German Medical Data Sciences 2024
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI240854

Extending the TOP Framework with an Ontology-Based Text Search Component

Franz MATTHIES[a,1], Christoph BEGER[a], Ralph SCHÄFERMEIER[a], Konrad HÖFFNER[a], and Alexandr UCITELI[a]
[a] *Institute for Medical Informatics, Statistics and Epidemiology, Leipzig University*
ORCiD ID: Franz Matthies https://orcid.org/0000-0001-7196-506X

Abstract. Introduction: Constructing search queries that deal with complex concepts is a challenging task without proficiency in the underlying query language – which holds true for either structured or unstructured data. Medical data might encompass both types, with valuable information present in one type but not the other. **Method:** The TOP Framework provides clinical practitioners as well as researchers with a unified framework for querying diverse data types and, furthermore, facilitates an easier and intuitive approach. Additionally, it supports collaboration on query modeling and sharing. **Results:** Having demonstrated its effectiveness with structured data, we introduce the integration of a component for unstructured data, specifically medical documents. **Conclusion:** Our proof-of-concept shows a query language agnostic framework to model search queries for unstructured and structured data.

Keywords. Document Analysis, Search Engine, Knowledge Bases, Classification

1. Introduction

Search engines play an important role not only in our day-to-day life but also in the scientific community for researchers as well as for clinical practitioners. While some data is stored in a structured format, readily accessible with specific query languages, a significant portion remains hidden in unstructured natural language. This natural language data, though valuable, requires more advanced techniques to unlock its potential. Even for structured data, a sufficient understanding of the query language and the exact schema of the data in question is paramount to gather sensible information.

In [1] we devised a way around the issue of having domain experts learn a particular query language when dealing with phenotype algorithms (for identifying, e.g., patient cohorts for studies) by separating the modeling of such an algorithm from the implementation. The reasoning goes that this procedure may very well be applied to the querying of unstructured data, too. Therefore, we integrated a text query component into our TOP framework by adhering to the same design philosophy and utilizing the existent intuitive graphical user interface to allow users to model, reuse, build upon, and share search concepts that describe a medical scenario for which it might simply be feasible to utilize relevant textual data. In contrast to the phenotyping component, the text component does not require the development of phenotype models, but only the creation

[1] Corresponding Author: Franz Matthies, Mailing address: Härtelstraße 16-18, 04107 Leipzig, Germany, email: franz.matthies@imise.uni-leipzig.de

or import of so-called Search Ontologies representing relevant search concepts and terms (labels) of a domain (refer to section 2.2)

A more pressing use case relates to so-called adverse drug events (ADE), which were for instance a main research topic of the POLAR project [2], where a clinically relevant event (for instance a 'fall' event) might or might not be causally linked to a particular medication when occurring in a temporal succession. Those events, while hurting the patients, might also result in high costs for the healthcare system but are not easily trackable and therefore preventable in e.g., patient-related clinical documentation. This is because for instance the event is poorly represented by medical code systems like ICD-10 (e.g., in the case of 'delirium') (see [1] for a short overview) or is not represented at all like for the 'fall' event – for the latter only the consequences of such an event might be encodable (e.g., 'fracture'). Finding instances of such events in all their possible expressions by querying textual data might therefore help prevent them.

There is certainly no lack of search engines even for specific domains [3–5] (an additional short overview for the life sciences is given in e.g. [4]) – not least "[t]he exponential growth of scientific publications in the life science domain has inspired a wide range of information retrieval services over the last decade". [4] However, these specialized engines often rely on pre-processing steps like gene normalization or clinical term detection, making them unusable for languages or domains lacking such training resources. Another hindering factor might be the aforementioned accessibility of the query language in question. EMERSE, for instance, shows an impressive evolution in tackling the latter by ultimately employing a semantically based query recommendation feature. [5–7]

Our approach to document search also tackles the accessibility issue by utilizing Search Ontologies (see section 2.2) and provides a means of filtering the query result by building upon Concept Graphs (see section 2.3). With this we want to be as independent as possible from outside resources but at the same time provide a framework with which an intuitive and comprehensive search is possible for practitioners of various fields as well as researchers. In the following we will outline our methodological background and give an exemplary walkthrough of a search based on a very simple 'fall-event' model and a small document set.

