© 2019 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI190820

Using openEHR Archetypes for Automated Extraction of Numerical Information from Clinical Narratives

Maximilian ZUBKE[a,b,1], Oliver J. BOTT [b] and Michael MARSCHOLLEK [a]
[a] *Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Hannover, Germany*
[b] *University of Applied Sciences Hannover, Hannover, Germany*

Abstract. Up to 80% of medical information is documented by unstructured data such as clinical reports written in natural language. Such data is called unstructured because the information it contains cannot be retrieved automatically as straightforward as from structured data. However, we assume that the use of this flexible kind of documentation will remain a substantial part of a patient's medical record, so that clinical information systems have to deal appropriately with this type of information description. On the other hand, there are efforts to achieve semantic interoperability between clinical application systems through information modelling concepts like HL7 FHIR or openEHR. Considering this, we propose an approach to transform unstructured documented information into openEHR archetypes. Furthermore, we aim to support the field of clinical text mining by recognizing and publishing the connections between openEHR archetypes and heterogeneous phrasings. We have evaluated our method by extracting the values to three openEHR archetypes from unstructured documents in English and German language.

Keywords. Text Mining, information extraction, machine learning, openEHR

1. Introduction

A vast amount of medical documentation is represented as narrative text. From a technical perspective, such data is unstructured because the information it contains cannot as easily be selected by a computer program as the content of structured information models such as HL7 or openEHR. However, clinical reports written in natural language are an attractive documentation model for physicians; besides the high flexibility and expressiveness of natural language, the comfortable usage is an additional advantage of this type of documentation: for example, a text can easily be produced with a speech-to-text application which is used by a radiologist during the analysis of an X-ray image. We therefore expect that this type of documentation will increase.

Furthermore, patient treatment is often time-critical, so that manual reading of a report should not take too much time. For this reason, there is the requirement that

[1] Corresponding Author, Maximilian Zubke, Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Carl-Neuberg-Str. 1, 30625 Hannover, Germany; E-Mail: maximilian.zubke@plri.de

clinical information systems process this type of data properly and extract the contained information automatically. We present a proposal for the integration of natural language processing (NLP) into the established development and governance infrastructure of openEHR. The aim of our research for this paper was to automate the extraction of the numeric values required by openEHR archetypes from narrative documents based on archetype specific language models.

1.1. openEHR

openEHR is a model driven, open-platform approach for electronic health records, which is used to achieve semantic interoperability across institutions. The information is described by unified, constraint-based data structures called archetypes, which are stored, versioned and distributed within the openEHR clinical knowledge manager (CKM) [1]. All archetypes are technically based on a common reference model that specifies data types and structures [2]. Similar to other medical documentation technologies, openEHR archetypes are developed and maintained in context of a worldwide respectively nationwide governance process, which supports the creation of unified and agreed data models. The life cycle of an archetype typically consists of its initial modelling, the upload into the CKM, one or several reviews by experts from other sites and finally the publication into the CKM database (Figure 1). After publication, for example, the use of an archetype "blood pressure" guarantees interoperability between institutions. [3–5]

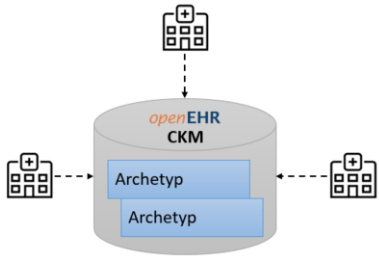

Figure 1. The CKM as a central database for agreed data structures (archetypes), designed to ensure a uniform description of electronic health records.

1.2. Related Work

Manual information extraction is often a cumbersome and error-prone task, which is the reason of the existence of a dedicated research field called (NLP) which aims to devise suitable solutions. There is a lot of research activity in the field of NLP, often beyond the medical context, but only a few publications involve openEHR. Altogether, there are only three papers about the combination of NLP and openEHR. In [6] a NLP pipeline for extracting values to openEHR archetypes is evaluated by a concrete application in pediatrics. The pipeline is based on manual concept modelling and information extraction is done by rules. Also the NLP pipeline presented in [7] uses a rule based approach evaluated for the specific use case diabetes. Due to the idea of using openEHR archetypes as information extraction templates, [7] is comparable to our work. Our contribution, however, proposes to substitute the expensive rule engineering by the utilization of existing metadata, optional extended by weakly supervised machine learning. A further difference is that we focus on the development

of a general, use case-independent approach for the extraction of numerical values from clinical transcripts using openEHR archetypes. The authors of [8] used supervised machine learning to automatically convert text data from HL7 FHIR or MIMIC-III to openEHR.

