AI-Driven Automation of Medical Keyword Extraction and ICD10 Code Association in Gastroenterology Records

Said El Khoury[1], Carl Angelo Mikael[1], Santa El Helou[2], Tina Yaacoub[1], Elio Mikhael[2], Youssef Abou Boutros[2], Ali Mansour[3], Cesar Yaghi[1,2]

[1] Saint Joseph University of Beirut, Lebanon.
{said.khoury, carlangelo.mikael}@net.usj.edu.lb
{tina.yaacoub, cesar.yaghi}@usj.edu.lb
[2] Gastroenterology service, Hotel Dieu de France, Beirut, Lebanon.
{santaelhelou, eliomikhael7, abouboutros123}@gmail.com
[3] Lab-STICC, UMR 6285 CNRS, ENSTA Bretagne, 29806 Brest, France.
{ali.mansour}@ensta-bretagne.fr

Abstract. This article investigates the extraction of medical keywords from French-language gastroenterology records, manually written by healthcare professionals at the Department of Gastroenterology, Hôtel Dieu de France University Medical Center in Beirut, and their association with ICD10 codes to enhance medical technology. Using Named Entity Recognition (NER) techniques and the DR BERT-CASM2 model in conjunction with the Medkit library in Python, we aimed to automate the processes of extraction and association from the French clinical data. We achieved high accuracy, yielding promising results and offering a structured and standardized approach to medical record management. This study lays the foundation for future advancements in automated medical record analysis, paving the way for more efficient and accurate diagnosis and treatment, particularly in French-speaking medical environments.

1 Introduction

The discharge summary (DS) is a critical document that records a patient's hospital stay from admission to discharge, providing vital information for subsequent patient follow-up, either within the same hospital or in a different healthcare setting. Efficient and accurate transmission of diagnostic results, treatment plans, complications, consultations, pending tests, and post-discharge follow-up arrangements is crucial for ensuring continuity of care. Delays or inaccuracies in communication among healthcare providers post-discharge can detrimentally impact treatment continuity, patient safety, satisfaction levels of both patients and clinicians, and the effective utilization of resources [1].

The International Classification of Diseases, 10th Revision (ICD-10), developed by the World Health Organization (WHO), provides standardized identification numbers for various health terms, including signs and symptoms, disease names, procedures, and abnormal findings. These codes facilitate global health information standardization for mortality and morbidity statistics, clinical care, research, disease analysis, healthcare management, outcome monitoring, and resource allocation [2].

Physicians utilize medical records to identify the principal diagnosis and associated comorbidities, converting these into ICD-10 codes to secure governmental financial support. Errors in this process, such as omitted ICD-10 codes or misinterpretations, compromise the accuracy and completeness of medical records. This necessitates extensive labor for rechecking ICD-10 entries, leads to inaccurate epidemiological reports, and results in inadequate reimbursement for hospital management [3].

To mitigate the challenges associated with manual ICD-10 processing, computer technologies can assist or automate the ICD-10 mapping process in clinical practice. For instance, Obeid et al. developed machine learning models trained on discharge summaries to identify patients with cirrhosis, achieving a precision of 0.965 and a recall of 0.978 [4]. Shaalan et al. proposed an ensemble model incorporating various clinical data sources, achieving notable classification accuracies for both inpatient and outpatient datasets, although it relied on a limited data scope from Maharaj Nakorn Chiang Mai Hospital [5]. Additionally, a collaborative residual learning model focusing solely on prescription data demonstrated substantial performance improvements in multi-label classification tasks, highlighting the potential for streamlined coding processes, albeit with limitations in capturing comprehensive clinical scenarios [6]. Complementing these efforts, Ghasemi and Lee employed unsupervised feature selection methods to identify key ICD-10 codes from a large coronary artery disease patient cohort, with Concrete Autoencoder methods excelling in feature reconstruction and mortality prediction, though the study's applicability may be confined to specific patient populations [7].

