Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240794

Leveraging Rule-Based NLP to Translate Textual Reports as Structured Inputs Automatically Processed by a Clinical Decision Support System

Akram REDJDAL[a,1], Natallia NOVIKAVA[a], Emmanuelle KEMPF[a,b], Jacques BOUAUD[a] and Brigitte SEROUSSI[a,c,d]
[a] *Sorbonne Université, Université Sorbonne Paris Nord, INSERM, LIMICS, Paris, France*
[b] *AP-HP, Henri Mondor, Department of Medical Oncology, Creteil, France*
[c] *AP-HP, Hôpital Tenon, Paris, France*
[d] *APREC, Paris, France*
ORCiD ID: Akram Redjdal https://orcid.org/0000-0003-3141-5463

Abstract. Using clinical decision support systems (CDSSs) for breast cancer management necessitates to extract relevant patient data from textual reports which is a complex task although efficiently achieved by machine learning but black box methods. We proposed a rule-based natural language processing (NLP) method to automate the translation of breast cancer patient summaries into structured patient profiles suitable for input into the guideline-based CDSS of the DESIREE project. Our method encompasses named entity recognition (NER), relation extraction and structured data extraction to systematically organize patient data. The method demonstrated strong alignment with treatment recommendations generated for manually created patient profiles (gold standard) with only 2% of differences. Moreover, the NER pipeline achieved an average F1-score of 0.9 across the main entities (patient, side, and tumor), of 0,87 for relation extraction, and 0.75 for contextual information, showing promising results for rule-based NLP.

Keywords. Natural language processing, Clinical decision support systems, Ontologies, Breast cancer, Structured data extraction.

1. Introduction

Multidisciplinary tumor boards (MTBs) are pivotal in fostering collaborative decision-making in breast cancer management, yet their effectiveness is under scrutiny amidst challenges of staffing shortages and increased workload [1]. In recent years, clinical decision support systems (CDSSs) have emerged as promising computer-based tools to leverage patient-specific data and provide personalized treatment strategies. DESIREE is a European project that focused on developing a multimodal CDSS for primary breast cancer patient management. A Breast Cancer Knowledge Model (BCKM) ontology was built to structure breast cancer-related knowledge used by the guideline-based decision

[1] Corresponding Author: Akram REDJDAL; E-mail: redjdalakram300@gmail.com.

support system (GL-DSS) of DESIREE [2]. However, DESIREE didn't tackle interoperability issues with electronic health records (EHRs) and health professionals had to manually input patient data, which is time-consuming and prone to errors.

Given that about 80% of EHR data is in textual format [3], integrating the GL-DSS with EHRs necessitates the utilization of natural language processing (NLP) methods to enable the automated extraction of relevant information from textual data. In recent years, NLP techniques have shifted from rule-based to machine learning (ML) methods that have proven to be more effective [4]. While ML methods offer efficiency, they raise concerns about environmental impact and explainability issues, making rule-based approaches, less data-intensive and more explainable, attractive again. In this paper, we present a rule-based NLP approach, allowing the automatic creation of structured patient profiles from breast cancer patient summaries (BCPSs) to be input to the GL-DSS.

2. Methods

2.1. BCPSs and structured data representation

BCPSs are textual documents providing the information necessary for making breast cancer therapeutic decisions, including personal and family histories, disease history, clinical examination findings, tumor characteristics, radiology, and pathology results, TNM classification, and care plan proposals. We worked on BCPSs sourced from APHP-EDS, the data warehouse of AP-HP University hospitals in Paris, France.

The BCKM ontology is structured to replicate the entity-attribute-value (EAV) generic model for data modeling and the integration of concepts related to the breast cancer domain, based on three main entities, the patient, the breast side, and the lesion [5]. Given that the ontology comprises over 1,600 classes to cover the entire domain, we narrowed our focus to characteristics commonly observed in BCPSs, such as age and menopausal status for patient attributes, clinical node status, multifocality, or BI-RADS (Breast Imaging-Reporting And Data System) score for side attributes, histology, grade, or size information for tumor attributes. These selected features served as the groundwork for the subsequent phases of the study.

2.2. Named entity recognition pipeline

We first developed an annotation scheme in collaboration with two domain experts (BS and JB) to comprehensively capture the features previously selected for each entity within a randomized sample of 30 BCPSs. Then, we developed a named entity recognition (NER) algorithm adhering to the defined annotation scheme. The algorithm was designed to identify and tag the selected attributes and their values within BCPSs, by using a combination of tools. EDS-NLP (https://aphp.github.io/edsnlp/latest/), proposed by APHP-EDS, was used to extract dates, ICD10 codes and drug information, ClarityNLP (https://github.com/ClarityNLP) was used to extract tumor size and TNM staging, and the rest of attributes were extracted using handmade regular expressions (RegEx). The NER algorithm was then further improved to extract contextual information including negations, family history mentions and hypotheses (e.g., a tumor is detected but there is no pathology result to confirm the cancer histology). Finally, rules to refine regular expressions and specific patterns were built in accordance with the

annotation scheme to ensure accurate identification and tagging of such contextual information within BCPSs.

