Health Informatics Meets eHealth
G. Schreier et al. (Eds.)

© 2016 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.
doi:10.3233/978-1-61499-645-3-93

SEMCARE: Multilingual Semantic Search in Semi-Structured Clinical Data

Pablo LÓPEZ-GARCÍA[a,1], Markus KREUZTHALER[a], Stefan SCHULZ[a], Daniel SCHERR[b], Philipp DAUMKE[c], Kornél MARKÓ[c], Jan A. KORS[d], Erik M. van MULLIGEN[d], Xinkai WANG[e], Hanney GONNA[e], Elijah BEHR[e], Ángel HONRADO[f]
[a]Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria; [b]Division of Cardiology Department of Medicine, Medical University of Graz, Austria; [c]Averbis GmbH, Freiburg, Germany; [d]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands; [e]St. George's University of London, United Kingdom; [f]SYNAPSE Research Management Partners S.L.

Abstract. The vast amount of clinical data in electronic health records constitutes a great potential for secondary use. However, most of this content consists of unstructured or semi-structured texts, which is difficult to process. Several challenges are still pending: medical language idiosyncrasies in different natural languages, and the large variety of medical terminology systems. In this paper we present SEMCARE, a European initiative designed to minimize these problems by providing a multi-lingual platform (English, German, and Dutch) that allows users to express complex queries and obtain relevant search results from clinical texts. SEMCARE is based on a selection of adapted biomedical terminologies, together with Apache UIMA and Apache Solr as open source state-of-the-art natural language pipeline and indexing technologies. SEMCARE has been deployed and is currently being tested at three medical institutions in the UK, Austria, and the Netherlands, showing promising results in a cardiology use case.

Keywords. Information Storage and Retrieval; Health Records, Personal; Data Mining

1. Introduction

Clinical data in electronic health records (EHRs) have a promising potential in many areas of healthcare [1]: to monitor and improve healthcare delivery [2], to identify disease mechanisms [3], to enhance drug safety [4], and to facilitate patient recruitment for clinical trials [5]. For example, only 18% of clinical trials in Europe and 7% in the US meet their patient enrolment quotas on time, causing delays in bringing new drugs to market. Exploiting patient-level data can also optimize clinical research in several ways, e.g. by enabling the definition of appropriate study designs or ensuring that inclusion/exclusion criteria map to an existing patient population [6].

Cardiovascular disease, in particular ischemic heart disease, is the leading cause of death in the developed world [22]. Almost half of all ischemic heart disease related deaths occur suddenly and are due to abnormal heart rhythms caused by a diseased heart muscle. The primary biomarker used to identify patients at high risk of sudden death is

[1] Corresponding Author: Pablo LÓPEZ-GARCÍA, Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Auenbruggerplatz 2, 8036 Graz, Austria, E-Mail: pablo.lopez@medunigraz.at

the ejection fraction, which is an imaging-based measure of the strength of cardiac contraction. This value may be defined in a quantitative manner expressed as a percentage value, or a qualitative manner described as normal, mild, moderate, or severe impairment. Other important biomarkers for risk stratification include ECG measures such as presence of cardiac rhythm abnormalities, QRS and QT duration, symptoms of breathlessness and transient loss of consciousness. A relevant blood based marker is creatinine, a surrogate for kidney function. A prototypical example of a combination of clinical criteria of interest would be: *"Patients older than 60 years with an ejection fraction less than 35 and a QRS duration ranging between 120 and 130 who are taking Heparin or Dalteparin"*.

Most use cases, such as the one above, have a similar goal in common: to identify patient cohorts based on existing data and clinical criteria like age, gender, diagnosis, signs, symptoms, and lab results. There are two main difficulties for achieving this goal: Firstly, processing the data: large parts of patient-level data are still scattered in heterogeneous resources and stored in unstructured or semi-structured form as free text [7]. Furthermore, at each institution these texts can be available in a different natural language, additionally complicated by medical language idiosyncrasies (e.g., ambiguous terms, acronyms, compounds, derivations, spelling variants, spelling errors, jargon, telegram style), and be mixed with quantitative data (e.g., lab results, drug dosages, dates/times) [8]. A second difficulty is the establishing of the clinical criteria: the system needs to enable users to precisely express the selection criteria, which requires optimised user interfaces.

In this paper we present SEMCARE [9], a recently-finished EU-funded project involving three medical institutions, aimed at helping to solve these problems by providing a multi-lingual platform (English, German, and Dutch) that allows users to express complex queries and get relevant search results from unstructured data in form of free texts.