2. Methods

2.1. Data

Any document or derivative thereof (e.g., concept graphs/clusters) that are shown or referenced in the following sections are taken from GRASCCO[2] – a synthetic (and therefore publicly shareable) corpus of 63 German clinical documents. [8] To be better suited for an English publication and to test the cross-lingual capabilities of the TOP Framework, we translated all documents with the DeepL API.

2.2. Search Ontology (SON)

The Search Ontology (SON) is a promising approach for ontology-based modeling of complex search queries, as demonstrated in previous studies. [9–11] The SON method

[2] https://zenodo.org/records/6539131 (accessed June 24, 2024)

is generic and can be applied in any domain by building domain-specific search ontologies (e.g., searching for patients who have suffered a fall).

A search ontology models search concepts (e.g., fall) and search terms (e.g., “fall”, “trip”, “stumble”, “tumble”, etc.). Search concepts are organized hierarchically and are designated or denoted by search terms (labels, strings, keywords, synonyms). A distinction is made between single and composite search concepts. Composite concepts describe a search by specifying an abstract expression that is based on further concepts and defined search operators. For example, certain injuries (such as a ‘fracture of the femur’ or ‘intracranial injury’) may also indicate a fall. In this case, a search concept can be introduced for each corresponding injury[3] and a composite concept (e.g., fall search) can be defined whose expression is based on all relevant concepts (e.g., “Fall OR Fracture of the femur OR Intracranial injury”). A specific search query (in a specific query syntax, such as Lucene query syntax) is generated at runtime by an appropriate generator from the expressions and terms of the concepts (taking into account the hierarchy, i.e., sub-concepts and their terms).

The modular architecture of a search ontology, the reusability of the ontology parts, and its relatively simple structure make it applicable even for non-ontological domain experts without excessive effort and in-depth knowledge of the resulting query syntax. SON-based modeling of complex queries enables the precise definition of the desired topic, reduces the effort and errors of manual query specification, and minimizes irrelevant search results.

2.3. Concept Graphs

In a previous publication, we proposed a new method “that utilizes general-purpose basic text analysis components and state-of-the-art transformer models to represent a corpus of documents as multiple graphs” [12] where the individual nominal phrases are represented as nodes which in turn are connected by edges depending on their semantic proximity. As a result, there would be a number of graphs according to the different conceptual groups found in a text corpus. By utilizing these concept graphs as a basis for generating document representations for a clustering task and comparing the results to another recent, well-performing approach, we provided first strong evidence that they are indeed able to very well capture information about documents and how they relate to each other – in particular for documents that are from domains and/or languages that are underserved with regard to NLP models for extracting information, since the generation of concept graphs doesn’t necessarily rely on specifically trained or fine-tuned models from a distinct domain.

We then touched on a potential use case where these graph structures could serve as a tool to enhance the initial search of relevant documents in a search engine and facilitate result filtering. By utilizing the rich semantic information encoded within the concept graphs, the search process can be optimized to yield more relevant and tailored results to users.