2.3. Structured data extraction

After developing the NER algorithm to annotate attributes and their values, we built a structured data extraction pipeline to capture the significant patient information within BCPSs. This pipeline is made of several steps as illustrated in Figure 1. First, we identified the different sections within BCPSs using section markers or keywords based on both domain knowledge and how BCPS are usually structured. Once sections were identified, the NER algorithm was applied to extract relevant attributes and values along with the contextual information. Additionally, date mentions were extracted for each section to build the patient timeline. Finally, a relation extraction algorithm was developed to identify meaningful connections between attributes and their entities. Two main relations were targeted: the ***hasSide*** relation to specify the laterality (left, right, or bilateral) of side attributes, and the ***isAttributeOf*** relation between tumor attributes and the tumor entity (especially useful in case of multifocal breast cancers). The goal of the relation extraction algorithm is to divide each section into sentences, then identify the side mention and/or the lesion mention in each sentence. The algorithm relates all tumor attributes in each sentence with the lesion mention in that sentence and links all the side attributes to the side mention. The two relations are represented in the BCKM ontology as object properties that link attributes to their entities.

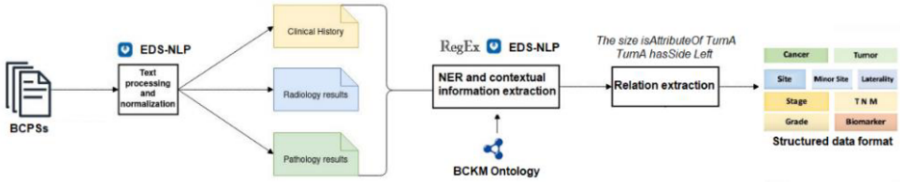

Figure 1 Structured data extraction pipeline.

2.4. Automatic creation of BCKM-conformant patient profiles

The next and final step was to map the extracted attributes to their corresponding concepts in the BCKM. This mapping process enabled the automatic creation of patient profiles that fit the GL-DSS format from data extracted from BCPSs. The workflow is as follows. Once significant patient data is extracted from specific sections, a postprocessing step is performed to resolve conflicts and prioritize most important information (e.g., if clinical nodes status is positive in the MRI and negative in palpation, the algorithm prioritizes the positive status). Finally, the inference module validates data completeness, ensuring the GL-DSS receives all necessary information, such as TNM classification and histologic type. The extracted data is formatted into a csv table, in compliance with HL7 FHIR standards for clinical data exchange and is used as input for existing algorithms developed within the DESIREE project that convert the table into a BCKM-translated patient instance, expressed as triplets in the N3 format [2].

2.5. Evaluation

The NLP algorithm was assessed according to two different methods performed on a sample of 50 randomly selected BCPSs, distinct from the 30 used for the annotation

scheme development. First, the annotations provided by the NER algorithm on the 50 BCPSs were reviewed by an expert oncologist (EK) and inaccuracies were analyzed. Average F1-scores and number of attribute mentions were computed for the three main entities (patient, side, lesion), contextual information, and relation extraction.

Subsequently, to assess the complete NLP pipeline, we *automatically* created patient profiles from the same set of 50 BCPSs, to be input in the GL-DSS, resulting in a set of recommendations (R^{auto}). This set of recommendations was compared to the recommendations provided by running the GL-DSS on the patient profiles *manually* created by a medical informatics intern with oncology expertise (NN), and considered as the gold standard (R^{GS}). Recommendations from R^{auto} and R^{GS} were automatically categorized as "Identical," "Different," or "Comparable," where the latter encompassed partial matches or instances where recommendations from one method were subsumed by those from the other (e.g., set 1 [Lumpectomy], set 2 [Lumpectomy OR Oncoplasty]).

3. Results

The NER algorithm achieved an average F1-score of 0.90, 0.89, and 0.91 for the lesion entity on 2,038 mentions, the side entity on 2,187 mentions, and the patient entity on 226 mentions, respectively. An average F1-score of 0.75 was computed for the contextual information extraction on 686 mentions and of 0.87 for the relation extraction on 2,021 mentions. The code for the NER and structured data extraction is available on GitHub (https://github.com/akramRedjdal/FR_BreastCancer_NER).

Figure 2 displays the results of the comparison of R^{auto} and R^{GS}. Thirty recommendations from R^{auto} were identical to R^{GS} (three of them were cases were the GL-DSS did not produce any recommendation), 19 recommendations were comparable in R^{auto} and R^{GS}, and one case was completely different with no R^{GS} produced.