2. Method

Our method consists of three steps:

1. Recognition of numerical values within a narrative document
2. Identification of archetype fields which require a numerical value
3. Assigning recognized numerical values to identified archetype fields

For step 3 we identified two cases:

- Case 1: Assignment is obvious and can be done without any user interaction
- Case 2: Assignment is hard to recognize (e.g. due to unknown abbreviations), weakly supervised machine learning is recommended

The overall workflow is visualized in Figure 2. Details about the processing steps are explained in the following subsections.

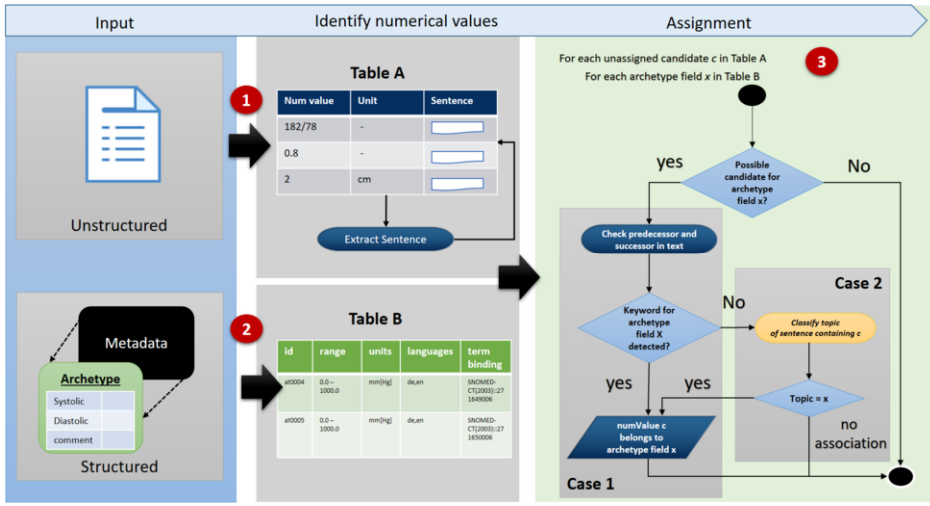

Figure 2. Overview of our method. Case 2 can be seen as optional extension of Case 1. Without Case 2, the proposed method works fully unsupervised.

2.1. Recognizing numerical values

To assign numerical values from clinical transcripts to openEHR archetypes, we initially use regular expressions to localize and extract the document contents of interest. Thereby we differ between integers, floats, intervals and ratios. Single numbers may be part of an interval or ratio, e.g. the ratio "120/70" consists of the two integers 120 and 70. To handle this we use a two-step approach: First, we collect all single numbers, intervals and ratios as mentioned above, next we repeat the recognition of numerical values based on the result of the previous step. This way we recognize single values of intervals and ratios without losing the semantic composition. After the

recognition of numerical values, we extract the associated sentence from the document. Finally, this step generates Table A of Figure 2, where each row contains a numerical value, the sentence of occurrence and, if available, the associated unit. The sentence will be further analyzed by the processing step described in 2.3.

2.2. Identifying numerical described values of an archetype

An archetype is a constraint-based data structure which describes a certain domain concept by the composition of well-defined data fields: The archetype "blood pressure" contains one field for the systolic and another field for the diastolic blood pressure. Each of these fields is connected with a data type like "Quantity" for measurements or "Text" for narrative values. In addition, there are restrictions that define the range of allowed values or the unit for a particular entry within the archetype. Using this metadata, we can easily recognize the parts of a given archetype to be filled by numerical values.
Figure 3 shows an overview of the metadata to a single field of an openEHR archetype.