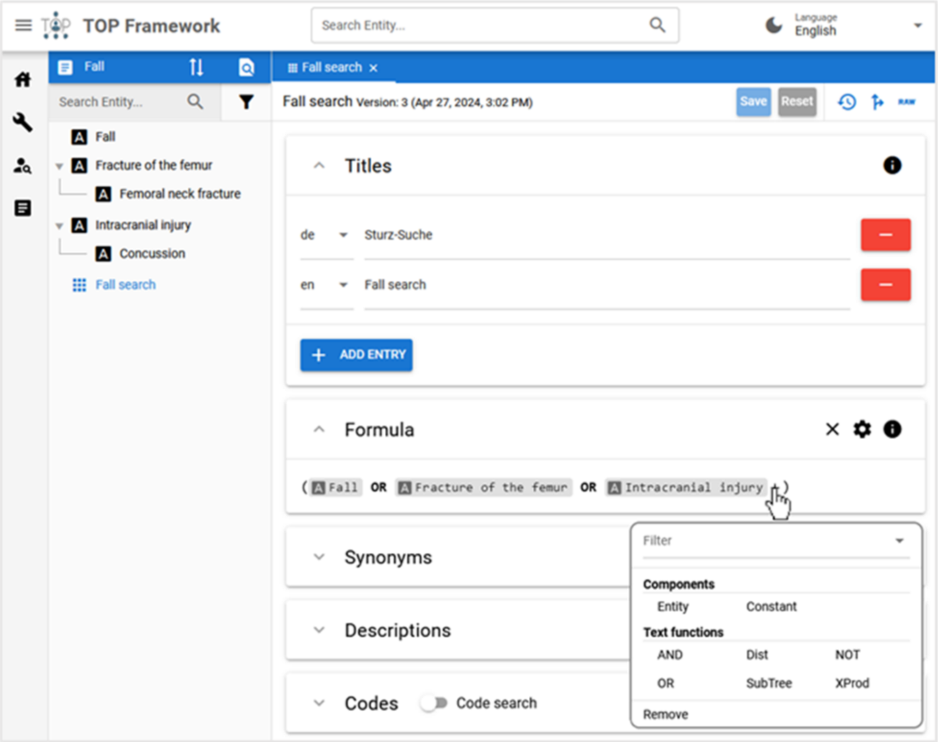
2.4. TOP Framework

In another recent publication, we showed that the TOP Framework is well suited for modeling phenotype algorithms in a manner that is doable by domain experts without

[3] If required, a complete hierarchy of injuries can be integrated.

intricate knowledge of any one particular query syntax (e.g., SQL) to detect for instance the adverse event delirium. [1] This is done by providing a web interface to the users where they can build these models with the help of graphical elements.

We extended this functionality to support the creation of the aforementioned concept graphs, as well as their integration into a Neo4j graph database[4]. We streamlined the API and the underlying code base for the creation of the former[5] and connected it to the TOP Framework's backend so that the graphs can be created and to a small degree curated from the frontend. The source documents might be gathered from a document server (e.g., an Elasticsearch[6] instance, which is the supported method at the moment). The specifications for that can be configured according to one's needs. As with the phenotypic queries, the ontologically modeled search queries (see section 2.2) are translated into the proper query language by an adapter. [13] As long as there are some sentence-transformer models for a particular language available (see section 2.3), it is feasible to use the additional cluster features of the TOP Framework. But even without it, the document search component is available since Search Ontologies are language agnostic. The framework therefore has an inbuilt cross-lingual support.



TOP Framework

Fall
- Fall
- Fracture of the femur
  - Femoral neck fracture
- Intracranial injury
  - Concussion
- Fall search

Fall search Version: 3 (Apr 27, 2024, 3:02 PM)

