A Semantic Platform for Information Retrieval from E-Health Records

Harsha Gurulingappa[1,2], Bernd Müller[1], Martin Hofmann-Apitius[1,2], and Juliane Fluck[1]

[1]Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
[2]Bonn-Aachen International Center for Information Technology (B-IT)
Dahlmannstraße 2, 53113 Bonn, Germany
Corresponding Author: harsha.gurulingappa@scai-extern.fraunhofer.de

Abstract

Electronic patient health records encompass valuable information about patient's medical problems, diagnoses, and treatments offered including their outcomes. However, a problem for medical professionals is an ability to efficiently access the information that are documented in the form of free-text. Therefore, the work presented here exhibits an information retrieval platform for efficient processing of e-health records. The system offers facilities for keyword searches, semantic searches, and ontological searches. An open evaluation during the TRECMED 2011 demonstrated competitive results.

1. Introduction

In the medical domain, recent progress in the research and development along with advancement in patient healthcare technologies has resulted in generation of enormous amount of data in various forms. Amongst them, free-text denote one important data resource due to their abundant existence, rapid rate of generation, as well as valuable information enclosed [(Meystre et al., 2008)]. Complex assumptions, interpretation of novel findings or contradictions, and hypotheses are often expressed using a natural language in free-text. Especially in the medical domain, a major portion of the patient clinical observations, including radiology reports, operative notes, and discharge summaries that are recorded as narrative text (dictated and transcribed, or directly entered into system by care providers) [(Demner-Fushman et al., 2009)]. The study of literature enables the identification of novel facts, hypotheses, new connections between the events occurring at different levels (i.e. from microscopic to physiological) and drives the generation of new ideas and clinical decision support.

However, the goal is hard to achieve by reading all the documents since the size of bibliographic space is extremely huge. The enormous growth of literature resources has urged the development of domain specific informatics tools in order to support the analysis of huge amount of unstructured information. Therefore information retrieval and information exaction approaches have gained popularity since over a decade. In the context of medical research, the information retrieval includes identifying the patient records from hospital repositories that can best answer a physician's question of interest. This task is not trivial due to the existence of various denominations of medical concepts in textual records. For example, a drug name "Aspirin" can be represented in terms of several synonyms, brand names, or abbreviations.

Considering the ambiguity inherent to the medical literature, identification of medical concepts in documents followed by retrieval has demonstrated success in the past years [(Schulz et al., 2008)]. Tagging the concepts and mapping them to standard database entries normalizes different forms of the same concept to one standard form. This helps to overcome the problems associated with synonyms, acronyms and morphological variants in text. However, an availability of comprehensive and domain specific dictionaries as well as consistent named entity recognition techniques are preconditions for such approaches.

In order to address this challenge, the TRECMED 2011 provides an experimental platform for open development, evaluation, and comparison of approaches for efficient information retrieval from e-health records. TRECMED provides a collection of de-identified e-health records from various hospital sources. For a given set of expert formulated questions (also referred to as *topics*), the task is to retrieve sets of records from the collection that can best answer the questions.

The work reported here presents the participation of Fraunhofer SCAI in the TRECMED 2011 challenge. The core of our framework contains approximately 100,000 de-identified e-health records pre-indexed with medical concepts, and relationships. Different query formulation strategies such as manual searches, semantic searches, ontological searches, and their combinations have been systematically performed and evaluated. The following sections give detailed information about the work strategy and the results obtained.

2. Methods

The dataset used for TRECMED 2011 contains 101,711 e-health records from University of Pittsburgh NLP repository[1]. The dataset is composed majorly of radiology reports constituting nearly 50% of the total dataset followed by history and physical exam reports,

[1]http://nlp.dbmi.pitt.edu/nlprepository.html

emergency department reports, and so forth. Altogether, 35 expert-formulated topics were provided and the task was to retrieve sets of records from the collection that can best answer the topic questions. An example of topic question is *find patients with gastroesophageal reflux disease who had an upper endoscopy*.

2.1. Pre-Processing

The TRECMED collection contains 101,711 reports. A notion of "Visit" defines all the reports corresponding to a patient's consult to the hospital. In the current dataset, the smallest visit corresponds to one report and the largest visit corresponds to 418 reports. Mapping between the reports and visits were provided in prior[2]. An official evaluation criteria required participants to return sets of visits for different topics. The pre-processing step combined multiple reports to their representative visits without changing the semantic structure of visits. For example, if a visit contains two radiology reports and two discharge summaries, after report-to-visit merging the final visit would have one radiology report section that is a combination of two constituent radiology reports and one similarly generated discharge summary section. The report-to-visit merging resulted in 17,198 visits that were subjected to further processing. Each visit contains 9 free-text sections that are formed by constituent reports. The sections are complaint (COMP), radiology reports (RAD), history and physical exams (HP), emergency department reports (ER), progress notes (PGN), discharge summaries (DS), operative reports (OP), surgical pathology reports (SP), and cardiology reports (ECHO).