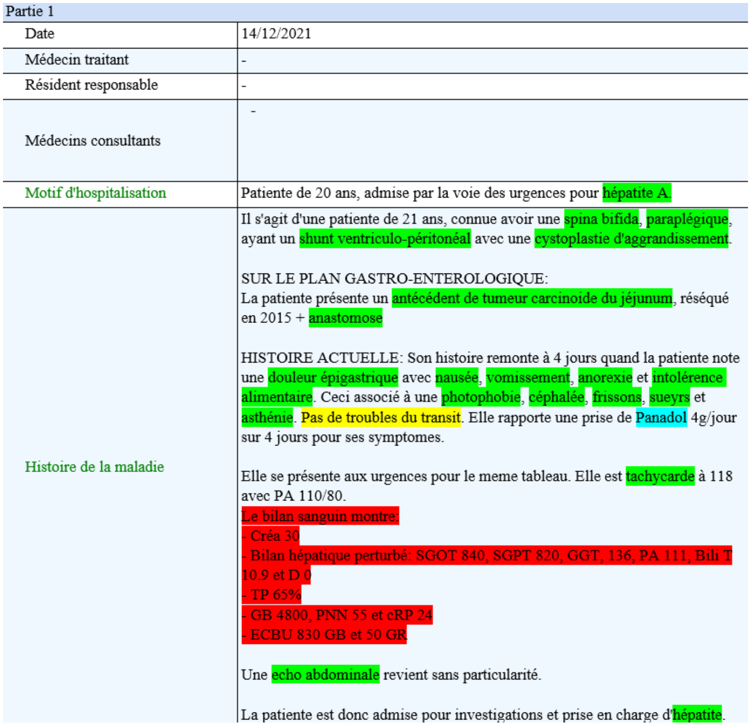
Building upon these foundations, our study introduces a novel approach to automate the extraction of medical keywords and their association with ICD-10 codes specifically within the domain of gastroenterology. Utilizing a French dataset from the Department of Gastroenterology at Hôtel Dieu de France University Medical Center in Beirut, we employ advanced Named Entity Recognition (NER) techniques and the DR BERT-CASM2 model in conjunction with the Medkit library in Python. This approach aims to enhance the accuracy and efficiency of medical record management by providing a structured and standardized method for keyword extraction and ICD-10 code association in French. Our system demonstrates high accuracy in identifying relevant medical terms and offers promising results, paving the way for further advancements in automated medical record analysis in Lebanon and other French speaking countries.

The remainder of this paper is divided as follows. Section 2 describes the system model while section 3 discusses the results. Section 4 provides the conclusion and future works.

2 System Model

2.1 Dataset

The dataset collected from the Department of Gastroenterology at Hôtel Dieu de France University Medical Center in Beirut consists of 21 medical report detailing the hospitalization and treatment of patients admitted to the hospital emergency room. A team of four doctors helped extract and label sentences corresponding to relevant ICD-10 codes, yielding a dataset with a total of 258 labeled rows. Fig. 1 shows an example of a case where the sentences highlighted in green are extracted and coded by the doctors, blue highlights correspond to medications that will be stored and ignored, yellow highlights indicate sentences with negations, and red sections, related to blood and scanner tests are ignored by the doctors.



Partie 1	
Date	14/12/2021
Médecin traitant	-
Résident responsable	-
Médecins consultants	-
Motif d'hospitalisation	Patiente de 20 ans, admise par la voie des urgences pour hépatite A.
Histoire de la maladie	Il s'agit d'une patiente de 21 ans, connue avoir une spina bifida, paraplégique, ayant un shunt ventriculo-péritonéal avec une cystoplastie d'aggrandissement. SUR LE PLAN GASTRO-ENTEROLOGIQUE: La patiente présente un antécédent de tumeur carcinoide du jéjunum, réséqué en 2015 + anastomose HISTOIRE ACTUELLE: Son histoire remonte à 4 jours quand la patiente note une douleur épigastrique avec nausée, vomissement, anorexie et intolérence alimentaire. Ceci associé à une photophobie, céphalée, frissons, sueyrs et asthénie. Pas de troubles du transit. Elle rapporte une prise de Panadol 4g/jour sur 4 jours pour ses symptomes. Elle se présente aux urgences pour le meme tableau. Elle est tachycarde à 118 avec PA 110/80. Le bilan sanguin montre: - Créa 30 - Bilan hépatique perturbé: SGOT 840, SGPT 820, GGT, 136, PA 111, Bili T 10.9 et D 0 - TP 65% - GB 4800, PNN 55 et cRP 24 - ECBU 830 GB et 50 GR Une echo abdominale revient sans particularité. La patiente est donc admise pour investigations et prise en charge d'hépatite.

Fig. 1. Part of a clinical case.

