pHealth 2024
M. Giacomini et al. (Eds.)

doi:10.3233/SHTI240066

Exploring Negated Entites for Named Entity Recognition in Italian Lung Cancer Clinical Reports

Domenico PAOLO [a,1], Alessandro BRIA [b], Carlo GRECO [c], Marco RUSSANO [d], Sara RAMELLA [c], Paolo SODA [a] and Rosa SICILIA [a]
[a] *Unit of Computer Systems & Bioinformatics, Università Campus Bio-Medico di Roma, Italy*
[b] *Department of Electrical and Information Engineering, University of Cassino and Southern Latium, Italy*
[c] *Operative Research Unit of Radiation Oncology, Fondazione Policlinico Universitario Campus Bio-Medico, Italy*
[d] *Operative Research Unit of Medical Oncology, Fondazione Policlinico Universitario Campus Bio-Medico, Italy*

Abstract. This paper explores the potential of leveraging electronic health records (EHRs) for personalized health research through the application of artificial intelligence (AI) techniques, specifically Named Entity Recognition (NER). By extracting crucial patient information from clinical texts, including diagnoses, medications, symptoms, and lab tests, AI facilitates the rapid identification of relevant data, paving the way for future care paradigms. The study focuses on Non-small cell lung cancer (NSCLC) in Italian clinical notes, introducing a novel set of 29 clinical entities that include both presence or absence (negation) of relevant information associated with NSCLC. Using a state-of-the-art model pretrained on Italian biomedical texts, we achieve promising results (average F1-score of 80.8%), demonstrating the feasibility of employing AI for extracting biomedical information in the Italian language.

Keywords. EHRs, deep learning, NER, trasformer, NSCLC

1. Introduction

Non-small cell lung cancer (NSCLC) stands as the most prevalent form of lung cancer and remains the leading cause of cancer-related fatalities worldwide [1]. Electronic Health Records (EHRs) have become the standard repository for the vast amount of information related to the comprehensive and complex oncology care process for NSCLC patients [2]. Being able to mine this information represents a precious opportunity to support oncology research for personalised health. Besides the significant challenge due to the unstructured nature of the texts, there have been several efforts using Natural Language Processing (NLP) techniques within the biomedical

[1] Corresponding Author: Domenico Paolo, domenico.paolo@unicampus.it

domain [3], and more specifically Named Entity Recognition (NER), which aims at the automatic recognition and classification of biomedical *entities*. Such entities can be individual words or phrases within a text that pertain to predefined biomedical categories, referred to as *entity types*, which provide specific clinical information, e.g. diagnosis, patient health status, therapy, etc.. In this scenario, the state-of-the-art is represented by large-scale pre-trained models based on the Bidirectional Encoder Representations from Transformers (BERT) architecture [4]. To the best of our knowledge there is few work in applying BERT-based models for NER to the specific field of NSCLC: existing work often incorporates clinical reports of lung cancer patients into broader datasets, primarily focusing on extracting general oncological information [5,6]. Only two contributions tackle the NER task with the aim of extracting specific information about lung cancer patients from their clinical reports [7,8]. Although both achieve impressive performance, the literature still lacks adapting transformer-based models for the biomedical domain in less-resourced languages as Italian. In our previous work [9] we propose a first attempt to fill this gap, introducing a set of 25 clinically relevant entity types related to NSCLC patients' stage, therapy, tumour position, comorbidities, etc. We apply a pre-trained trasformer-based architecture [10] to perform NER on the manually annotated clinical reports of a cohort of 257 patients, validating the solution against two other pre-trained models.

Even though our previous work led to promising results, the proposed set of 25 entity types does not take explicitly into account the information related to negations. Indeed, clinical reports often include information about the negation or absence of a particular entity (e.g. absence of a symptom), which is proven to be relevant in automated clinical information extraction [11]. Hence, in this work we improve the previous approach presenting an enlarged set of 29 clinical entity types that encompass *negated entities*, defined as such expressions that refer to the absence of an entity. Further to this we conduct a comparative analysis applying the approach in [9] on the enlarged manually annotated dataset, considering both scenarios with and without the inclusion of negated entities.