2. Methods

In order to meet the defined use cases also targeting different languages to which the search scenario is applied, the SEMCARE architecture is constituted by the following components: **Extract-Transform-Load (ETL), Data Integration:** this component transforms heterogeneous data sources (structure and content) into standardized input for further processing (see "Text Mining" component). Data sources might consist of plain text, PDF, Microsoft Word documents, or HL7 messages. **Data Storage:** a structured copy of the data is stored in an i2b2 star schema [10], an efficient and well-established relational clinical data model that allows third party applications to be used for data analysis and visualization. **Anonymization:** a de-identification service can be applied to ensure patient and carer privacy. This component was not used in our case, as the project required that all clinical texts were previously anonymized. **Terminology Management:** this component allows users to browse and administrate (add, edit, and delete concepts and terms) the biomedical terminologies used by the text mining component. **Text Mining:** this component constitutes the core of the semantic middleware. Its main functionality is to provide extended structured information (annotations) using the biomedical terminologies available in the system, and open source Natural Language Processing (NLP) (Apache UIMA [11, 25, 26]) and indexing (Apache Solr [12])

technologies. **Search and Analytics:** a semantically enriched web-based front end for semantic search and analysis of the processed data.

All NLP pipeline components were embedded within UIMA to exploit a flexible and adaptable configuration framework and constitute of the following main components: a tokenizer (Lucene Snowball) and a sentence detector (rule based); Word de-compounding based on morphosemantic analysis [23]; Concept mapping via a customised version of the UIMA Concept Mapper annotator acting on the terminologies explained above; Negation detection based on NegEx [24], as well as a dictionary-based abbreviation and acronym resolver.

3. Results

SEMCARE was installed in three hospitals: St. George's Hospital (London, UK), Erasmus Medical Center (Rotterdam, Netherlands), and the University Hospital of Graz (Austria). Each installation was deployed in isolation from the clinical information system. With approval by the local ethics committees, documents in English, Dutch, and German were extracted from the operational hospital EHR systems, and only limited access was given to authorized staff. All documents were anonymized before being accessed by SEMCARE. The semi-structured discharge letters were transferred to the SEMCARE search server (Solr) in a batch process, obtaining an enhanced index exploiting an NLP pipeline via UIMA. Standard biomedical terminologies were used (SNOMED CT [13] and ICD10 [14] for English; ICD10-GM [15], LOINC [16], and ABDAMED [17] for German).

In addition, German and Dutch terminologies were extended using semi-automatic translation techniques. These terminologies enable semantic search by mapping synonymous terms to concepts, which are related by hierarchical links.

Figure 1 shows the graphical user interface of SEMCARE providing the services to end users, with its three main constituting parts: A **Query builder** (top) allows users to express complex queries graphically in an accessible manner. The listing of **Search**

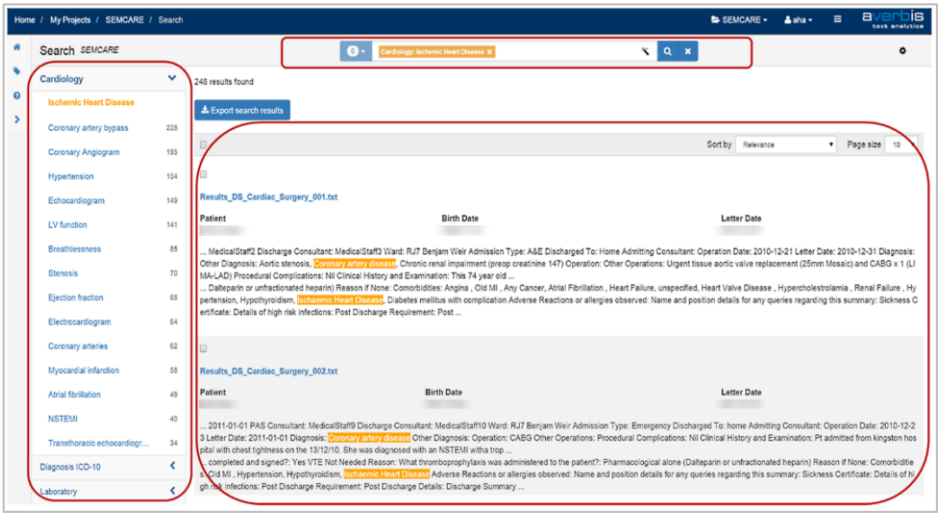

Figure 1. Graphical user interface.