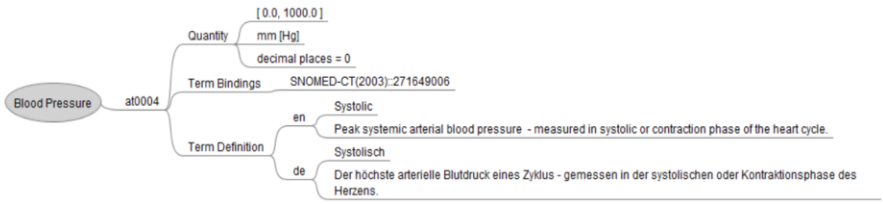

Figure 3. Metadata for the field with the id at0004 of archetype "blood pressure".

Combining all these metadata, the numerical values to be extracted from a narrative report can be described precisely. The collection of relevant archetype fields including constraints and multilingual descriptions as well as optional term bindings represents the second step of our algorithm.

2.3. Assigning numerical values to archetype fields

Finally, the extracted numerically described information resulting from step 1 has to be assigned to the archetype fields selected by the algorithm described in section 2.2.

Due to the existence of very complex formulations, we decided to create two variants of step 3: A relatively simple but fully automated version and a second, more advanced variant based on a neural network resulting from a weakly supervised, active learning process. The use of two differing variants can be substantiated by the following cases.

2.3.1. Case 1: Explicit phrasing

The archetype field name and the allowed value occur directly after each other or are separated only by words that are not nouns. The field name is considered as a noun. This assignment works well for texts such as the following example: "*...body temperature of 24°...".*

Furthermore, we maintain a keyword list, which contains archetype name, archetype keywords, archetype description (tokenized), field name, field description

(tokenized) and, if available, keywords of linked terminology concepts (i.e. from SNOMED CT). Additionally, the keyword list is extended by automatic generation of abbreviations and subwords using common string operations. For example, from “blood pressure” the abbreviations “bp” and “blpre” or for “body temperature” the subword “temperature” will be generated. If the field name was not found close to the numerical value, we repeat the search with these keywords. This case is completely based on matching patterns derived from the archetype.

2.3.2. Case 2: Ambiguous phrasing caused by use of contextual knowledge

Due to the metadata from the archetype, which we consider as a-priori knowledge, the assignment can work unsupervised as mentioned above. However, there are possible phrasings, which cannot be processed this way due to a lack of information. We call this missing information “contextual knowledge”, which is demonstrated in Table 1.

Table 1. Example texts containing numerically described information.

#	Text	Contextual knowledge
1	“ blood pressure 137/60,”	common notations of separated values (“137/60” instead of “systolic=137” and “diastolic=60”)
2	“T 100.9 BP 136/86 Pulse 117 RR 12 98% sat on room air”	contextual abbreviations (“T” for body temperature)
3	“HEART: Shows bradycardia at 53.”	diagnoses instead of the underlying observation (“bradycardia at 53” instead of “pulse 53”)

To handle such more complex assignments, we have integrated a topic detection step, with archetype fields as the topics. This classifier determines whether a certain numerical value (represented by a word w occurring in a given text sequence at position i) belongs to a certain archetype field by calculating the probability shown in equation 1.

$$P\left(w_i = < num > = topic_j \,\middle|\, w_{i-1} = \text{"is"}, w_{i-2} = \text{"}today\text{"}, w_{i-3} = \text{"}bp\text{"}, \dots\right) \quad (1)$$

Thereby, the probability increases when keywords from the archetype field corresponding to $topic_j$ occur close to the numerical value located at position i of the token sequence from the previous extracted sentence. We have realized this step by an active learning, recurrent neural network [9] built on fastText [10]. This network learns actively, because it only needs a minimum of labeled data, and these data records are selected by the learning algorithm. The required labels are entered by an interactive question-answering approach. Thereafter the remaining unlabeled records are clustered based on their similarity [11] to the interactively labeled records. In the best case, the user only has to answer one question regarding a certain numerical value.

After preparing a training set of sample sentences or sample text snippets, the neural network learns the probability threshold to equation 1 which indicates the semantic meaning of a certain numerical value.