The rest of the manuscript is organised as follows. Section 2 presents the materials used in our study and section 3 provides the description of the proposed approach. Section 4 outlines the final validation results. Finally, section 5 provides concluding remarks.

2. Materials

In this study, we employed the CLARO dataset [9], consisting of 257 patients diagnosed with stage III and IV NSCLC. The population was enrolled under two different approvals of the Ethical Committee and written informed consent was obtained from all patients. The authors confirm that all ongoing and related trials for this intervention are registered. For each patient we gathered oncological visits and radiotherapy visits dated before therapy start, resulting in 758 different documents. This corpus is manually annotated using the BIO tagging format with 25 entity types validated by domain experts and encompassing information related to cancer description (i.e. type, stage, metastases, abnormalities, position, progression, morphology, histology, TNM classification), therapy (i.e. type, duration, dosage, drugs, line, medication frequency, exams), patient personal information (i.e. smoking habit,

Table 1. Performance average ± standard deviation over the different datasets.

Dataset	Performance Metrics		
	P	R	F1
Original	0.816 ± 0.109	0.873 ± 0.082	0.843 ± 0.095
29 entities	0.776 ± 0.139	0.846 ± 0.100	0.808 ± 0.121
26 entities	0.808 ± 0.108	0.867 ± 0.082	0.835 ± 0.095

familiarity, comorbidity, weight, height, visit date, pain NRS scale, symptoms, general events).

3. Methods

The proposed pipeline consists of three main steps: (i) corpus generation; (ii) model training; (iii) model validation. First, we enlarge the annotation of the dataset presented in Section 2 in order to encompass negated entities. A negated entity is the mirror of an entity type that represents its absence: hence, the original 25 entity types would generate other 25 entity types that represent their negation, or absence. However, examining the clinical reports we were able to actually retrieve only 4 novel entity types denoting the absence of symptoms, comorbidity, focal anomaly or of an exam. This process results in a novel set of 29 entity type labels and in a new annotated corpus. Second, after preprocessing with sentence detection and tokenization, we use the new dataset for fine-tuning the Italian biomedical checkpoint in [10]. The fine-tuning process was conducted using the hyperparameters validated in [9]: batch size of 8, 12 epochs, dropout rate of 0.1, embedding size of 768, Adam optimizer and learning rate of $5 \cdot 10^{-5}$. We also cope with the class imbalance present in our real-world dataset, employing the focal loss function [12]. By increasing the impact of misclassified examples, focal loss enables the model to pay more attention to the minority classes, thus alleviating the effects of class imbalance. Third, we validate our model with a 10-fold cross-validation technique *per patient*, meaning that reports of the same patient are entirely included in one fold. This approach mitigates potential biases due to the presence of same patient's reports in both training and test folds. Then we aggregate the predicted scores of all test folds into a single set to derive an overall measure of F1-score (F1), Precision (P), and Recall (R). This evaluation employs a stringent criterion at the entity level, considering an entity correctly predicted only if all system-labeled tokens exactly match the ground truth tokens. For the sake of brevity we omit further details that the interested reader can find in [9].

4. Results

Table 1 shows the average and standard deviation results computed among all entity types using the original 25-entities dataset and the novel 29-entities dataset for NER. The introduction of negated entities clearly impacts the overall performance of the model. Including more entity types results in a higher number of classes to recognize, thereby increasing the complexity of the NER problem and contributing to a decline in performance. To further investigate this phenomenon we performed the same

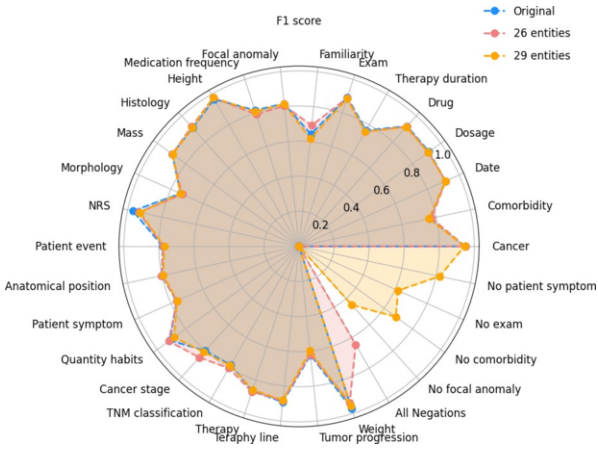

Figure 1. F1-score measured for each entity type across the three different datasets.

