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

doi:10.3233/SHTI240371

Semantic Mapping of Named-Entities in openEHR Templates and Ad-hoc Generation of Compositions

Nektarios LADAS[a,b,1], Stefan FRANZ[a], Dominik WOLFF[a], Alina REHBERG[a], Michael MARSCHOLLEK[a,b] and Matthias GIETZELT[a,b]
[a] *Peter L. Reichertz Institute for Medical Informatics, TU Braunschweig and Hannover Medical School, Germany*
[b] *GeMTeX Consortium of the German Mecial Informatics Initiative, Germany*
ORCiD ID: Nektarios Ladas https://orcid.org/0000-0001-5918-8384

Abstract. Integration of free texts from reports written by physicians to an interoperable standard is important for improving patient-centric care and research in the medical domain. In the context of unstructured clinical data, NLP Information Extraction serves in finding information in unstructured text. To our best knowledge, there is no efficient solution, in which extracted Named-Entities of an NLP pipeline can be ad-hoc inserted in openEHR compositions. We therefore developed a software solution that solves this data integration problem by mapping Named-Entities of an NLP pipeline to the fields of an openEHR template. The mapping can be accomplished by any user without any programming intervention and allows the ad-hoc creation of a composition based on the mappings.

Keywords. Data Integration, Natural Language Processing Information Extraction, openEHR, Interoperability

1. Introduction

To extract and integrate important information found in unstructured texts of medical reports of physicians to an interoperable standard, first, it is required to apply the Natural Language Processing (NLP) method of Information Extraction (IE), which aims to extract entities known as Named-entities (NE). I.e., the extraction of NE is a sub-method of Information Extraction (IE) named Named-entity recognition (NER) [1] and has the task of detecting words and assigning them to specific fields.

Integration of unstructured texts in an interoperable standard is not an easy task, and can be quite challenging, and is dependent on the text complexity and semantics of the source document and the mapping process to the destination dataset.

There is constant development in the area of NER through the application of Deep and Machine Learning, and, more recently, through the application of Large Language Models [2]. Despite this, there is still a gap in a efficient method that inserts extracted

[1] Corresponding Author: Nektarios Ladas, Hannover Medical School, PLRI-OE 8420, Carl-Neuberg-Straße 1, 30625 Hannover, Germany; E-mail: ladas.nektarios@mh-hannover.de.

entities of the pipeline direct into data models that are represented by an interoperability standard, such as openEHR or HL7 FHIR.

Our main research aim is to design and develop a two-fold method in which NE(s) are dynamically mapped without recompiling to openEHR elements, and a process in which compositions are then ad-hoc generated in a destination repository.

We have applied this approach to various extractions in several medical use cases; to demonstrate our research, we use the rather complex example in molecular genetic reports (MGR), since such reports can contain more than one genetic variant that needs to be assigned to multiple archetypes. Archetypes in openEHR are: "*models defining possible arrangements of data that correspond to logical data points and groups for a domain topic; a collection of archetypes constitutes a library of re-usable domain content definition elements*" [3].

1.1. Related Work

Concerning the mapping of NE(s) and the creation of data in an openEHR template, Wulff et al. describe a method, where the mapping requires programming through a developer with Java knowledge [5]. Another relevant work is the publication by Zubke et al.; in which numerical values from texts are assigned to fields of an archetype. Although they suggest it as a proposal for the integration of IE pipelines into openEHR, the work is limited to numerical values and does not imply other data types of information [4].

2. Methods

As part of the German Medical Informatics Initiative, at the Hannover Medical School's Medical Data Integration Center, we integrate data from various sources in the openEHR standard. One of our tasks is to integrate information through IE from reports written by physicians in free-form text into an openEHR repository. We have developed a platform, that implements two processes: the first process is the mapping of the NE(s) to their corresponding openEHR elements that can be edited manually by a user, and the second process is a method in which the results of the pipeline are directly parsed and ad-hoc inserted into an openEHR data model.

A requirement for the platform is that the output results of the pipeline are in a pre-defined JSON format. Table 1 shows an example of a part of a MGR and the associated output of the IE pipeline. Elements contained in an archetype, that are generated *n* times, are assigned with an incremental index.

Table 1. Section of example text of an MGR containing a gene variant and the results of the pipeline extraction

Fragment of example text of a MGR in German	Extracted NER(s) in JSON output

KRAS
chr12:25398284:NM_033360.4:Exon2:c.35G>T:p.G12V
18,2%
……..
Qualität der DNA-Sequ. sehr gut
	{
	"c_Chromosome[0]": "chr12",
	"c_GenomPosition[0]": "25398284",
	"c_transcriptGene[0]": "KRAS",
	"c_AlleFrequencey[0]": "18.2"
	"c_Sequencing[0]": "DNA"
	}

In the above example, the type of analyzing (DNA or RNA), comes in the text after the gene variants are written, which is in Englisch translated “Quality of DNA-Sequence. Very good”, and is contained in the openEHR archetype “Genomic variant Result” and describes the applied sequencing type and analyzing quality. The information is found and extracted later in the text, and without an appropriate mapping and parsing of the NE(s), the data might not be correctly inserted, and the archetype will not contain the data that are related to a specific sequencing analysis.

Through the Medblocks tool [4], the contents of an openEHR template are translated into a simplified XML format. Templates in openEHR are “*models of content corresponding to use-case specific data sets, constituted from archetype elements*” [3]. The resulting XML format consists of the main data paths for each item and their openEHR field names. Following, we edit the fields and assign their corresponding NE manually in the attribute “nlpTag”.