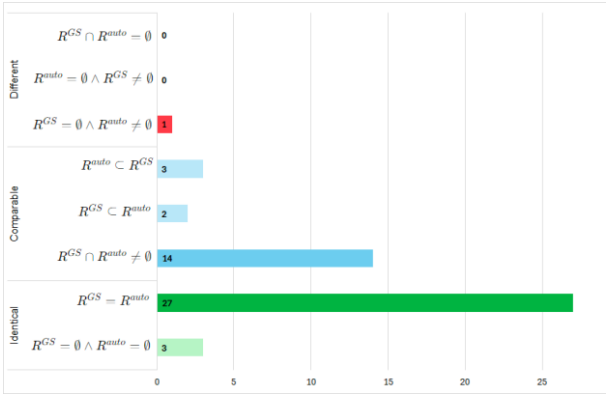

Figure 2 Evaluation of the whole NLP pipeline and the GL-DSS.

4. Discussion and Conclusions

The rule-based NER algorithm achieved an average F1-score of 0.9 for the three main entities, which underscores the efficiency of rule-based methods in extracting significant patient data from BCPSs. However, we observed a notable decrease in F1-score for

contextual information extraction (0.75), specifically for family history and antecedents, indicating a potential area for improvement by integrating machine learning techniques to improve contextual understanding. In comparison with existing systems like RUBY [6], which employs a combination of deep learning (DL) and rule-based algorithms for information extraction in breast cancer reports, the method we proposed shows competitive performance, and in some cases was superior, in identifying key entities. This underscores the capacity of rule-based NER to rival or surpass more elaborate systems regarding accuracy and precision. During the evaluation of the complete NLP pipeline, 60% (30/50) of treatment recommendations in R^{auto} matched the gold standard R^{GS}, with three instances showing no recommendation from both methods due to information missing in the BCPSs. In cases classified as comparable (19/50), R^{auto} encompassed R^{GS} on two instances due to oncoplasty and chemotherapy proposals (proposed by R^{auto} and not by R^{GS}) based on tumor sizes inaccurately interpreted by the NER algorithm (the sizes were actually distances between tumors in the text). Conversely, R^{GS} encompassed R^{auto} on three instances, highlighting two missed chemotherapy recommendations due to incorrect tumor size identification, and one missed lumpectomy contraindication due to an unobserved patient history. These findings underscore the critical need for accurate size extraction and confirms the need for better context analysis in NER approaches. Finally, one case had completely different recommendations, with the automated method erroneously recommending a treatment based on a misinterpretation of a family history of invasive cancer considered as a patient disease, pinpointing again a weakness in context detection. The limitations of this work come from the size of samples. The generalizability needs to be proven (only breast cancer domain and BCPSs coming from one MTB).

In conclusion, with more than 66% (30 + 3) of recommendations compliant with the GS, this work highlights the strengths of rule-based NLP methods in extracting data from BCPSs and points out the needs for better context understanding, suggesting the incorporation of machine learning methods for this specific task. Yet, results suggest that simple to process, sustainable, and explainable rule-based systems could be as effective as complex to process, and environmentally impactful, black box deep learning models.

Acknowledgements: The authors thank Judith Leblanc and the Clinical Research Platform Paris-East team for their support, and the Assistance Publique–Hôpitaux de Paris (AP-HP) clinical data warehouse (EDS AP-HP) team for providing the data for this study.

References

[1] Soukup T, Lamb BW, et al. Cancer multidisciplinary team meetings: impact of logistical challenges on communication and decision-making. BJS Open. 2022 Jul 7;6(4):zrac093. doi: 10.1093/bjsopen/zrac093.
[2] Bouaud J, et al. Implementation of an ontological reasoning to support the guideline-based management of primary breast cancer patients in the DESIREE project. Artif Intell Med. 2020 Aug;108:101922.
[3] Raghavan P, Chen JL, Fosler-Lussier E, Lai AM. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? AMIA Jt Summits Transl Sci Proc.2014 PMID: 25717416.
[4] Banda JM, et al. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. Annu Rev Biomed Data Sci. 2018 Jul;1:53-68.PMID: 31218278
[5] Bouaud J, Guézennec G, Séroussi B. Combining the Generic Entity-Attribute-Value Model and Terminological Models into a Common Ontology to Enable Data Integration and Decision Support. Stud Health Technol Inform. 2018;247:541-545. PMID: 29678019.
[6] Schiappa R, et al. RUBY: Natural Language Processing of French Electronic Medical Records for Breast Cancer Research. JCO Clin Cancer Inform. 2022 Jul;6:e2100199. doi: 10.1200/CCI.21.00199.