3. Patient Demography Identification

Patient demography identification task identifies patient's age and gender indicated within the visit. An age-identifier was developed that is a rule-based and regular-expression based system for the identification of de-identified age groups mentioned in visits. The system finally classifies a visit as *child*, *teen*, *adult*, or *elder*. Identifying the age within patient visits is not a trivial task since a visit may contain ages of patient's relatives such as son, father, mother, etc. Manually crafted rules were applied to filter out ages of non-patients and an evaluation of the system was internally performed that indicated superior results. Visits with ambiguous multiple age groups information were classified into multiple age groups respectively. For example, the visit *ge4U9SGxaDRw* defines the patient as *teen* and *adult*. As a result of age identification, 9185 visits were classified as *adult*, 5747 as *elder*, 581 as *teen*, 273 as *child*, and 3248 had no age information. A gender-identifier was developed that is a rule-based and regular-expression based system for identification of patient's gender mentioned in visits. The system finally classifies a visit as *male* or *female*. The gender-identifier recognizes gender-specific nouns and pronouns such as male, female, she, her, etc. and based on the frequency of gender mentions it classifies a visit. Visits with ambiguous gender information were classified into both gender categories. As a result of gender identification, 8034 visits were categorized as *male*, 6916 as *female*, and 2248 visits had no gender information.

[2] http://www-nlpir.nist.gov/projects/trecmed/2011/tm2011.html

Type of relation	No. of occurrences
LOCATION-OF	151,225
PROCESS-OF	58,443
TREATS	26,816
IS-A	20,417
PART-OF	20,228

Table 1: Top five frequently occurring types of relationships in TRECMED visit collection.

4. Concept and Relation Identification

Different tools were applied for the recognition of concepts and relations in visits. Concept and relation identification was performed on all free-text sections of visits.

MetaMap was applied for the identification of UMLS concepts in visits. UMLS contains over 100 semantic classes of concepts such as the anatomy, physiology, disorder, and many more. All classes of UMLS concepts recognized by MetaMap were used.

SemRep (Semantic Knowledge Representation)[3] is a tool for the identification of relations in any arbitrary text. SemRep identifies relationships between UMLS concepts in text within the sentences. Types of relations that SemRep identifies is pre-defined by the UMLS. Table 1 shows top five types of frequently occurring relationships. Altogether, 30 different types of relationships were identified in TRECMED visit collection.

ProMiner was used along with pre-processed dictionaries for the identification of named entities (referred to as *concepts*) in text. The dictionaries used for concept identification can be broadly categorized as dictionaries for medical problems, treatments, and diagnostic tests. Dictionaries used and the information they contain are as follows:

MedDRA provides a comprehensive terminology for medical problems such as signs, symptoms, diseases, adverse effects, syndromes, and many more. The curated version of applied MedDRA dictionary contains 15,436 entries with nearly 55,000 synonyms.

MeSH-Disease provides a comprehensive terminology for medical problems covered by the "C" sub-hierarchy of MeSH. However, MeSH is hierarchically organized into 14 levels and provides facilities for ontological searches. The curated version of applied MeSH-Disease dictionary contains 4597

[3] http://skr.nlm.nih.gov/

entries with nearly 4,500 synonyms.

DrugBank covers names and synonyms of drugs including their brand names, systemic names and registry codes. The curated version of applied DrugBank dictionary contains 6826 entries with nearly 64,500 synonyms.

ATC[4] provides a coverage of pharmacological, therapeutic, and chemical class names. Examples include terms such as *adrenergic antagonist*, *anti-bacterial agent*, *Prostaglandin*, etc. Synonyms of ATC terms were extracted from the UMLS. Mappings exist between ATC and DrugBank entries within the DrugBank database. Curated ATC dictionary contains 658 entries with nearly 3,500 synonyms.

MeSH-Diagnostic provides a comprehensive terminology for diagnostic tests covered by the "E" sub-hierarchy of MeSH. Applied MeSH-Diagnostic dictionary contains 2,548 entries with nearly 22,000 synonyms.

A CRF-based system was trained over manually annotated concepts in approximately 800 e-health records provided by the I2B2 challenge 2010[5]. The system was trained for the recognition of medical problems, treatments, and tests in e-health records (referred to as CRF-Prob, CRF-Treat, CRF-Test respectively). Concepts recognized by the CRF were morphosyntactically normalized[6]. Table 2 shows counts of different types of concepts and relations occurring in the TRECMED dataset.