2.2 System Design

The proposed system flowchart is shown in Fig. 2. The system starts by reading the specified medical case from the database, then filtering it by removing non-relevant sections (Bilan, Scanner). It then tokenizes the text into sentences using

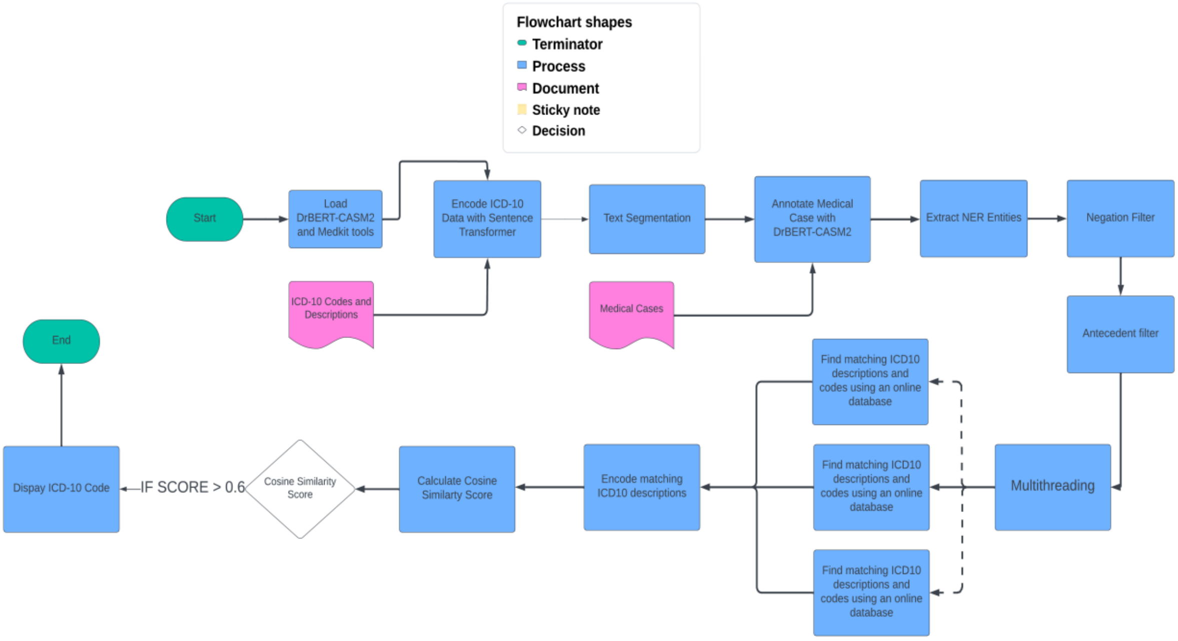

Flowchart shapes
Terminator
Process
Document
Sticky note
Decision
Start
Load DrBERT-CASM2 and Medkit tools
Encode ICD-10 Data with Sentence Transformer
Text Segmentation
Annotate Medical Case with DrBERT-CASM2
Extract NER Entities
Negation Filter
ICD-10 Codes and Descriptions
Medical Cases
Antecedent filter
End
Find matching ICD10 descriptions and codes using an online database
Find matching ICD10 descriptions and codes using an online database
Find matching ICD10 descriptions and codes using an online database
Multithreading
Encode matching ICD10 descriptions
Calculate Cosine Similarty Score
Cosine Similarity Score
IF SCORE > 0.6
Dispay ICD-10 Code

Fig. 2. Conceptual flowchart.

SentenceTokenizer and further segments sentences into syntagmas, meaningful linguistic units, using SyntagmaTokenizer.

Next, the HFEntityMatcher is employed with the Dr. BERT model (medkit/DrBERT-CASM2) to perform named entity recognition (NER) on the syntagmas, identifying medical entities related to problems based on pre-trained models and rules. The main purpose of the HFEntityMatcher is to recognize and extract named entities, such as medical conditions, medications, procedures, and anatomical terms, from text documents. It leverages advanced deep learning techniques and model architectures to achieve high accuracy and robustness in entity recognition tasks. In the context of medical text processing, the HFEntityMatcher is specifically trained on biomedical text datasets (CASM2) to handle domain-specific entities and terminology effectively.

Following entity recognition, the code implements a negation detection system using the Negation Detector from Medkit. This system identifies instances where medical entities are negated in the text (e.g., "aucun," "sans") and marks them as negated entities. Regular expressions (regex) play a crucial role in detecting negation patterns within the text, reducing the time taken by the system to process all entities. For example, the regex pattern \bpas\s*d['e]\b is used to match French negation expressions like "pas de," "pas d'," "pas d'e," etc., indicating the absence or negation of a condition or entity. Similarly, other regex patterns such as \bsans\b, \bne\s*semble\s*pas, \b[l']?absence\b, \b(?:aucun)\b, and \b(?:exclue)\b are employed to detect negation cues like "sans" (without), "ne semble pas" (does not seem), "l'absence" (the absence), "aucun" (none), and "exclue" (excluded) respectively. In other terms, when a keyword is detected with a negation keyword in the same syntagma, the detected medical keyword is labelled as negated. The same process is taken into consideration when filtering entities preceded by the word “antecedent” or medical terms referring to previously contracted illnesses.