The input is a sequence of vectors describing the words in the sentence to classify. These vectors are delivered by pertained neural language model [12–14]. Due to the

integration of this character embeddings, our assignment algorithm is able to deal with latent features such as:

- synonyms and hard abbreviations (T means body temperature)
- sequence patterns (information value 120/90 always appears after the title to the corresponding archetype field, never vice versa)
- typos (Blod pressure means blood pressure)

In the following we call such neural networks archetype-specific language models. An archetype-specific language model is trained to recognize narrative descriptions of archetype related document contents.

3. Results

We have evaluated our method with the 3 openEHR archetypes written in the first column of Table 2. With each archetype, we processed 75 documents in English and 75 in German language. Both corpora also include documents that do not contain any value to the archetype at all. For details see Table 2.

Table 2. Corpus overview: artefacts per document

Content	English corpus (n=75)	German corpus (n=75)
Num values (min/avg/max)	1/20/81	3/23/69
Blood pressure documented	49/75 (65%)	52/75 (69%)
Body temperature documented	47/75 (63%)	45/75 (60%)
Pulse documented	53/75 (71%)	50/75 (67%)

Thus, we consider our assignment algorithm as a composition of two classification problems:

1. Is the archetype contained in the narrative document? (Measured by Recall)
2. Which value(s) belong(s) to the archetype? (Measured by Precision)

The English texts were taken from a public collection of medical transcripts[2], the German documents were provided by a German hospital. Unfortunately, the collection of anonymized documents from the German hospital was made available for internal research only and may neither be published in full nor in part.

Both corpora have been labeled for the three mentioned openEHR archetypes by the authors. The result of the evaluation using this gold standard is documented in Table 3.

Table 3. Results of evaluation: case1 in brackets, followed by 10-fold cross validation of case2

Archetype (#num values)	Precision	Recall
English (75 documents)		
Blood Pressure (4)	(0.90) 0.90	(0.87) 1.00
Body Temperature (1)	(0.90) 1.00	(0.59) 0.96
Pulse / Heart Beat (1)	(0,86) 1.00	(0.65) 0.90
German (75 documents)		
Blood Pressure (4)	(0.90) 0.93	(0.95) 1.00
Body Temperature (1)	(0.98) 1.00	(0.74) 0.97
Pulse / Heart Beat (1)	(0.82) 1.00	(0.70) 0.90

[2] https://www.mtsamples.com/

We interpret these results as proof of concept for our approach. Further experiments are planned.

4. Discussion & Conclusions

Our approach shows that openEHR archetypes can simplify clinical text mining. Instead of defining keywords and constraints manually, like in [15], this information can be extracted from the archetype definitions.

Furthermore, we could observe, that keyword-based matching algorithms already yield remarkable results. However, there are possible phrasings, which cannot be processed this way due to a lack of information, which we call "contextual knowledge". To learn such associations, we introduced the idea of archetype-specific language models, which is visualized in Figure 4.

Indeed, there will be differences between data authors from different institutions. But we expect that there will be a finite, manageable amount of formulations for a particular information. Because the manual description of extraction rules is time-consuming and prone to errors, we nevertheless recommend to use machine learning for this task.