4.1. Assertion Classification on Medical Problems

For classification of assertions made over medical problems, the ConText program [(Harkema et al., 2009)] was used. ConText program contains three separate modules for the identification of negation, temporality, and experiencer information provided over mentions of medical problems in text. The negation module identifies any negations made over medical problems. The temporality module classifies a medical problem as *history*, *recent*, or *hypothetical*. Similarly, the experiencer module identifies if a medical problems occurs in the patient or patient's relatives (such as father, mother, son, etc.). Context program was applied to identify negations, temporalities, and experiencer information made over mentions of problems mapped to MedDRA, MeSH-Disease, UMLS (*Disorder* semantic-type), and CRF-Prob. The negation and experiencer modules were applied as-is whereas the *history* and *hypothetical* rules associated with temporality module were modified. Examples of such modifications include removal of patterns such as *reported*, *complains*, and *presented* that asserts a medical problem as *history*. Similarly, modifications associated with *hypothetical* assertions include removal of patterns such as *as needed*, *come back for*, and so forth. Using the experiencer module, problem mentions were classified as *in-patent* or *not-in-patient*. Several instances exist where a medical problem can attain multiple assertions. For example, in the sentence *His father had no history of hypertension*, the medical problem *hypertension* belongs to *history*, *negation*, and *not-in-patient*. Table 3 shows counts of assertions made over medical problems identified by different concept identification approaches. Nearly 30% to 35% of medical problems recognized by different techniques are negated and this indicates the importance of negation identification in patient health records.

5. Indexing

Free-text fields of TRECMED visits including demographies as well as medical concepts, and relationships occurring in different sections of visits were indexed with SCAIVIEW [(Hofmann-Apitius et al., 2008)]. SCAIView is a high performing and scalable Information Retrieval (IR) system based on Lucene[7]. It provides a framework for indexing several gigabytes of document data and to quickly perform complex searches over text as well as concepts. Free-text in the form of stemmed tokens appearing in different sections of patient visits were indexed. Meta-data such as ICD-9CM codes appearing in the *admit-diagnosis* and *discharge-diagnosis* fields of visits were expanded before indexing. Concepts and relations occurring in different sections of visits were indexed separately. For example, the current index allows searching for the keyword *diabetes* or the MeSH concept *Diabetes Mellitus* (MeSH-ID:D003920) in *discharge summary* (DS) sections of visits. Figure 1 illustrates the workflow adapted for indexing the TRECMED records. The system allows keyword searches, semantic searches, and ontological searches. For a given query, the system retrieves a ranked list of patient visits from the index.

6. Querying and Retrieval

Various search strategies were experimented and Lucene BM25F[8] was applied as a scoring function to measure the similarity between visits and the query. Descriptions of runs and the underlying query formulation and search strategies are discussed in the following subsections.

6.1. MEDRUN1

MEDRUN1 serves as a baseline run where queries were formed by manual extraction of key terms from the topic questions. Queries were formulated in a way to reflect knowledge-based human queries. This

[4] Anatomical Therapeutic Chemical classification system, http://www.whocc.no/atc_ddd_index/

[5] https://www.i2b2.org/NLP/Relations/Documentation.php

[6] http://www.ncbi.nlm.nih.gov/books/NBK9680/

[7] http://lucene.apache.org/java/docs/index.html

[8] http://nlp.uned.es/~jperezi/Lucene-BM25/

Concept/Rel.	No. of occurrences	No. of unique occurrences
UMLS	9,571,099	36,747
Relations	342,712	82,833
MedDRA	1,298,729	4,605
MeSH-Disease	1,144,267	2,239
DrugBank	239,258	902
ATC	38,140	157
MeSH-Diagnostic	406,711	939
CRF-Prob	1,657,912	294,038
CRF-Treat	630,256	76,341
CRF-Test	632,404	47,836

Table 2: Counts of different types of concepts and relations occurring in TRECMED dataset. Total number of occurrences (column 2) and number of unique occurrences after normalization (column 3) are reported.

	Negation	History	Hypothetical	Not-in-patient
UMLS	609,193	224,077	833	21,447
MedDRA	460,117	164,413	787	14,572
MeSH-Disease	377,913	149,822	749	13,497
CRF-Prob	563,682	192,375	1,029	15,341

Table 3: Counts of assertions made over medical problems.

run provides a rationale for the comparison of performance of semantic and ontological searches against knowledge-based human searches.

