intervene with the strategic sampling approach, the score increases dramatically until it reaches a saturation range. There are two reasons for this increase in performance: (i) the training set contains many more sentences with a larger number of clinical entities and relationships, and (ii) the distribution of the different classes in the training set when strategic sampling is applied is more uniform than in the training set when random sampling is applied (Fig. 2). Figure 4 visualizes the entity classes, which benefit the most from the intervention with strategic sampling. Again, the saturation range is reached quite fast after the intervention.

References

1. Wang, Y. et al. Clinical information extraction applications: A literature review. J. Biomed. Inform. 77, 34–49 (2018).
2. Sheikhalishahi, S. et al. Natural language processing of clinical notes on chronic diseases: Systematic review. JMIR Med. Inform. 7, e12239 (2019).
3. Kreimeyer, K. et al. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. J. Biomed. Inform. 73, 14–29 (2017).
4. Wu, S. et al. Deep learning in clinical natural language processing: A methodical review. J. Am. Med. Inform. Assoc. 27, 457–470 (2020).
5. Pons, E., Braun, L. M., Hunink, M. M. & Kors, J. A. Natural language processing in radiology: A systematic review. Radiology 279, 329–343 (2016).
6. Maros, M. E. et al. Comparative analysis of machine learning algorithms for computer-assisted reporting based on fully automated cross-lingual radlex mappings. Sci. Rep. 11, 1–18 (2021).
7. Viani, N. et al. A natural language processing approach for identifying temporal disease onset information from mental healthcare text. Sci. Rep. 11, 1–12 (2021).
8. Chu, S. H. et al. An independently validated, portable algorithm for the rapid identification of copd patients using electronic health records. Sci. Rep. 11, 1–9 (2021).
9. Zeng, X., Linwood, S. L. & Liu, C. Pretrained transformer framework on pediatric claims data for population specific tasks. Sci. Rep. 12, 1–13 (2022).
10. Khurshid, S. et al. Cohort design and natural language processing to reduce bias in electronic health records research. NPJ Digit. Med. 5, 1–14 (2022).
11. Patel, T. A. et al. Correlating mammographic and pathologic findings in clinical decision support using natural language processing and data mining methods. Cancer 123, 114–121 (2017).
12. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).