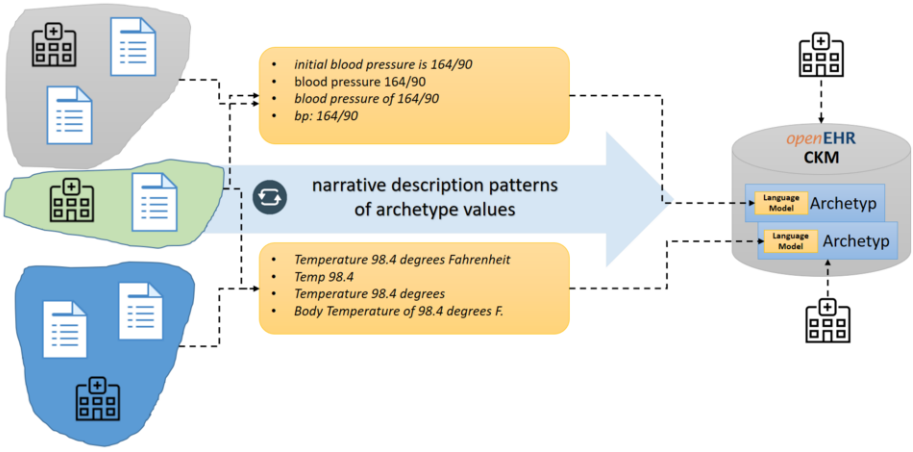

Figure 4. Language models as parts of archetype definitions to support the integration of narrative described information into openEHR based patient records. A language model is initially generated by a hospital, optional extended by one or many further hospital(s) and finally made available to all hospital with access to the openEHR CKM.

4.1. Using Templates instead of generic Archetypes

The approach in this paper expects an openEHR archetype, which represents a concrete medical concept (e.g. Blood Pressure) including its semantic constraints. However, there exists a second kind of archetypes: generic archetypes (e.g. Laboratory Test Result) whose specific constraints are concretized during the integration into a certain openEHR template. Also templates can be published, shared and reused over the

openEHR CKM. Hence, our approach can also be used to generate language models for an openEHR template instead of an archetype.

5. Acknowledgements

The research for this paper was done within the project "HiGHmed" as part of the German Medical Informatics-Initiative (MI-I). The project is funded by the German Federal Ministry of Education and Research (BMBF, grant id: 01ZZ1802C).

References

[1] Ocean Informatics / Ocean Health Systems, openEHR Clinical Knowledge Manager, https://www.openehr.org/ckm/ [cited 2019 February 22].

[2] T. Beale , Archetypes: Constraint-based Domain Models for Future-proof Information Systems, *Eleventh OOPSLA workshop on behavioral semantics: serving the customer* (2002),16-32.

[3] H. Leslie and S.B. Ljosland, Peer Review of Clinical Information Models: A Web 2.0 Crowdsourced Approach, *Stud Health Technol Inform* **245** (2017), 905–909.

[4] A. Wulff, B. Haarbrandt and M. Marschollek, Clinical Knowledge Governance Framework for Nationwide Data Infrastructure Projects, *Stud Health Technol Inform* **248** (2018), 196–203.

[5] B. Haarbrandt, B. Schreiweis, S. Rey et.al., HiGHmed – An Open Platform Approach to Enhance Care and Research across Institutional Boundaries, *Methods Inf Med* **57** (2018), 66-81.

[6] A. Wulff, M. Mast, and M. Marschollek, Automatisierte Extraktion entscheidungsrelevanter Informationen aus Anamneseberichten der pädiatrischen Intensivmedizin, *German Medical Science GMS Publishing House* (2018), doi: 10.3205/18gmds024.

[7] I. Nikolova, G. Angelova et al., Medical archetypes and information extraction templates in Automatic Processing of Clinical Narratives. *International Conference on Conceptual Structures* (2013), 106-120.

[8] A. Roehrs, C.A. da Costa et al., Toward a model for personal health records interoperability, *IEEE journal of biomedical and health informatics* **23** (2019), 867-873.

[9] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural computation* **9** (1997), 1735–1780.

[10] A. Joulin, E. Grave et al., Bag of Tricks for Efficient Text Classification, Computing Research Repository (2016), arXiv preprint: 1607.01759.

[11] A. Huang, Similarity measures for text document clustering, *Proceedings of the sixth new zealand computer science research student conference*, Christchurch, New Zealand **4** (2008), 9-56.

[12] T. Mikolov, K. Chen, et al., Efficient estimation of word representations in vector space, *Computing Research Repository* (2013), arXiv preprint:1301.3781.

[13] E. Grave, P. Bojanowski et al., Learning word vectors for 157 languages, *Computing Research Repository* (2018), arXiv preprint: 1802.06893.

[14] P. Bojanowski, E. Grave et al., Enriching word vectors with subword information, *Transactions of the Association for Computational Linguistics* **5** (2017), 135–146.

[15] M. Zubke, Classification based extraction of numeric values from clinical narratives, *Proceedings of RANLP Workshop on Biomedical Natural Language Processing* (2017), 24-31.