Titles

de Sturz-Suche

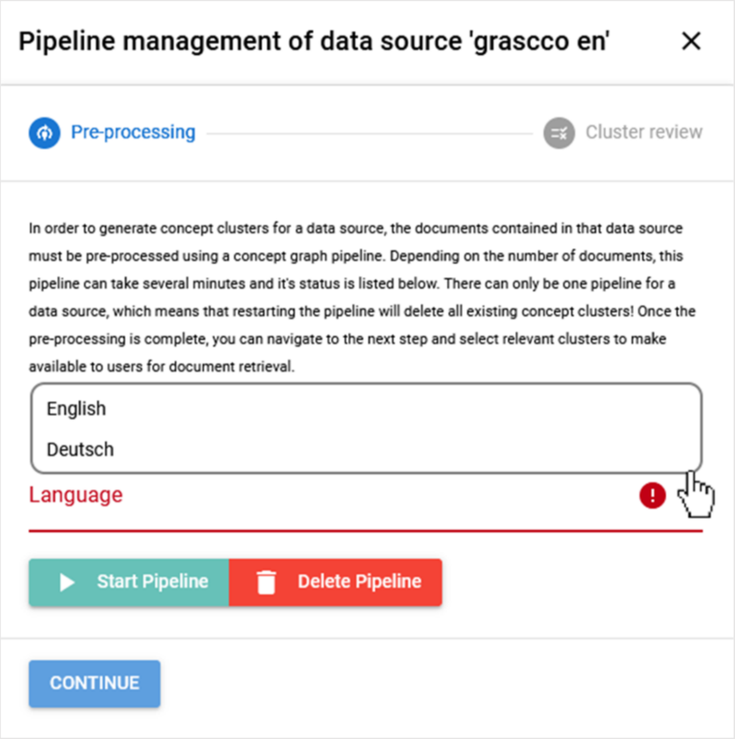
en Fall search

Formula

(Fall OR Fracture of the femur OR Intracranial injury)

Figure 1. SON construction: not shown in this figure is that e.g., the Single Concept 'Fall' comprises also synonyms like 'trip', 'stumble', 'tumble', etc. as well as German equivalents. The same goes for the other Single Concepts. The relatively simple Composite Concept selected here can then be used for the search (simple because the individual concepts are just connected by OR).

[4] https://neo4j.com/product/neo4j-graph-database/ (accessed June 24, 2024)
[5] https://github.com/Onto-Med/concept-graphs (accessed June 24, 2024)
[6] https://www.elastic.co/elasticsearch (accessed June 24, 2024)



Pipeline management of data source 'grascco en'

In order to generate concept clusters for a data source, the documents contained in that data source must be pre-processed using a concept graph pipeline. Depending on the number of documents, this pipeline can take several minutes and it's status is listed below. There can only be one pipeline for a data source, which means that restarting the pipeline will delete all existing concept clusters! Once the pre-processing is complete, you can navigate to the next step and select relevant clusters to make available to users for document retrieval.

Figure 2. Screenshot of the TOP Framework pre-processing dialog. New concept graph pipelines are initiated within the framework by selecting a data source and a supported language.