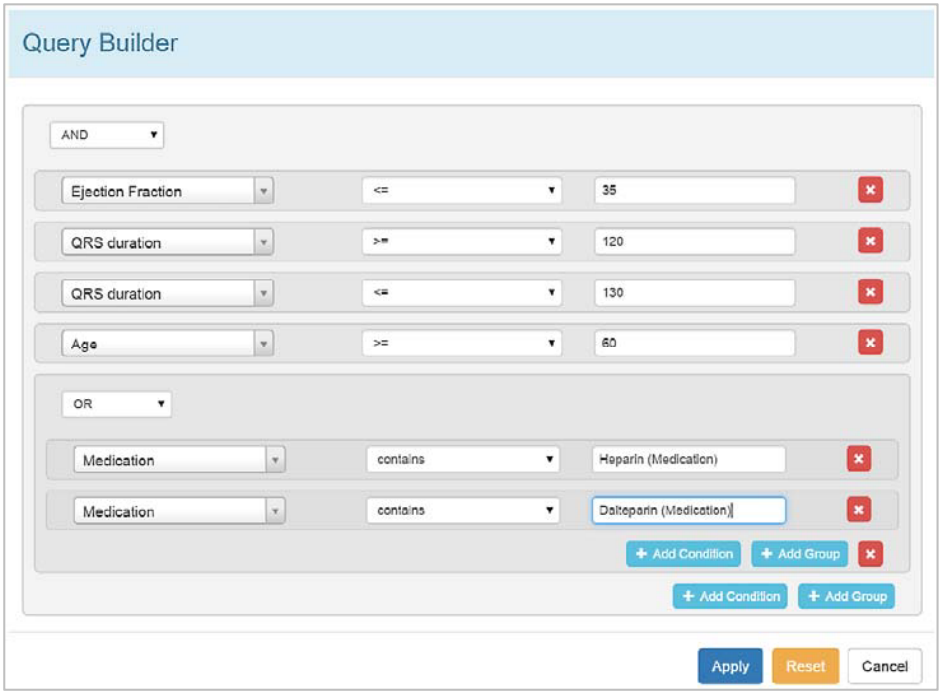

Figure 2. Query builder.

results (centre) shows all the (unstructured) clinical notes matching the criteria from the query. **Facets** (left) permit a quick overview, grouping, and filtering of results based on relevant semantic axes [27] like Disorders, Drugs or Lab, according to the scope of the source terminologies and their subdivisions.

Figure 2 shows the input window when the query builder is clicked, and how to express the cardiology use case presented in the introduction: *"Patients older than 60 years with an ejection fraction less than 35 and a QRS duration ranging between 120 and 130 who are taking Heparin or Dalteparin"*.

Figure 3 shows a search result matching the criteria, and Figure 4 shows the result conveniently summarized for the end user in a structured way.

4. Discussion and Related Work

Most existing clinical data are stored in unstructured form, mainly in EHRs. The difficulties in analysing and capturing semantics from free texts have been studied since

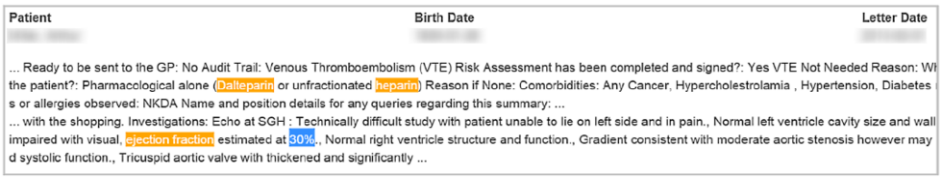

Figure 3. A specific search result matching the input query. Entities and values are marked in the text.

Ejection Fraction	30
QRS duration	125
Diagnosis ICD-10	Oedema Essential (primary) hypertension Diabetes mellitus Living alone Pain Pure hypercholesterolaemia Diabetes mellitus with complications
Cardiology	Ischemic Heart Disease Aortic stenosis Breathlessness Coronary Angiogram Echocardiogram Hypertension Ejection fraction QRS duration

Figure 4. Structured summary of the search result shown in the previous figure.

long ago in the field of computational linguistics, and several well-established conferences are devoted to the matter (e.g. ACL, EACL, SIGIR). However, clinical NLP is still a relatively small sub-area, especially when compared to other biomedical text-mining areas that focus on literature abstracts instead of EHR content. One of the main reasons for the early stage of clinical NLP research is the difficulty of accessing and sharing real patient data due to privacy concerns, which itself leads to a lack of gold standards or references for evaluation.

This is gradually changing within the English speaking community, particularly in the United States, as scientific challenges have been established to foster clinical NLP developments, like the i2b2 [10], SemEval [18], or TREC [19] medical records track challenges. However, the situation in other languages is still very preliminary due to the lack of a joined effort to foster research for processing clinical narratives.

To improve the situation in Europe and languages other than English, SEMCARE adds to previous EU-funded projects, like EHR4CR [20] or MANTRA [21].