The code initializes a Sentence Transformer model named 'FremyCompany/BioLORD-2023-M'. Sentence Transformers are neural network-based models that generate high-dimensional vector representations (embeddings) for input sentences. The specific model used ('FremyCompany/BioLORD-2023-M') is tailored for biomedical text and is designed to encode medical sentences into meaningful embeddings. Then, the system process entities with the help of a Python function that takes a list of entities and a set of syntagmas as input. It aims to merge adjacent entities if they form a phrase found in any of the syntagmas, recursively processing the merged entities until no further merging is possible. In other words, the function makes sure that no other technical keyword is divided into medical keyword.

Then detected abbreviations are replaced by their respective definition, and specific terminologies, used by the medical team, by their respective definition or synonyms recognized in the medical world.

A four-multithreading process is built to help examine the extracted entities on an online database to determine the best matching ICD10 codes in a faster and more efficient way. On each thread, we looked for the specific keyword in the "Aide au Codage CIM-10", an online program that functions as a synonym dictionary or online database for ICD10 codes, with a very broad alphabetical index based on the CIM-10, which represents the French version of the ICD10 syllabus. This step was integrated so that the system narrows down the possible ICD10 codes matches that are relevant to the extracted keywords. Any keyword that does not match with any ICD10 on the third-party website is stored so that if it occurs another time, the system will ignore it, reducing by 15% the average time taken to finish a case.

Cosine similarity scores are calculated between the encoded matching descriptions and the previously encoded ICD10 descriptions. Then ICD10 codes with their corresponding descriptions are displayed based on the highest cosine similarity scores. However, if the score is less than 60%, a dictionary search will be conducted on the designated entity. The cosine similarity score is calculated once again based on the new provided definition. In case the score is still less than 60% or in case of not finding a suitable definition, the system will automatically end the process and move on to the next entity. This process facilitates the association of clinical NER entities with relevant ICD10 codes, providing potential solutions for medical coding tasks.

A Python function coordinates a complex process of retrieving, filtering, and matching medical keywords with their corresponding ICD10 codes. The system then iterates through the detected keywords. If a keyword is labeled as “antecedent”, 2 ICD10 codes are extracted: the first one representing the diagnosis as it is, and the second one representing the diagnosis without taking into account the “antecedent” label. This approach was implemented in order to improve drastically the accuracy, even though the process will take longer. This is due to the fact that there isn’t a fixed standard on what ICD10 code is needed for words labeled as “antecedent”.

Another function is designed to streamline the retrieval of ICD10 codes associated with specific medical keywords, contributing to the broader context of medical record analysis automation. Utilizing Selenium's WebDriver capabilities, this function operates within a web-based environment, interacting with elements on web pages to extract targeted information efficiently. It employs explicit waits (WebDriverWait) to handle asynchronous page loading and dynamic content, enhancing reliability during interactions with web elements. Upon locating the search input box using XPath, the function inputs the provided keyword, triggers a search operation by simulating a key press (Enter), and waits for the resulting table to load. Exception handling mechanisms are incorporated to address potential issues, such as element visibility delays or unexpected page behaviors, ensuring robustness in data retrieval processes. The function then parses the retrieved table, extracting ICD10 codes and their corresponding descriptions. This systematic approach, coupled with error handling strategies, enhances the function's resilience and accuracy in extracting essential medical coding information from online databases.

One last function begins by validating the availability of essential data, including ICD10 codes and descriptions. It then calculates the similarity scores between a given element and the description list using cosine similarity, which leverages semantic understanding to connect keywords to relevant ICD-10 codes, even if they don't precisely match. This opens doors to capturing broader associations, potentially improving recall by retrieving related codes not directly reflected in the keywords. Cosine similarity is then used to compare the matched ICD10 descriptions taken from the “Aide au Codage” program with the initially extracted keyword. The one with the highest score is chosen and its code is extracted.

3 Results and Discussions

The evaluation of the system involved processing anonymous medical cases and comparing the extracted output with the labels provided by medical professionals. The primary metric for these tests was accuracy, defined as the ratio of correctly extracted ICD-10 codes to the expected ICD-10 codes. Additionally, the time required to process each case was measured, as the goal is to aid doctors in automating and accelerating the ICD-10 matching process.