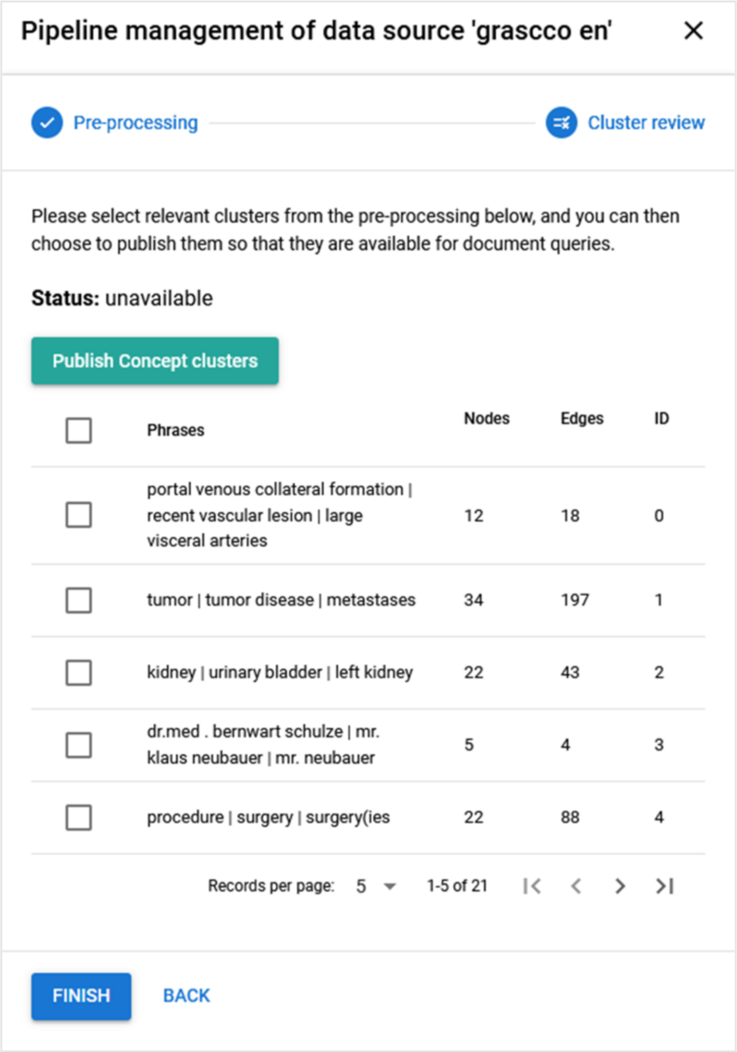
3. Results

In the following, we want to exemplify the basic workings of the document search component of the TOP framework and what one can expect for instance from a relatively small corpus (see section 2.1). The underlying Concept Graphs for the concept clusters described in 3.2 don't rely on domain-specific language models, and the Search Ontology used in 3.1 can be modeled for any language.

3.1. Document Search

As described in section 2.2 Search Ontology (SON), a document search query is generated by so-called Search Ontologies. The TOP Framework allows for an intuitive modeling thereof by keeping related search concepts (single or composite) in a concept repository. Composite concepts can then be constructed by selecting functions (such as And, Or, Not, Dist, SubTree, XProd) and their elements. Elements can be either other concepts (Entity) from within the repository or, in case of Dist, a number (Constant) that allows for fuzzy search in case of one-word concepts or proximity search in case of concepts that have more than one word. Each concept can also have one or more codes associated with it which are backed by a Terminology/Ontology server.

The terms associated with a code are then resolved during query runtime. Figure 1 exemplifies this for a ‘Fall’ repository and a composite concept ‘Fall search’.
Since all concepts can have entries for multiple languages, the same repository can be used for search queries on different document corpora with varying languages, providing that the title and synonyms have an appropriate entry in the respective language that is selected on query runtime. That makes the models of a search concept highly reusable. This even works across different sites that host a TOP Framework instance, since a modelled SON just refers to terms by their word form and potentially codes.



Pipeline management of data source 'grascco en'

Please select relevant clusters from the pre-processing below, and you can then choose to publish them so that they are available for document queries.

Status: unavailable

□	Phrases	Nodes	Edges	ID
□	portal venous collateral formation | recent vascular lesion | large visceral arteries	12	18	0
□	tumor | tumor disease | metastases	34	197	1
□	kidney | urinary bladder | left kidney	22	43	2
□	dr.med . bernwart schulze | mr. klaus neubauer | mr. neubauer	5	4	3
□	procedure | surgery | surgery(ies	22	88	4

Figure 3. After a pipeline has finished, users can review the resulting clusters and proceed by selecting only relevant ones.

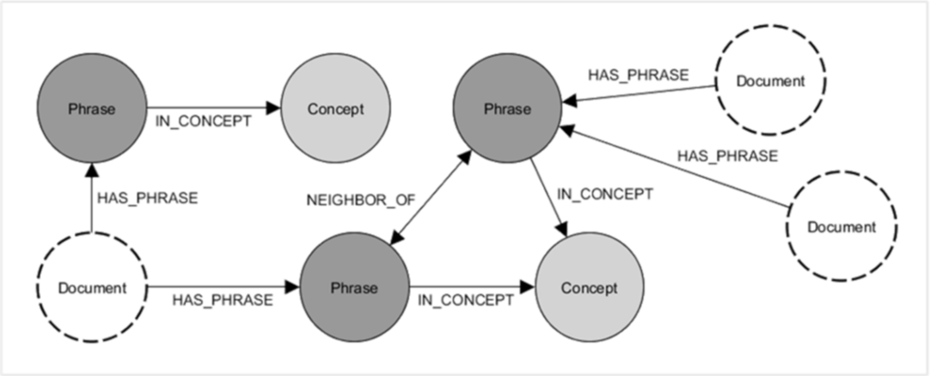


Figure 4. Exemplary graph of a finalized cluster derived by the concept graph component. The graph is stored in Neo4j.