6.2. MEDRUN2

MEDRUN2[9] applies semantic search strategy to search for UMLS concepts and relations in the index. MetaMap and SemRep programs were applied for the identification of UMLS concepts and relations in topic questions. Automatically identified concepts and relations in topic questions were used for searching in the concept and relation fields of the index. Examples of SemRep found relations in the topic-116: *Patients who received methotrexate for cancer treatment while in hospital* are:
a. (C0920425) Cancer Treatment USES (C0025677) Methotrexate
b. (C0025677) Methotrexate ADMINISTERED_TO (C0030705) Patients

Information about demographies and sections to be searched were extracted from the topic questions. For example, in the topic-110: *Patients being discharged from hospital on hemodialysis*, the system would search in discharge summary (DS) sections of visits with a higher priority in comparison to rest of the sections. A higher priority was assigned to necessary sections by duplicating them in the query. In visits, the concepts referring to medical problems that are negated or occurs as family status were omitted during search. No difference was made when searching for problem concepts occurring as *history* or *recent* event. Nevertheless, the system allows searching for negated concepts, concepts referring to family members, and concepts indicating *history*, *hypothetical* or *recent* events.

6.3. MEDRUN3

MEDRUN3 applies semantic search strategy to search for ProMiner and CRF identified concepts in the index. ProMiner and pre-trained CRF were applied for the identification of concepts in topic questions. Automatically recognized concepts in topic questions were applied for querying in the concept space of the index. Information about demographies and sections to be searched were extracted from the topic questions. Problem concepts that are negated, historical, or indicating family status were processed as described during MEDRUN2.

6.4. MEDRUN4

MEDRUN4 applies ontological search strategy to search for ProMiner and CRF identified concepts in the index. ProMiner and pre-trained CRF were applied for the identification of concepts in topic questions. Automatically recognized concepts in topic questions were applied for querying in the concept space of the index. For the MeSH-Disease and MeSH-Diagnostic concepts, hyponyms (also referred to as child concepts) of the concepts present in topic questions were also used during querying. For example, in the topic-113, MeSH concept *Adenocarcinoma* has several hyponyms such as *Endometrioid Carcinoma*, *Hepatocellular Carcinoma*, and many more. Information about demographies and sections to be searched were extracted from the topic questions. Problem concepts

[9]Our officially submitted runs SCAIMED1, SCAIMED2, SCAIMED3, and SCAIMED4 are similar to MEDRUN1, MEDRUN2, MEDRUN3, and MEDRUN4 in terms of the underlying search methodologies except minor differences in assertions indexed over medical problems. Runs SCAIMED5, SCAIMED6, and SCAIMED7 were generated by merging the results of SCAIMED1&2&3, SCAIMED2&3&4, and SCAIMED1&2&3&4 respectively.

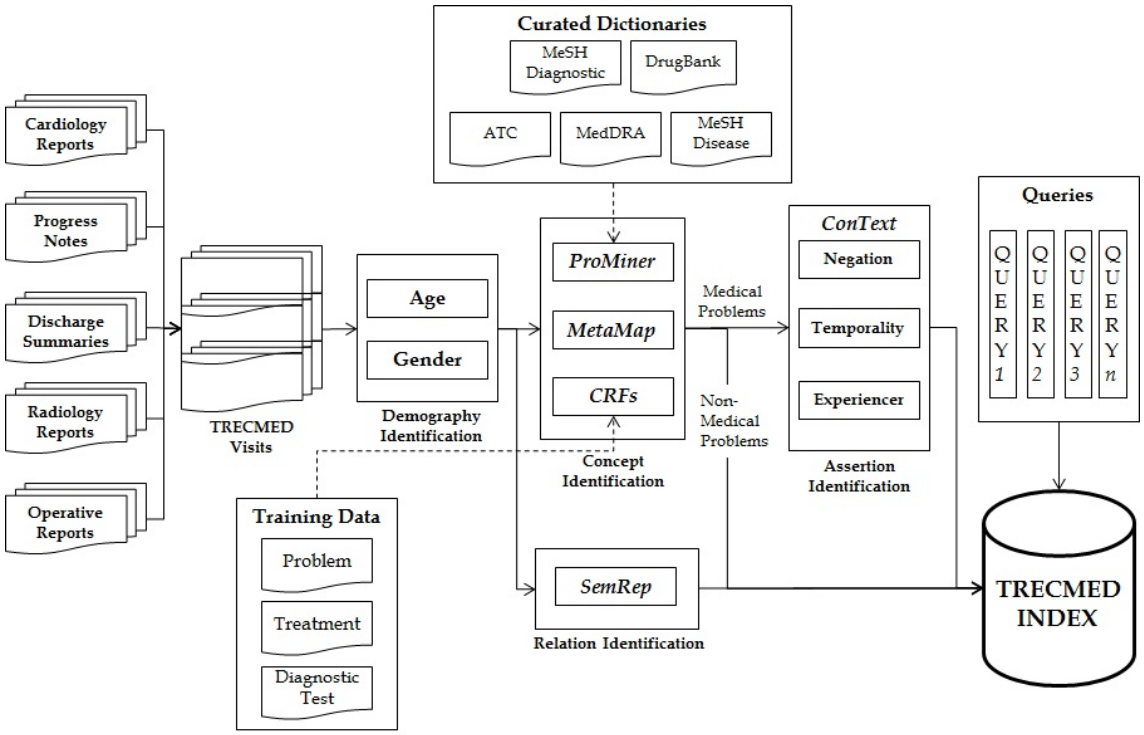

Figure 1: Illustration of the workflow adapted for indexing the TRECMED records.

that are negated, historical, or indicating family status were processed as described during MEDRUN2.