5. Conclusion and Future Work

The vast amount of existing clinical data in EHR systems promises great potential. However, several challenges are still to be solved for data to be useful in real applications. A major obstacle is given by the fact that important content is only available as unstructured texts; with clinical texts being especially difficult to process, due to language idiosyncrasies in different natural languages, as well as the limited coverage of domain-specific terminology resources for all languages other than English.

The SEMCARE system addresses these difficulties by providing a multi-lingual platform for performing queries on unstructured medical data. It was deployed and is being tested at three medical institutions in Europe, with documents in English, German, and Dutch. The core technologies in SEMCARE are a solid basis of adapted biomedical terminologies, and a state-of-the-art NLP pipeline.

Project participants are now testing SEMCARE at their institutions and have just finished a more complete evaluation that includes NLP performance, identifying high-

risk individuals for cardiology use cases. A user satisfaction survey has shown that the system excels by very good response times and enables users to get a quick insight into retrospective patient cohorts, which can be further exploited for e.g. hypothesis generation, rare disease detection or high-risk patient profiling. SEMCARE's semantic retrieval approach, powered by terminologies and enriched by synonyms that reflect clinicians' language preferences, guaranteed an efficient exploration of large document spaces with good retrieval quality.

As future work, we plan to fine tune the platform, expand the use of SEMCARE to other domains, and publish the evaluation results in detail.

Acknowledgments

The SEMCARE platform was developed under the European Union's Seventh Framework Program for research, technological development, and demonstration activities (grant no. 611388).

References

[1] P.B. Jensen, L.J Jensen, and S. Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics* **13**(6) (2012), 395–405.
[2] P.J. Embi et al. Development of an electronic health record-based Clinical Trial Alert system to enhance recruitment at the point of care. AMIA Annual Symposium Proceedings. Vol. 2005. American Medical Informatics Association, 2005.
[3] I.S. Kohane. Using electronic health records to drive discovery in disease genomics. *Nature Reviews Genetics* **12**(6) (2011), 417–428.
[4] M.J. Schuemie, et al. Using electronic health care records for drug safety signal detection: a comparative evaluation of statistical methods. *Medical Care* **50**(10) (2012), 890–897.
[5] T. van Staa et al. Pragmatic randomised trials using routine electronic health records: putting them to the test. *BMJ* **344**:e55 (2012).
[6] J. Powell and B. Iain. Electronic health records should support clinical research. *Journal of Medical Internet Research* **7***(1):*e4, (2005).
[7] S.M. Meystre, et al. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform* **35** (2008), 128–44.
[8] W.W. Chapman et al. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *JAMIA* **18**(5) (2011), 540–543.
[9] SEMCARE, http://semcare.eu, last access: 10.02.2016.
[10] I2B2, https://www.i2b2.org/, last access: 10.02.2016.
[11] Apache UIMA, https://uima.apache.org/, last access: 10.02.2016.
[12] Apache Solr, http://lucene.apache.org/solr/, last access: 10.02.2016.
[13] SNOMED CT, http://www.ihtsdo.org/, last access: 10.02.2016.
[14] ICD10, http://www.who.int/classifications/icd/en/, last access: 10.02.2016.
[15] ICD10-GM, https://www.dimdi.de/static/en/klassi/icd-10-gm/, last access: 10.02.2016.
[16] LOINC, https://loinc.org/international/german, last access: 10.02.2016.
[17] ABDAMED, http://www.wuv-gmbh.de/abdata-pharma-daten-service/datenangebot/abdamed/, last access: 10.02.2016.
[18] SEMEVAL, http://alt.qcri.org/semeval2016/, last access: 10.02.2016.
[19] TREC, http://trec.nist.gov/, last access: 10.02.2106.
[20] EHR4CR, http://www.ehr4cr.eu/, last access: 10.02.2106.
[21] MANTRA, https://sites.google.com/site/mantraeu/, last access: 10.02.2106.
[22] T.A. Gaziano et al. (2010). Growing epidemic of coronary heart disease in low-and middle-income countries. Current problems in cardiology, **35**(2), 72-115.
[23] Markó, K., Schulz, S., & Hahn, U. (2005). MorphoSaurus Design and Evaluation of an Interlingua-based, Cross-language Document Retrieval Engine for the Medical Domain. *Methods Archive*, **44**(4), 537-545.

[24] Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., & Buchanan, B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, **34**(5), 301-310.
[25] Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association, **17**(5), 507-513.
[26] http://ohnlp.sourceforge.net/MedKATp/, last access: 10.03.2016
[27] Natarajan, K., Stein, D., Jain, S., & Elhadad, N. (2010). An analysis of clinical queries in an electronic health record search utility. International journal of medical informatics, **79**(7), 515-522.