3.2. Concept Clusters

With a data server set up that holds the documents and a connection to the framework, the process of generating suitable concept graphs from the documents can be started by selecting a language as shown in Figure 2 − at the moment we provide reasonable basic default settings for two languages, namely English and German. The complete concept graph creation process that is described in detail in [12] is then performed automatically. However, an option to adjust these by providing an appropriate configuration is easily implementable – either because one is dissatisfied with the resulting graphs, or one wants to adapt it to additional languages. For now, we decided against it, to not overwhelm the user.

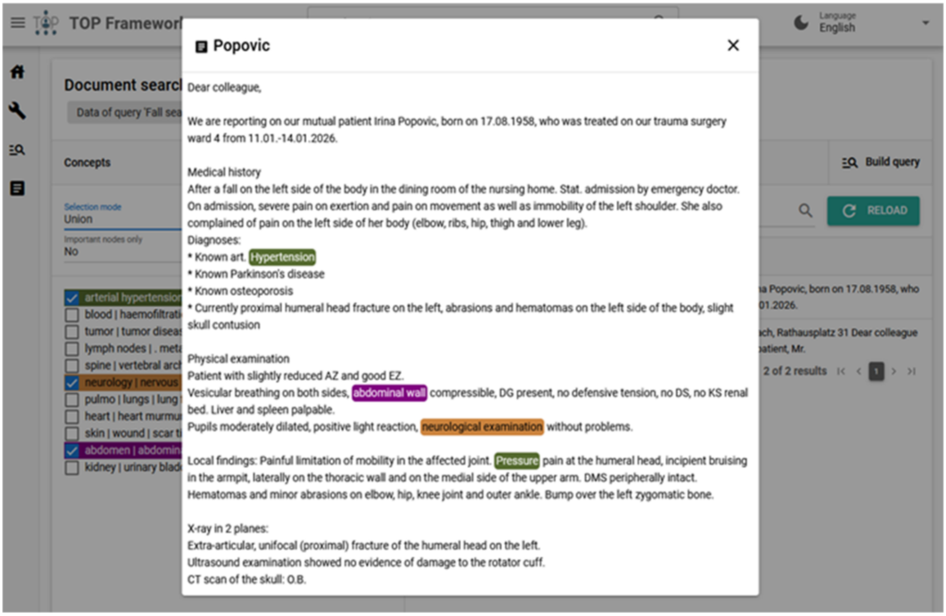
After the pre-processing step was successful, some basic information about the resulting concept graphs can be reviewed and those selected are published to the Neo4j database. For instance, Figure 3 shows that the fourth concept (from 21 in total) is relatively small in comparison with only five nodes and (probably more significant) seems to deal with names only, which might not be relevant for filtering search results.

The selected concepts, which are per definition disconnected from each other as each one subsumes a distinct set of phrases, are then stored in the graph database where they are linked by their documents, which is exemplified by Figure 4: a document has a variety of phrases [HAS_PHRASE] and they in turn belong to a concept [IN_CONCEPT]. Each phrase can have a number of edges denoting its neighbors [NEIGHBOR_OF] depending on their semantic similarity and a certain cut-off point defined during the graph creation step.

At this point, to distinguish between the methodological background – our concept graphs – and the actual entities that can be worked with in our framework via the graph database, we refer to the latter as concept clusters. Having published the relevant ones, they are then usable as a filter mechanism for the document collection – be it the whole document collection connected to by the adapter specification (e.g., an Elasticsearch instance), or just a subset of documents that was retrieved by a document search query with a Search Ontology (see section 2.2) as was exemplified in the previous subsection. Figure 5 for instance shows this by using one of three filter methods: either EXCLUSIVE which allows for only one concept to be chosen, INTERSECTION which filters the documents such that, by virtue of their phrases, each one must be connected to each selected concept, or UNION which shows all documents that are connected to any of the selected concepts. In particular, this figure shows a further filtering of the five result documents we would have gotten, when applying the exemplified Search Ontology from Figure 1 on the GRASSCO corpus (only two remain when selecting the three concepts that deal with the arterial system, neurology, and the abdomen).