6.5. Run Combinations

Visits retrieved during two or more runs amongst MEDRUN1, MEDRUN2, MEDRUN3, and MEDRUN4 were systematically merged. If a *Visit* occurs in more than one run, its final score was computed using $\sum \frac{BM25F(Visit_i)}{Rank(Visit_i)}$ where i indicates the run.

7. Results

7.1. Performance Evaluation

In information retrieval, along with the relevance of the retrieved documents, the order in which they are presented is important. For example, a system that returns maximum relevant documents within top N documents is worthier than the system that returns maximum relevant documents within middle N documents. Therefore, performances of the submitted runs were evaluated using the Binary Preference score (bpref) as primary metric and the precision at R (R-Prec) as secondary metric.

7.2. Evaluation Results

The reported results are based on the bpref and R-prec scores. Table 4 shows the results of preliminary runs. Table 5 shows the results of run combinations where visits retrieved during different runs were merged according to Section 6.5. Table 6 shows results of the

Run-ID	bpref	R-Prec
MEDRUN1	0.4852	0.3218
MEDRUN2	0.4470	0.2909
MEDRUN3	**0.5503**	**0.3966**
MEDRUN4	0.5333	0.3774

Table 4: Performance of preliminary runs using various search strategies.

impact of age, gender, CRF concepts, and relations on the semantic search.

From Table 4, semantic search in the concept space generated by ProMiner and CRFs achieved good results with bpref score of 0.5503. Results of semantic search with dictionary concepts and CRF-identified concepts considerably outperformed rest of the preliminary runs (means without any post-processing). Searching with MetaMap and SemRep identified UMLS concepts and relations showed poor results. This indicates potential false recognitions that these systems may perform during concept or relation identification. Results of ontological search (MEDRUN4) performed better than manual searching but poorer than a normal semantic search. One potential reason for shortcomings of ontological search is that MeSH was used as a primary hierarchy for hyponym extraction. For several MeSH concepts such as *cancer* (in topic-116) or *colonoscopy* (in topic-113), MeSH provides hundreds of hyponym concepts organized at various levels of hierarchy. It may be fuzzy for topic

Run Description	bpref	R-Prec
MEDRUN1 + MEDRUN3	0.5732	0.5455
MEDRUN2 + MEDRUN3	0.5410	0.3796
MEDRUN3 + MEDRUN4	0.5517	0.3920
MEDRUN1 + MEDRUN2 + MEDRUN3	0.5658	0.3949
MEDRUN2 + MEDRUN3 + MEDRUN4	0.5487	0.3981
MEDRUN1 + MEDRUN3 + MEDRUN4	**0.5767**	**0.4088**
MEDRUN1 + MEDRUN2 + MEDRUN3 + MEDRUN4	0.5746	0.4079

Table 5: Performance of different run combinations.

Run Description	bpref	R-Prec
MEDRUN3	0.5503	0.3966
MEDRUN3 (excl. Age)	0.5505	0.3954
MEDRUN3 (excl. Gender)	0.5499	0.3934
MEDRUN3 (excl. Assertions)	0.5356	0.3793
MEDRUN3 (incl. Relations)	0.5494	0.3969

Table 6: Impact of age, gender, assertions, and relations on semantic search.

evaluators (coming from medical backgrounds) to accept certain hypernym/hyponym concept relations as described in MeSH.

Post-processing by merging the retrieved visits from different runs showed substantial improvement in overall performance (*see* Table 5). Merging the retrieval results of text search, semantic search, and ontological search outperformed rest of the runs in terms of bpref and R-prec scores. This indicates the success of the applied function for merging the visits retrieved from different runs.