experiments collapsing negated entities in only one entity type, resulting in a 26-classes problem. The results are shown in the last row of Table 1. Although they are still lower than those on the original dataset, they outperform the system trained on 29 entity types, confirming the previous consideration.

Despite these results allow for quite evident conclusions on the overall performance of the proposed approach, a deeper analysis is needed. Figure 1 shows a radar plot of the F1-score performance metric for each entity type and each dataset. To maintain brevity, other metrics are excluded, but the following considerations apply to them as well. Looking at the original 25 entity types there is no difference in performance between models, except for a small difference in 'Cancer stage', 'NRS' and 'Familiarity'. This implies that the performance on the original entity types remains unaffected by the introduction of new ones. The observed overall decline might then be attributed to the system's challenge in identifying negated entities. We suspect that this difficulty may arise from the infrequent occurrence of negated entities in clinical reports, exacerbating the issue of imbalance on a broader scale.

5. Conclusions

This study demonstrates the effectiveness of Named Entity Recognition in extracting clinical information from Italian clinical reports of NSCLC patients. The extensive range of identified entities offers a comprehensive overview of both present and absent relevant characteristics of a patient's health status. Given the exponential growth of clinical data, automated and efficient information extraction methods are increasingly essential. While further analysis is needed for downstream use in automated prognostic tasks, this work already provides a tool for clinicians to quickly retrieve pertinent information from unstructured documents. This has the potential to significantly enhance patient care and clinical decision-making, laying the foundation for future research and applications in predictive tasks and personalized health.

Acknowledgment

Domenico Paolo is a Ph.D. student enrolled in the National Ph.D. in Artificial Intelligence, XXXVIII cycle, course on Health and life sciences, organized by Università Campus Bio-Medico di Roma. This work was partially supported by PRIN 2022 MUR 20228MZFAA-AIDA (CUP C53D23003620008, CUP H53D23003480006). Resources are provided by NAISS and SNIC at Alvis @ C3SE, partially funded by SRC through grant agreement no. 2022-06725 and no. 2018-05973.

References

[1] Lung cancer-non-small cell: Statistics, 2022.
[2] Pranjul Yadav et al. Mining electronic health records (ehrs) a survey. *ACM Computing Surveys (CSUR)*, 50(6):1–40, 2018.
[3] Sunyang Fu et al. Clinical concept extraction: a methodology review. *Journal of biomedical informatics*, 109:103526, 2020.
[4] Jacob Devlin et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
[5] Yuting Hu et al. Named entity recognition for chinese biomedical patents. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 627–637, 2020.
[6] S Mithun et al. Development and validation of deep learning and bert models for classifi- cation of lung cancer radiology reports. *Informatics in Medicine Unlocked*, page 101294, 2023.
[7] Huanyao Zhang et al. A novel deep learning approach to extract chinese clinical entities for lung cancer screening and staging. *BMC Medical Informatics and Decision Making*, 21(2):1–12, 2021.
[8] OSWALDO SOLARTE PABÓN et al. A deep learning approach to extract lung cancer information from spanish clinical texts. *Available at SSRN 4049602*.
[9] Domenico Paolo et al. Named entity recognition in italian lung cancer clinical reports using transformers. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 4101–4107. IEEE, 2023.
[10] Tommaso Mario Buonocore et al. Localizing in-domain adaptation of transformer-based biomedical language models. *Journal of Biomedical Informatics*, 144:104431, 2023.
[11] Hermenegildo Fabregat et al. Negation-based transfer learning for improving biomedical named entity recognition and relation extraction. *Journal of Biomedical Informatics*, 138:104279, 2023.
[12] Tsung-Yi Lin et al. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.