4. Discussion

As part of the ongoing development of the TOP Framework, several key enhancements are being explored to further refine the search capabilities. We already provide an integration with terminology servers, but the phrases in the graph database should be grounded on their entries, too. We also want to utilize the concept graphs not only for filtering of result sets, but also for ontology learning techniques to improve and generate composite concepts for a SON. Furthermore, an interesting direction is to integrate a cross-lingual feature into the graph database. Since all nodes (document, concept, phrase) live in the same space, one could think of adding a relation [TRANSLATION] that connects phrases from different languages or two concepts from different language sources that share the same notion.



Dear colleague,

We are reporting on our mutual patient Irina Popovic, born on 17.08.1958, who was treated on our trauma surgery ward 4 from 11.01.-14.01.2026.

Medical history
After a fall on the left side of the body in the dining room of the nursing home. Stat. admission by emergency doctor. On admission, severe pain on exertion and pain on movement as well as immobility of the left shoulder. She also complained of pain on the left side of her body (elbow, ribs, hip, thigh and lower leg).
Diagnoses:
- Known art. Hypertension
- Known Parkinson's disease
- Known osteoporosis
- Currently proximal humeral head fracture on the left, abrasions and hematomas on the left side of the body, slight skull contusion

Physical examination
Patient with slightly reduced AZ and good EZ.
Vesicular breathing on both sides, abdominal wall compressible, DG present, no defensive tension, no DS, no KS renal bed. Liver and spleen palpable.
Pupils moderately dilated, positive light reaction, neurological examination without problems.

Local findings: Painful limitation of mobility in the affected joint. Pressure pain at the humeral head, incipient bruising in the armpit, laterally on the thoracic wall and on the medial side of the upper arm. DMS peripherally intact. Hematomas and minor abrasions on elbow, hip, knee joint and outer ankle. Bump over the left zygomatic bone.

X-ray in 2 planes:
Extra-articular, unifocal (proximal) fracture of the humeral head on the left.
Ultrasound examination showed no evidence of damage to the rotator cuff.
CT scan of the skull: O.B.

Figure 5. One of the two remaining documents from the initial result set for a "Fall search"-SON (c.f. Figure 1), that shows a fall event (here "After a fall on the left side of the body […]") after filtering it with Concept Clusters and the UNION selection mode: all phrases that are contained in the selected concept(s) will be highlighted in the document view.

But first and foremost, we would need to extend our current proof-of-concept with a proper qualitative evaluation. However, for this we would need a bigger (annotated, or at least manually classified) document collection that comprises potential target hits for e.g., the "fall event" use case which would reasonably show the benefits of our framework in terms of usability, reusability, as well as in metrics common to information retrieval, e.g., precision and recall. Due to our participation in the GEMTEX-Project [14], where currently the de-identification step of all documents takes place, we hope to get access to a valuable data set on which we can run an extensive analysis in the future. Even more so that experts from various fields like pharmacology or neurology are taking part as well.

5. Conclusion

Having designed the TOP Framework to facilitate the modeling of and collaboration on Phenotype Algorithms as well as sharing them, we extended its features by adding a document search component, which demonstrates the feasibility to utilize the same conceptual vector for modeling queries on structured as well as unstructured data. It's therefore possible for domain experts to learn this one concept only – which is in addition simpler and aided by intuitive graphical elements, but backed by the same expressiveness of any common query language – to get answers to research or clinical questions from both data types. We also gave an outlook for a more in-depth analysis of our framework.

Declarations

Conflict of Interest: The authors declare that there is no conflict of interest.

Contribution of the authors: FM: writing of the manuscript; CB, RS, KH, AU: proofreading; FM, CB, AU: implementation & integration of the text-search component.

Funding: The 'Terminology and Ontology-based Phenotyping' (TOP) project was funded by the German Federal Ministry of Education and Research within SMITH TOP (reference number: 01ZZ2018) and within the GeMTeX project (reference number: 01ZZ2314D).