The impact of different factors such as age, gender, assertions, and relations on the semantic search was experimented (*see* Table 6). Excluding the age and gender information from the run MEDRUN3 resulted in slight decrease in bpref and R-Prec scores. A potential reason for the low impact of age on retrieval is that all the topic question addressing the ages of patients are associated with adults (*e.g.* Topic-114: *Adult patients discharged home with pallative care/home hospice.*) and the corpus contains over 85% visits belonging to adults (including elders as adults in Section 3.). There are only two topic questions addressing the gender of patients (*i.e.* both focussing on female) and medical conditions associated with these questions are *breast cancer* and *osteopenia* that are more common in females than in males. Relations contributed extremely little to the R-Prec score but the bpref declined. The potential reason for decrease in performance of the system with relations is that SemRep generates potential false positives with entity recognition and therefore the identified relations can hamper the performance of retrieval.

Differences in performance between the run MEDRUN3 and remaining runs without post-processing (i.e. MEDRUN1, MEDRUN2 and MEDRUN4) were analyzed over different topic questions. Figure 2, Figure 3, and Figure 4 shows analysis

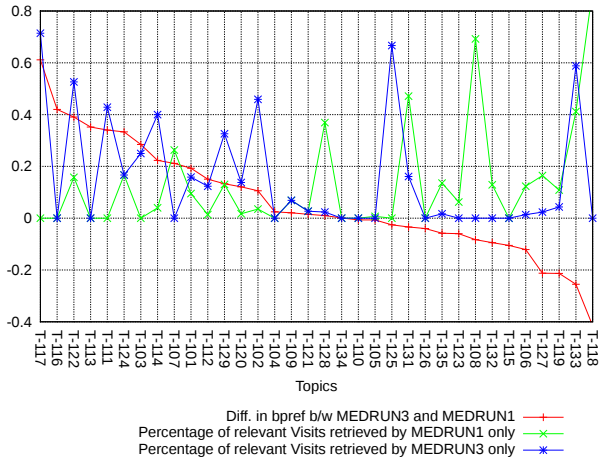

Figure 2: Differences in bpref scores between runs MEDRUN3 and MEDRUN1 for different topics. Percentage of unique relevant documents retrieved by both runs have been indicated.

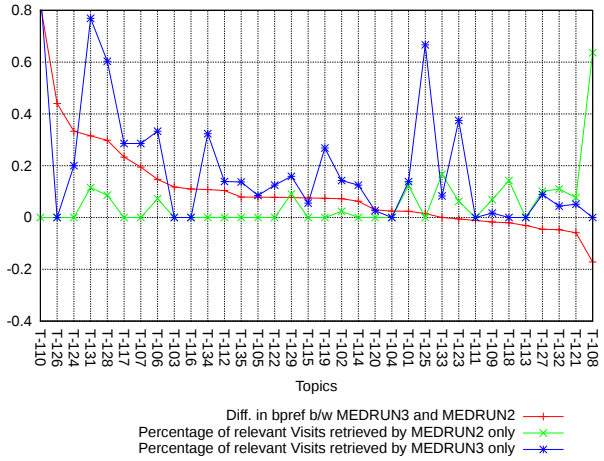

Figure 3: Differences in bpref scores between runs MEDRUN3 and MEDRUN2 for different topics. Percentage of unique relevant documents retrieved by both runs have been indicated.

of difference in results. Table 7 shows the counts of topics for which *no-difference*, *gain*, and *loss* were observed by comparison of the run MEDRUN3 with the rest.

Table 7 shows that semantic search in the concept

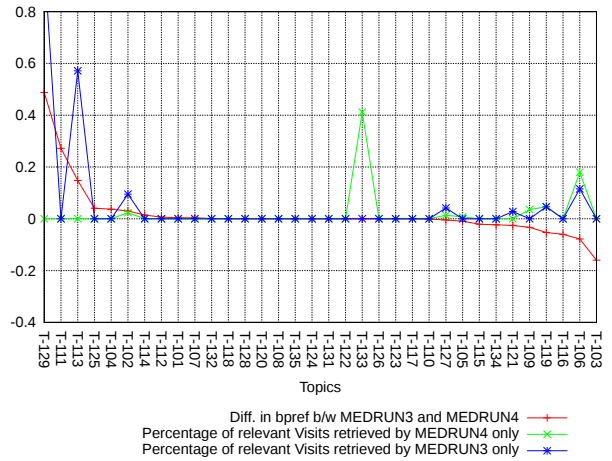

Figure 4: Differences in bpref scores between runs MEDRUN3 and MEDRUN4 for different topics. Percentage of unique relevant documents retrieved by both runs have been indicated.