The cases were categorized into easy, medium, and hard based on the number of ICD-10 codes they contained. The most challenging case had 36 ICD-10 codes to be matched. The system successfully extracted and matched 29 out of the 36 codes, achieving an accuracy of 80.5% and taking approximately 15.5 minutes to process. This case had the lowest accuracy and the longest processing time, exceeding the second longest by 10 minutes, which took 4 minutes and 32 seconds.

A medium-level case containing 18 ICD-10 codes was also tested. The first iteration, performed without the multithreading feature, took approximately 10 minutes and 30 seconds. In the second iteration, multithreading was enabled, reducing the processing time to 4 minutes and 32 seconds, marking a 57% decrease. Additionally, removing non-relevant keywords from the case further reduced the processing time by an additional minute.

All easy cases were successfully coded by the system. Overall, the system matched 211 out of 258 available ICD-10 codes, yielding an accuracy of 82%. Before implementing multithreading, the average time per word for the matching process was 24.8 seconds and after, the average time decreased to 10.7 seconds, affirming the 57% decrease in processing time.

One of the key challenges we faced was related to the triage process, specifically identifying and filtering the relevant medical keywords from the vast amount of clinical data. Defining an appropriate threshold for cosine similarity scores to ensure accurate ICD10 code matching also proved complex. Additionally, since our approach exclusively relied on open-source tools without the use of predefined or commercial solutions, we encountered limited documentation, which added to the complexity of development. Furthermore, using third-party services to verify ICD10 codes was time-consuming, impacting the overall efficiency of the system.

Despite these challenges, the benefits of the devised solution are significant. Our system serves as a cornerstone for the Lebanese healthcare system by providing a free, open-source tool for automating medical keyword extraction and ICD10 code association. This tool is particularly valuable for French-speaking countries, which can adopt it for their own medical record management needs, facilitating wider access to advanced medical technologies.

Conclusion

The project on automating keyword extraction and code association processes in medical technology, particularly focused on French gastroenterology records, has yielded valuable insights and achievements. One key learning is the effectiveness of Named Entity Recognition (NER) techniques combined with advanced models like DR BERT-CASM2 and Medkit library in Python for extracting medical keywords and associating them with ICD10 codes. This approach has demonstrated high accuracy in identifying relevant medical terms and codes, laying a solid foundation for automated medical record analysis. Moving forward, several recommendations and future improvements can enhance this solution. Firstly, continuous refinement of NER models and incorporating domain-specific knowledge can further boost accuracy and relevance in keyword extraction. Integration with real-time data sources and feedback mechanisms can ensure the system stays updated with evolving medical terminology and practices. Additionally, enhancing the user interface and accessibility of the solution can improve usability for healthcare professionals.

References

1. Chakravarthy, R., Shahid, M., Basha K M., Angadi SP., Sherikar N.: An Audit of Orthopaedic Discharge Summaries Comparing Electronic With Handwritten Summaries: A Quality Improvement Project. Cureus. 15(5) e39396

2. Horsky, J., Drucker, EA., Ramelson, HZ.: Accuracy and Completeness of Clinical Coding Using ICD-10 for Ambulatory Visits. AMIA Annu Symp Proc. 2018 (2017) 912–920

3. Isaradech, N., Khumrin, P.: Auto-mapping Clinical Documents to ICD-10 using SNOMED-CT. AMIA Jt Summits Transl Sci Proc. 2021 (2021) 296–304

4. Obeid, JS., Khalifa, A., Xavier, B., Bou-Daher, H., Rockeyy, DC.: An AI Approach for Identifying Patients With Cirrhosis. Journal of Clinical Gastroenterology (2023) 82-88

5. Shaalan, Y., Dokumentov, A., Khumrin, P., Khwanngern, K., Wisetborisut, A., Hatsadeang, T., Karaket, N., Achariyaviriya, W., Auephanwiriyakul, S., Theera-Umpon, N., Siganakis, T.: Ensemble model for pre-discharge icd10 coding prediction (2020)

6. Shaalan, Y., Dokumentov, A., Khumrin, P., Khwanngern, K., Wisetborisut, A., Hatsadeang, T., Karaket, N., Achariyaviriya, W., Auephanwiriyakul, S., Theera-Umpon, N., Siganakis, T.: Collaborative residual learners for automatic icd10 prediction using prescribed medications (2020)

7. Ghasemi, P., Lee, J.: Unsupervised Feature Selection to Identify Important ICD-10 Codes for Machine Learning: A Case Study on a Coronary Artery Disease Patient Cohort. 10.2196/preprints.52896 (2023)