TOP Framework: https://top.imise.uni-leipzig.de/

Tools

DeepL API, Deepl SE: https://www.deepl.com (accessed June 24, 2024)

References

[1] Beger C, Boehmer AM, Mussawy B, Redeker L, Matthies F, Schäfermeier R, et al. Modelling Adverse Events with the TOP Phenotyping Framework. Stud Health Technol Inform. 2023;307:69–77. doi:10.3233/SHTI230695.

[2] Scherag A, Andrikyan W, Dreischulte T, Dürr P, Fromm MF, Gewehr J, et al. POLAR – „POLypharmazie, Arzneimittelwechselwirkungen und Risiken" – wie können Daten aus der stationären Krankenversorgung zur Beurteilung beitragen? Präv Gesundheitsf. 2022. doi:10.1007/s11553-022-00976-8.

[3] Van Landeghem S, Hakala K, Rönnqvist S, Salakoski T, Van de Peer Y, Ginter F. Exploring Biomolecular Literature with EVEX: Connecting Genes through Events, Homology, and Indirect Associations. Adv Bioinformatics 2012;2012. doi:10.1155/2012/582765.

[4] Faessler E, Hahn U. Semedico: A Comprehensive Semantic Search Engine for the Life Sciences. In: Bansal M, Ji H, editors. Proceedings of ACL 2017, System Demonstrations, Vancouver, Canada: Association for Computational Linguistics; 2017, p. 91–6.

[5] Hanauer DA, Mei Q, Law J, Khanna R, Zheng K. Supporting information retrieval from electronic health records: A report of University of Michigan's nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE). J Biomed Inform. 2015;55:290–300. doi:10.1016/j.jbi.2015.05.003.

[6] Hanauer DA, Wu DTY, Yang L, Mei Q, Murkowski-Steffy KB, Vydiswaran VGV, et al. Development and empirical user-centered evaluation of semantically-based query recommendation for an electronic health record search engine. J Biomed Inform. 2017;67:1–10. doi:10.1016/j.jbi.2017.01.013.

[7] Wu DTY, Hanauer D, Murdock P, Vydiswaran VGV, Mei Q, Zheng K. Developing a Semantically Based Query Recommendation for an Electronic Medical Record Search Engine: Query Log Analysis and Design Implications. JMIR Form Res. 2023;7:e45376. doi:10.2196/45376.

[8] Modersohn L, Schulz S, Lohr C, Hahn U. GRASCCO — The First Publicly Shareable, Multiply-Alienated German Clinical Text Corpus. German Medical Data Sciences 2022 – Future Medicine: More Precise, More Integrative, More Sustainable! 2022;296:66–72. doi:10.3233/SHTI220805.

[9] Uciteli A, Kropf S, Weiland T, Meese S, Graef K, Rohrer S, et al. Ontology-based specification and generation of search queries for post-market surveillance. J Biomed Semant. 2019;10:9. doi:10.1186/s13326-019-0203-7.

[10] Uciteli A, Goller C, Burek P, Siemoleit S, Faria B, Galanzina H, et al. Search ontology, a new approach towards semantic search. In: Plödereder E, Grunske L, Schneider E, Ull D, editors. Informatik 2014, Bonn: Gesellschaft für Informatik e.V.; 2014, p. 667–72.

[11] Kropf S, Uciteli A, Schierle K, Krücken P, Denecke K, Herre H. Querying archetype-based EHRs by search ontology-based XPath engineering. J Biomed Inform. 2018;9:16. doi:10.1186/s13326-018-0180-2.

[12] Matthies F, Beger C, Schäfermeier R, Uciteli A. Concept Graphs: A Novel Approach for Textual Analysis of Medical Documents. Stud Health Technol Inform. 2023;307:172–9. doi:10.3233/SHTI230710.

[13] Beger C, Matthies F, Schäfermeier R, Kirsten T, Herre H, Uciteli A. Towards an Ontology-Based Phenotypic Query Model. Appl Sci. 2022;12:5214. doi:10.3390/app12105214.

[14] Meineke F, Modersohn L, Loeffler M, Boeker M. Announcement of the German Medical Text Corpus Project (GeMTeX). Stud Health Technol Inform. 2023;302:835–6. doi:10.3233/SHTI230710.