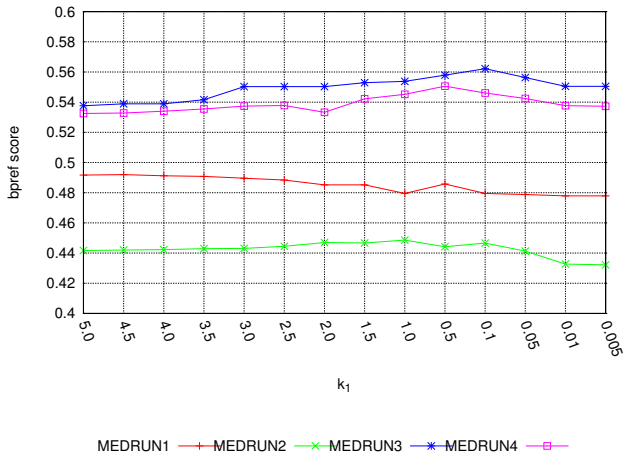

Figure 5: Performance of retrieval (bpref scores) for different values of k_1 for different runs.

Table 7: Counts of topics for which *no-difference*, *gain*, and *loss* were observed by comparison of the run MEDRUN3 with runs MEDRUN1, MEDRUN2, and MEDRUN4.

Runs	Gain	No. Diff	Loss
MEDRUN3 & MEDRUN1	19	0	15
MEDRUN3 & MEDRUN2	24	1	9
MEDRUN3 & MEDRUN4	11	12	11

space generated by in-house NER tools (i.e. ProMiner & CRFs) resulted in an improvement in retrieval performance over majority of topics in comparison to searching with keywords or in the UMLS space. Although, an overall quantitative comparison showed that semantic search can perform better than ontological search (see Table 4), from Table 4 it was clear that semantic and ontological searches can perform competitively depending on questions of interest.
Evaluation of the retrieval performance depends on several factors and they include:

- Number of highly-relevant or relevant versus number of irrelevant or unjudged documents retrieved.
- Relative ranking of relevant and irrelevant documents.
- Relative ranking of highly relevant and relevant documents.

From Figure 2, it can be observed that for topics 116, 113, 104 and 134, bpref scores with semantic search (*i.e.* MEDRUN3) were better than text search (*i.e.* MEDRUN1) but both runs retrieved exactly the same relevant visits with different ranking. On contrary for topic 107, although MEDRUN1 retrieved nearly 25% more relevant visits in comparison to MEDRUN3, the bpref score for MEDRUN3 was higher than MEDRUN1. Similarly for topic 133, although MEDRUN3 retrieved nearly 20% more relevant visits than MEDRUN1, the bpref score of MEDRUN1 was relatively better than MEDRUN3. This indicates that performances of retrieval depend on ability of system to fetch the relevant documents as well as rank the relevant ones with higher priority.

7.2.1. Parameter Optimization of BM25F and its Influence on Retrieval

The similarity scoring function BM25F can be tuned with two free parameters *i.e.* b and k_1. Experiments were performed using the run MEDRUN3 with different values of b and k_1. It was observed that altering the parameter b did not have any influence on the performance of retrieval whereas altering k_1 showed changes in the behavior of the retrieval. By default, BM25F uses k_1=2. Different values of k_1 were chosen between the values 0.005 and 5.0 that can be observed in Figure 5.
At the end of parameter optimization, the best result was obtained by MEDRUN3 with k_1=0.1 with bpref score of 0.5622 and R-prec score of 0.4107. The performances of different runs varied with changes in the parameter k_1. Although it was not possible to establish one global maximum value of k_1 that suits different runs, observations showed that searching in concept space (MEDRUN2, MEDRUN3, and MEDRUN4) favored lower k_1 values such as 0.1 to 0.5 whereas the text search favored higher values of k_1 like 4.0 to 5.0.

8. Error Analysis

Runs were analyzed in comparison to gold standard judgements by topic evaluators in order to understand common sources of errors. One potential reason for shortcomings of retrieval performance during the run MEDRUN2 was false positive concept identification by the MetaMap or SemRep programs. An example is *Topic 107: Patients with ductal carcinoma insitu (DCIS)*, where MetaMap identified several occurrences of *DCI*

in documents (that designates a place) as ductal carcinoma. MEDRUN3 utilized the concepts identified by ProMiner with acronym disambiguation strategy that helped in overcoming various false positive concept recognition that can hamper the performance of retrieval.

The author was able to identify cases where semantic search retrieved documents that were judged as *irrelevant* although they contained relevant information. An example is *Topic 117: Patients with Post-traumatic stress disorder*. The MEDRUN3 run retrieved the visit */6RlgeNinbY+* as one amongst the top 10 visits. This visit was judged as *irrelevant* but a manual investigation of the visit revealed the evidence that the patient had post-traumatic stress disorder. This is exemplified by the statements *The patient no longer works. He was trapped in a house fire several years ago and was extensively burned. He has post-traumatic stress disorder. He has been treated for depression*. Another example is *Topic 101: Patients with hearing loss*. The run MEDRUN3 retrieved the visit *D3PsCRkoq+R8* as one amongst the top 10 visits. This visit was judged as *irrelevant* by topic evaluators. Whereas a manual investigation revealed the evidence that the patient had hearing loss. This is exemplified by statements *Extremities: Negative for clubbing or edema. Skin: No rashes, nodules, or lesions. Neurological: He is awake and alert. His visual fields are intact. He has severe hearing loss, but is otherwise nonfocal*. Such evidences render the quality of human judgements during TRECMED questionable.