In Table 2, an example of a mapping of an NE to its relevant openEHR field is shown.

Table 2. Mapping NE to an openEHR field as an XML object.

Extracted NER in JSON output	XML openEHR Field
{ “c_transcriptGene[0]” : “KRAS“ }	<input path="…." label="Gen-Name (HGNC)" nlpTag="c_transcriptGene"> </input>

The “path” attribute is automatically generated by Medblocks and defines the destination field in openEHR flat format. Technically, Medblocks does not support the generation of cardinalities in the form of archetypes within a composition. We have developed a new XML node that functions as a wrapper within the elements of an archetype that automatically generates 1–to-n archetypes. In Table 3, an example of the archetype wrapper is demonstrated.

Table 3. Definition of an archetype wrapper.

Extracted NER(s) in JSON output	Archetype wrapper

{
"c_Chromosome[0]": "chr12",
"c_GenomPosition[0]": "25398284",
"c_transcriptGene[0]": "KRAS",
"c_AlleFrequencey[0]": "18.2"
“c_Sequencing[0]”: “DNA”
}
	<archetype name=“Genomic Variant Result“ sourceTag=”c_Sequencing”>
	<input path="…."
	label="Gen-Name (HGNC)"
	nlpTag="c_transcriptGene">
	</input>
	…..
	</archetype>

Through the attribute “sourceTag” the parser (named NE Parser) generates a new archetype and maps the NE(s) values to the openEHR fields of the archetype. The NE Parser is the core function of the platform and is responsible for reading the NE(s) and then assigning them to their respective XML elements while creating the cardinalities.

After the mappings are entered for a specific template, they are stored on a mapping file. Whenever a document is processed from the IE pipeline, the results are then ad-hoc posted to a REST service. After the composition content is generated through the NE Parser, it is further posted by the same process to the openEHR repository server.

3. Results

We mapped the NE from the unstructured texts of MGR. The main openEHR template has 122 fields shared in 21 Archetypes [7]. In our location, an MGR can contain one or more genomic variants archetypes. Within an MGR, we have identified 16 relevant NE(s) to be inserted in 4 archetypes and 16 fields into the respective openEHR data model.

After we assigned the NE(s) based on our mapping method, we validated the pipeline with 50 MGR(s) independently by two reviewers; all compositions along with their respective archetypes and fields were correctly parsed and generated in the openEHR repository.

The archetype wrapper processed without any issues and each of the genetic variants is correctly stored under their respective type of analysis in the archetype.

4. Discussion

We have developed a platform that can be used as an intermediate process through a REST-API to automate the integration of data directly from the extraction NE of an NLP pipeline into an openEHR repository that is based on a specific template. The mapping process can be accomplished without the need for a developer. Each of the 50 documents used for validation, is inserted without any error in the openEHR repository. When we added another NE for the extraction, which originally consisted of 15 fields for the genomic variant archetype, the update of the mapping table of the new NE was accomplished without the need to recompile a source code and only by just adding the new component. This provides flexibility and easiness to custom edit the results of any template.

The suggested platform can be easily adapted on various IE pipelines for integrating the data into an openEHR repository but must have the same output structure required for the mapping. This can be a point of criticism for the suggested solution since various pipelines deliver their results in other structures, but the transformation of the results in a required output format should not be a major issue.

5. Conclusions

Through our platform, we developed a new method to integrate unstructured medical texts into our openEHR repository for various use cases.

This method can be applied as a main approach to help close the gap of integration between unstructured medical texts and interoperable datasets.

In the future, we plan to build a UI that allows for visual mapping without the manual editing of a user. After we gather more mapping information, we will use large language

models, fine-tune them, and check if they can help to replace or enhance the manual mapping.

Availability

The source code is available upon request.

Acknowledgments

This work was supported by BMBF within the GeMTeX (German Medical Text Corpus) Project as part of German Mecial Informatics Initiative under grant 01ZZ2314J.

References

[1] Sun P, Yang X, Zhao X, Wang Z. An Overview of Named Entity Recognition. 2018 International Conference on Asian Language Processing (IALP) [Internet]. 2018 [cited 2024 Feb 08];273-278. Availiable from: doi:10.1109/IALP.2018.8629225.
[2] Wang S, Sun X, Li X, Ouyang R, Wu F, Zhang T, et al. GPT-NER: Named Entity Recognition via Large Language Models [Internet]. arXiv.org. 2023 [cited 2024 Feb 08]. Available from: https://arxiv.org/abs/2304.10428.
[3] openEHR. Archetype Technology Overview [Internet]. 2019 [cited 2024 Feb 08]. Available from: https://specifications.openehr.org/releases/AM/latest/Overview.html.
[4] Zubke M, Bott JO, Marschollek M. Using openEHR Archetypes for Automated Extraction of Numerical Information from Clinical Narratives. Stud Health Technol Inform. 2019 Sep 3;267:156-163.
[5] Wulff A, Mast M, Hassler M, Montag S, Marschollek M, Jack T. Designing an openEHR-Based Pipeline for Extracting and Standardizing Unstructured Clinical Data Using Natural Language Processing. Methods Inf Med. 2020 Dec;59(2):64-78.
[6] Medblocks Documentation. Medblock UI [Internet]. 2024 [cited 2024 Feb 08]. Available from: https://medblocks.com/docs/medblocks-ui (Accessed: 08 February 2024).
[7] Ocean Health Systems. Clinical knowledge manager - Molecular genetic findings Template [Internet]. [cited 2024 May 15]. Available from: https://ckm.highmed.org/ckm/templates/1246.169.236.