From Figure 2, it can be observed that semantic search failed in several cases compared to text search. The best example for this scenario is topic-118 in Figure 2. For topic-118: *Adults who received coronary stent during admission*, the text search retrieved nearly 80% of relevant visits that were not retrieved by semantic search. The reason was during MEDRUN3, searching in the concept space was performed using the concept designating *coronary stent*, *coronary artery stent*, and so forth that did not successfully retrieve many relevant visits. A lot of visits mentioned coronary stents administered to patients that were explained descriptively. Example include visit *kwFRWomsN1Ly: Stenting at two sites of the vien graft of the right coronary artery and mid posterior descending artery with 2.5 mm drug-eluting stent*. Another example of such visit is *r3FTkt2ecEdg: stent placed in the first obtuse marginal branch of the circumflex coronary artery*. These are few examples of relevant visits that were retrieved by MEDRUN1 (text search) and not retrieved by semantic search. This exemplifies some limitations associated with semantic search when the coverage of semantic concept space is not very comprehensive.

MEDRUN4 that uses ontological search performed competitively in comparison to MEDRUN3. Although, the overall results of MEDRUN3 is better than MEDRUN4, Table 7 indicates 10 topics where ontological search performed poorer than semantic search. As mentioned perviously, MeSH was used as resource for ontology expansion and this may conflict with topic evaluator's *hypernym-hyponym* acceptability for evaluation. An example is *Topic 116: Patients who received methotrexate for cancer treatment while in hospital*. MEDRUN4 retrieved *PDfRzvZE904q* as one amongst the best 10 retrieved visits. This visit was judged as *irrelevant* by the respective topic evaluator. Upon manual investigation, this visit revealed evidenced that the patient suffered from *T-cell lymphoma* and the patient was administered *high dose methotrexate therapy* while in the hospital. T-cell lymphoma is a subtype of cancer that was likely to be not addressed during topic evaluation.

9. Conclusions

This work reports on a semantic framework for information retrieval in e-health records. Indexing the medical concepts and relations allows semantic searches and ontological searches in the concept space. The system also provides facilities to search for inter-related medical concepts. In addition, the performance of system with different search strategies has been systematically evaluated. Semantic search in the concept space indicated superior results in comparison to the conventional search with textual queries. The results of run with concept-based search outperformed rest of the preliminary runs without any post-processing with best bpref score of 0.5503. A strategic combination of results obtained from text search, semantic search, and ontological search yielded the highest scoring bpref score of 0.5767.

Currently, the performance of retrieval has been tested with 35 topics. In the future, it is necessary to evaluate the system using more questions. This minimizes the deviation of results from standard average and gives a better estimation of system's actual performance. The system with comprehensively indexed medical relationships may substantially enhance the search performance. Finally, the developed system is believed to help domain experts and medical professionals to carry out patient record searches more efficiently. This promotes evidence-based medicine and therefore improves the overall quality of patient care and safety.

10. Acknowledgements

This work is partly funded by Bonn-Aachen International Center for Information Technology (B-IT) Research School with the NRW state of Germany. Thanks to Dr. Roman Klinger for sharing the CRF implementation. Thanks to Heinz-Theodor Mevissen for his valuable contribution.

11. References

Dina Demner-Fushman, Wendy W. Chapman, and Clement J. McDonald. 2009. What can natural language processing do for clinical decision support? *J Biomed Inform*, 42(5):760–772, Oct.

Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. 2009. Context: an algorithm for determining negation, experiencer, and

temporal status from clinical reports. *J Biomed Inform*, 42(5):839–851, Oct.

Martin Hofmann-Apitius, Juliane Fluck, Laura Furlong, Oriol Fornes, Corinna Kolrik, Susanne Hanser, Martin Boeker, Stefan Schulz, Ferran Sanz, Roman Klinger, Theo Mevissen, Tobias Gattermayer, Baldo Oliva, and Christoph M. Friedrich. 2008. Knowledge environments representing molecular entities for the virtual physiological human. *Philos Transact A Math Phys Eng Sci*, 366(1878):3091–3110, Sep.

S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*, pages 128–144.

Stefan Schulz, Philipp Daumke, Pascal Fischer, Marcel Müller, and Marcel Lucas Müller. 2008. Evaluation of a document search engine in a clinical department system. *AMIA Annu Symp Proc*, pages 647–651.